FUSION OF MULTIMODAL INFORMATION FOR MULTIMEDIA INFORMATION RETRIEVAL

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TURGAY YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

FUSION OF MULTIMODAL INFORMATION FOR MULTIMEDIA INFORMATION RETRIEVAL

submitted by **TURGAY YILMAZ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Adnan Yazıcı _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Özgür Ulusoy
Computer Engineering Department, Bilkent University _____

Prof. Dr. Adnan Yazıcı
Computer Engineering Department, METU _____

Prof. Dr. Ahmet Coşar
Computer Engineering Department, METU _____

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU _____

Assist. Prof. Dr. İsmail Sengör Altıngövde
Computer Engineering Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TURGAY YILMAZ

Signature :

ABSTRACT

FUSION OF MULTIMODAL INFORMATION FOR MULTIMEDIA INFORMATION RETRIEVAL

Yılmaz, Turgay
Ph.D., Department of Computer Engineering
Supervisor : Prof. Dr. Adnan Yazıcı

September 2014, 237 pages

An effective retrieval of multimedia data is based on its semantic content. In order to extract the semantic content, the nature of multimedia data should be analyzed carefully and the information contained should be used completely. Multimedia data usually has a complex structure containing multimodal information. Noise in the data, non-universality of any single modality, and performance upper bound of each modality make it hard to rely on a single modality. Thus, multimodal fusion is a practical approach for improving the retrieval performance. However, two major challenges exist; 'what-to-fuse' and 'how-to-fuse'. In the scope of these challenges, the contribution of this thesis is four-fold. First, a general fusion framework is constructed by analyzing the studies in the literature and identifying the design aspects of general information fusion systems. Second, a class-specific feature selection (CSF) approach and a RELIEF-based modality weighting algorithm (RELIEF-MM) are proposed to handle the 'what-to-fuse' problem. Third, the 'how-to-fuse' problem is studied, and a novel mining and graph based combination approach is proposed. The approach enables an effective combination of the modalities represented with bag-of-words models. Lastly, a non-linear extension on the linear weighted fusion approach is proposed, by handling both of the 'what-to-fuse' and 'how-to-fuse' problems together. We have conducted comprehensive experiments on CalTech101, TRECVID 2007, 2008, 2011 and CCV datasets with various multi-feature and multimodal settings; and

validate that our proposed algorithms are efficient, accurate and robust ways of dealing with the given challenges of multimodal information fusion.

Keywords: Multimodal fusion, multimedia information retrieval

ÖZ

ÇOĞULORTAM BİLGİ ERİŞİMİ İÇİN ÇOK KİPLİ BİLGİNİN BİRLEŞTİRİLMESİ

Yılmaz, Turgay
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Adnan Yazıcı

Eylül 2014, 237 sayfa

Çoğulortam verilerine etkili bir erişim, verideki mantıksal içerik üzerine bina edilir. Mantıksal içeriğin çıkarılması için, çoğulortam verisi dikkatlice analiz edilmeli ve bilgi verinin içerdiği tüm bilgi kullanılmalıdır. Çoğulortam veriler, içinde çok kipli bilgi barındıran karmaşık bir yapıya sahiptir. Verideki gürültü, herhangi bir tekil kipin genelgeçer bilgi içerememesi ve her kipin performans üst limiti sebebiyle, herhangi bir kipten sağlanacak bilgiye güvenmek mümkün değildir. Bu yüzden, bilgi erişimi işleminin performansını artırmak için çok kipli bilginin birleştirilmesi kullanışlı bir yöntem olarak ortaya çıkmaktadır. Fakat, bu yöntemle ilgili olarak iki temel zorluk bulunmaktadır; 'ne' ve 'nasıl' birleştirilmeli. Verilen bu zorluklar kapsamında, bu tezin katkıları dört başlık altında incelenebilir. İlk olarak, literatürdeki çalışmaları incelenerek ve genel bilgi birleştirme sistemlerinin tasarım kriterleri saptanarak genel bir birleştirme çerçeveleri ortaya konmuştur. İkinci olarak, 'ne' birleştirilmeli problemini çözmek amacıyla, sınıfa özgü öznitelik seçim (CSF) yöntemi ve RELIEF-tabanlı bir kip ağırlıklandırma algoritması (RELIEF-MM) önerilmiştir. Üçüncü olarak, 'nasıl' birleştirilmeli problemi ele alınıp, madencilik ve çizge tabanlı yeni bir yöntem önerilmiştir. Bu yöntem kelime torbaları modeliyle temsil edilen kiplerin etkili bir şekilde birleştirilmesini sağlamaktadır. Son olarak, bahsedilen iki problem birlikte ele alınarak, doğrusal ağırlıklandırmalı birleştirme üzerine, doğrusal olmayan bir ilave yapılmıştır. CalTech101, TRECVID 2007, 2008, 2011 and CCV veri kümelerinde çeşitli çok öznitelikli ve çok kipli ayarlar ile kapsamlı deneyler yapılmış, ve önerilen

algoritmaların belirtilen problemlerin çözümünde verimli, etkin ve sağlam yöntemler olduğu ortaya konmuştur.

Anahtar Kelimeler: Çok kipli birleştirme, çoğulortam bilgi erişimi

To my wife

ACKNOWLEDGMENTS

The writing of the acknowledgments is something makes one feel as getting closer to finishing. It was really a long journey, yet enjoyable! It took 6.5 years, and covered the experiences of two countries, two universities and three companies. Therefore, the completion of this thesis has been possible with the help of many people, to whom I would like to express my heartfelt thanks.

First of all, I would like to express my inmost gratitude to my supervisor Prof. Dr. Adnan Yazıcı for his guidance, insight throughout the research and trust on me. It is an honor for me to share his knowledge and wisdom. I'm also very thankful to him for giving me the freedom to explore different research directions, and his support and guidance in non-academical issues. This certainly has given me a great experience, a broad perspective, and presented opportunities I would not have otherwise enjoyed.

I am also grateful to my thesis monitoring committee members, Prof. Dr. Özgür Ulusoy and Assist. Prof. Sinan Kalkan, for their excellent feedback, questions and suggestions which had an important influence on this thesis to be shaped into its current form. I would also like to thank other jury members that gave valuable comments on my dissertation. They are Prof. Dr. Ahmet Coşar, Assoc. Prof. Pınar Karagöz and Assist. Prof. İsmail Sengör Altıngövde.

An important portion of this dissertation has been completed in Tokyo. I am very thankful to Prof. Dr. 喜連川 優 and the members of his lab (especially Rage Uday Kiran, 横山 大作, 高 文梁, 鍛治 伸裕, 吉永 直樹 and 豊田 正史), from the University of Tokyo, for their help and hospitality during my studies as a visiting researcher there. It was an honour and rewarding experience for me to be a part of Kitsuregawa Lab. Having been in Tokyo for almost two years, I often got asked how I could handle the culture of Japan. It would be really hard, without the assistance and friendship of Ali Cevahir. In addition, the assistance of İsmail, Erkan, Hasan, and others are very much appreciated.

Being a member of Multimedia Database Research Group at METU, I would also thank all members of our research group. Although we usually do not share the same working environment, and meet frequently, we managed to publish several research papers with Elvan Gülen, Dr. Yakup Yıldırım, Assoc. Prof. Murat Koyuncu and Utku Demir.

I've worked in three different companies through my dissertation period; Havelsan, Rakuten Institute of Technology and Türksat. I'm thankful to my managers and colleagues for their support and toleration during my research. Among them, Hüseyin Özgür Tan has an important place, and deserves special thanks and gratitudes for his support and priceless friendship.

Throughout my research period, I've conducted some joint studies with Assoc. Prof. Kemal Akkaya, Assist. Prof. Fatih Şenel and Dr. Hakan Öztarak, in a different research domain from mine. I'm thankful to them for the experience and perspective they gave me. It was really rewarding and pleasant for me to work with them.

I'm also thankful to TÜBİTAK, which has financially supported my studies, in the scope of the projects 106E012 and 109E014.

The love, self-sacrifice and patience of a family is great to have. I would like to express my sincerest thanks to my parents, my younger brother Erdoğan, his wife Fadime, my parents-in-law and Burak. They never had an idea about what this dissertation is about, usually interested in when I will complete the thesis instead of how I do it, and could not fully understand what took so long. Yet, the endless encouragement and support they provided has always been a fantastic help and is very much appreciated.

Finally I would like to give my inmost gratitude and thanks to my wife Eda, who certainly has the biggest contribution by taking me to Tokyo and helping me to focus exclusively on my thesis. Not only the Tokyo part, but the whole period of my dissertation proceeded with her great support, understanding and love. Though she had to be a Ph.D. widow for a long time, and we could not spend our spare times as we want, I know that she always has the understanding and respect on me. So, this thesis is dedicated to her, with great love and gratitude.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

AD	Average Distance
AHP	Analytical Hierarchy Process
ANP	Analytical Network Process
AP	Average precision
ARM	Association Rule Mining
ASR	Automatic Speech Recognition
AVG	Simple Averaging
AVRORA	AVR Simulation and Analysis Framework
BoW	Bag-of-Words
BR	Binary Relevance
CalTech101	California Institute of Technology, Dataset for 101 Categories
CBIR	Content based information / image retrieval
CCV	Columbia Consumer Video Database
CFS	Correlation Based Feature Selection
CL	Color Layout
CMOS	Complementary metal–oxide–semiconductor
COTS	Commercial off-the-shelf
CS	Color Structure
CSF	Class-specific feature selection
CSh	Contour Shape
DC	Dominant Color
DoG	Difference of Gaussian
ED	Euclidean Distance
EDH	Edge Direction Histogram
EH	Edge Histogram
ES	Exhaustive Search
Exh-CC	Class-common Exhaustive Search
Exh-CS	Class-specific Exhaustive Search

FG	Fusion Gain
FIM	Frequent Itemset Mining
FLH	Frequent Local Histograms
GA	Genetic Algorithms
GCM	Grid-based Color Moment
GT	Gabor Texture
GUI	Graphical User Interface
GWT	Grid-based Wavelet Texture
HT	Homogeneous Texture
ICA	Independent Component Analysis
IMG	Independent Modality Grouping
IPO	Input-Process-Output
JDL/DFG	Data Fusion Group in Joint Directors of Laboratories
k-NN	k nearest neighbor
LoG	Laplacian of Gaussian
LP	Label Power Set
LPC	Linear Predictor Coefficients
LSA	Latent Semantic Analysis
LWA	Linear Weighted Averaging
MAP	Mean average precision
MAX	Maximum Selection
MD	Minimum Distance
MFCC	Mel-frequencies cepstrum coefficients
MIN	Minimum Selection
MLE	Mixture of Local Experts
MPEG	Moving Picture Experts Group
MST	Minimum Spanning Tree
MT	Machine Translation
NB	Naïve Bayes
NIR	Near Infra Red
NWA	Non-linear Weighted Averaging
NWA-BCl	Best Class Accuracy Based NWA
NWA-Bco	Best Common Accuracy Based NWA
NWA-CB	Convergence Based NWA

OCR	Optical Character Recognition
PCA	Principle Component Analysis
PPT	Pruned Problem Transformation
RA	Random Assignment
RBF	Radial Basis Function
RELIEF	(Not an abbreviation, the name of the algorithm)
RELIEF-MM	RELIEF for Multimedia Data
RGB	Red Green Blue
RS	Region Shape
SBE	Sequential Backward Elimination
SC	Scalable Color
SFS	Sequential Forward Selection
SIFT	Scale-invariant feature transform
STIP	Spatio-Temporal Image Features
SURF	Speeded Up Robust Features
SVM	Support Vector Machines
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TREC	Text Retrieval Conference
TRECVID	TREC Video Retrieval Evaluation
WVSN	Wireless Visual / Video Sensor Networks
XM	eXperimantation Model Software
ZCR	Zero Crossing Rate
ZCRE	Zero Crossing Rate and Energy

CHAPTER 1

INTRODUCTION

" The hand of one fell on the trunk: he said 'This creature is like a water-pipe'.
The hand of another touched its ear: to him it appeared to be like a fan.
Since another handled its leg, he said, 'I found the elephant's shape to be like a pillar'.
Another laid his hand on its back: he said 'Truly, this elephant was like a throne' "

The elephant in the dark[1]

Increase in the use of digital images and videos in recent years has shown the need for modeling and querying the multimedia data. Free-browsing and text-based retrieval of previously annotated data are not enough due to the limitations for querying. Therefore, developing techniques for the retrieval of multimedia data based on the semantic content has attracted many researchers [73, 118]. However, the gap between the low level features and the semantic content of the multimedia data makes the semantic content extraction issue a challenging problem, and impedes achieving consistently high retrieval accuracies in many potential "real-world" applications. In order to handle such impediment, one or more of the routes below may be followed [24]:

- Developing superior content extraction methods than the currently available ones
- Optimization of components (preprocessing, feature extraction, classification, etc.) of currently available methods
- Fusion of multiple classifiers, features/modalities or information sources

[1] A poem from Rumi (Masnavî 3:1267-1270). Similar stories are also famous in various religious traditions like Sufi, Buddhist, Hindu and Jain lore.

Development of new extraction methods and making optimization on the available ones are traditional ways of dealing with the given problem. The usual solution approach in such studies includes an experimental assessment of alternative designs (by means of the utilized features, classifiers, etc.), and building the solution on the best design alternative. However, different design alternatives can provide complementary information about the patterns to be extracted [61]. Thus, fusing multiple information sources, classifiers or features is a popular approach for semantic content extraction, in the last few decades.

In addition to the route followed for semantic content extraction, the content of the multimedia data is another crucial aspect. In order to extract the semantic content from the multimedia data more effectively, the nature of the data must be examined carefully and information contained in the data should be used as completely as possible. The multimedia data usually has a complex structure containing multimodal information (i.e. audio, visual and/or textual modalities). In the context of semantic content extraction, using a single modality may not be enough to obtain a successful retrieval solution, because of the potential noise in sensed data, non-universality of any single modality and the performance upper bound of each modality [98]. In addition, each modality abstracts multimedia data from a different aspect. Thus, different modalities complement each other [51]. Eventually, fusing multimodal information in multimedia data improves the retrieval performance.

1.1 The Problem

As stated above, information fusion is an effective way of improving the retrieval performance. However, combining a set of modalities, features, classifiers or information sources includes several difficulties. In order to understand such difficulties clearly, we can make an analogy with a committee of experts [30]. Assume a committee of experts trying to give decisions on the issues they asked. Considering that the experts in the committee may have different backgrounds and expertise areas, each of them could give different decisions. Then, how does such a committee arrive at a final decision? What is the decision-making process of the committee? How can the final decision of the committee be arrived? Is voting a good way or does it neglect the

experts' differences in skills? Should every expert give a vote on every subject or should the authorities of experts be limited according to their expertise area and skills? In addition, is there a way to decide whether any of the experts is faking or really an expert? Should we believe every expert without any questioning? Assume we found some ways to solve above problems. Should we always trust these solutions? Or should we keep in mind that any of the experts can change his mindset, or some experts can have progress on their expertise while some others lose their abilities and expertises in time?

The analogy makes us distinguish the mostly known, thereof mostly studied, problems of the fusion systems. The analogy represents two major problems; (i) selecting appropriate committee members and determining their effect on the final decision, based on the given problem, (ii) finding an appropriate mechanism to combine the decisions of the selected committee members. Actually, these are the problems pointed out by many researchers in the information fusion domain. The major two problems of information fusion, which have not been adequately addressed yet and are still attractive research areas [4, 98, 143] are:

- *What to fuse*: Determining the best components for fusion,

- *How to fuse*: Finding the best fusion methods.

In addition to the above given problems, another important issue is the lack of a well defined general framework for information fusion [63, 64, 143]. The literature on information fusion shows that several disparate research areas utilize information fusion [63] and a big number of studies from these research areas try to find some optimum solutions to the above given problems [70]. Nevertheless, almost all of the solution are ad-hoc strategies [2] and does not present a general framework which defines all factors affecting the fusion process [47]. Building a general framework for information fusion helps us see the big picture of the whole fusion process and identify which variables are effective during fusion.

1.2 Scope & Contributions

The scope of this dissertation includes the construction of a general fusion framework by performing a literature survey, and the above given two core issues of information fusion in the context of multimedia information retrieval. For the *What to fuse* problem, firstly, a class-specific feature/modality selection approach is proposed. Then, this approach is extended into RELIEF-based feature/modality weighting algorithm. For the *How to Fuse* problem, a novel mining and graph based combination approach which enables an effective combination of the modalities represented with Bag-Of-Words models is proposed. In addition, a non-linear weighted averaging approach, which attacks the *What to Fuse* and *How to Fuse* problems together, is proposed. Below, each of these work items are described in brief.

The thesis study is started with a literature survey, primarily analyzing the information fusion literature and identifying the design aspects of a general information fusion system. The analysis enabled us to propose a general framework which helps to represent a big picture for information fusion systems. In the framework, each design aspect is accepted as an affecting variable for the *Fusion Process*. In accordance with the problem definition, the process is composed of two primary tasks; *Defining What to Fuse* and *Defining How to Fuse*. In addition, another task named *Defining Fusion Scenario* takes place before the *Fusion Process*. *Defining Fusion Scenario* task is based on the inputs of fusion and helps to define overall architecture of the *Fusion Process*. Besides, the task of *Defining What to Fuse* requires an effective and efficient selection of the fusion elements and is configured with following parameters: *Selection of Sources* and *Fusion Strategy*. After performing this task, *Selected Fusion Elements* are obtained, which are related with the variables *Content Representation* and *Normalization of Sources*. Finally, *Defining How to Fuse* task is performed, which defines how we combine the selected elements. The task is configured with the following variables: *Fusion Level*, *Fusion Method*, *Operation Modes* and *Synchronization*.

After identifying the design aspects of fusion systems, two core problems of fusion are focused. As a first step, the *What to Fuse* and *How to Fuse* problems are considered together, and an effective fusion architecture is investigated by regarding the most frequently utilized approaches in the literature. The most frequently utilized approach

is the Linear Weighted Fusion [37, 133, 145], due to its simplicity and reasonable performance despite its simplicity. However, it suffers from the performance upper-bound of linearity and dependency on the selection of weights. In this context, the study is focused on two questions, considering two important deficiencies of Linear Weighted Fusion: (i) Can we find a method which is as simple as the linear weighted fusion and can exceed the performance upper bound of linear weighted fusion? (ii) Can we find a method which is less-dependent on the selection of the weights? Aligned to these aspects, a 'simple' alternative for linear combination is introduced, which is a non-linear extension on it. The approach is based on the Analytical Network Process [109], which is a popular approach in Operational Research, but never applied to multimodal information fusion before. The approach benefits from two major ideas; interdependency between classes and dependency of classes on the features. The proposed method is evaluated with Columbia Consumer Video (CCV) Database by using multimodal features of SIFT, MFCC and STIP. Experiments demonstrate that proposed approach outperforms linear combination and other simple approaches, moreover it is less-dependent on the selection of weights.

Secondly, the problem of *What to fuse* is studied in the context of multimodal information fusion. As a contribution, a class-specific feature/modality selection (CSF) approach for the fusion of multiple features/modalities is proposed. In order to eliminate the high-dimensionality of multiple features and provide efficient querying over the multimedia documents, a dissimilarity based approach is used. The class-specific features are captured through a training phase, in which the class-specific features are determined by using the representativeness and discriminativeness of features for each class / concept. The calculations of representativeness and discriminativeness are based on the statistics on the dissimilarity values of the training data. The proposed approach is firstly evaluated in a multi-feature setting by using the CalTech 101 dataset with 8 MPEG-7 visual features and compared with the retrieval performance of single features, simple combination approaches and exhaustive search approach. Then several experiments in a multimodal environment are conducted by using TRECVID 2007 dataset with 3 visual, 2 audio and 1 textual modalities. Lastly, the proposed approach is utilized for efficient feature selection and combination in a Wireless Video Sensor Networks application [94]. The results obtained from all three test configurations show

that proposed class-specific feature selection approach is an effective and efficient feature selection method.

Thirdly, the proposed CSF approach is extended and converted into a RELIEF algorithm extension, due to the similarities between CSF and RELIEF. The RELIEF algorithm is considered one of the most successful weighting algorithms [105] and in which the calculations are based on the distances between training samples. Yet, there exists no usage of the RELIEF algorithm for multimodal feature selection in multimedia retrieval, to the best of our knowledge. Employing the RELIEF algorithm for multimodal feature selection on multimedia data enables us to identify some weaknesses of the algorithm in the following major issues: class-specific feature selection, complexities with multi-labeled data and noise, handling unbalanced datasets, and using the algorithm with classifier predictions. Considering the characteristics of multimedia data and multimedia retrieval systems, the original RELIEF algorithm is extended for the given issues, and the *RELIEF for multimedia data* (RELIEF-MM) algorithm is proposed. RELIEF-MM employs an improved weight estimation function, which exploits the representation and reliability capabilities of modalities, as well as the discrimination capability, without any increase in the computational complexity. The comprehensive experiments conducted on TRECVID 2007, TRECVID 2008 and CCV datasets validate RELIEF-MM as a timely-efficient, accurate and robust way of modality weighting for multimedia data.

Lastly, but not least, the problem of *How to fuse* is studied, in order to propose an effective combination approach. The most popular and effective methods in multimedia analysis studies in the last decade are based on the use of local parts / features in multimedia documents and employing Bag-of-Words (BoW) approaches. Thus, the last part of the thesis is focused on combining the Bags of Words obtained from different modalities. Most of the currently available studies focus on combining the BoWs with early or late fusion schemes [50, 52, 81, 122]. However, most of the studies do not use intramodal and intermodal relations effectively [10]. In order to combine all available information provided by any single modality, correlations within a modality and correlations between different modalities; we propose a novel mining and graph based combination approach. In order to combine all available information effectively, the classification outputs of each single modality, intramodal

process and intermodal process are combined with a late fusion approach. For the late fusion, a linear weighted averaging approach is utilized with the weights generated by using RELIEF-MM algorithm. Throughout the intramodal process, the *word*s of each modality and the correlation between these *words* are converted into a graph representation, and then the meaningful *phrase*s are extracted by using these *word*s. In order to extract the *phrase*s, the most together occurring k number of words are extracted from the constructed *word* graph. The intermodal process is similar with the intramodal process. Differently, in intermodal process, the correlation between the extracted phrased of different modalities are calculated and converted into a graph representation. Then, the *multimodal phrase*s are extracted from the graph. Both of these processes end up with using the extracted *phrase*s for classification. Experiments conducted on TRECVID 2011 dataset with visual, audio and text modalities provide promising results.

1.3 Organization

In Chapter 2, an introduction to the basic concepts used in this dissertation is given. The chapter first gives a brief definition of pattern recognition and classification. Then, the concept of information fusion is described with a brief definition, history, reasons for fusion and expectations from fusion. Lastly, the multimodality concept is defined.

Chapter 3 presents the literature survey on information fusion and proposes a general framework representing the big picture for designing an information fusion system.

In Chapter 4, the Non-Linear Weighted Averaging based fusion approach, which is based on Analytical Network Process, is introduced. The chapter includes a brief description on Analytical Network Process, related work, the description of the proposed approach, experiments and the evaluation.

In Chapter 5, the class-specific feature selection approach is described. The chapter first describes multi-feature modeling in dissimilarity space. Then the class-specific feature selection approach is given in detail. After the description of the approach, experiments with multi-feature and multimodal settings are presented. In addition, the utilization of the approach in a Wireless Video Sensor Networks application is given.

In Chapter 6, the RELIEF-MM algorithm for modality weighting is presented. The chapter first given a detailed related work, as well as the description of the original RELIEF family algorithms. Then, the proposed algorithm is given in detail. Lastly, the experiments conducted on TRECVID 2007, TRECVID 2008 and CCV datasets are presented with evaluations.

Chapter 7 presents the mining and graph based fusion approach for combining the Bags of Words obtained from different modalities. The chapter includes gives a brief definition on the approaches frequently utilized by the state-of-the-art studies. Then, after a discussion on alternative approaches for combining the Bags of Words, the proposed approach is described and the experiments conducted are presented.

In Chapter 8, a demo application for multimedia information retrieval is presented. Throughout the chapter, first the need for such an application is discussed, and the application is presented with several screenshots. Lastly, an evaluation on the demo application is given.

Lastly, Chapter 9 provides a broad summary and conclusion on the dissertation, as well as the future work.

CHAPTER 2

BACKGROUND INFORMATION

In this chapter, the fundamental concepts about the study are presented. Considering the thesis title, the study includes issues on Pattern Recognition, Classification, Data/Information Fusion (with a Pattern Recognition point of view) and multimodality. Corresponding sections are presented in this chapter. The major goal of this chapter is to give the readers some brief information about the building stones of the proposed study.

2.1 Pattern Recognition and Classification

In machine learning, pattern recognition is defined as;

> *"The assignment of some sort of output value (or label) to a given input value (or instance), according to some specific algorithm."* [140]

Classification (or categorization) is a subset of pattern recognition that attempts to assign input values to some predefined pattern classes. A pattern class is a collection of similar (not necessarily identical) objects. Classes are defined by using class samples, which are any of training/learning samples, prototypes or paradigms.

More formally, classification can be defined as assigning an input s_i to a class c_j by approximating a function $\phi^{'} : S \times C \rightarrow \{T, F\}$ by maximizing the coincidence of $\phi^{'}$ with the actual classification ϕ, where $S = s_1, s_2, .., s_m$ is the set of inputs, $C = c_1, c_2, .., c_n$ is the set of classes and $\{T, F\}$ are boolean values that defines whether the classification is true or false [114].

A classifier (or learner) is defined as the algorithm performing the classification. More concretely, the function implementing the classification algorithm can be called as a classifier. Classifiers accept instances of classes as input and tries to determine the correct class of it. During classification process, instances are represented with features. Features can have different types of values; categorical/nominal (i.e. "male" or "female"), ordinal (i.e. an ordered set: "large", "medium" or "small"), integer-valued, real-valued, etc. [138]. As the output, a simple classifier can give the label of the corresponding class. More complicated classifiers can give score values for the classification results. Furthermore, it is possible to return a ranked list of probable classes according to calculated score values.

Classification is a supervised (learned) procedure, that means a classifier learns classifying inputs by using a training set of class instances. Besides, there exists another subset of pattern recognition which is an unsupervised procedure, called as Clustering. Since clustering enables only grouping instances according to their similarities and cannot give labels to these groups; it is out of our scope in this study.

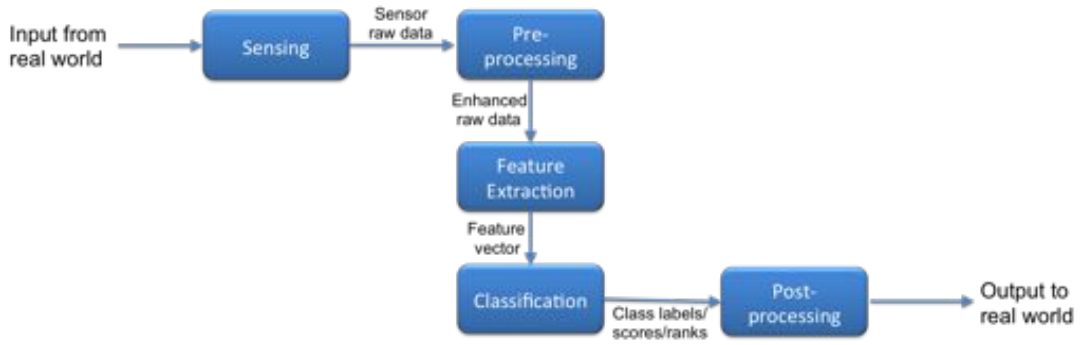


Figure 2.1: A Typical Classification Process

A typical pattern recognition process is summarized in Figure 2.1. Considering the figure, the process starts with the perception of some input from the real world via some hardware called sensors. A sensor converts physical inputs (i.e. sounds or images) into signal data. Then, this raw data of sensors are preprocessed. The preprocessing step can include enhancement and segmentation operations that makes the raw data more easily processable and removes unnecessary parts of it. After preprocessing, a feature extraction step is employed and several important properties of the real world input that are useful for classification are extracted by using the sensor data. Afterwards,

features are used for classification. The classification results with a class label, a score or a ranked list. Lastly it is possible to have an enhancing post-processing mechanism on the results.

2.2 Data/Information Fusion

Data/information fusion is the process of combining data/information from multiple sources in order to infer new results that may not be resulted by any of the single sources or obtain more efficient and accurate results than any of the single sources. A typical fusion system is illustrated in Figure 2.2.

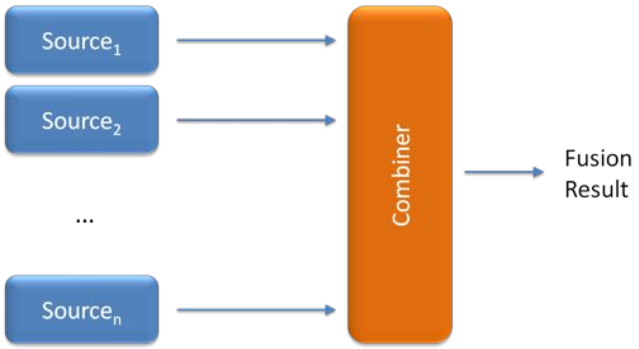


Figure 2.2: A Typical Fusion System

In a fusion system like in Figure 2.2, the 'source' can be any of the followings: sensor, feature, modality, classifier, information resource (dataset, library, different real world situations etc).[1] [2] Explanations for combining each source type are given below, with corresponding examples:

- Sensors: The fusion system combines outputs of multiple sensors. For instance, combining outputs of a RGB and a NIR camera.

[1] Meanwhile, it should be noted that the fusion scope of pattern recognition or multimedia retrieval systems do not involve the fusion studies in sensor level, in principle. Studies in these areas deal with the features of the objects as the lowest level, not with the signal inputs or raw sensor data. Thus, no details will be given for sensor fusion although some introductory information is presented.

[2] Since the scope of this study is "Fusion of Multimodal Information in Multimedia Information Retrieval", the thesis usually refers fusion process as combining modalities or features. Considering the fact that the source for a fusion system can be any of sensor, feature, classifier or information resource; these phrases can also be interchangeably used with the phrase "source". Also, any of these source types is applicable for a generic fusion structure.

- Features: The fusion system combines instances of different features. For instance, combining color and shape features to recognize an object.

- Modalities: The fusion system combines data of different modalities. This case is not much different from multiple features case, considering that modalities consist of several related features. An example for multimodality is combining face and hand-shape modalities in a biometrics system.

- Classifiers: The fusion system combines multiple different classifiers. Creating different classifiers can be don by several ways. In [30] Duin et al. presents methods of generating different classifiers. Some of them include using different classification algorithms, having different algorithm-parameter choices, performing different initializations. Examples for these cases, respectively, are as follows:

  - combining results of a Bayesian classifier and a decision tree for the same inputs
  - combining k-NN classifiers with different number of neighbors
  - combining differently-initialized neural network classifiers

- Information resources: The fusion system can combine information of different datasets/instances/situations. In other words, the system can combine multiple outputs (in time-based manner) of a single sensor, different instances of a single feature, classifiers trained with different datasets (classification algorithm, parameters, initializations, etc. are the same) or different instances of single modality. Corresponding examples are:

  - combining satellite images of an area that are taken at different times
  - combining shape features of a person such that the shapes are from different perspectives to obtain a more robust shape recognition
  - combining two decision-tree classifiers that are trained over different datasets
  - combining two face recognizers of a biometrics system where one of them is trained in good light conditions, the other one is trained for dark situations

A more imposing definition is presented by JDL/DFG[3] [63, 74]:

> *"Information fusion is an Information Process dealing with* ***the association, correlation, and combination of data and information*** *from* ***single and multiple sensors or sources*** *to achieve* ***refined estimates of parameters, characteristics, events, and behaviors*** *for observed entities in an observed field of view"*(Figure 2.3)

The JDL/DFG introduced a model of data fusion that identifies levels of fusion processing, types of fusion functions, and candidate algorithms for performing fusion. This model and related techniques have been applied to several non-military applications such as environmental modeling, control of complex systems and medical applications, as well as their primary scope of military domain [74]. Nevertheless, the detailed studies of the JDL/DFG group is beyond our scope, considering that the group studies on military-based domain and consequently their aspect of fusion is mostly on the sensor fusion.

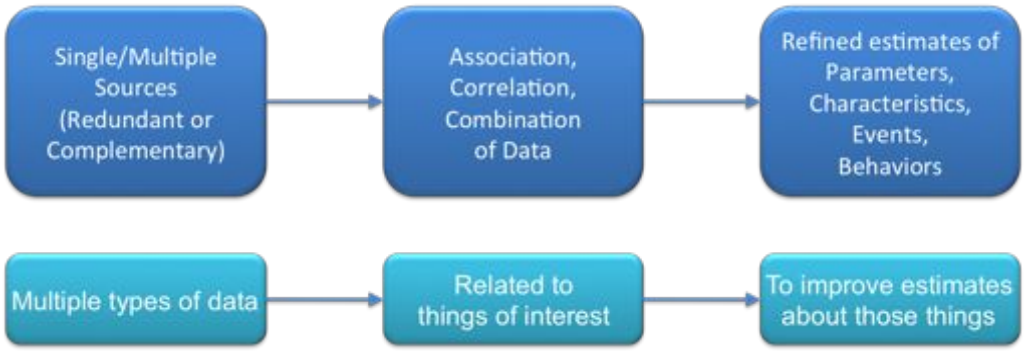


Figure 2.3: JDL/DFG Definition of Fusion

Information fusion is utilized in a vast number of research areas, some of which are pattern recognition, information retrieval systems, geospatial information systems, cheminformatics, bioinformatics, wireless sensor networks, biometrics systems. In different research areas information fusion has different meanings [20]. In applied sciences, engineering and military applications, information fusion is mostly identified with sensor fusion. Studies aim to combine data of multiple sensors or multiple data instances of a single sensor. In pattern recognition and machine learning, information

[3] JDL/DFG is the Data Fusion Group in Joint Directors of Laboratories. JDL/DFG is established in 1984, under U.S. Department of Defense.

fusion is mostly explained with combining classifiers to increase classification accuracy or handling classifier ensembles according to their outputs by resampling the input. The distinction between the research areas originates from what they work on. The areas working on input signals of systems prefer sensor fusion, whereas the working on features and feature-based recognition or retrieval choose classifier fusion. However, as mentioned in the Section 2.2.1, recent studies tend to relax such distinction and extend their working area to more generic information fusion.

One more remarkable issue on the information fusion is the diversity of studies even in classifier combination, the pattern recognition point of view on information fusion. Due to the diversity of studies, there are many different names in the literature [69]: combination of multiple classifiers, classifier fusion, mixture of experts, committees of neural networks, consensus aggregation, voting pool of classifiers, dynamic classifier selection, composite classifier system, classifier ensembles, divide-and-conquer classifiers, pandemonium system of reflective agents, change-glasses approach to classifier selection.

2.2.1 A Brief History

In the literature of pattern recognition, previously, main effort focused on designing one good classifier. Then it is argued that building a number of classifiers with low dimensionality and high performance, and combining them could achieve more successful results. Thus, information fusion studies has begun [144]. The roots of the fusion studies can be found in the neural network literature, as early as 1960's [63, 64].

Early approaches, until the beginning of 2000's, aimed to combine results of multiple classifiers [61]. They did not consider multimodality [63], moreover many of them studied combining multiple classifiers on only single feature [61]. Later on, several studies found out that using a different classifiers for different feature gives better results [31, 61, 67].

After 2000, the area of information fusion has become more attractive, even a conference series named "International Workshop on Multiple Classifier Systems" started in 2000.

In the contemporary approaches, the sense of fusion began to extend from combining classifiers to combining information, where information can be any of features, modalities, classifiers or data sources. Correspondingly, the importance of multimodality issue has increased [98, 123]. In addition, studies on dependency/correlation of sources has begun to increase in this decade, whereas most of the studies had preferred independent sources in the earlier years [62, 64].

2.2.2 Reasons for Fusion

Information fusion primarily aims at having more efficient and accurate result by combining currently available (multiple) systems/components. Some of the reasons for not relying on a single source are given below:

- Fusion of complementary sources provides a more complete representation of the world. Fusion of redundant (cooperative or competitive) sources reduces the uncertainty and increases the robustness [63].
- A practical benefit of fusion is that it lowers unreliable sources. It cannot be known during design time how each feature, modality or data source performs in real world environments and which of them are the most reliable. So, by fusion, dependency on any of the sources can be decreased [63].
- Noise in the sensed data causes inefficiencies in recognition. Having multiple sources can decrease the effect of noise. For instance; consider person recognition. The same face under changing lighting condition appear more differently, then different faces can be captured. Designing multimodal system not based on only face traits can resolve the problem [46, 98].
- None of the sources is universal for the recognition problem; each of them have a usage area. While no single source is perfect, a combination of them should ensure wider coverage of usage area, hence improving accessibility. For instance, consider person recognition, again. An iris recognition system may be unable to obtain the iris information of a person with long eyelashes, dropping eyelids or certain pathological conditions of the eye. Thus, designing a multimodal system not only based on iris recognition can increase the usability [46, 98].

- Each single system or source has an upper bound on system performance. The recognition performance of any single system cannot be continuously improved by tuning the feature extraction, classifier, or some other steps. There is an implicit upper bound on the number of distinguishable patterns by using a determined methodology and features. Thus, no single source or methodology is totally perfect. Integrating them can give better results [46, 98, 144].

2.2.3 Expectations

There are several expectations from a fusion system to be more efficient, effective and practicable.

- Fusion should increases robustness and performance of classification [20, 63, 128].
- A correct fusion system should be at least as effective as any of its parts [63].
- A fusion system should be flexible (it should handle any new sources blindly) [15].
- A fusion system should be fast (online learning should be possible) [15].

2.3 Multimodality

The meaning of the word "modality" in our usages comes from the domain semiotics. In semiotics, the definition of modality is as follows:

> *"A modality is a particular way in which the information is to be encoded for presentation. It refers to a certain type of information and/or the representation format in which information is stored."* [139]

According to the definition, modality is a vague concept that can be concretized in different ways. Considering our domain of pattern recognition and information retrieval, the building stone to represent the objects is the features. At one extreme, each of the features can be treated as separate modalities. At the other extreme, all of the features can be treated as one modality [143]. A mid-point can be grouping the

features according to some criteria. For instance it can be regarded that the extracted features of a video object can be grouped according to their media-source: formulating visual, audio, and caption modalities with related features. However, each of these modalities can be expanded. For instance; even visual data can be defined with several modalities like color, shape, texture, face, etc.

The multimedia community employs one the above given approaches for the selection of modalities. But there is no absolute evidence on which of these feature compositions yield the optimal result [143].

In the literature of fusion, the multimodality issue mostly occurs in the multimedia and biometrics domains. The multimedia domain needs multimodality to express the complex structure of multimedia data including content from different media-sources. The biometrics domain tries to recognize humans based upon their physical and behavioral traits. Since recognition with only a single trait is not enough, the studies regard each trait as a different modality for fusion.

In this study, in principle, each different media source (i.e. visual, audio and text) is taken as a different modality. We also consider that different features from the same media source, but containing a significant amount of complementary information, should be regarded as different modalities. For instance 'motion' related features are extracted from the visual part of the video, however, we accept motion as a different modality. In addition, the 'color', 'shape', 'texture' features extracted from the visual media source abstract the information from different aspect, thus we accept these features as different modalities, wherever necessary. In brief, the criteria of being a modality for this study is to have a different aspect of representing the data and a significant amount of complementary information with other modalities.

CHAPTER 3

THE BIG PICTURE & LITERATURE SURVEY

In this chapter, a literature survey on the information fusion systems is presented. The survey identifies the design aspects of a general information fusion system. In addition, a general framework which helps to represent a big picture for information fusion systems, is proposed. The chapter gives detailed descriptions of the affecting variables of fusion systems and references to the state-of-the-art studies.

3.1 General Framework for Fusion

There is a vast number of studies utilizing information fusion, but each individual study describes the fusion method in its context of theory. A detailed literature survey can provide insight to understand what factors have an effect on the success of the fusion. However, a general framework that combines all variables having effect on the fusion results into a unified view is still missing. In addition, traditional work on multimodal fusion is mostly heuristic-based and ad-hoc solutions. Studies usually solve the problem empirically and then justify the solution theoretically. Hence, construction of a general framework is a crucial contribution for the information fusion literature. In this section, a formal representation of the general framework for fusion is introduced with the variables / factors affecting the success of the fusion system.

Our proposed framework is illustrated in Figure 3.1. The framework is presented with an Input-Process-Output (IPO) model, where Fusion Input is converted to Fusion Output through a Fusion Process. Fusion Input is multiple of any source like sensors, features, classifiers or information resources. Fusion Output is the combined informa-

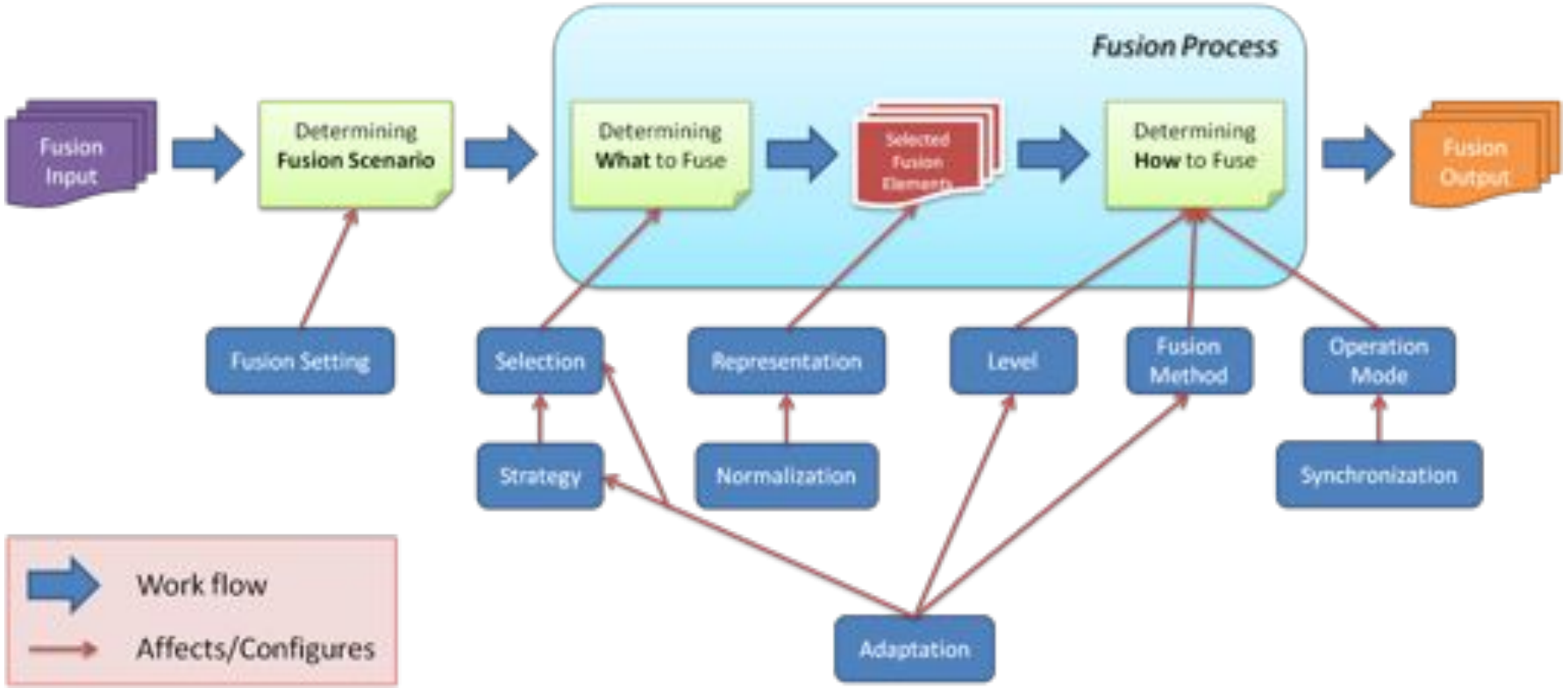


Figure 3.1: General Framework for Fusion

tion. Fusion Process is the core of the architecture that handles the combination of sources. The process is composed of two primary tasks; 'Defining What to Fuse' and 'Defining How to Fuse'. In addition, another task named 'Defining Fusion Scenario' takes place before the Fusion Process.

'Defining Fusion Scenario' task bases on the Fusion Inputs and helps to define overall architecture of the Fusion Process. The task includes evaluating the sources that are provided as Fusion Input; and constituting a fusion architecture that defines multiples of which source types are handled during the Fusion Process. For instance, considering a multi-feature case, we want to combine 3 different types of features. Here, we can consider both feature and classifier source types. Thus it is possible to create the following scenarios:

- Scenario 1: (Multi-feature, single-common-classifier) Create the same type of classifier for each feature separately, process each feature in the classifier, combine results of the classifiers
- Scenario 2: (Multi-feature, single-unique-classifier) Create a separate (in different type) classifier for each feature, process the feature in the classifier, combine results of the classifiers
- Scenario 3: (Multi-feature, multi-classifier) Create some defined number of classifiers for each feature, process the feature in each of the classifiers, combine results of all classifiers

The number of possible scenarios depends on the source types provided as the Fusion Input. The 'Fusion Setting' variable defines which scenario is applied for the Fusion Process.

After defining a fusion scenario, two primary problems of information fusion should be handled; selecting what to fuse and finding out how to fuse. 'Defining What to Fuse' task requires an effective and efficient selection of fusion elements. The 'Selection of Sources' and 'Fusion Strategy' variables define how to select them. After performing this task, 'Selected Fusion Elements' are obtained. There are two important variables affecting the 'Selected Fusion Elements' for the flexibility and the processing time of the overall system: 'Content Representation' and 'Normalization of Sources'. The

second task, 'Defining How to Fuse', requires constituting a fusion approach in order to obtain maximum gain from the selected fusion elements. 'Fusion Level', 'Fusion Method', 'Operation Modes' and 'Synchronization' variables define the required approach with the several crucial aspects.

3.1.1 Fusion Setting

Fusion setting defines and determines which source types will be combined in the designed system. As mentioned in Section 2.2, the source types are; sensors, features, classifiers and information resources. The selection of each type can be either single or multiple. Using a single source (i.e. single sensor, single feature, single classifier, etc.) type means there will not be a fusion process for that source type. On the other hand, having multiple source types (i.e. multiple sensors, multiple features, multiple classifiers, etc.) means performing a fusion for that source type. It is clear that at least one of these source types should be multiple, in order to have a fusion system. Also, it is possible to have more than one configuration item as multiple but such a case increases the complexity of the system. Therefore, the studies in the literature usually set one of the configuration items as multiple and make others single for the fusion experiments [46, 106].

Generating all possible combinations with appropriate single and multiple selections gives all possible fusion scenarios. In addition, the number of scenarios can be increased by incorporating relations between these source types (i.e. having a common or unique classifier for each feature in multi-feature, single-classifier scenario).

One 'Fusion Setting' related issue in the literature is the discussion of "*Selection* or *Fusion*[1]" [67] that is analyzed in the section for Fusion Strategy (Section 3.1.3). *Selection* refers to a configuration that each classifier involved in the combination process is experienced on some local area of feature space. Besides, *Fusion* is a configuration which all the classifiers are equally experienced on the whole feature space. In a fusion setting aspect, *Selection* refers to multi feature setting and *Fusion* is a multi classifier setting. Many studies in the literature ([9,31,67,69,96]) has attended

[1] This definition of *"Fusion"* should not be compared, confused or supposed same with our general scope of fusion.

this discussion and experimental evidences showed that combining multiple features is more beneficial than combining multiple classifiers.

Independently from above discussion, in [31], Duin et al. compare a multi-feature setting (using a classifier for each feature) with a multi-classifier setting (different classification algorithms for the same features) and experimentally show that multi-feature setting gives better results than multi-classifier setting. Also, they conclude that a multi-feature and multi-classifier setting (different classification algorithms for different features) is much better.

3.1.2 Selection of Sources

'Selection of Sources' variable enables deciding which sources are the best and should be selected for fusion. Selection is a critical issue that directly affects the performance of the fusion system. The selection can be either a hard selection which picks some of the features and leaves others out, or a soft selection that determines effect of each source to the fusion (like weighting). There are several consideration points for selecting sources.

The most important consideration on the selection of the sources is the contribution of them to the fusion result. It is crucial to determine how much gain a source can provide or find out whether including a source affects positively or negatively. Thus, some evaluation methods are necessary to understand which sources are better. For instance, during sensor fusion, available sensors should be evaluated according to their noise level and cost of computation [63]. In [98], Poh et al. argue that quality and reliability of sources can help selection of sources. In addition, they give mathematical formulations of some example quality measures. However, they leave the reliability measure of sources as an open issue. In [120], Snidaro et al. present a quality metric for sensor selection, so that fusion process is dynamically regulated with the performance of the sensors. In addition they give a review of effective image quality evaluation methods. These methods can be beneficial for feature/modality selection. In [17], Callan et al. introduce an efficient way of resource selection and present some quality and reliability measures, which can also be useful for selection of sources.

Another important consideration point is the compliance of the sources with the designed fusion system, actually the selected fusion algorithm. The dependency/independency of sources should be carefully analyzed and compliant ones with the fusion method should be used. Not taking this issue into consideration can lead to dramatic decreases and inconsistencies in the performance of the system. Thus 'Fusion Strategy' variable has an important effect on the 'Selection of Sources' variable. This issue is analyzed in Section 3.1.3 in detail.

The 'Selection of Sources' variable contains three major attributes:

- Selection Type: Selection type can be either static or dynamic. In static selection, the selection is performed once during the construction of the system or during the training phase. In dynamic selection, sources are selected during the running of the system. A dynamic selection mostly refers to an 'Adaptation' capability. Thus, 'Adaptation' variable is an affecting variable on the 'Selection of Sources'.
- Context Relation: A selection procedure can be either context sensitive or insensitive. A context sensitive procedure perform selection depending on currently available conditions and information. For example; performing different selection schemes for different categories of objects or different video categories. Besides, a context insensitive procedure behaves equally for all conditions and always results with the same selection scheme.
- Selection Method / Metric: The method / metric defines selection procedure.

3.1.3 Fusion Strategy

In a fusion system, the sources, which are input for the combiner, can either be complementary or redundant sources. 'Fusion Strategy' of a fusion system determines how the system behaves the input, as complementary or redundant sources. Either case can be beneficial for the fusion process. Complementary sources reflect different sides of the problem domain like different feature spaces or uncorrelated data sources. Fusion of complementary sources provides a more complete representation of the world and resolves ambiguity and incompleteness. Besides, redundant sources can be cooperative or competitive, that provides data on the same side of the problem domain.

Fusion of redundant sources provides reduced uncertainty and increased robustness. Also it improves accuracy and reliability [63, 72]

A fusion system can prefer any of these two, or both of them at the same time. However an important restriction exists. Complementarity and redundancy are two contradictory cases for sources. More clearly, complementary sources refer to independent inputs, whereas redundant sources mean dependent inputs. Usually, the mathematical models do not function with both of these two models. Each method defines their input as either independent or dependent. Still, it is possible to use both of these two types of information in a complex system that employs more than one mathematical model.

In the literature, two main directions on strategy exist for fusion [64]:

- Fusion of independent(complementary) information: Two sub-types exist.
  - By assuming independency: This is the approach of early years' studies. Using methodologies just assume that the inputs are independent. Its success is based on the simplicity and some good luck. The approaches using this assumption are usually late fusion approaches, since it is not possible to find out dependencies of features/modalities. For this usage, it is obvious that violation of independence hurts the success of fusion.
  - By creating independency: Independency is obtained with the help of some independence analysis methods. The approaches that can apply this step are usually feature level fusion, considering that the dependencies can be obtained by analyzing features/modalities. Applying such a step before fusion, guarantees independence and enables more robust systems.
- Fusion of dependent(redundant) information: Using information of dependent sources can be obtained by exploiting statistical dependencies between features/-modalities. As mentioned, before feature level fusion is required for such a process.

Although several studies insist on using dependent sources and exploiting relationships between features / modalities [62–64, 98], with current evidences and experiments it is not possible to say any of these approaches is superior.

Complementarity and redundancy have more different names in the literature except the ones given above (independent and dependent). In the literature of "Multiple Classifiers", using complementary and redundant sources is named differently: *Selection* for complementary strategy and *Fusion*[2] for redundant strategy [141]. *Selection* refers to a configuration that each classifier involved in the combination process is experienced on some local area of feature space. In *Selection*, when a feature vector is submitted for classification, related classifier has the authority. Often, but not necessarily, more than one classifier can have the authority. Besides, *Fusion* is a configuration which all the classifiers are equally experienced on the whole feature space. For any feature vector, all classifiers are taken into account for decision. [67]

Another important discussion topic on the use of dependent sources is the order of dependency [64]. Most of the studies dealing with dependency issue assume the dependency is bivariate and use linear transformation methods like Principle Component Analysis (PCA), Independent Component Analysis (ICA), Factor Analysis, etc. to create independency or exploit linear statistical dependencies. However, it is not known whether higher order dependencies exist in between the features/modalities. In [64], Kludas et al. claim that information interaction (an information-theoretic dependence measure which is multivariate, high dimensional) is superior to the traditional (bivariate) dependence measures. But their experimental studies have not verified this theoretical idea.

3.1.4 Content Representation

Diversity of combined features/modalities causes complexity and difficulty in fusion and learning. Each feature can have its own feature space, dimensionality, feature value types (i.e. continuous, symbolic, etc.), feature value boundaries, etc. This heterogeneity of features/modalities causes the learning and fusion systems to have complex setups. But still, it is possible to have a homogeneous representation of the features. For a fast (providing online learning) and flexible (handling any new features/modalities blindly) fusion system, we should have a homogeneous representation of the involved features, that is regardless of the intrinsic dimensionality and scale of each feature/modality. [15]

[2] This definition of *"Fusion"* should not be compared, confused or supposed same with our general scope of fusion.

When the heterogeneity of the representation is discussed, a normalization mechanism is required on the features/modalities. Thus, 'Normalization of Sources' variable in the general framework has an effect on the 'Content Representation'. Details on normalization of sources in order to use a heterogeneous representation in a fusion system is described in Section 3.1.5

Another important issue for a fusion system, related to the content representation, is when the preferred representation is constructed. To obtain an efficient system, the representation construction of the classifier model should be offline, i.e. during training. For instance, considering an information retrieval system which uses some number of features for recognition, recognition information (related to the used features) of training instances should be extracted and indexed in an appropriate representation during the training phase. Then, during query phase, such information can be used easily and fast. Not doing so causes extraction of this information during query phase, which results in a very slow retrieval system.

In [15], Bruno et al. analyze the studies in the literature at three different representations:

- **Feature-based Representation:** Feature-based representation is a straightforward approach and mixes heterogeneous vectors of various dimensions and scales. In order to use such representation, different dimensionalities should be projected and different scales should be normalized. This causes a complex setup for the fusion system. Thus, the fusion system becomes very dependent on the parameter settings of the currently used features and less flexible for adding new features to the system. A way of handling various dimensions and scales can be the conversion of included modalities into a single modality and representing the features in a unimodal approach. In [121], Snoek et al. give design of such system for multimodal video processing: Visual and auditory modalities can be converted into textual modality by using some Optical Character Recognition (OCR) and speech recognition methods.

- **Similarity-based Representation:** Similarity-based representation uses similarity or distance values of features for representing data. Using similarities make the fusion system independent from the intrinsic dimensionality of the

features. It should be noted that calculating the similarity and distance values are performed via some similarity functions. In similarity-based representation, the system still becomes dependent to the scales of the used similarity and distance values. So the similarity values should be normalized before combining them. Very often, different scales of expert outputs make them non-comparable. Normalization is a way to make them comparable, which is discussed in Section 3.1.5. In [16], Bruno et al. utilize similarity-based representation. Also, in [15] and [16], they give a list of other studies utilizing this representation.

- **Preference-based Representation:** Preference-based representation is one step ahead of the similarity-based one. The problem of different scales of feature similarities can be solved by using preference-based approach. In this representation, ranks of the features according to their similarities are held. It provides a dimensionality-independent and scale-independent system, which can be defined as fully homogeneous. But it should be noted that it causes a problem of combining several ranked lists. In [15], Bruno et al. utilize preference-based representation. However, finding out different preferences is an open research issue.

3.1.5 Normalization of Sources

During a fusion process, in order to utilize the fusion elements, all of them should have values in the same value types (i.e. continuous, symbolic, etc.), boundaries, scale etc. However, usually they are represented in different types, boundaries and scales. 'Normalization of Sources' variable is used to configure such requirement.

In the literature there are several normalization techniques. In [46], Jain et al. systematically study the effects of different normalization techniques. They give definitions of several normalization techniques and perform experiments on them. They study on the following normalization methods: Min-max, Decimal scaling, z-score, Median and MAD, Double Sigmoid, $tanh$-estimators and bi-weight estimators.

Results of their experiment show that Min-max, z-score and $tanh$-estimators methods followed by a simple sum fusion are superior to other techniques. Min-max and z-score

are sensitive to outliers, whereas $tanh$-estimators is robust and efficient. If location and scale parameters are known, Min-max and z-score methods can be preferred for efficiency. Otherwise, $tanh$-estimators method should be preferred.

In addition, in [98], Poh et al. give some other useful way of transforming outputs into a common domain for comparison. They suggest transforming outputs into probability or log-likelihood ratio domain, and giving successful examples on transforming into log-likelihood domain.

3.1.6 Fusion Level

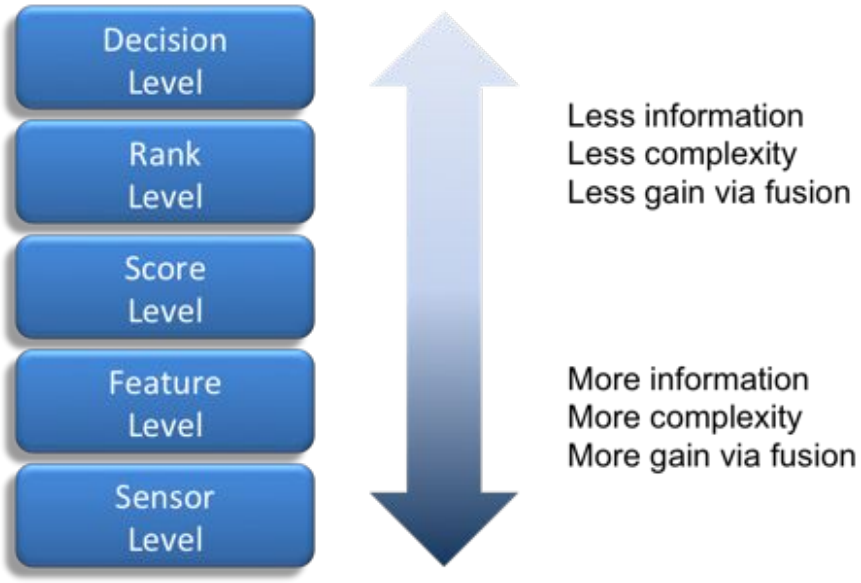


Figure 3.2: Fusion Levels

In the literature, the fusion process is performed at 5 different levels [9, 46, 63, 98, 113, 124, 128, 144]. Figure 3.2 gives an illustration of the fusion levels. The selected level for the fusion differentiates the information available for fusion and computation complexity of the system. At the lower levels, more information is available but using such detailed information causes a computationally complex system; whereas at the higher levels, less information is provided for the fusion operation and it is easier to combine them. Besides, lower levels provide more gain via fusion than the higher levels, due to the usable information available. The levels of fusion will be discussed in the following subsections.

5 different levels of fusion are also grouped at two higher classes according to when the classification of the features is performed: Early Fusion and Late Fusion.

Early Fusion is the fusion performed before the classification. The available information before classification is still considerably much –less from signal-based fusion, but still much more than late fusion– since the system can obtain the unprocessed (unclassified) data. It is possible to exploit relations in between the data (i.e. features) but also it is hard and computationally expensive to combine them. In addition, the combination can result in a high-dimensional data, the curse of dimensionality problem. Then, training of such a system requires a lot of training data. Having more training examples may create a risk of over-fitting data. Thus, Early Fusion is an effective but computationally complex and risky way of fusion. [63, 124]

A typical feature-based early fusion scheme is presented in Figure 3.3. Firstly features of the sample are extracted. Then features are directly combined without any classification or recognition process. Such combination (i.e. concatenating them) is a difficult task. After combination, a learning process is performed. [123, 124]

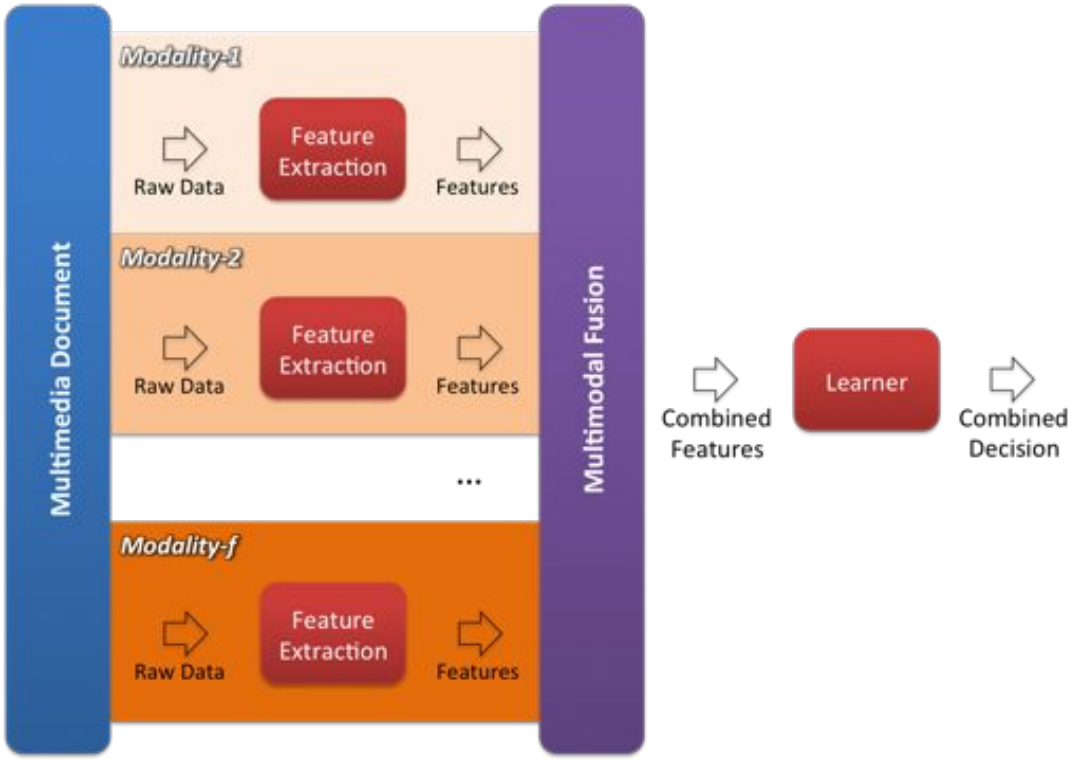


Figure 3.3: General Early Fusion Scheme

Besides, Late Fusion is the fusion performed after the classification. Contrary to early fusion, this fusion type is simpler since it uses a processed and interpreted data by the classification. However, the information available is very limited, which means there is a potential loss of correlations between features/modalities. Thus, late fusion is a less effective but computationally much better way of fusion. [63]

A typical late fusion scheme is given in Figure 3.4. Firstly features of samples are extracted and learning (classification) process of each modality/feature is performed separately. Then, results of these classifications are combined into an appropriate representation. Final decision of fusion is obtained either by a second-level learner or a simple aggregation method. [123]

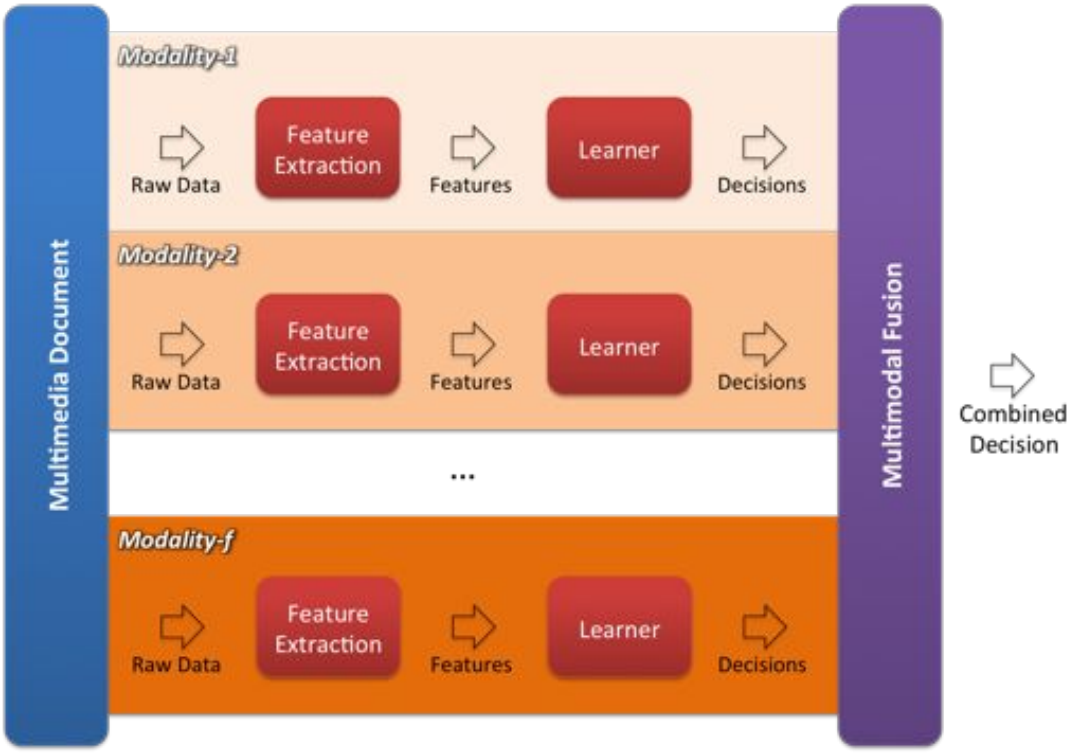


Figure 3.4: General Late Fusion Scheme

It should be noted that, whole process of the late fusion process can be computationally-expensive considering that it contains many classifiers, although the combination process is much simpler than that of the early fusion.

Considering feature-based fusion (classifier combination) studies in the literature, using correlations of features is not very popular and frequently studied, as discussed

in Section 3.1.3. Using early fusion without obtaining correlation gain is not beneficial. Also simplicity of late fusion makes it attractive. Therefore, most of the studies in the literature prefer late fusion [123]. Furthermore, these studies usually reach at good results with late fusion. In [123], Snoek et al. compare late fusion methods with early fusion methods (without exploiting correlations) experimentally and concludes that late fusion is superior to the early fusion.

3.1.6.1 Sensor Level

Sensor level fusion is the fusion level having the richest source of information, but the fusion process is the most complex one. Dealing with sensor data requires much more effort than other fusion levels because of the extensiveness of the data. Additionally the data includes noise. [98]

Sensor level fusion includes;

- combining multiple sensors (i.e. Image fusion with RGB and NIR sensors)
- combining multiple snapshots with a single sensor (i.e. Image fusion by using images taken at different times with single sensor)

3.1.6.2 Feature Level

In feature level fusion, the feature sets originating from multiple feature extraction algorithms are combined into a single feature set. The combination method requires appropriate feature transformation, reduction and normalization strategies due to the differences in feature spaces and types. The primary benefit of feature level fusion is the detection of correlated feature values, which improves the recognition accuracy. [98]

Besides, the feature level fusion may have several drawbacks [63, 124] due to;

- The ‘curse of dimensionality’ caused by dealing with several features,
- Different feature spaces and types of features,

- Computational expensiveness caused by the transformation, dimensionality reduction and normalization procedures to solve the problem of differences in feature spaces and types,
- Needing a lot of training data in order to perform dimensionality reduction and normalization procedures,
- Risk of over-fitting data caused by using a lot of training data.

3.1.6.3 Score Level

In score level fusion, the classification process results with match scores and match scores of multiple classifiers are fused during the combination process. The score level fusion is the mid-point among fusion levels. Although it is not possible to extract correlation information, some valuable information still exists. Also ease of accessing and processing match scores (compared to lower levels) makes score level fusion more interesting. Thus, fusion at this level is the most commonly discussed approach in literature [98].

Despite the ease of accessing such valuable information, using it still requires some challenge. Different sources can have different intervals of matching score, so fusion process should handle the variance in the intervals of scores (i.e. normalization on the match scores).

3.1.6.4 Rank Level

In rank level fusion, the usable information for fusion becomes less and only ranks for the classification results are available. The fusion process combines rank outputs of multiple classifiers. Using rank lists as inputs to the fusion makes the fusion process much simpler since using rank lists do not require a normalization process and the rank lists of different sources are directly comparable. So, in this level of fusion it is simpler to implement a fusion system than the score level.

Still, rank level fusion has a problem to deal with: Combining multiple rank lists (from multiple sources) without any score information requires a rank aggregation technique.

3.1.6.5 Abstract Level

Abstract level fusion is the highest level fusion. In this level of fusion, the least information is available for fusion: the final recognition decisions of classifications. This level is the simplest to implement a fusion system. However, the gain obtained via fusion is the minimum, compared to the other levels of fusion.

This level of fusion is only suitable for combining COTS systems since most of the COTS recognition systems provide access only to the final recognition decision [98].

3.1.7 Fusion Methodology

Fusion Methodology defines which algorithm is used for combining sources. The literature includes a lot of methods proposed and experimented for fusion. However, a superior fusion algorithm has not been accepted by the researchers [143]. It is difficult to predict whether a combination is superior, so no clear preference of one combination method is compromised [67].

The algorithms utilized in the literature can be analyzed in two groups according to whether they have a learning step: Non-trained and Trained methods [9, 30, 46, 47, 61, 130, 143, 145]. Non-trainable (Combination/Linear/Fixed Rule) methods bases on linear aggregation and voting methods. Mostly used ones are product aggregation, sum aggregation, minimum selection, maximum selection, median selection, majority voting, concatenation, weighted average aggregation, linear combination. Success of these methods is based on their simplicity and they are usually preferred due their simplicity. Besides, Trainable (Classification/Learning-based) methods contains more complicated classification algorithms and requires a training step in order to obtain a model for the classification. Mostly utilized methods include Bayesian networks, neural networks, Gaussian mixture models, factor graphs, decision templates, genetic algorithms, adaptive weighting, borda count, logistic regression, belief functions, Dempster-Shafer techniques, fuzzy integrals, bagging, boosting, random subspaces, k-nearest neighbor, decision trees, support vector machines and label ranking.

It is possible to group the algorithms in several different manners. For instance,

in [121] Snoek et al. group the studies as knowledge-based and statistical approaches. Knowledge-based approaches use (predetermined or trained) knowledge base rules, whereas statistical approaches prefer statistical correlations between sources and classes.

An important issue in fusion methodology is the relation between fusion level and methodology. Fusion level directly affects the choice of fusion algorithm. Some of the methods can be used in all levels, whereas most of them can be used for only one level. The relations between the methods and the levels are given in Table 3.1 [15, 46, 47, 61, 62]. Note that this table is an exhaustive one, not a complete table.

As seen on the table, the Trainable - Score Level section contains the most of the methodologies. In fact, this is not because of the usability of the methods, but most of the studies deal with score level fusion, therefore the number of algorithms in that section is more than others.

The table also shows that there are so many methods available to be used for fusion. Although there is not clear superiority of one fusion method [67], several studies argue and experimentally show that learning-based (trainable) methods are better than the non-trainable ones [30, 31, 46, 124, 130, 143]. However, there exist some counter-examples. In [46] Jain et al. show that combination approach is better than some classification approaches (decision trees and linear discriminant analysis).

3.1.8 Operation Modes

A fusion system can operate in one of three different modes: serial mode, parallel mode, or hierarchical mode. Operation modes in fusion defines whether the sources will be used incrementally (serial), at once (parallel) or combination of these two (hierarchical).

In a serial architecture (Figure 3.5(a)), fusion is performed at more than one step. At each step, one new source is fused with the result of the previous fusion step. The output of one source is typically used to narrow down the number of possible results before the next source is used. Therefore, multiple sources of information do not have to be acquired simultaneously. Further, a decision could be made before acquiring all

Table 3.1: Relation of Fusion Algorithms with Fusion Levels

	Trainable	**Non-trainable**
Abstract Level	Knowledge-base Rules Boosting Neural Networks Gaussian Mixture Models Label Ranking	Majority Voting Product (AND) Aggregation Sum (OR) Aggregation
Rank Level	Borda Count Logistic Regression Dempster-Shafer Rank Boost	Highest Rank
Score Level	Adaptive Weighting Logistic Regression Bagging Fuzzy Integrals Dempster-Shafer Belief Functions Random Subspaces Decision Templates Genetic Algorithms Neural Networks Gaussian Mixture Models Factor Graphs K-Nearest Neighbor Decision Trees Support Vector Machines	Product Aggregation Sum Aggregation Minimum Selection Maximum Selection Median Selection Weighted Average Linear Combination
Feature Level	Latent Semantic Analysis Probabilistic LSA Canonical Correlation Analysis	Concatenation Weighted Average

the sources. This can reduce the overall recognition time and make the system less dependent on each of the sources. [106]

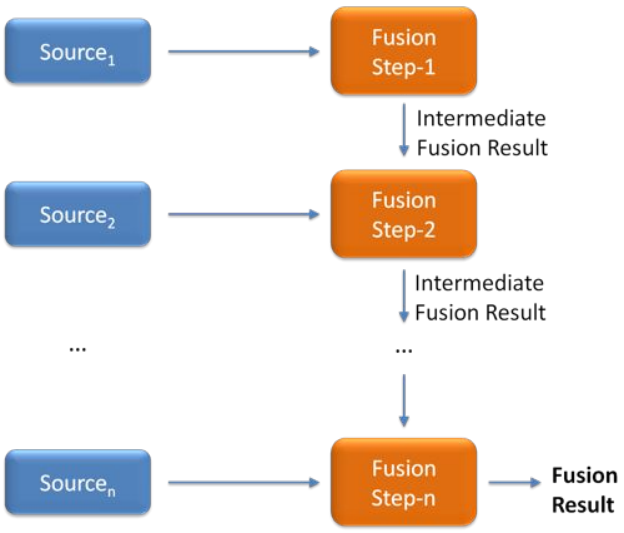
In a parallel scheme (Figure 3.5(b)), all of the sources are ready at the fusion time and fused at once. The information from multiple sources is used simultaneously in order to perform recognition.

In a hierarchical scheme (Figure 3.5(c)), both of serial and parallel types are employed. At each step, some of the sources are fused in a parallel way and the result is forwarded to the next fusion step.

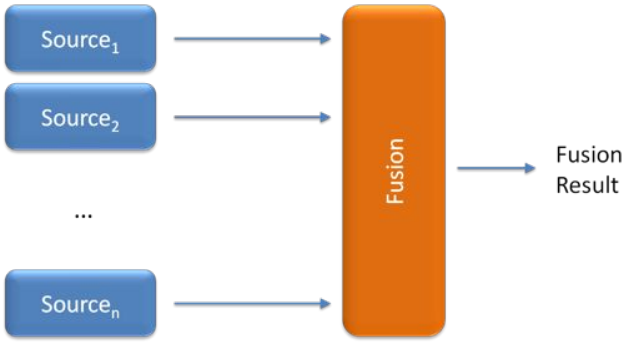
A special sub-type of serial mode can be an *Iterated* architecture, where the same source used for a number of times (serially). Such architecture can provide the system from unnecessary operations and make the system runnable on low-performance hardwares by performing the fusion process in some defined number of steps.

While the parallel fusion strategy is the most commonly used for information fusion in the literature [98, 121], there are several advantages of serial fusion. It offers the possibility of making reliable decisions with only a few sources, leaving only difficult samples to be handled by the remaining sources [98]. Also using a serial architecture can be beneficial in the systems that obtain their data sequentially. A video data is a good example for such a situation. In [121], Snoek et al. give some of the studies performing iterated fusion are; [6, 84, 125], which are processing video data incrementally.

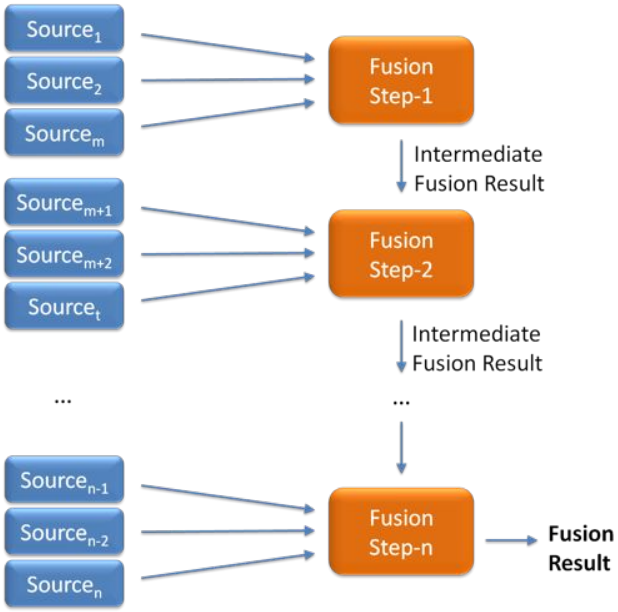
An important issue in operation modes is the processing of different modalities for a multimodal resource with a parallel operation mode. To perform the fusion operation all of the modalities should be ready at the time of fusion. Such requirement exposes the need of synchronization (or alignment) of the modalities according to each other. For instance, assume a video data having text, audio and visual modalities. Extracted information from each of these modalities should be aligned in order to perform fusion in a correct way. The alignment is configured by the 'Synchronization' variable, which is introduced in Section 3.1.9



(a) A Typical Serial Fusion Scheme



(b) A Typical Parallel Fusion Scheme



(c) A Typical Hierarchical Fusion Scheme

Figure 3.5: Fusion Schemes for Types of Operation Modes

3.1.9 Synchronization

An important aspect for multimodal fusion is the synchronization and alignment of the different modalities according to each other so that all modalities have a common time-line. For instance, assume a video data having text, audio and visual modalities. Extracted information from each of these modalities should be aligned in order to perform fusion in a correct way. In the literature, usually timestamps on the modalities are utilized and timestamps of the secondary modalities are converted to the timestamp of primary modality [121]. However, timestamps are not always available on the modalities. In such cases, domain and application specific, ad-hoc solutions are applied.

3.1.10 Adaptation

Choosing the best sources to combine and best combination method are crucial and difficult tasks that directly affect the performance of the fusion system. Since most of the solutions are ad-hoc strategies; it is crucial to have a careful analysis on the sources and methodologies [2]. However, careful analysis means careful setup on inputs which makes the system dependent on specific conditions. Also, it cannot be known during design time how each feature, modality or data source performs in real world environments and which of them are the most reliable. Solution to this dependency is to make the fusion system adaptive.

Having an adaptive system requires making the system adaptable to changing data environment for choosing the best sources to combine and best combination method. Such capability can be achieved by re-configuring the 'Selection of Sources', 'Fusion Strategy', 'Fusion Level' and 'Fusion Methodology' variables in the fusion architecture.

[47], Jain et al. summarize the adaptive and non-adaptive fusion methodologies used in the literature. Some of the adaptive fusion techniques are; Adaptive weighting, Mixture of local experts (MLE), Hierarchical MLE, Associative switch. Some of the non-adaptive fusion techniques are; Voting, Sum Aggregation, Product Aggregation, Minimum Selection, Maximum Selection, Mean Selection, Borda Count, Logistic

Regression, Dempster-Shafer, Fuzzy Integrals, etc. However, these lists are constructed using the currently available studies in the literature. It is always possible to convert non-adaptive methods to some adaptive version, with some extra effort and new methodology.

3.2 Open Issues In Fusion

The best way to define the problems in the information fusion research area is to investigate the currently available studies. The problems worked on and the future directions pointed out are exactly the problems of the area. Below, the open research issues depicted from the literature are presented.

The problems of information fusion is based on selecting the best sources to combine and finding the best way to combine them. In [143], Wu et al. state such problems as arguing two core issues have not been adequately addressed yet: (i) How to determine best modalities? (ii) How to best fuse them? In [68], Kuncheva addresses the problems similarly: (i) Choosing a suitable combination method is a difficult problem, (ii)It is not known how to use a data set and which classifier to select with it. In addition, Kuncheva comments on the research area that combining classifiers is an promising area and still, there are many experimental and heuristic studies to be offered. Likewise, in [98] Poh et al. claim that there is a huge space of different fusion architectures that has not been explored yet. In [2] Arevalilli et al. highlight a crucial point; ad-hoc (experimental or heuristic) strategies requires careful analysis for choosing the best method to combine features. Such judgment leads us to the adaptability issue that eases and relaxes the process of careful analysis. However, in [98] Poh et al. state that adaptability is an open research issue and requires some well-defined quality and reliability measures of sources and fusion methods.

In [62], Kludas gives some different important points as fusing dependent sources and predicting the performance improvement by fusing different modalities/sources/samples. Kludas mentions that most of the studies in the literature deals with independent sources. In [145], Yan et al. study on the second consideration of Kludas (in a limited domain of rank aggregation) and state the open problems as; (i) What are the limits for

combination having the scores of each different source? (ii) Is linear(non-trainable) combination sufficient? (iii) How the scores of sources should be normalized?

Interestingly, in 1992, in [144] Xu et al. list most of these problems as new problems to be studied in the literature: (i) Is it possible to determine recognition rate theoretically instead of experimentally? (ii) Current assumption on fusion is that individual classifiers are independent. It is necessary to develop approaches for dependent classifiers. (iii) How many classifiers/features are appropriate for fusion?

Beyond these problems, another crucial open research issue on consensus is the lack of a general theoretical framework for information fusion. In [143], Wu et al. state that traditional work on multimodal fusion is mostly heuristic-based and lacks theories to answer questions on selecting modalities and fusion methods. In [64], Kludas et al. complain about the fact that a general theoretic framework is still missing, although information fusion is an independent research area over last decades. In [63], they summarize the current state of the literature as follows: A vast number of disparate research areas utilize information fusion, but they describe the fusion methods in their context of theory. Also, in multimedia, relation between basic features and content description is limited, namely the semantic gap, so the fusion problem is solved empirically then justified theoretically. Thus, a general formal theoretical framework is missing for information fusion. In [63], they highlight an important point: Due to the lack of a formal theoretical framework and ambivalent fusion results in several studies in the literature; there exists a vibrant discussion on the theoretical achievable performance improvement boundaries of fusion system compared to single source systems.

Besides, in [67], Kuncheva appreciates currently available studies on a theoretical framework but finds them immature since they are only for special cases, usually assuming independent classifier outputs. Yet, she founds currently available heuristic and ad-hoc solutions useful as a pre-phase towards a more general theory of classifier combination. Similarly, in [47], Jain et al. state that there exists only a few theoretical explanations on classifier combination and most of them apply to simplest schemes under rather restrictive assumptions. Actually, current studies are still lacking to state a general theoretical framework. For instance; in [61], Kittler et al. introduce a

theoretical framework but only for some special combination methods like product, sum, min, max and majority voting. In [145], Yan et al. present a theoretical framework for average precision boundaries.

In brief, the following list identifies the problems in information fusion research area. Actually, this list is in accordance with the variables given in Section 3.1:

- Lack of a general theoretical framework.
- Determining limiting theoretical upper bounds of performance.
- How to determine best sources?
- How to best fuse them?
- Dealing with dependent sources.
- Normalization problems of different sources.
- Well-defined quality and reliability measures for selection and adaptability of sources.

CHAPTER 4

NON-LINEAR WEIGHTED AVERAGING[1]

Linear combination is a popular approach in information fusion due to its simplicity. However, it suffers from the performance upper-bound of linearity and dependency on the selection of weights. In this chapter, we introduce a 'simple' alternative for linear combination, which is a non-linear extension on it. The approach is based on the Analytical Network Process, which is a popular approach in Operational Research, but never applied for fusion before. The approach benefits from two major ideas; interdependency between classes and dependency of classes on the features. Experiments conducted on CCV dataset demonstrate that our proposed approach outperforms linear combination and other simple approaches, moreover it is less-dependent on the selection of weights.

4.1 Overview

Combining the information gathered from multiple modalities is an empirically validated approach to increase the retrieval accuracy [4]. Among the various combination methods that have been proposed, most frequently utilized approach is the Linear Weighted Fusion (or Linear Combination) [37, 133, 145], due to its simplicity and reasonable performance despite its simplicity. Some other well-known methods are as follows: Majority Voting, Support Vector Machines, Bayesian Inference, Dempster-

[1] This chapter was published as [149].

Shafer, Neural Networks, Decision Templates and Borda Count [4].

When compared with the linear combination, these approaches are; (a) either has a simple design as the linear combination but worse/equal in performance, (b) or better in performance but require complex training setups in order to obtain an adequate performance. Moreover, the approaches in the latter group are usually not limited to linear approximations. So, it can be argued that the use of linearity in combiner design causes a performance upper bound on retrieval accuracy. A detailed analysis on the performance limits of linear combiners can be found in [145]. Besides, another important drawback with the linear combiners is the high dependency of the combiner performance on the selection of the weights. However, the selection of the optimal weights is one of the important issues that have not been adequately addressed yet in the fusion domain [4, 98].

Aligned to above given issues, we would like to investigate for a combination approach which (i) is as simple as the linear weighted fusion, (ii) can achieve the performance upper bound of linear weighted fusion, and (iii) is less-dependent on the selection of the weights. Through this study, we resemble the multimodal fusion problem to the real-life multi-criteria decision making problem in Operations Research domain and would like to introduce two popular approaches, Analytical Hierarchy Process (AHP) [108] and Analytical Network Process (ANP) [109]. AHP is a linear solution approach having the same principles with the linear weighted averaging method. However, ANP is a quite different solution that extends the linear weighted averaging method into a non-linear one, and has never been applied in the information fusion domain before. Thus, in this study, we adapt and extend the calculation approach and parameters of ANP for multimodal fusion. We show that it can be utilized as a 'simple', 'non-linear' and 'less-weight-dependent' way of fusion, which overcomes the problems listed above. We evaluate the approach by using the Columbia Consumer Video (CCV) dataset against several different approaches and obtain convincing results. Moreover, we empirically show that non-linear weighted averaging makes the accuracies less dependent on the selection of weights.

4.2 Linear Weighted Averaging and AHP

We focus on a score-based late fusion scheme, with a setting that each classifier is dedicated for a single feature (as given in Figure 4.1). Each of the classifiers performs multi-class classification and outputs of classifiers are homogeneous, giving score values for the same set of retrieval classes. Also, assume that we have m number of retrieval classes and n number of features.

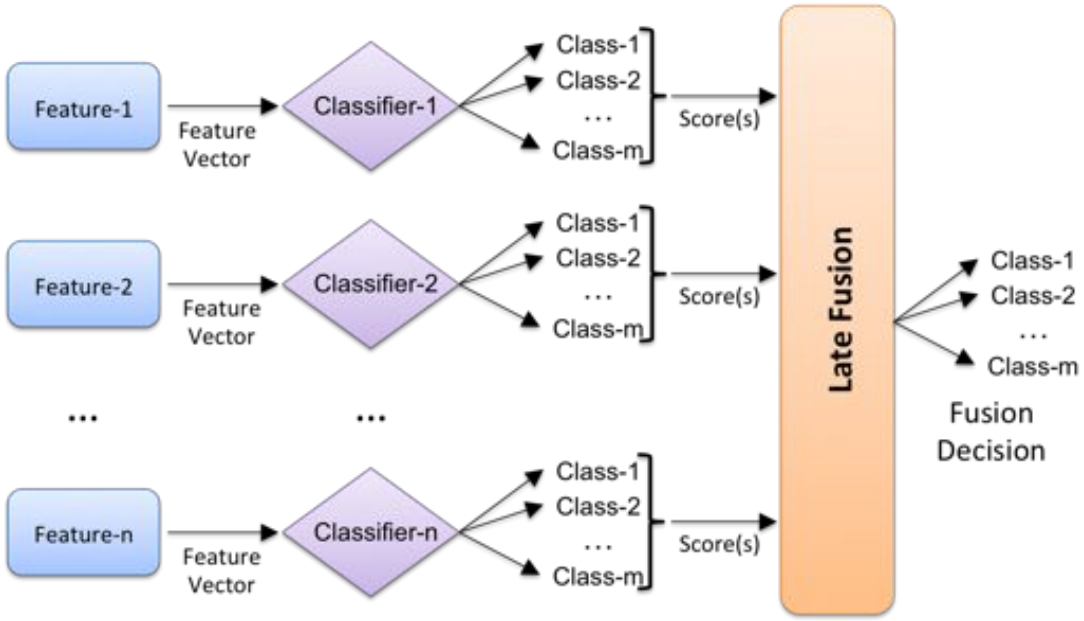


Figure 4.1: A Score-based Late Fusion Scheme

In a such a fusion architecture, the outputs of several classifiers are aggregated in order to make a final decision. In linear weighted fusion methods, the information obtained from multimodal features is combined by assigning some particular weight for each modality and performing a summation or product operation to combine. Considering a summation preference, the final decision is calculated by;

$$\mathbf{S}_L = \mathbf{D}\mathbf{W}_D \ , \tag{4.1}$$

where $\mathbf{D}$ is a $m \times n$ matrix, containing the output scores of classifier in each column; $\mathbf{W}_D$ is a n-sized vector, containing the weights of each feature; and $\mathbf{S}_L$ is a m-sized vector, containing the combined decision scores for each retrieval class.

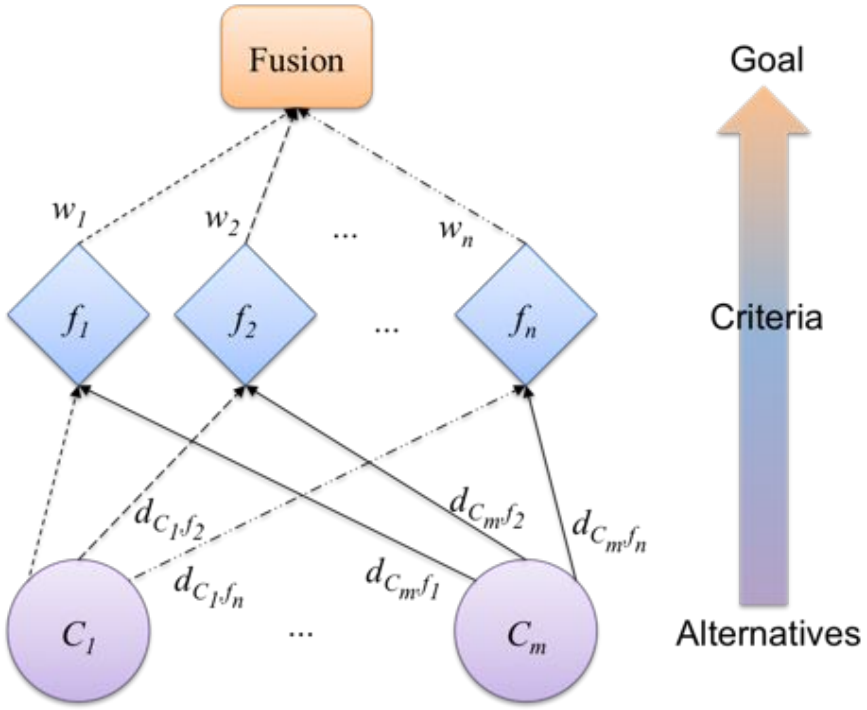


Figure 4.2: AHP Decision Hierarchy

AHP presents Equation 4.1 with a more concrete representation. First the multi-criteria decision making problem is modeled with a simple hierarchical model consisting of a *goal*, *criteria* and *alternatives* nodes. Figure 4.2 presents a hierarchy for the multimodal information fusion problem with m number of classes and n number of features. Here, it should be noted that the edges between nodes are unidirectional, as a result of being a 'hierarchy'. In order to find the combined decisions, the total of alternative path lengths from each *alternative* to *goal* is calculated, where a path length is the product of the values on the edges along the path. A detailed description of AHP can be found in [108].

A crucial step in this approach is the determination of weights, which directly affects the fusion performance. An optimal solution is not guaranteed without an exhaustive search in the feature space. However, several heuristic solutions can be applied. As the most simplistic case, the weights of features can be selected equally ($w_i = 1/N$) which is also called *Simple Averaging*. Furthermore, some well-known heuristics are RELIEF [59], Information Gain [43] and Gain Ratio [101]. In this study, we utilize RELIEF and exhaustive search for experimental purposes. We use also a random

weight selection approach to show the effect of weight selection.

4.3 Non-linear Weighted Averaging and ANP

ANP is a generalization of AHP and created with a consideration that many decision problems cannot be modeled with a simple hierarchy because they can involve interactions/dependencies of the included nodes [109]. Thus, ANP proposes to model the decision problem with a network which allows to define bidirectional transitions between the nodes. A network model, which is designed for the multimodal fusion problem with m number of classes and n number of features, is given in Figure 4.3. Combined decision calculation is similar with AHP. However in ANP, the number of alternative paths is more than AHP, even indefinitely many, considering the possible bidirectional transitions between the nodes.

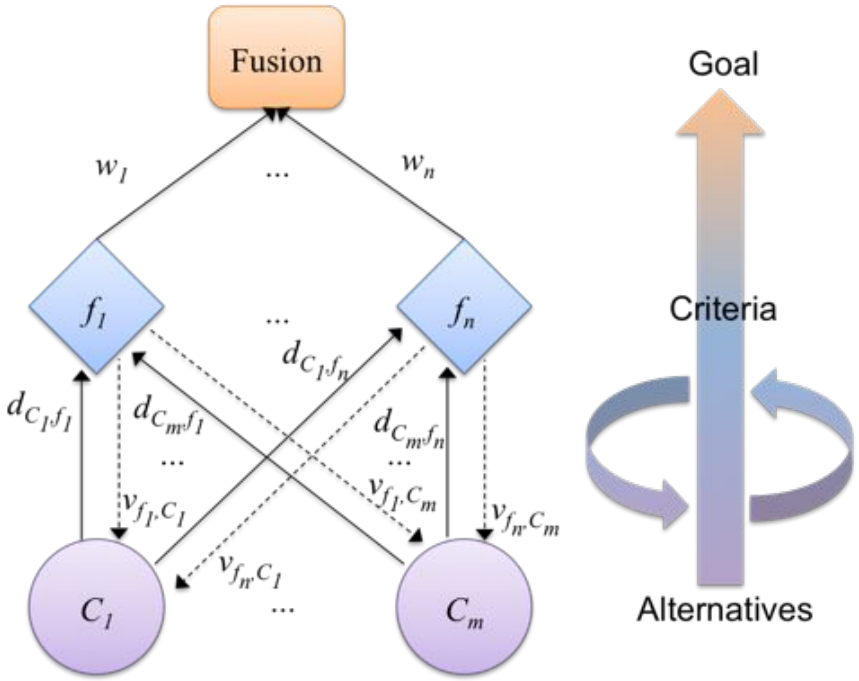


Figure 4.3: ANP Decision Network

Considering the ANP approach, we can extend the linear weighted averaging approach

into a non-linear approach by employing an additional weight factor.

$$\begin{aligned} \mathbf{S}_N &= \mathbf{W}_I \mathbf{S}_L \ , \\ &= \mathbf{W}_I (\mathbf{D}\mathbf{W}_D) \ , \end{aligned} \tag{4.2}$$

where $\mathbf{W}_D$ represents the direct weights, which are the traditional feature weights as used in linear weighted averaging. Besides, $\mathbf{W}_I$ is used for the indirect weights, which can be described by incorporating two crucial ideas, in a multimodal fusion problem: (i) interdependency between the retrieval classes, and (ii) class-specific feature selection. The former idea provides exploiting the interdependencies between classes and benefit from the correlation as a weighting factor. In order to obtain the correlation between the classes, outputs of the classifiers are utilized. The correlation between the classifier outputs are usually ignored by many of the late fusion approaches, and only the corresponding output score of each classifier with the retrieval class is used during combination. For instance, in linear weighted averaging, the fusion result for C_1 is calculated by using only the scores for C_1 of each classifier. To exploit the interdependency, we incorporate all score outputs of all classifiers while performing fusion. Furthermore, the latter idea is based on the dependency of classes on the features. Although feature weighting methods usually propose solutions such that the resulting feature set is selected independent of the classes, defining feature weights that are specific to each class is an intuitive and promising approach [151]. For instance, in a multimodal scenario of multimedia data, the audio features are more useful for a $MusicPerformance$ class, whereas it is better to utilize visual modality for detecting a $Beach$ occurrence. In order to obtain class-specific feature weights, the feature weight calculation methods can be used separately for each feature, in a one-against-all fashion.

Considering these two ideas, the indirect weights $\mathbf{W}_I$ are calculated as;

$$\mathbf{W}_I = (\mathbf{D}\mathbf{V})^i \ , \tag{4.3}$$

where $\mathbf{D}$ is a $m \times n$ matrix, containing the output scores of classifier in each column; and $\mathbf{V}$ is a $n \times m$ matrix, containing the class-specific weights. In $\mathbf{V}$, each column holds the feature weights for a retrieval class. Considering that the product $\mathbf{DV}$ provides a square matrix, any power of this term is applicable. It should be noted that having $\mathbf{D}$ in the calculation of $\mathbf{W}_I$ and using powers provide 'non-linearity' into the solution. In

addition they provide an implicit feature weighting estimation capability and make the solution 'less-dependent' on the weights $\mathbf{W}_I$ and $\mathbf{V}$. The resulting $\mathbf{W}_I$ contains linear combination results by using own class-specific features on the diagonal and linear combination results by using the class-specific weights of other classes as the rest. Thus, the final non-linear weighted averaging formulation is as follows;

$$\mathbf{S}_N^i = (\mathbf{DV})^i(\mathbf{DW}_D) \; . \tag{4.4}$$

In order to obtain the most appropriate value of i, we focus on three solutions: First one is based on the converging characteristic of Equation 4.4. Solution is converting Equation 4.4 into a general eigenvalue problem at the convergence point. However, it is not guaranteed to obtain the best fusion performance for the converged $\mathbf{S}_N$ value. Second solution is searching for the i value between 1 and convergence-based i value, which gives the best accuracy, via a training set. For the third one, the class-specific approach is mentioned again and it is argued that it is most likely to see the i value being different for each class. Thus, the i value is optimized for each class separately, similarly with the second approach.

4.4 Experiments

The experiments are carried out on the Columbia Consumer Video (CCV) Database [55], based on the semantic retrieval of classes. The dataset contains multimodal features –visual (SIFT), audio (MFCC), motion (STIP)– of 9,317 videos for 20 semantic classes listed on Table 4.1. The dataset is equally divided into training and test sets. Feature details can be found in [55]. To measure the retrieval accuracy, Average Precision (AP) and Mean Average Precision (MAP) metrics are used.

As the first test, non-linear weighted averaging method (NWA) is compared against; (i) Single features, (ii) Simple combination; Simple Averaging (AVG), Minimum Selection (MIN), Maximum Selection (MAX), (iii) Learning based combination; Naive Bayes (NB), Support Vector Machines (SVM), (iv) Linear weighted averaging (LWA) methods. For the feature weight selection of LWA and NWA, a RELIEF based feature weighting is used. For the NWA calculation, the 'best class accuracy' based approach is preferred. During all tests, first a classification process is performed with

Table 4.1: Accuracy comparisons. The best result for each category is highlighted in bold.

	SIFT	STIP	MFCC	AVG	MIN	MAX	NB	SVM	LWA	NWA
Basketball	66.95%	63.37%	44.65%	73.11%	67.16%	70.04%	22.87%	72.55%	69.35%	**75.89%**
Baseball	40.30%	18.38%	9.17%	43.15%	31.18%	39.00%	24.39%	46.37%	46.01%	**48.91%**
Soccer	49.29%	39.18%	17.59%	53.68%	49.65%	47.63%	25.40%	54.98%	53.98%	**58.26%**
IceSkating	81.18%	65.82%	16.18%	81.37%	71.27%	79.79%	73.63%	83.77%	82.90%	**85.32%**
Skiing	76.85%	60.27%	29.73%	74.31%	68.47%	72.03%	64.77%	**78.50%**	75.27%	78.18%
Swimming	68.84%	53.80%	15.35%	68.89%	56.70%	65.65%	57.30%	70.65%	71.30%	**72.61%**
Biking	36.85%	23.52%	11.36%	39.52%	29.90%	36.75%	32.68%	41.35%	38.73%	**42.76%**
Cat	34.24%	23.82%	17.40%	39.37%	33.94%	34.19%	41.65%	**41.75%**	35.27%	40.02%
Dog	25.48%	27.64%	22.10%	37.80%	35.07%	31.03%	9.92%	39.00%	28.81%	**42.99%**
Bird	17.40%	14.12%	17.63%	26.60%	22.81%	22.97%	16.34%	26.21%	19.86%	**28.80%**
Graduation	31.58%	22.09%	12.44%	36.23%	36.66%	28.28%	26.80%	40.05%	35.34%	**44.94%**
Birthday	33.32%	15.38%	35.94%	49.43%	41.39%	41.27%	45.92%	47.04%	40.54%	**55.53%**
Wed.Reception	18.65%	22.54%	12.41%	24.15%	**27.65%**	20.29%	2.98%	22.39%	17.37%	26.22%
Wed.Ceremony	35.20%	32.88%	35.04%	50.79%	**58.64%**	40.83%	37.86%	54.39%	38.74%	55.63%
Wed.Dance	56.68%	47.61%	28.01%	61.19%	54.52%	54.95%	46.45%	61.17%	59.53%	**66.62%**
MusicPerf.	48.20%	37.75%	56.71%	65.74%	60.51%	61.27%	61.68%	67.90%	53.77%	**68.87%**
NonMusicPerf.	45.21%	53.23%	29.78%	59.61%	51.77%	54.50%	11.79%	53.22%	53.31%	**64.60%**
Parade	48.71%	39.19%	25.62%	58.85%	56.82%	51.26%	46.13%	58.58%	55.17%	**65.33%**
Beach	69.99%	47.49%	37.34%	71.41%	64.16%	67.97%	3.83%	74.02%	71.83%	**75.43%**
Playground	44.59%	30.26%	23.83%	51.30%	49.62%	43.72%	51.11%	52.28%	49.51%	**57.90%**
MAP	46.48%	36.92%	24.91%	53.32%	48.39%	48.17%	35.18%	54.31%	49.83%	**57.74%**

SVM classifiers, then the results of these classifications are combined. The multi-class classification with SVM is performed with a one-against-all approach. When needed, Naive Bayes implementation of MatLab Statistics Toolbox and LibSVM [18] are used. In Table 4.1, the APs of each class and the MAPs are presented for each combination approach. In Figure 4.4, the MAPs of all approaches are visually compared.

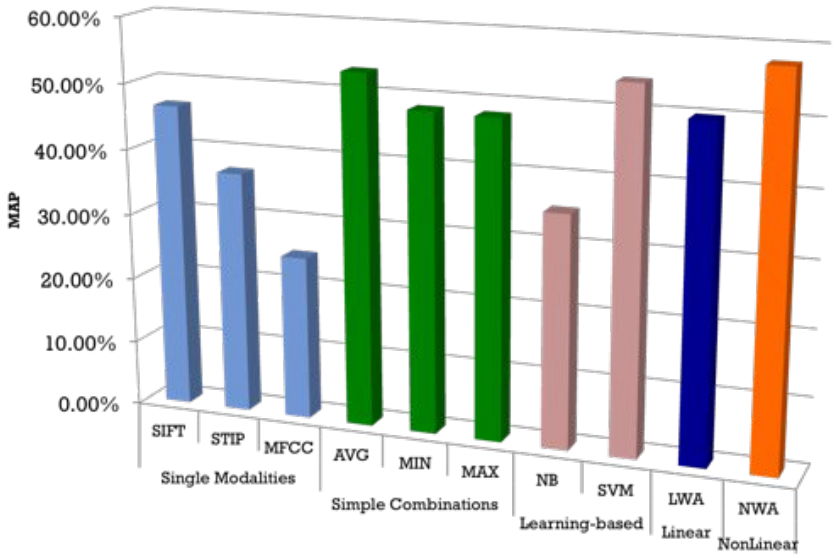


Figure 4.4: MAP comparisons

As a secondary test, LWA and three NWA calculation approaches, which are convergence-based (NWA-CB), best common accuracy (NWA-BCo) and best class accuracy (NWA-BCl), are compared against three different feature weighting methods: Random, RELIEF and Exhaustive Search. The comparison is presented in Table 4.2 and Figure 4.5 For the 'Random' weighting approach, a random feature weighting process is repeated 1000 times, and the minimum (Rand-Min) and the mean (Rand-Avg) values obtained is presented in the table.

Table 4.2: LWA, NWA vs. Weighting Methods

	Rand-Min	Rand-Avg	RELIEF	Exh.Search
LWA	30.135%	47.618%	49,829%	57.783%
NWA-CB	55.139%	56.944%	57.734%	57.734%
NWA-BCo	56.031%	57.082%	57.740%	57.783%
NWA-BCl	56.242%	57.287%	57.741%	57.966%

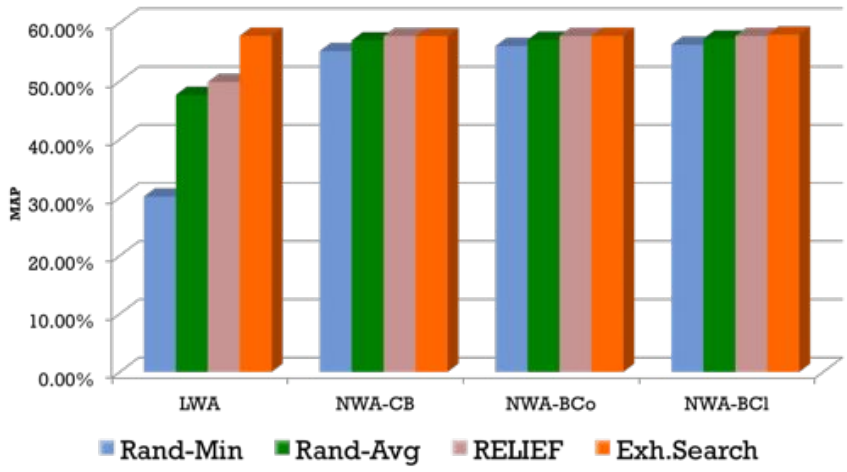


Figure 4.5: LWA, NWA vs. Weighting Methods

Considering the results given in Table 4.1 and Table 4.2, NWA easily achieves the performance upper-bound of linearity and outperforms all other approaches. Simple methods like MIN and MAX seems not adequate for fusion, since they lack the advantage of combining multiple features; though they perform better than the best of the single features. Besides, the AVG method, which is also a linear approach with equal weights, is more accurate than LWA. This is the result of a probable deficiency of RELIEF method to assign weights. However, NWA eliminates such deficiency and obtains the best accuracy values despite the use of RELIEF weights. Thus, the most crucial evaluation is the superiority of NWA solutions on LWA, independent from the feature weights. In addition, particularly focusing on Table 4.2, NWA seems to be less dependent on the selection of weights than the LWA method and can provide reasonably good results even with a worse selection of feature weights. A last comment on this table can be the slight but robust increase in the accuracy by the extensions made on the NWA-CB.

4.5 Evaluation of Fusion System Design

Considering the general fusion framework proposed in Section 3.1, an evaluation of the fusion architecture described in this chapter is given below. Having a ‘multi-modal, multi-classifier’ fusion scenario and focusing on a non-linear weighting solution, the

proposed approach contributes to both ‘What to Fuse’ and ‘How to Fuse‘ problems. Below, how each affecting factor is handled through the proposed solution is described.

- Fusion Setting: The approach combines multiple modalities, each of the modality being a different feature. Before combination, the data of each modality is classified with a separate classifier, and the results of the classifiers are combined.
- Selection of Sources: The approach uses a static feature weighting scheme based on the non-linear weighting. Since the approach enables being less-dependent on the selection of weights, the weights can be calculated with any weighting mechanism. The approach requires the assignment of two types of weights; (i) direct weights, which are the traditional weights for each feature (ii) indirect weights, which are based on the product of class-specific feature weights and output scores from classifiers.
- Fusion Strategy: The approach focuses on the use of complementary information for fusion.
- Content Representation: A feature-based representation is preferred. During the tests, bag-of-words (BoW) based features are utilized, thus the representation can be also be accepted as BoW based.
- Normalization of Sources: The fusion inputs are classifier outputs, where each of them lays in between $[0, 1]$. Thus, a normalization process is not applied on the fusion inputs.
- Fusion Level: The approach is a late fusion approach.
- Fusion Methodology: In this study, a new fusion methodology is proposed. The approach is non-linear weighted averaging, which is an extension on the linear averaging approach.
- Operation Modes: The mode for operation is a parallel scheme.
- Synchronization: The utilized dataset provides synchronized features from different modalities, based on the video start / end intervals. Thus, an additional synchronization is not required.

- Adaptation: The approach is not fully adaptive since a small dependency still exists on the static feature weights. However dependency on these weights is limited, and the indirect weights which are calculated during fusion has more effect. Thus, the approach is accepted as 'almost-adaptive'.

4.6 Remarks

In this chapter, an ANP-based non-linear weighted averaging method is introduced for the multimodal fusion problem. The method extends linear weighted fusion with two crucial ideas; interdependency between classes and dependency of classes on the features. The approach is tested on CCV dataset in a multimodal fusion scenario. The results demonstrate that introduced non-linear weighting approach is superior to linear combination as well as the other basic approaches and is less-dependent on the selection of weights.

CHAPTER 5

CLASS-SPECIFIC FEATURE SELECTION[1]

In this chapter, a class-specific feature/modality selection mechanism is introduced. Throughout the chapter, firstly the approach is presented in detail. Then the evaluations of the approach are given. The approach is firstly designed for combining multi-features of images, and evaluated in a multi-feature setting by using the CalTech101 dataset with 8 MPEG-7 visual features. The approach is compared with the retrieval performance of single features, simple combination approaches and exhaustive search approach. Then it is also applied to a multimodal setting by using TRECVID 2007 dataset with 3 visual, 2 audio and 1 textual modalities. Lastly, the proposed approach is utilized for efficient feature selection and combination in a Wireless Video Sensor Networks application.

5.1 Overview

CBIR systems aim to retrieve pictures from large image repositories according to the needs of the users [26]. In CBIR systems, images are usually modeled with a set of low level features, such as color, texture or shape, from which underlying similarity

[1] Section 5.1 through 5.4 of this chapter was published as [151]. Section 5.6 was published as [94].

functions are used to perform queries [2].

The ultimate goal of designing CBIR systems is to achieve the best possible retrieval accuracy. To achieve high accuracy on a retrieval task, traditional approaches prefer creating superior low level features than the currently available ones, or optimization of them [24, 61]. However, the noise in sensed data, non-universality of any single low level feature and performance upper bounds prevent relying on a single feature [98]. Furthermore it has been observed that the sets of patterns misclassified by the different methodologies would not necessarily overlap and complementary information provided by different features improves the performance [61]. In the information fusion literature, fusing multiple features is an empirically validated approach for increasing the retrieval performance [31, 61, 67, 144].

Dealing with multiple features entails processing intrinsic high dimensionality of each feature and handling heterogeneous dimensions / scales of different features. Modeling the CBIR system to operate in feature space (storing image features in the database) makes the system struggle with the heterogeneity of different features and prevents it from being fast and flexible [15], which refer to a fast retrieval operation and the system can handle any new features blindly, respectively. A CBIR operating in feature space is not fast since similarity calculation is done at query-time. Also, it cannot be flexible either, considering that handling a new feature requires renewing the system for processing the dimensionality and scale of the new feature. Therefore, an alternative approach, that regards the fastness and the flexibility issues, is modeling the system in dissimilarity space. In accordance with the ideas of [86, 97] for representing images with dissimilarities, Bruno et al. [16] present fusing multiple features in dissimilarity space. In dissimilarity space, the images in the database are represented with the dissimilarity values to prototype objects of the particular image categories. Thus, the retrieval operation is faster and adding new features to the system is easier as long as the distance function is available at once for processing the images in the database.

Beyond the representation problem of images, another crucial issue is to find out the features that are more beneficial for fusion. This problem, namely the feature selection problem, tries to determine which subset of features yield to an optimal result. In [47], Jain et al. group widely-used techniques with a general aspect of

view: exhaustive search, branch-and-bound search, best individual features, sequential forward/backward selection, sequential forward/backward floating search. The methods except the exhaustive search provide computationally more efficient ways of finding an optimal set, however, exhaustive search guarantees to find the optimal solution. For each of these methods, selection criteria during forward/backward selection operations can differ; information gain, previously-defined quality metrics or the complexity can be a consideration. With a more specific view on the problem, some of the recent approaches in the information fusion literature can be listed as: Finding principal/independent components [64, 143], selecting the most coherent and less complex features according to the heterogeneity issue [63], calculating the information gain obtained [5, 58] and defining quality and reliability metrics on features [98, 120].

Although there are many different approaches for the selection of features, all of them have a common preference: The selection process is independent of the category (semantic meaning) of the images. However, considering the idea that different features can be more effective, representative and discriminative for different image categories, using a category dependent feature selection approach can be more beneficial.

Here, a class-specific feature selection approach for the fusion of multiple features is proposed. In order to eliminate the high-dimensionality of multiple features and provide efficient querying over the images, we prefer a dissimilarity based approach. To learn the class-specific features, we carry out a training phase. During the training, the class-specific features are determined by using the representativeness and discriminativeness of features for each image class. The calculations of representativeness and discriminativeness are based on the statistics on the dissimilarity values of training images.

5.2 Multi-Feature Modeling in Dissimilarity Space

The literature of information fusion agrees on the idea that combining multiple features enhances the efficiency. However, how to combine such information is still a research topic. One of the discussed issues is the representation of images. In feature based representation, an image is usually represented with a multi-dimensional feature vector

and having multiple features causes dealing with multiple of such multi-dimensional feature vectors, each having different dimensions and scales. Handling the complexity of different dimensions and scales of different features makes the CBIR system more dependent on the currently available features and less flexible to new features. In [15], Bruno et al. discuss these issues in detail. Still, a more crucial flaw for feature-based representation is the inefficiency of the fast querying capabilities. Having features in the database requires calculating the similarities of related images for every query task.

A more convenient way is the dissimilarity based representation [15, 32, 86, 97]. In dissimilarity based representation, feature values are not stored in the database; instead the dissimilarity values of images are stored. Thus, the CBIR system does not need to deal with the intrinsic dimensionality of features to combine them. In addition, a query task is simpler; it does not require similarity calculations for each query. The dissimilarity values of images are calculated once, before including the image into the CBIR system. To calculate the dissimilarity values, the dissimilarity functions of each feature are utilized. Hence dissimilarity-based representation is a more flexible and fast way of representing the images in a CBIR system employing multiple features.

In dissimilarity based representation, the dissimilarities between each image couple is not necessary. Instead, the dissimilarities of the images in the image database with prototype images of the system are enough (Figure 5.1). The number of prototype images is quite smaller than the size of the image database. Usually, the prototype images are grouped according to their image classes (semantic meanings of images) in order to meet semantic query requirements. In a multi-feature CBIR system, such distance values between the images in the image database and the prototype images should be stored separately for each feature.

More formally, assuming that $F = \{f_1, f_2, ..f_k\}$ is the set of features available for the CBIR system having k number of features, $C = \{c_1, c_2, .., c_m\}$ is the image database having m number of images, $P = \{P_1, P_2, ..P_n\}$ is the set of prototype image classes containing n number of image classes, each prototype image class is $P_i = \{p_1^i, p_2^i, ..p_t^i\}$ where number of prototype images is t and t is not necessarily the same in all prototype image classes; the multi-feature CBIR system has following

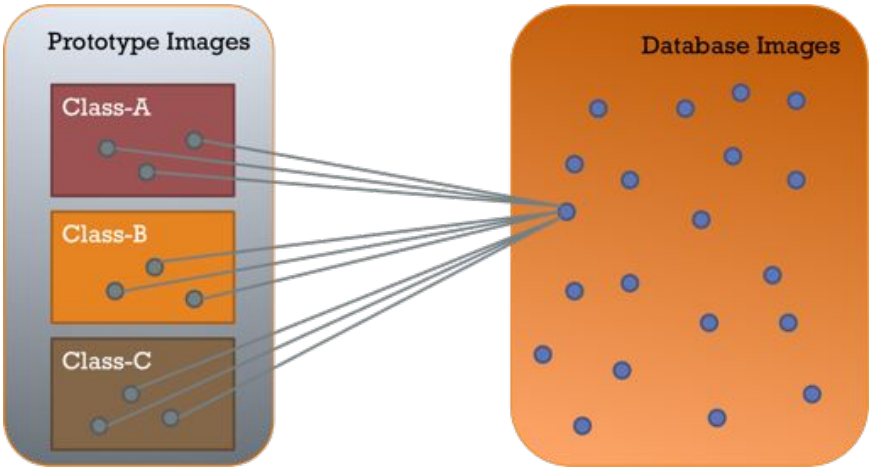


Figure 5.1: Dissimilarity based representation

distance-based representation for each image class i and feature f:

$$D_f^i = \begin{pmatrix} d_f(c_1, p_1^i) & d_f(c_1, p_2^i) & \cdots & d_f(c_1, p_t^i) \\ d_f(c_2, p_1^i) & d_f(c_2, p_2^i) & \cdots & d_f(c_2, p_t^i) \\ \vdots & \vdots & \vdots & \vdots \\ d_f(c_m, p_1^i) & d_f(c_m, p_2^i) & \cdots & d_f(c_m, p_t^i) \end{pmatrix}, \tag{5.1}$$

where $d_f(x, y)$ is the dissimilarity between the database image x and the prototype image y for feature f.

A semantic query (for instance "Find pictures of cars") executed in this kind of CBIR system is handled as follows: The distance matrices of D_f^is are evaluated, where i is the class of 'car' images and $f \in F$. First, for each matrix, prototype aggregation with a predefined algorithm is performed and an aggregated distance vector that represents the distances of all images in the image database to the 'car' semantic image class is obtained. Then k number of distance vectors, each representing a different feature, are combined with a feature selection algorithm. The combination of k number of distance vectors results with a single distance vector which shows the distances of all database images to the 'car' class.

In this study, we propose a class-specific feature selection approach for the feature selection problem stated above. The prototype aggregation problem is beyond the scope of this study. However, two different basic aggregation methods (minimum and average) are utilized during the empirical study in order to see the effect of using

different aggregation techniques.

5.3 Exploiting Class-specific Features

In CBIR systems, as mentioned in Section 1, a particular feature or a common set of features is usually used to compare the query image with the database images. In these systems, the features are selected to represent the problem domain. However, if the size of the database and/or the diversity of image collection is increased, these methods fail to give satisfactory results. Specifically, using the same features for different domains and types of objects yields unsatisfactory results. Finding a solution to the problem is quite simple: using different features for different object types. For example, shape features are more important than color features for a 'car' object whereas a 'sea' object can be defined with color and texture features. Another example is presented in Figure 5.2 visually. A 'ball' object can be in any color but the shape is the 'ball' is always circle. However, a 'sky' object can be in any shape, but is always 'blue'. Besides, both shape and color are important for a 'banana' object.

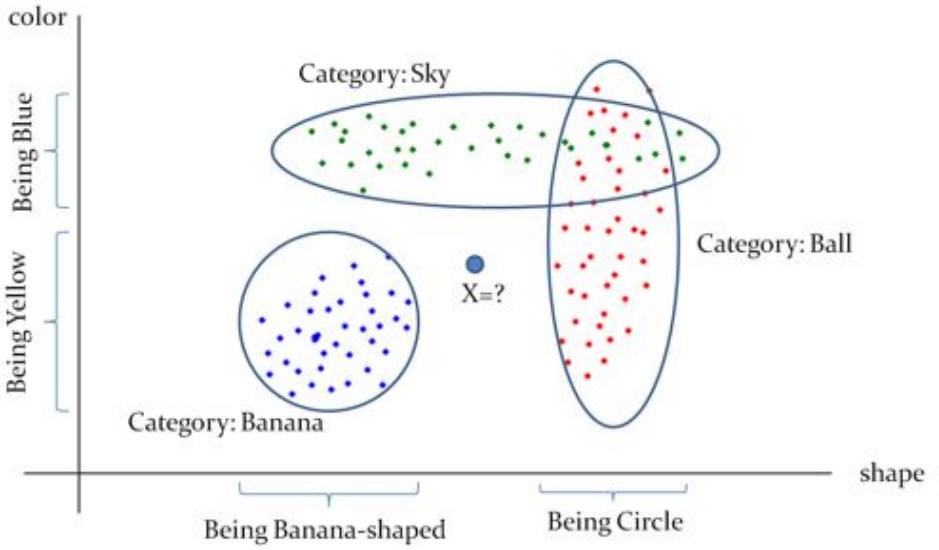


Figure 5.2: Examples for Class-specific Features

To describe the approach more formally, assume an image database having images from 2 semantic classes. It is assumed that class C_1 contains n_1 number of images and C_2 contains n_2 number of images in the database. Also, it is assumed that the images of

class C_1 can be defined better with color features and the images of C_2 can be defined better with shape features. If this database is used in a CBIR system that compares images according to only color features or shape features, the performance of the system is nearly 50% in terms of accuracy. If color features are used, the performance of the system is satisfactory for C_1, but not for C_2. To obtain a satisfactory performance for the whole system, different features should be used for different classes.

By considering this idea, in [134], Uysal et al. utilized an approach identifying the Best Representative Feature (BRF) for each object class, which maximizes the correct match in a training set. Similarly, in [127] Swets et al. propose to use Most Expressive Features and Most Discriminating Features. However, these approaches lack the advantages of fusing multiple features since they select only one feature for each class.

Besides, Jain et al. [45] apply the idea in biometrics domain. They propose combining multiple traits by selecting person-specific traits for recognition. However, they do not propose a feature selection methodology. They obtain the person-specific traits after an exhaustive search process on the training data.

In this study, we propose a class-specific feature selection mechanism by finding out the representative and discriminative features for each image class. Representative characteristics of features are calculated according to the dissimilarities of images within the same class, and discriminative characteristics are calculated according to the ability of features to distinguish between different image classes. Using these characteristics, the importance values of features for each image class are calculated as detailed below. The importance values of features for each category are also called the Class-Specific Features (CSF) index. The mechanism is based on statistical calculations over the dissimilarity values of all prototype images. Providing such prototype images can be considered as the training phase of the CBIR system. The CSF indices are used as the weights of the features during feature combination process.

5.3.1 Calculation of CSF Indices

To calculate the CSF indices, firstly the dissimilarity values of prototype images to each other is calculated and a dissimilarity matrix is obtained as $D_f^i(P)$ for each f,

similar to the one given in Section 2. Differently, $D_f^i(P)$ includes dissimilarities of prototype images in image class i to all prototype images of all image classes. $D_f^i(P)$ contains $n \cdot t$ rows and t columns.

$$D_f^i(P) = \begin{pmatrix} d_f(p_1^1, p_1^i) & d_f(p_1^1, p_2^i) & \cdots & d_f(p_1^1, p_t^i) \\ d_f(p_2^1, p_1^i) & d_f(p_2^1, p_2^i) & \cdots & d_f(p_2^1, p_t^i) \\ \vdots & \vdots & \vdots & \vdots \\ d_f(p_t^1, p_1^i) & d_f(p_t^1, p_2^i) & \cdots & d_f(p_t^1, p_t^i) \\ \vdots & \vdots & \vdots & \vdots \\ d_f(p_1^n, p_1^i) & d_f(p_1^n, p_2^i) & \cdots & d_f(p_1^n, p_t^i) \\ d_f(p_2^n, p_1^i) & d_f(p_2^n, p_2^i) & \cdots & d_f(p_2^n, p_t^i) \\ \vdots & \vdots & \vdots & \vdots \\ d_f(p_t^n, p_1^i) & d_f(p_t^n, p_2^i) & \cdots & d_f(p_t^n, p_t^i) \end{pmatrix}. \tag{5.2}$$

After obtaining the dissimilarity matrices $D_f^i(P)$ for each feature and image class, dissimilarity values of each image category in each matrix are aggregated both column-wise and row-wise. Thus, the mean and standard deviation vectors are obtained as follows;

$$\mu(D_f^i(P)) = \begin{bmatrix} \mu_f^{i,1} \mu_f^{i,2} & \cdots & \mu_f^{i,n} \end{bmatrix}^T, \tag{5.3}$$

$$\sigma(D_f^i(P)) = \begin{bmatrix} \sigma_f^{i,1} & \sigma_f^{i,2} & \cdots & \sigma_f^{i,n} \end{bmatrix}^T. \tag{5.4}$$

Here, $\mu_f^{i,j}$ denotes the mean of dissimilarities from all images in class i to all images in class j for feature f. Also, $\sigma_f^{i,j}$ denotes the corresponding standard deviation.

To obtain the CSF indices, four important parameters are extracted from the above given vectors of $\mu(D_f^i(P))$ and $\sigma(D_f^i(P))$:

- Mean of Class ($\mu_f^{i,i}$): $\mu_f^{i,i}$ is the average dissimilarity value of a class to itself, for a particular feature f. Mean of Class is a representative characteristic for features. For a selected class, the features with lower dissimilarity values represent the image class better. Thus, the CSF index is inversely proportional to the mean of the category.

- Standard Deviation of Class ($\sigma_f^{i,i}$): $\sigma_f^{i,i}$ is another important representative property. For any class, a feature with small standard deviation entails close image-to-image dissimilarity values within the class. Such a feature can be

considered as a better feature. Thus, the CSF index is inversely proportional to the standard deviation of an image class.

- Standard Mean Distance to Other Classes (δ_f^i): Standard mean distance to other classes is a discriminative feature which is calculated by using the dissimilarities of a class to other classes. It is calculated as follows:

$$\delta_f^i = \sqrt{\frac{\sum_{j=1}^{n} (\mu_f^{i,i} - \mu_f^{i,j})^2}{n}} \,, \tag{5.5}$$

  where n is the number of image classes. This calculation gives us the average dissimilarity of an image class i to all other classes. Thus, having a greater dissimilarity means better discrimination among all categories, which means that the CSF index is directly proportional to δ_f^i.

- Correctness Ratio (ω_f^i): Although the three parameters given above are important and provide good representation and discrimination, the issue of correctness of the feature is not considered. It is important for a feature to give the lowest dissimilarity values for the images in a class which is the same with the class of the query images. Correctness ratio of a particular feature f can be defined as what percentage of the means in a $\mu(D_f^i(P))$ vector are larger than the mean value of the class i ($\mu_f^{i,i}$). As the correctness ratio decreases, the representation ability decreases, which means that the CSF index is directly proportional with the correctness ratio.

Considering the effects of the above parameters, the CSF index of a particular feature f on a particular image class i is calculated using the formula below;

$$CSF_f^i = \frac{(1 - \mu_f^{i,i}) \cdot \delta_f^i \cdot \omega_f^i}{\sigma_f^{i,i}} \,. \tag{5.6}$$

5.3.2 Normalization on Dissimilarities

As mentioned before, CBIR system having dissimilarity-based representation does not need to deal with the intrinsic dimensionality of features to combine them. However, different scales of different features are still a problem to be solved. Different scales of the values contained in the features causes dissimilarity values to be in different scales.

In the literature, there are several normalization methods to handle the different scales of multiple features [46]: Min-max, decimal scaling, z-score, median, double sigmoid, tanh estimators, bi-weight estimators. In [46], Jain et al. empirically show that min-max, z-score and tanh estimators methods are superior. Also they note that the simplest method (min-max) would suffice when the minimum and maximum values are known. Min-max normalization transforms values from a known (or estimated) range $[min, max]$ into $[0, 1]$ range with the following basic formulation: $x' = (x - min)/(max - min)$. Considering that we have the prototype images and dissimilarity values of prototype images to themselves, it is easy to find the minimum and maximum dissimilarity values for each feature. Thus the min-max normalization approach is preferred in this study.

5.4 Evaluation of CSF in Multi-Feature Setting

To demonstrate the validity of the proposed approach, a number of experiments are carried out. For the experiments, the CalTech 101 image dataset [34] is used. It contains pictures of objects belonging to 101 categories. During the tests, all of the 101 classes in the dataset are used. Randomly selected 10 images for each class, hence a total of 1010 images, are treated as the prototype images. For the query purposes, randomly selected 20 images for each class and a total of 2020 images are employed the image database. In addition, as the features to be combined, 8 visual features of MPEG-7 [78] in three types are utilized: Color descriptors of Color Layout(CL), Color Structure(CS), Dominant Color(DC), Scalable Color(SC); Shape descriptors of Contour Shape(CSh), Region Shape(RS); Texture descriptors of Edge Histogram(EH), Homogeneous Texture(HT). The dissimilarities of the images for these features are calculated by using the MPEG-7 reference software (eXperimentation Model, XM) [83].

The tests are mainly performed on semantic retrieval of images; the semantic classes are queried over the image database. The images are fetched and sorted according to the dissimilarity values. To measure the retrieval accuracy, *Precision*, *Recall*, *Average Precision(AP)* and *Mean Average Precision(MAP)* metrics are used. *Precision* is the fraction of retrieved images that are relevant to the search, whereas *Recall* is the ratio

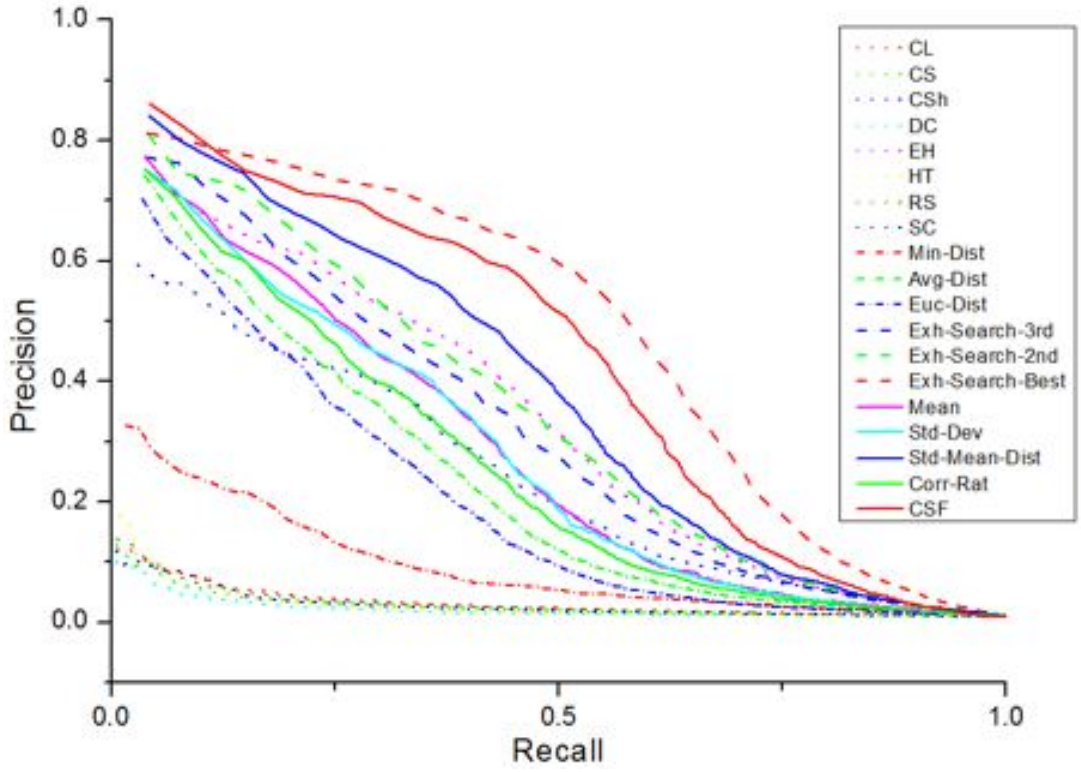


Figure 5.3: Precision-Recall Graph for Semantic Retrieval

of the number of relevant images retrieved to the total number of relevant images in the collection. The *AP* is the sum of the precision at each relevant hit in the retrieved list, divided by the minimum between the number of relevant documents in the collection and the length of the list. Considering that image collection in our test contains 2020 images, AP is measured at 2020. *MAP* is the *AP* averaged over several image classes. In other words, the *AP* of each image class is calculated separately, then the *MAP* is found by averaging them.

As the primary test, the accuracy of the proposed method on semantic retrieval is measured. In order to perform a detailed comparison, this test is executed in four steps. As the first step, the retrieval accuracies of each single feature is calculated. For the second step, following simple combination approaches are tested: Minimum Distance(MD), Average Distance(AD), Euclidean Distance(ED). The combined dissimilarity is obtained by selecting the minimum dissimilarity (distance) in MD, averaging all available dissimilarities in AD and calculating an Euclidean distance on the available dissimilarities in ED. For the third step, feature selection by an exhaustive search approach is

applied and the combined dissimilarity is calculated by averaging the dissimilarities of resultant features from the feature selection. An exhaustive search for feature selection requires calculating all combinations of available features, 2^8 cases in total for our test. Considering that performing an exhaustive search during each query is not applicable due to the time cost, the selection process is executed once on the prototype images. Then, 10 best selections (ES[1-10]) are found and semantic retrieval test is performed for each of these 10 feature selections. As the last step, the approach proposed in this study is performed for feature selection. Calculated CSF indices are used to combine the dissimilarity values with a weighted-sum approach. Not only the CSF index, but also the four parameters of the CSF are tested separately in order to see which one is more influential. In Figure 5.3, the Precision-Recall graphs of these methods are given. In addition, the AP of some sample categories, MAP of Best 10, 20, 50 and all 101 categories are presented in Table 5.1. Also, how many times each method has the best score and mean ranks of each method are included in the table. The results given in the table are visualized in Figure 5.4.

Considering the test results, it is observed that obtaining an increase in the accuracy requires a good selection on the features. Simple methods like MD, AD and ED are not enough for selection. MD lacks the advantages of combining multiple features whereas AD and ED always combine all of the features and are affected by the unfavorable features. Besides, the exhaustive search guarantees to find the optimal feature selection by evaluating all possible combinations. Therefore ES1 outperforms the other methods. However the ES[1-10] ranking obtained at the training phase is not the same during the querying. For instance, ES5 performs better than ES2, ES3 and ES4. Such situation is caused by difference between training and query images. Although it is not observed in this test conditions, it could be possible that the best combination obtained during the training phase do not give best results during querying. It is possible to handle such incompliance by executing the exhaustive search during each query, but it causes time inefficiency.

On the other hand, our proposed method of CSF gives successful accuracy results that are very close to the best selection in total and even better for one fourth of the image classes. Regarding that the results of the best selection in ES can be considered as the upper-bound for the retrieval task, the CSF method can be qualified as a robust

Table 5.1: Semantic Query Results

		electric guitar	saxophone	inline skate	stop sign	revolver	Best-10	Best-20	Best-50	All-101	# of Best	Mean Rank
Single Features	CL	0.013	0.035	0.127	0.376	0.023	0.406	0.259	0.129	0.073	0	22.6
	CS	0.028	0.135	0.037	0.388	0.021	0.361	0.241	0.119	0.066	0	23.0
	CSh	0.865	0.766	0.725	0.253	0.361	0.841	0.743	0.542	0.339	3	13.9
	DC	0.020	0.033	0.055	0.427	0.019	0.258	0.171	0.088	0.050	0	23.6
	EH	0.895	0.874	0.633	0.827	0.928	0.924	0.855	0.667	0.424	6	10.1
	HT	0.006	0.063	0.159	0.235	0.029	0.304	0.210	0.112	0.063	0	23.0
	RS	0.097	0.120	0.153	0.624	0.114	0.354	0.233	0.121	0.070	0	22.4
	SC	0.016	0.065	0.057	0.654	0.064	0.352	0.229	0.116	0.066	1	22.8
Simple	MD	0.401	0.253	0.190	0.253	0.733	0.703	0.550	0.318	0.176	0	19.1
	AD	0.029	0.614	0.734	0.863	0.615	0.813	0.716	0.493	0.310	1	14.2
	ED	0.014	0.563	0.704	0.792	0.542	0.766	0.677	0.457	0.284	0	15.9
Exh. Search	ES1	**0.958**	0.964	0.870	0.856	0.970	0.963	0.927	**0.806**	**0.563**	**36**	**5.0**
	ES2	0.841	0.919	0.763	0.917	0.830	0.895	0.820	0.630	0.418	3	9.6
	ES3	0.565	0.951	0.831	0.985	0.898	0.923	0.828	0.616	0.400	1	11.1
	ES4	0.928	0.960	0.806	0.880	0.797	0.923	0.862	0.693	0.459	5	8.0
	ES5	0.934	0.916	0.844	0.910	**0.973**	0.927	0.872	0.704	0.484	7	6.9
	ES6	0.815	0.885	0.720	0.811	0.794	0.867	0.780	0.609	0.405	4	10.5
	ES7	0.641	0.968	**0.911**	0.981	0.959	0.932	0.855	0.663	0.441	6	8.7
	ES8	0.587	0.916	0.785	0.935	0.854	0.896	0.797	0.594	0.387	2	11.3
	ES9	0.578	0.972	0.844	0.979	0.852	0.927	0.845	0.634	0.409	3	10.6
	ES10	0.746	0.942	0.886	0.981	0.841	0.926	0.864	0.714	0.482	8	7.1
Proposed	μ	0.583	0.700	0.762	0.893	0.653	0.834	0.747	0.556	0.359	2	12.3
	σ	0.174	0.878	0.786	0.959	0.803	0.876	0.787	0.567	0.362	1	11.5
	δ	0.867	0.887	0.835	0.961	0.959	0.942	0.866	0.683	0.458	4	7.4
	ω	0.315	0.617	0.734	0.862	0.640	0.817	0.722	0.518	0.333	1	13.1
	CSF	0.955	**0.981**	0.889	**0.987**	0.957	**0.970**	**0.928**	0.769	0.521	24	5.5

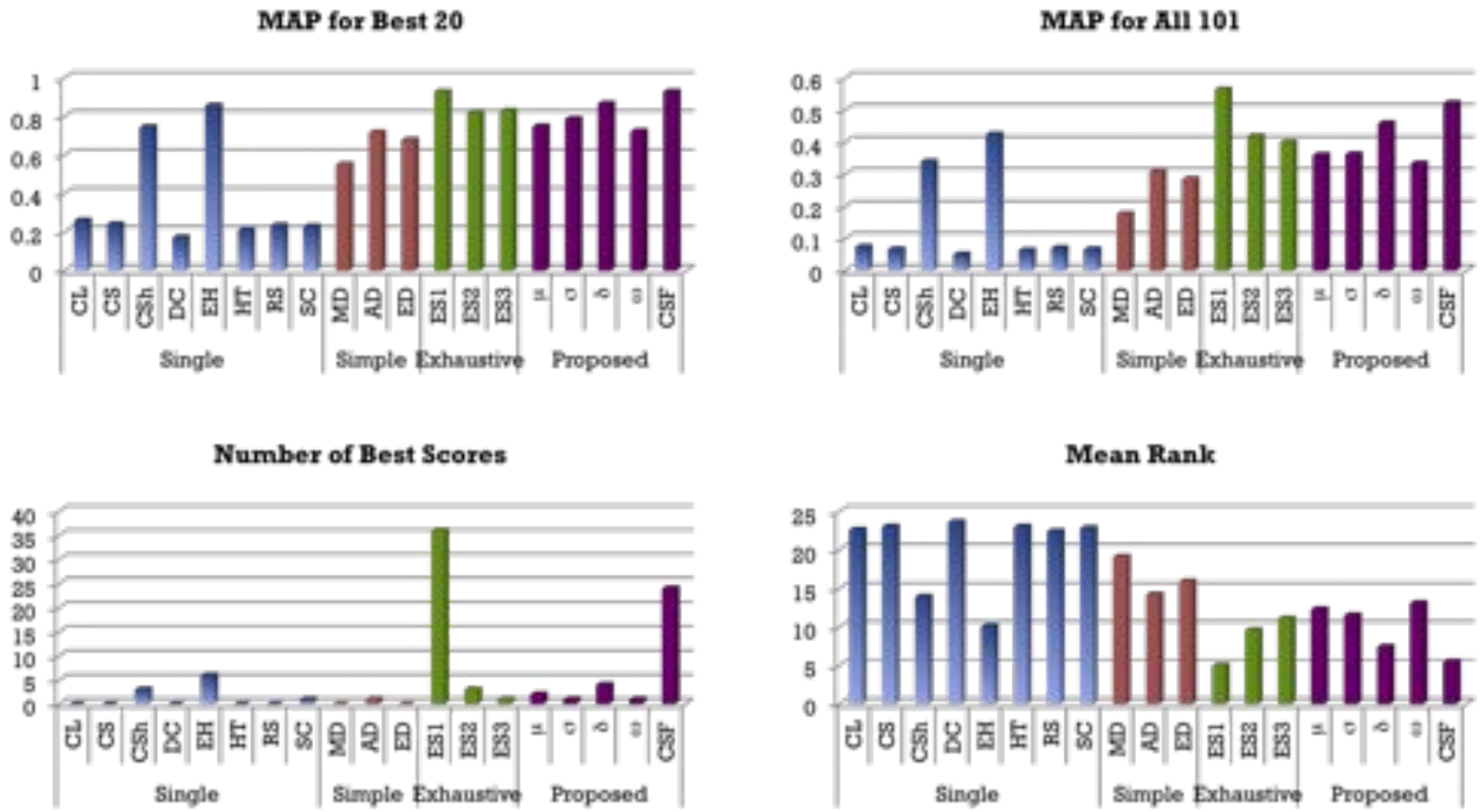


Figure 5.4: Evaluation of Semantic Query Results

and successful approach. In addition, our claim of exploiting class-specific features can be supported by the results of ES method. Different feature combinations in ES selections perform better in different image classes, which results different classes requires the use of different features.

Another observation on the results is the superiority of δ parameter of CSF approach among other parameters. Therefore, it can be stated that the discriminativeness characteristics of features are more effective than the representativeness.

An important discussion for combining multiple features is the independency of features. Using complementary features with the methods requiring independent inputs can cause a decrease in the accuracies. Therefore, many studies exist in the information fusion literature that performs an independence analysis [64]. In this empirical study, the features utilized are not fully independent. It is previously stated that simple methods like MD, AD and ED are not successful enough for the selection task. One important reason in their inefficiency is the fact that they cannot eliminate complementary information and the violation of independence assumption decreases their performance. However, the ES and CSF approaches enable selecting different combinations and eliminates complementary features.

As mentioned in Section 2, a prototype aggregation is necessary to combine the dissimilarities of multiple prototypes. Although prototype aggregation is beyond our scope, a secondary test is performed to show the effect of prototype aggregation. During the first test, *averaging* is used for aggregation. In this test, the previous test is repeated with a *minimum* aggregation method. The comparison of two methods is given in Figure 5.5. It is clearly shown that *averaging* is superior than *minimum*. However, these two are very simplistic methods and there are better ways of exploiting the information included in the prototypes.

As the last test, the time complexities of our proposed method and exhaustive search are compared. The query execution times of these two approaches are quite the same since querying includes only a weighted/unweighted summation of several features. However, the execution times for the training phases, which are carried out in order to find out the optimal set of features, differ much. Time complexity of exhaustive search is $O(m^2 \cdot 2^n)$ where m is the total number of prototype images and n is the

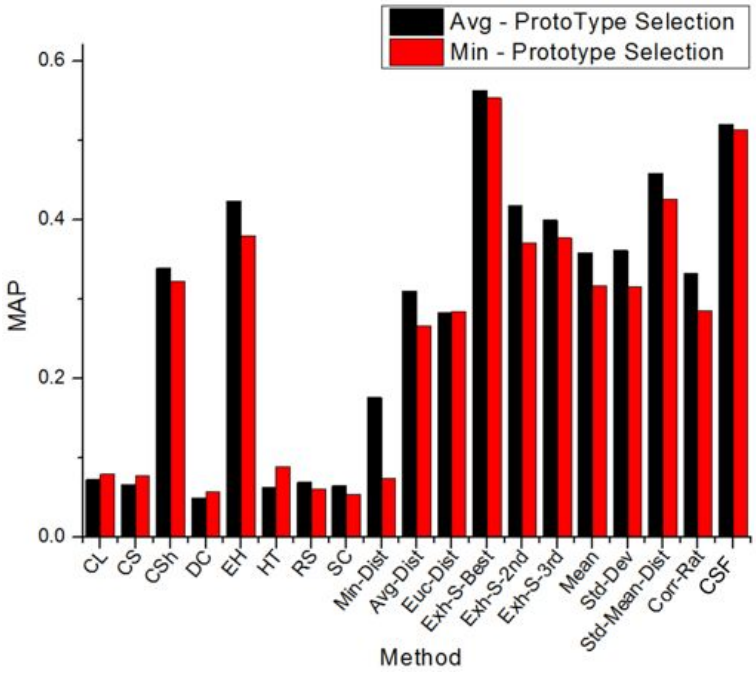


Figure 5.5: Comparison of Average and Minimum Aggregation Methods

number of features. Whereas, time complexity of our proposed method is $O(m^2 \cdot n)$. Time-measurements obtained in this test validated these theoretical definitions. Results are given in Table 5.2. The results show us that CSF approach is 50 times better than the ES approach, in our case. If the number of features increases, execution time for ES could be worse.

Table 5.2: Execution Times for Training Phases

	Total Execution Time
Exhaustive Search	1,049,652 msec
CSF Calculation	19,802 msec

5.5 Evaluation of CSF in Multimodal Setting

After evaluating the CSF mechanism in multi-feature setting and resulting that it is a timely-efficient, accurate and robust way of feature selection, it is decided to evaluate the validity of the CSF mechanism with a multi-modal setting. So, some additional experiments are performed with a multi-modal setting on multimedia data. In this

section, the tests with a multimodal setting are presented in detail.

5.5.1 Test Setup

Evaluations on multimodal setting are based on the international benchmark for video information retrieval TRECVID. TRECVID is a popular workshop in video information retrieval, proposing a large corpora of videos which are manually annotated. Considering that the shot segmentations and labeling are available, video datasets provided by TRECVID are attractive widely-used. In our tests, TRECVID 2007 is considered [91]. TRECVID 2007 corpus is composed of 100 hours of multilingual video, roughly equally divided into training and test sets. The development data comprises 110 videos and 30.6 GB, whereas the test data is 109 files and 29.2 GB. The annotations on the TRECVID 2007 dataset is provided in a multi-label manner, which means each shot can contain more than one label. The distribution of shots according to labeled concepts is presented in Table 5.3. A performance comparison of TRECVID 2007 participants and further details can be found in [91].

In our setup, for shot segmentation, the outputs of common shot reference is used as the video shots. The dataset contains 21,532 reference shots for training and 18,142 reference shots for test. In the experiments, we used the 20 semantic concepts which were selected in TRECVID 2007 evaluation. During the tests, the shots are considered as individual and independent documents, which no contextual information or interaction is taken into account between shots.

Considering a multi-modal setting; visual, audio and textual features are extracted from the videos. For visual features, one key frame per shot is adopt and the middle frame for each shot is selected as the key frame. For audio features, entire audio of each shot is processed. For the textual features, the automatic speech recognition (ASR) and Machine Translation (MT) texts, which are provided by TRECVID, are employed.

For visual modalities, 8 visual features of MPEG-7 [78] in three types are utilized: Color descriptors of Color Layout(CL), Color Structure(CS), Dominant Color(DC), Scalable Color(SC); Shape descriptors of Contour Shape(CSh), Region Shape(RS);

Table 5.3: Shot counts for each concept type in TRECVID 2007 dataset

	Training			Test			Total
Class	Multi-labelled	Single-labelled	Total	Multi-labelled	Single-labelled	Total	
Airplane	19	39	58	23	124	147	410
Animal	140	705	845	27	224	251	2192
Boat_Ship	237	60	297	56	110	166	926
Car	141	531	672	103	332	435	2214
Charts	60	55	115	3	61	64	358
Computer_TV-screen	232	297	529	29	177	206	1470
Desert	25	42	67	5	21	26	186
Explosion_Fire	9	37	46	8	44	52	196
Flag-US	7	5	12	2	4	6	36
Maps	25	92	117	4	89	93	420
Meeting	221	521	742	34	673	707	2898
Military	205	225	430	11	30	41	942
Mountain	45	79	124	21	75	96	440
Office	337	794	1131	37	173	210	2682
People-Marching	69	201	270	14	58	72	684
Police_Security	147	108	255	23	66	89	688
Sports	19	263	282	22	102	124	812
Truck	79	47	126	88	128	216	684
Waterscape_Waterfront	380	522	902	80	209	289	2382
Weather	25	9	34	1	5	6	80
Total	2422	4632	7054	591	2705	3296	20700

Texture descriptors of Edge Histogram(EH), Homogeneous Texture(HT). These features are grouped in three modalities according to the type they belong. The feature extraction and distance calculation tasks are performed by using the MPEG-7 reference software (eXperimentation Model, XM) [83].

As the audio features, Linear Predictor Coefficients(LPC), Zero Crossing Rate (ZCR), Energy and Mel-frequencies cepstrum coefficients (MFCC) are used. The features are included into the test as two modalities according to their dimensionalities; LPC, ZCR and Energy in one modality and MFCC in another modality. The feature extraction is performed by using the Yaafe toolbox [12]. The distance measure used is the Euclidean distance.

For the textual modality, the term frequency inverse document frequency (TF-IDF) weights [112] are calculated as features. During calculation, no stop-word filtering or preprocessing is done. For the distance calculation, Cosine similarity metric is used.

Therefore, the final list of modalities as follows: Visual-color, Visual-shape, Visual-texture, Audio-Simple, Audio-Complex, Textual.

Similar to the evaluation in Section 5.4, the tests are performed on semantic retrieval of images; the semantic classes are queried over the video (shot) database. The shots are fetched and sorted according to the similarity values. The similarities are calculated in three different ways of prototype aggregation:

- *minimum* prototype aggregation, by using 1 prototype instance for each class
- *k-minimum* prototype aggregation, by using 20 prototype instances for each class
- *averaging* prototype aggregation, by using all prototype instances of each class

It should be noted that the 1-prototype and all-prototype configurations correspond to the *minimum* and *averaging* prototype selection/aggregation approaches evaluated in Section 5.4, respectively. So, here a new prototype selection approach is also evaluated: *k-minimum* (or prototype selection with k=20. In our tests, a separate classifier is created for each modality, so a total of 6 classifiers are used.

To measure the retrieval accuracy, *Average Precision(AP)* and *Mean Average Precision(MAP)* metrics are used. The *AP* is the sum of the precision at each relevant hit in the retrieved list, divided by the minimum between the number of relevant documents in the collection and the length of the list. Regarding the evaluation rules of TRECVID, AP is measured at 2000. *MAP* is the *AP* averaged over several image classes. In other words, the *AP* of each image class is calculated separately, then the *MAP* is found by averaging them.

In order to perform a detailed comparison, the tests of each classifier configuration are executed in four steps. As the first step, the retrieval accuracies of each single feature is calculated. For the second step, modality selection by an exhaustive search approach is applied and the combined dissimilarity is calculated by averaging the dissimilarities of resultant features from the modality selection. An exhaustive search for modality selection requires calculating all combinations of available modalities, 2^6 cases in total for our test. The tests are performed for each combination. It should be noted that, exhaustive search is performed in a way that the modality selection is done independent of the classes and selected modalities are applied on all of the classes. For the third step, a well-known and widely used feature selection algorithm, RELIEF-F [66], is tested. As the last step, the CSF approach is performed for modality selection. Not only the CSF index, but also the four parameters of the CSF are tested separately in order to see which one is more influential. Different from the CSF evaluations in Section 5.4, the CSF formulation is updated as follows;

$$CSF_f^i = \frac{(1-\mu_f^{i,i}) \cdot (\delta_f^i)^v \cdot \omega_f^i}{\sigma_f^{i,i}} \,, \tag{5.7}$$

and the effect of different v values is observed.

As the fusion approach, a late fusion with a simple linear weighting approach is preferred for simplicity. Thus, during the exhaustive search, RELIEF-F and CSF steps, calculated weights are used to combine the similarity/dissimilarity values with a weighted-sum approach.

5.5.2 Test Results

Table 5.4, Table 5.5 and Table 5.6 present the AP of all concept types and the MAP

values, for each classifier configuration listed above. Also, in Table 5.7, a general comparison of all tests are given in terms of MAPs. In Figure 5.6, a visual comparison of MAP values for each classifier configuration is presented. In Figure 5.7, the effect of v parameter in CSF formula is illustrated.

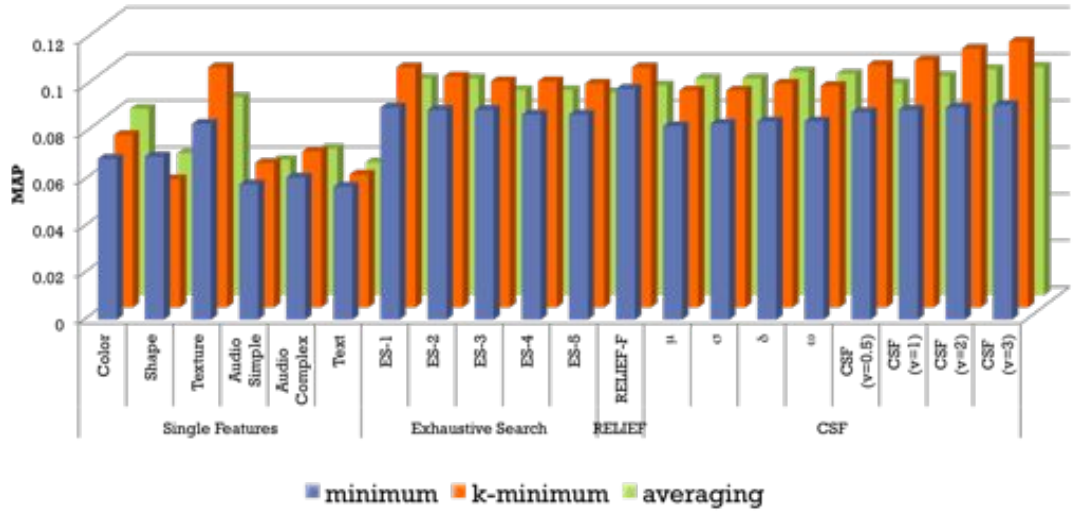


Figure 5.6: Comparison of MAPs for each classifier configuration

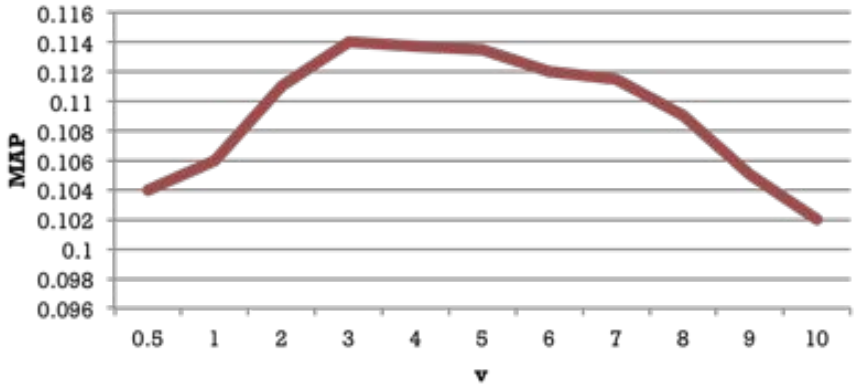


Figure 5.7: Effect of v on retrieval

5.5.3 Evaluation and Discussion

Considering presented test result, following evaluations can be done:

- Combination of different modalities give better results than the single modalities.

Table 5.4: Semantic Query Results (*minimum* prototype aggregation)

	Single Features						Exhaustive Search					Relief-F	CSF							
	Visual Color	Visual Shape	Visual Texture	Audio Simple	Audio Complex	Textual	ES-1	ES-2	ES-3	ES-4	ES-5	Relief-F	Mean	StdDev	MeanDist	CorrRat	CSF (v=0.5)	CSF (v=1)	CSF (v=2)	CSF (v=3)
Airplane	0.080	0.330	0.094	0.036	0.029	0.023	0.068	0.088	0.062	0.048	0.053	**0.377**	0.042	0.045	0.046	0.046	0.050	0.052	0.072	0.089
Animal	0.082	0.101	**0.142**	0.071	0.094	0.120	0.132	0.137	0.114	0.109	0.132	0.127	0.123	0.127	0.123	0.126	0.127	0.128	0.127	0.121
Boat_Ship	0.051	0.077	**0.092**	0.066	0.067	0.045	0.074	0.071	0.075	0.078	0.075	0.079	0.063	0.068	0.066	0.066	0.069	0.068	0.071	0.076
Car	0.140	0.072	0.162	0.135	0.163	0.149	0.169	0.170	**0.193**	0.189	0.169	0.163	0.170	0.173	0.172	0.173	0.175	0.178	0.183	0.187
Charts	0.047	0.027	0.032	0.025	**0.048**	0.026	0.041	0.044	0.042	0.041	0.040	0.032	0.021	0.020	0.025	0.021	0.025	0.026	0.028	0.030
Computer_TV-screen	0.092	0.082	0.099	0.082	0.074	0.068	0.125	0.119	0.096	0.096	0.124	**0.158**	0.140	0.142	0.133	0.134	0.135	0.131	0.131	0.136
Desert	0.012	0.005	0.020	0.020	0.008	0.008	0.022	0.027	0.016	0.015	0.019	0.020	0.019	**0.029**	0.015	0.015	0.022	0.014	0.009	0.008
Explosion_Fire	0.026	0.032	0.030	0.028	0.016	0.012	0.056	0.027	**0.074**	0.063	0.055	0.031	0.014	0.013	0.013	0.016	0.014	0.014	0.014	0.013
Flag-US	0.006	0.000	0.003	0.002	0.002	**0.024**	0.005	0.008	0.007	0.005	0.005	0.001	0.002	0.002	0.003	0.002	0.003	0.003	0.004	0.005
Maps	0.039	0.049	**0.078**	0.031	0.070	0.036	0.041	0.049	0.046	0.037	0.041	0.033	0.021	0.022	0.018	0.037	0.038	0.038	0.038	0.039
Meeting	0.312	0.309	0.324	0.264	0.242	0.247	0.403	0.375	0.427	**0.439**	0.403	0.365	0.410	0.388	0.392	0.398	0.407	0.412	0.405	0.392
Military	0.041	0.020	0.023	0.021	0.008	0.011	0.061	0.054	0.025	0.028	0.062	0.022	0.039	0.037	0.052	0.039	0.046	0.056	0.076	**0.097**
Mountain	0.024	**0.068**	0.054	0.023	0.026	0.036	0.032	0.038	0.029	0.028	0.032	0.054	0.031	0.037	0.039	0.042	0.037	0.038	0.041	0.043
Office	0.064	0.023	0.086	0.089	0.098	0.071	0.096	0.088	0.116	0.122	0.097	0.023	0.116	0.116	0.114	0.111	**0.124**	0.122	0.118	0.113
People-Marching	0.055	0.015	0.071	0.028	0.023	0.028	0.105	0.110	0.086	0.085	0.106	0.132	0.099	0.110	0.138	0.122	0.146	**0.153**	0.150	0.136
Police_Security	0.062	0.029	0.051	0.028	0.029	0.029	0.057	0.062	**0.070**	0.068	0.058	0.067	0.065	0.060	0.054	0.059	0.069	0.067	0.060	0.064
Sports	0.035	0.016	0.045	0.034	0.037	**0.063**	0.042	0.036	0.037	0.039	0.042	0.015	0.039	0.047	0.050	0.039	0.036	0.035	0.033	0.034
Truck	0.075	0.031	0.095	0.060	0.079	0.060	0.089	0.092	0.091	0.087	0.089	**0.105**	0.084	0.088	0.087	0.088	0.092	0.092	0.092	0.092
Waterscape_Waterfront	0.113	0.110	**0.171**	0.103	0.110	0.076	0.153	0.152	0.161	0.156	0.154	0.166	0.156	0.157	0.150	0.168	0.164	0.158	0.156	0.153
Weather	0.018	0.001	0.017	0.004	0.002	0.002	0.050	**0.060**	0.029	0.037	0.008	0.002	0.003	0.004	0.005	0.006	0.012	0.016	0.012	0.008
MAP	0.069	0.070	0.084	0.058	0.061	0.057	0.091	0.090	0.090	0.088	0.088	**0.099**	0.083	0.084	0.085	0.085	0.089	0.090	0.091	0.092
MAP Rank	17	16	13	19	18	20	3	5	7	9	10	**1**	15	14	12	11	8	6	4	2
Number of Best Scores	0	1	**4**	0	1	2	0	1	3	1	0	3	0	1	0	0	1	1	0	1
Mean Rank	12.8	14.5	8.0	16.6	14.8	16.0	8.1	7.5	**7.4**	8.8	8.5	9.2	11.9	10.6	11.6	10.5	8.2	8.0	8.5	8.7

Table 5.5: Semantic Query Results (*k-minimum* prototype aggregation)

	Single Features						Exhaustive Search					Relief-F	CSF							
	Visual Color	Visual Shape	Visual Texture	Audio Simple	Audio Complex	Textual	ES-1	ES-2	ES-3	ES-4	ES-5	Relief-F	Mean	StdDev	MeanDist	CorrRat	CSF (v=0.5)	CSF (v=1)	CSF (v=2)	CSF (v=3)
Airplane	0.103	0.032	0.114	0.034	0.030	0.023	0.114	0.103	0.123	0.108	0.048	**0.377**	0.050	0.056	0.082	0.064	0.094	0.114	0.147	0.159
Animal	0.067	0.108	0.181	0.055	0.171	0.095	0.181	0.160	0.120	0.117	**0.199**	0.124	0.106	0.107	0.100	0.106	0.110	0.109	0.105	0.099
Boat_Ship	0.055	0.059	0.095	0.062	0.072	0.044	0.095	0.088	0.080	0.079	0.077	**0.101**	0.078	0.090	0.082	0.080	0.082	0.082	0.086	0.090
Car	0.136	0.080	0.217	0.165	0.172	0.137	0.217	0.206	0.187	0.198	0.212	0.172	0.222	0.231	0.229	0.231	**0.231**	0.228	0.222	0.209
Charts	0.027	0.019	0.030	0.025	0.019	0.028	0.030	**0.031**	0.026	0.020	0.020	0.030	0.011	0.011	0.014	0.011	0.015	0.020	0.024	0.028
Computer_TV-screen	0.077	0.081	0.110	0.063	0.078	0.077	0.110	0.101	0.108	0.107	0.116	**0.124**	0.110	0.097	0.108	0.101	0.108	0.107	0.105	0.103
Desert	0.013	0.009	0.028	0.004	0.006	0.007	0.028	0.017	**0.039**	0.027	0.015	0.004	0.007	0.007	0.007	0.007	0.005	0.005	0.005	0.005
Explosion_Fire	0.012	0.018	0.018	0.011	0.016	0.010	0.018	0.015	0.022	0.013	0.017	**0.031**	0.012	0.011	0.013	0.013	0.012	0.012	0.013	0.013
Flag-US	0.005	0.002	0.002	0.001	0.003	**0.013**	0.002	0.000	0.001	0.006	0.002	0.001	0.002	0.003	0.001	0.002	0.003	0.002	0.001	0.001
Maps	0.024	0.032	0.157	0.037	0.052	0.035	0.157	**0.166**	0.045	0.040	0.136	0.032	0.024	0.031	0.027	0.058	0.087	0.084	0.097	0.095
Meeting	0.335	0.368	0.365	0.362	0.270	0.268	0.365	0.422	0.296	0.396	0.381	0.407	**0.494**	0.463	0.474	0.474	0.489	0.493	0.486	0.463
Military	0.072	0.015	0.060	0.022	0.007	0.009	0.060	0.040	**0.133**	0.128	0.064	0.033	0.039	0.043	0.072	0.046	0.049	0.060	0.075	0.091
Mountain	0.029	0.024	0.057	0.022	0.030	0.047	0.057	**0.058**	0.042	0.040	0.051	0.057	0.030	0.032	0.033	0.034	0.039	0.041	0.044	0.046
Office	0.087	0.031	0.122	0.088	0.122	0.084	0.122	0.134	0.064	0.110	0.125	0.023	0.138	0.120	0.116	0.112	0.157	0.174	**0.181**	0.177
People-Marching	0.064	0.020	0.090	0.033	0.026	0.031	0.090	0.082	0.156	0.131	0.101	0.157	0.154	0.178	**0.193**	0.175	0.188	0.193	0.179	0.167
Police_Security	**0.065**	0.028	0.034	0.031	0.029	0.033	0.034	0.039	0.042	0.051	0.037	0.059	0.061	0.051	0.050	0.052	0.064	0.064	0.063	0.061
Sports	0.041	0.017	0.037	0.038	0.039	**0.064**	0.037	0.041	0.040	0.037	0.035	0.015	0.055	0.048	0.046	0.044	0.036	0.034	0.032	0.032
Truck	0.065	0.055	**0.127**	0.075	0.088	0.053	**0.127**	0.102	0.091	0.102	0.105	0.126	0.094	0.098	0.099	0.098	0.102	0.104	0.106	0.105
Waterscape_Waterfront	0.125	0.105	**0.198**	0.096	0.110	0.078	**0.198**	0.174	0.174	0.160	0.184	0.183	0.165	0.174	0.163	0.195	0.189	0.186	0.185	0.182
Weather	0.086	0.002	0.007	0.014	0.004	0.001	0.007	0.002	**0.155**	0.071	0.001	0.002	0.003	0.004	0.006	0.006	0.008	0.016	0.064	0.150
MAP	0.074	0.055	0.103	0.062	0.067	0.057	0.103	0.099	0.097	0.097	0.096	0.103	0.093	0.093	0.096	0.095	0.104	0.106	0.111	**0.114**
MAP Rank	16	20	6	18	17	19	6	8	9	10	11	5	14	15	12	13	4	3	2	**1**
Number of Best Scores	1	0	2	0	0	2	2	3	3	0	1	**4**	1	0	1	0	1	0	1	0
Mean Rank	12.60	16.00	**6.05**	16.25	14.00	15.30	**6.05**	9.20	9.20	9.55	9.25	9.65	11.15	10.85	10.45	10.45	8.60	8.10	7.80	8.40

Table 5.6: Semantic Query Results (*averaging* prototype aggregation)

	Single Features						Exhaustive Search					Relief-F	CSF							
	Visual Color	Visual Shape	Visual Texture	Audio Simple	Audio Complex	Textual	ES-1	ES-2	ES-3	ES-4	ES-5	Relief-F	Mean	StdDev	MeanDist	CorrRat	CSF (v=0.5)	CSF (v=1)	CSF (v=2)	CSF (v=3)
Airplane	0.112	0.037	0.093	0.068	0.040	0.023	0.185	0.157	0.101	0.169	0.114	**0.377**	0.072	0.081	0.098	0.080	0.119	0.129	0.138	0.135
Animal	0.066	0.078	**0.106**	0.071	0.078	0.059	0.069	0.070	0.064	0.088	0.062	0.068	0.067	0.086	0.071	0.086	0.071	0.067	0.067	0.066
Boat_Ship	0.112	0.071	0.098	0.049	0.058	0.043	0.087	0.111	0.082	**0.113**	0.079	0.109	0.083	0.088	0.092	0.092	0.091	0.094	0.100	0.104
Car	0.189	0.181	0.196	0.138	0.182	0.115	0.210	0.224	**0.232**	0.215	0.224	0.182	0.222	0.227	0.230	0.224	0.227	0.226	0.228	0.230
Charts	0.008	0.025	0.024	0.027	0.019	**0.029**	0.009	0.011	0.011	0.015	0.009	0.024	0.010	0.013	0.015	0.014	0.016	0.018	0.022	0.023
Computer_TV-screen	0.059	0.073	**0.096**	0.049	0.072	0.080	0.086	0.094	0.090	0.086	0.086	0.096	0.089	0.077	0.086	0.081	0.091	0.093	0.087	0.089
Desert	0.013	0.010	0.013	0.005	0.007	0.007	0.013	**0.019**	0.012	0.019	0.009	0.005	0.009	0.011	0.011	0.011	0.006	0.006	0.006	0.006
Explosion_Fire	0.011	0.026	0.030	0.013	0.015	0.010	0.014	0.020	0.014	0.029	0.011	**0.031**	0.017	0.017	0.026	0.028	0.023	0.027	0.028	0.028
Flag-US	0.005	0.002	0.002	0.001	0.003	**0.013**	0.002	0.006	0.006	0.001	0.003	0.001	0.002	0.003	0.001	0.002	0.003	0.002	0.001	0.001
Maps	0.020	0.021	0.034	0.030	**0.051**	0.032	0.016	0.020	0.024	0.025	0.020	0.030	0.027	0.031	0.031	0.036	0.026	0.026	0.028	0.029
Meeting	0.376	0.264	0.395	0.352	0.300	0.328	**0.464**	0.389	0.431	0.271	0.450	0.357	0.393	0.325	0.366	0.281	0.422	0.428	0.430	0.422
Military	0.034	0.024	0.031	0.023	0.010	0.015	0.038	0.049	0.027	**0.062**	0.025	0.025	0.031	0.041	0.048	0.043	0.034	0.036	0.039	0.041
Mountain	0.044	0.030	0.057	0.025	0.030	0.023	0.046	0.054	0.043	0.047	0.041	**0.057**	0.037	0.037	0.041	0.040	0.049	0.052	0.055	0.057
Office	0.130	0.083	0.135	0.083	0.117	0.140	0.184	0.156	0.209	0.108	0.218	0.023	0.200	0.133	0.145	0.091	0.210	**0.222**	0.221	0.213
People-Marching	0.054	0.046	0.067	0.027	0.048	0.020	0.094	0.094	0.065	**0.114**	0.067	0.077	0.074	0.080	0.095	0.080	0.086	0.087	0.095	0.089
Police_Security	0.050	0.030	0.032	0.039	0.026	0.036	**0.065**	0.053	0.050	0.037	0.053	0.049	0.046	0.033	0.043	0.033	0.050	0.050	0.050	0.049
Sports	**0.048**	0.039	0.036	0.026	0.037	0.036	0.035	0.040	0.041	0.041	0.042	0.015	0.039	0.042	0.040	0.041	0.041	0.040	0.040	0.040
Truck	0.078	0.101	0.118	0.070	0.085	0.052	0.092	0.102	0.097	0.102	0.090	**0.118**	0.095	0.103	0.103	0.103	0.095	0.097	0.100	0.101
Waterscape_Waterfront	0.161	0.076	0.139	0.066	0.083	0.071	0.155	0.179	0.154	**0.181**	0.137	0.161	0.120	0.151	0.148	0.154	0.155	0.159	0.164	0.173
Weather	0.036	0.002	0.003	0.003	0.003	0.001	0.004	0.019	0.006	0.034	0.004	0.002	0.004	0.005	0.008	0.007	0.009	0.014	0.046	**0.070**
MAP	0.080	0.061	0.085	0.058	0.063	0.057	0.093	0.093	0.088	0.088	0.087	0.090	0.082	0.079	0.085	0.076	0.091	0.094	0.097	**0.098**
MAP Rank	14	18	11	19	17	20	4	5	8	9	10	7	13	15	12	16	6	3	2	**1**
Number of Best Scores	1	0	2	0	1	2	2	1	1	**4**	0	**4**	0	0	0	0	0	1	0	1
Mean Rank	11.00	14.75	8.75	16.30	14.40	15.50	10.35	**6.75**	9.75	7.65	11.65	10.05	12.50	10.05	8.85	9.95	8.65	8.00	7.20	7.75

Table 5.7: Semantic Query Results, General Comparison of Different Classifiers

	1-nn	k-nn	c-nn	svm
Visual-Color	0.069	0.074	0.080	0.081
Visual-Shape	0.070	0.055	0.061	0.061
Visual-Texture	0.084	0.103	0.085	0.092
Audio-Simple	0.058	0.062	0.058	0.060
Audio-Complex	0.061	0.067	0.063	0.065
Textual	0.057	0.057	0.057	0.047
ES-1	0.091	0.103	0.093	0.102
ES-2	0.090	0.099	0.093	0.102
ES-3	0.090	0.097	0.088	0.102
ES-4	0.088	0.097	0.088	0.102
ES-5	0.088	0.096	0.087	0.101
Relief-F	**0.099**	0.103	0.090	0.102
Mean	0.083	0.093	0.093	0.102
StdDev	0.084	0.093	0.093	0.103
MeanDist	0.085	0.096	0.096	**0.103**
CorrRat	0.085	0.095	0.095	0.101
CSF (v=0.5)	0.089	0.104	0.091	0.101
CSF (v=1)	0.090	0.106	0.094	0.099
CSF (v=2)	0.091	0.111	0.097	0.097
CSF (v=3)	0.092	**0.114**	**0.098**	0.096

However, selection of modalities is a critical issue. A wrong selection can lead to worse results than the best of the single modalities.

- Considering the results of single features, it can be observed that the performances of modalities vary in different classes. For instance "Visual-Texture" modality gives the best results for "Animal" class, whereas "Textual" modality is better for "Flag-US" class. Such results validate the base idea of CSF, which is different classes can be represented better with different modalities.

- A similar evaluation with the above can be done by considering the exhaustive search results. The results of exhaustive search supports the claim of exploiting class-specific features. Different feature combinations in exhaustive search selections perform better in different classes, which results different classes requires the use of different features.

- Although the exhaustive search should guarantee to find the optimal feature selection by evaluating all possible combinations, it lacks the use of class-specificity. As mentioned in the test setup, exhaustive search generate modality selection sets that are common for all classes. However, RELIEF-F and CSF can find the informative modalities for each class separately. Thus, RELIEF-F and CSF obtains better results than exhaustive search.

- CSF gives successful accuracy results against RELIEF-F.

- As declared in Section 5.4, the superiority of Standard Mean Distance (δ) parameter of CSF among other parameters is still valid. Also, updating the CSF formulation by getting vth power of Standard Mean Distance has a positive effect on the accuracy results. As the v value increases, the accuracy increases. So, the discriminativeness characteristics of features is the most effective parameter. In essence, the calculation of RELIEF-F is very similar to the Standard Mean Distance and the accuracy results of RELIEF-F is reasonably high.

- As mentioned in Section 5.4, an important discussion for combining multiple features is the independency of features. Using complementary features with the methods requiring independent inputs can cause a decrease in the accuracies. In this study, the modalities utilized are not fully independent. RELIEF-F

approach is known to be good at handling features with high dependencies. Also, exhaustive search can handle the dependency issue since it tries to find the optimal solution. The test results show that CSF approach is as successful as these two approaches at eliminating complementary features and selecting the most informative ones.

- Prototype aggregation method have a direct effect on the accuracy.. As evaluated in Section 5.4, the *averaging* approach is superior to the *minimum*. In this test setup, *k-minimum* approach is included. According to the results, *k-minimum* performs superior than both of *averaging* and *minimum*, whereas *averaging* is still better than *minimum*. The reason why *k-minimum* is better is quite clear; selecting the k prototypes with minimum distances to the query instance prevents the negative effect of noisy prototypes. Thus, a successful prototype selection mechanism is crucial for such a setup and classifier.

- The AP values of different classes can dramatically change. This is not because of the success of classifiers or fusion mechanism; but the unbalanced dataset. As presented in Table 5.3, the number of training and test instances fluctuate among classes excessively. When the AP values of classes and instance counts analyzed in detail, it can be observed that change in the number of training instances does not affect the performance so much. However, the number of test instance counts directly affects the success in each of the classes. This is probably because of the noisy instances in the test set. When the number of test instances of a class increases, the ratio of noisy instances decreases, so the performance of the class increases. However, the evaluation in [91] is different. They argue that the successful classes are the ones which are the extensively studied ones. But, such an evaluation is not applicable for our tests, since we have not performed any special research on any of the classes.

- Considering the accuracy results of TRECVID 2007 participants, the results of proposed approach obtains are very successful and close to the accuracies of the most successful participants.

5.6 A Utilization of CSF in Wireless Video Sensor Networks

In this section a utilization of CSF approach for efficient feature selection and combination in an Wireless Video Sensor Networks application, is presented. Considering that the study is not directly towards the Ph.D. topic, a brief summary is presented as well the test results.

5.6.1 Overview

Wireless Visual Sensor Networks (WVSNs) have started to receive a lot of attention very recently due to their potential to be deployed flexibly in various outdoor applications with lower costs [1]. Such networks deploy a large number of image/video sensors [41, 102] with different capabilities and can collect/process multimedia data. Typical applications of WVSNs include multimedia surveillance, target tracking, habitat monitoring, intrusion detection and health care delivery [1].

In such applications, battery operated image/video camera sensors are deployed to acquire different viewpoints of the occurring events. One of the major problems in these surveillance and target tracking applications is to classify the detected objects accurately. If the detected objects are classified appropriately on site, then the central decision unit, i.e. the sink, may be alarmed effectively. This is very crucial given that these applications are geared for security and safety. For instance, given a power plant surveillance application, built with wireless camera sensors and used to detect the intruders, only human intruders or more specifically only non-worker human intruders may be alarmed to the guards. In other situations, such as in case of an animal or employed worker in power plant, no alarm may be necessary.

In order to perform an accurate object classification, an effective set of features should be selected for classification and and a robust classifier should be constructed. Although there exist lots of features and classifiers in the literature for visual object classification [27, 48, 119]; the important point for WVSN applications is to employ those features and classifiers which are lightweight in terms of processing, energy, time and storage as well as their accuracy in classification. Also real-time applicability is crucial considering that the classification process is performed on the sensor, at the

time object is detected. Another important requirement is the flexibility of the system for adding new features and object classes in order to be able to extend the classifier for recognizing new classes and make it applicable for other domains.

In this study, we choose two simple –but effective– features; shape and velocity of the detected objects. As the classifier, we employ the Genetic Algorithms (GA) based classifier proposed in [152]. Actually, the classifier is designed as a Minimum Distance classifier empowered with a GA-based approach by employing a GA-based prototype selection mechanism. Minimum Distance classification approach provides lightweight solution with its low-complexity in processing and time. Besides, GA provides increase in accuracy and lowering the storage requirement by enhancing the prototypes in the classifier model and including a probabilistic knowledge. In addition, the classifier utilizes the Class-Specific Features (CSF) [151] in order to relate the prototype classes with the most representative and discriminative features for them. The experiments show that the classifier can classify the most usual object types such as human or vehicle effectively in our typical surveillance application with lower costs in terms of energy, time and storage.

5.6.2 Experimental Evaluation

This section includes the experiment setup, metrics and the results.

5.6.2.1 Experiment Setup and Performance Metrics

For the experiments, we assume a power plant surveillance application scenario. In this scenario, when an intrusion occurs at the area under surveillance, the detected objects are classified at the camera sensors. The classification is performed as a multi-class choice with 3 classes: $Human$, $Vehicle$ and $Animal$. For the camera sensor experiment data, the Caltech 101 image dataset [34] (for $Vehicle$ and $Animal$ classes) and search results from Google Image Search (for $Human$ class) are used by formatting them into the CmuCam3 [107] output format. The CalTech101 dataset does not contain $Vehicle$ and $Animal$ classes. So, images from several different classes in Caltech 101 dataset are regrouped according to these classes. The dataset is divided

Figure 5.8: Sample images from test dataset

into three sets: First-Training, Second-Training and Test. The number of images is determined as 10 for each class in each of the training sets and 20 for each class in the test set. Sample images from our constructed dataset are given in Figure 5.8.

As mentioned above, the CSF mechanism [151] is applied in order to find representative and discriminative features for each object class. CSF mechanism gives weights of each feature for each class. Acquired weights are given in Table 5.8. According to these weights, it has been observed that $Velocity$ is the dominant feature for all classes. However the effect of it is more for $Animal$ than the other two classes.

Table 5.8: CSF Weights

	$Shape\_Ratio$	$Velocity$
$Human$	0.371051	0.628949
$Vehicle$	0.342217	0.657783
$Animal$	0.130904	0.869096

We have considered two metrics:

- Classification Performance: This metric shows the performance in estimating the class of the intruder. The bigger is the performance, the better is the quality of the approach.
- Energy Overhead: This constitutes total energy in processing and transmitting the frames (if needed). Our goal is to minimize this overhead in order to maximize the lifetime of the cameras.

5.6.2.2 Performance Results

Classification Performance

Under given test setup, the results given in Table 5.11 and Table 5.12 are obtained. We compare the results with a previous study in [95] which uses a user advanced fuzzy membership sets, Table 5.9 and Table 5.10. The classifier performs multiple labeling by providing fuzzy membership values in the range [0,1] for each class. So, in order to measure the precision values, the class with the highest membership value is taken as the classification result.

Table 5.9: Confusion Matrix for [95]

		Prediction		
		$Human$	$Vehicle$	$Animal$
Actual	$Human$	20	0	0
	$Vehicle$	0	19	1
	$Animal$	1	3	16

Table 5.10: Class Precisions for [95]

Class	**Precision**
$Human$	1.00
$Vehicle$	0.95
$Animal$	0.8
Total	**0.85**

Table 5.11: Confusion Matrix for Proposed Approach

		Prediction		
		$Human$	$Vehicle$	$Animal$
Actual	$Human$	20	0	0
	$Vehicle$	0	18	2
	$Animal$	0	0	20

Energy Overhead

In order to prove the efficiency of this algorithm, we have also performed experiments to assess the energy consumption on the camera sensor. We have used the AVR Simulation and Analysis Framework (AVRORA) to calculate energy costs [131].

Table 5.12: Class Precisions for Proposed Approach

Class	**Precision**
$Human$	1.00
$Vehicle$	0.90
$Animal$	1.00
Total	**0.97**

AVRORA is an emulator which can provide realistic results as if the approach is run on a typical CMOS sensor. It has built in functions that can compute the processing and communication costs.

We have used a baseline approach which processes the frames at a base-station and determines the location of the objects. In that case, the frames are sent to the base-station traveling through multiple hops (i.e., k). This is referred to as '*TraditionalMethod*' in the graphs. Our approach performs the localization and classification on site and does not send any data to the base-station. However, it may need to send an alarm (i.e., one simple message) to the base-station when an intruder is detected and located. The results are given in Table 5.13 and Table 5.14.

Table 5.13: Energy Costs for Different Tasks

Task	**Cost in Joule**
C: One-time CPU cost to process the frame to extract and classify the moving object	0.0220
M: Transmission cost of the whole frame for 1 hop	0.0700
T: Transmission cost of the alarm for 1 hop	0.0007
Taking the video data	Same for both cases

Table 5.14: Total Energy Costs in Joules

Process	**Traditional Method**	**Proposed Method**
For 1 Hop	$M = 0.0700$	$C + T = 0.0227$
For k Hops	$M * k = 0.0700 * k$	$C + T * k = 0.022 + 0.0007 * k$

The results for varying k (*Hop Count*) values are depicted in Figure 5.9. As can be seen from this figure, energy overhead for our approach is constant and significantly smaller than the traditional method. We would like to note that in this experiment every moving object detection event is sent as an alarm to the sink. However, if we define some alarm criteria for the proposed method, the energy consumption would be further reduced (i.e., alarms are only sent when needed).

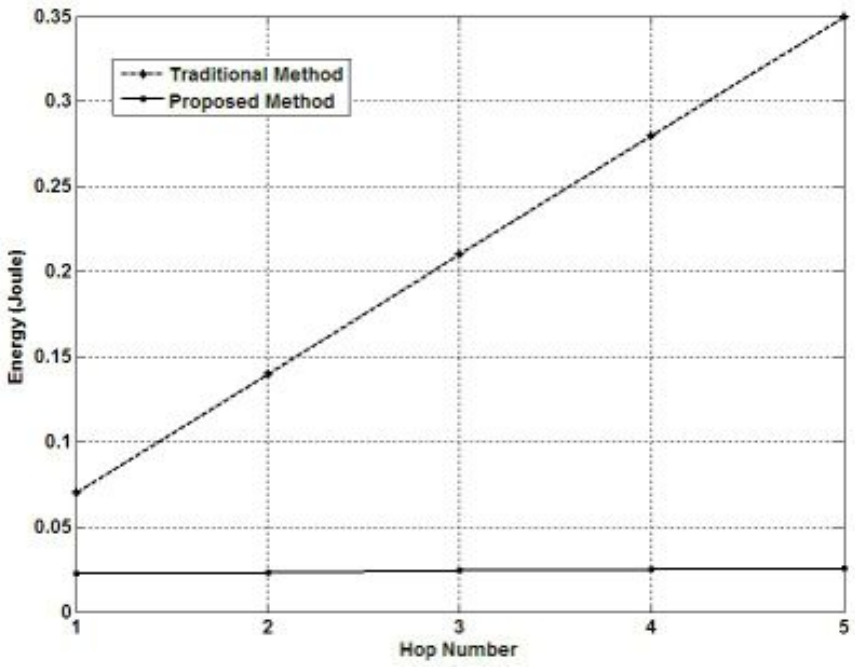


Figure 5.9: Energy Costs of Two Different Methods

5.6.2.3 Evaluation

In order to reduce the false alarms on detected objects in WVSNs applications, the detected objects can be classified at camera nodes to improve the quality of surveillance applications and also extend the lifetime of the network. In this study, we have presented a lightweight object classification approach which can work on-site at a camera sensor. The approach utilizes a minimum distance classifier enhanced by a genetic algorithm based prototype selection approach on top of two simple but effective features, which are the shape and velocity of the detected objects. The approach also benefits from the idea of exploiting class specific features.

The experimental evaluation has revealed that our approach can effectively classify typical objects, i.e. human, animal and vehicles, on a typical surveillance application with an error rate of 3% overall. We also assessed the energy overhead of our approach on the individual camera sensors. The energy consumption is significantly reduced compared to the cases where the classification is performed at the base station due to communication overhead. In the future, we plan to increase the number of features used in classification to further improve the classified object types.

5.7 Evaluation of Fusion System Design

Considering the general fusion framework proposed in Section 3.1, an evaluation of the fusion architecture described in this chapter is given below. The approach is based on a 'multi-modal, multi-classifier' fusion scenario and focuses on the 'What to Fuse' problem. Below, how each affecting factor is handled through the proposed solution is described.

- Fusion Setting: The proposed approach is utilized in two different fusion settings; (i) multi-feature and (ii) multi-modal. For both settings, the dissimilarity values between samples are combines. Thus the setting is can be accepted as 'multiple features / modalities, with multiple classifiers', considering that dissimilarity calculation is a very simple classification approach.
- Selection of Sources: The approach uses a static feature weighting scheme based on the proposed weight calculation formula. Yet, the weights have a context relation, since the approach is a class-specific feature selection approach.
- Fusion Strategy: The approach focuses on the use of complementary information for fusion.
- Content Representation: A dissimilarity-based representation is preferred.
- Normalization of Sources: Min-max normalization is applies to normalize the dissimilarity values between images / videos.
- Fusion Level: The approach is a late fusion approach, since the combination inputs are the dissimilarity values.

- Fusion Methodology: Considering that the focus of the study is the feature / modality selection, linear weighted averaging approach is utilized as the fusion methodology.
- Operation Modes: The mode for operation is a parallel scheme.
- Synchronization: A simple shot-based synchronization is applied.
- Adaptation: In the scope of the approach, adaptation capability is not considered.

5.8 Remarks

In this chapter, a class-specific feature selection approach for the fusion of multiple features is presented. In order to eliminate the high-dimensionality of multiple features and provide efficient querying over the images, a dissimilarity based approach is utilized. The class-specific features are determined by using the representativeness and discriminativeness of features for each image class. The calculations of representativeness and discriminativeness are based on the statistics on the dissimilarity values of training images.

The approach is firstly tested in a multi-feature setting by using the CalTech101 dataset with 8 MPEG-7 visual features. The approach is compared with the retrieval performance of single features, simple combination approaches and exhaustive search approach. Then it is also applied to a multimodal setting by using TRECVID 2007 dataset with 3 visual, 2 audio and 1 textual modalities. Lastly, the proposed approach is utilized for efficient feature selection and combination in a Wireless Video Sensor Networks application. The results obtained from these tests show that the proposed class-specific feature selection approach is an effective and efficient feature selection method.

Further study on this issue would be as follows: Employing prototype selection and aggregation methods within the proposed approach, utilizing proposed approach with a dissimilarity based classification mechanism and performing multi-modal feature selection obtained from video data.

CHAPTER 6

RELIEF-MM: AN EFFECTIVE MODALITY WEIGHTING APPROACH[1]

Fusing multimodal information in multimedia data usually improves the retrieval performance. One of the major issues in multimodal fusion is how to determine the best modalities. In order to combine the modalities more effectively, we propose a RELIEF based modality weighting approach, named as RELIEF-MM. The original RELIEF algorithm is extended for weaknesses in several major issues: class-specific feature selection, complexities with multi-labeled data and noise, handling unbalanced datasets, and using the algorithm with classifier predictions. RELIEF-MM employs an improved weight estimation function, which exploits the representation and reliability capabilities of modalities, as well as the discrimination capability, without any increase in the computational complexity. The comprehensive experiments conducted on TRECVID 2007, TRECVID 2008 and CCV datasets validate RELIEF-MM as an efficient, accurate and robust way of modality weighting for multimedia data.

6.1 Overview

Increase in the use of digital multimedia data in recent years has shown the need for multimedia retrieval systems. Retrieval of multimedia data is based on its semantic

[1] This chapter was published as [150]. In addition, a preliminary version of this chapter was published as [148].

content. In order to handle the semantic content effectively, the nature of the multimedia data should be examined and information contained in multimedia data should be used completely. The multimedia data usually has a complex structure containing multimodal information (i.e. audio, visual and textual modalities). Regarding the noise in sensed data, non-universality of any single modality and the performance upper bound of each modality, relying on a single modality may not be applicable [98]. Furthermore, it has been observed that the sets of patterns misclassified by different modalities do not necessarily overlap, and complementary information provided by different modalities improves recognition capability [61]. Since each modality abstracts videos from a different aspect, different modalities in multimedia data complement each other [51]. Thus, combining multimodal information usually improves the retrieval performance. However, there exist two major issues that have not been adequately addressed yet and are still attractive research areas [4, 98, 143]: (i) How to determine the best modalities? (ii) How best to fuse them? This study focuses on the first problem and presents a modality weighting approach in order to use the multiple modalities effectively.

The modality selection is a combinatorial search problem that aims to find the best subset of available modalities giving the highest accuracy. Such a computational problem can be solved to some extent by using a weighting strategy. Modality weighting is a generalization of the selection problem, where the modalities are ranked by assigning some weights in between $[0, 1]$ to each modality, instead of a binary selection. The use of weights enables some well-established optimization techniques and efficient algorithmic implementations to be employed [126]. Furthermore, a weighting strategy is a practical solution since the most frequently utilized fusion approach is the Linear Weighted Fusion [37, 133, 145], in which the combined decision is calculated as a weighted sum of the available modalities.

The previous studies on using multiple modalities can be categorized into three groups: (i) using all features/modalities by averaging them (ii) performing an empirical selection and (iii) determining the effectiveness of each feature with a weighting algorithm. Despite their wide usage among fusion studies, the first two are simplistic approaches; the first one treats all features as equally-likely although any of the features can be non-informative or redundant, whereas the second approach requires an empirical observation and manual selection based on the observation. On the other hand, the third

direction requires design of an efficient feature weighting algorithm, which proposes a polynomial time heuristic for the combinatorial explosion problem while dealing with multiple features.

Regarding the third direction, we focus on some adaptable solutions from *feature weighting* studies in the machine learning literature. However, the feature weighting solutions are not easily applicable to the modality weighting problem, considering the issues of (i) the intrinsic multi-dimensionality of modalities and (ii) the multivariate inputs of fusion systems. The former issue states that *feature weighting* methods give weights for each dimension of an input feature vector, whereas modality weighting methods assign weights to each modality, each of which is a multi-dimensional feature, by accepting each modality as a black-box. Besides, the latter is a more general issue in fusion systems. The inputs of a fusion system are not necessarily feature values. The prediction scores for different features / modalities are frequently combined in state-of-the-art fusion studies. An intuitive idea to discard these problems is to utilize a weighting approach that works in distance based metric space, instead of using a feature space. Utilizing a distance space solves the intrinsic dimensionality problem of multiple modalities by converting multi-dimensional feature values of a modality to a uni-dimensional distance value. Furthermore, it enables handling of the prediction scores after converting them into applicable dissimilarity values with appropriate conversion functions.

Among the existing *feature weighting* algorithms, we focus on the RELIEF algorithm [59], which is considered one of the most successful weighting algorithms and in which the calculations are based on the distances between training samples. Furthermore, according to the best of our knowledge, there exists no usage of the RELIEF algorithm for multimodal feature selection[2] in multimedia retrieval. The key idea of RELIEF is to iteratively estimate feature weights according to their ability to discriminate between neighboring samples. Employing the RELIEF algorithm for multimodal feature selection on multimedia data enables to identify some weaknesses of the algorithm, which have not been addressed before. Our solution is based on RELIEF-F, which is the multi-class extension of the basic RELIEF algorithm. We extend

[2] The final goal of this study is to select the effective modalities by weighting the available modalities and each modality is a multi-dimensional feature. Thus, from now on, the phrases 'modality selection', 'modality weighting' and 'multimodal feature selection' are used interchangeably.

RELIEF-F in the following aspects, considering the characteristics of multimedia data and multimedia retrieval systems:

(i) *Class-specific selection:* Multimedia retrieval is a multi-class problem with a high number of concepts / classes. One major drawback of RELIEF-F is that it generates weights in a class-common way, where the same feature weights are assigned for all concepts. However, each concept can be represented better with different features that are specific to that concept [137, 151]. Thus, it is important to use a class-specific modality weighting approach in the multimedia retrieval systems, in order to handle the high number of classes.

(ii) *Multi-labeled data:* Multimedia data is usually multi-labeled. However, the RELIEF-F algorithm cannot perform well when the training samples are multi-labeled. RELIEF-F estimates the weights of the features according to their ability to discriminate between different classes. Having multi-labeled samples causes the algorithm not to discriminate between classes effectively, due to the ambiguity produced by the samples associated with multiple concept types.

(iii) *Noisy data:* Multimedia data contains a vast amount of noise. However, the way RELIEF-F deals with noisy data is inadequate. Similar to the multi-label issue, noise in the samples hinders a correct discrimination between classes.

(iv) *Unbalanced data:* The training samples provided in multimedia datasets are usually unbalanced between classes. Although RELIEF-F applies k nearest neighbor approach to deal with the outlier data, an unbalanced dataset prevents RELIEF-F from eliminating outlier data effectively. Assuming that each class has approximately the same amount of noisy samples (as a ratio), using the same k for all classes makes the algorithm include more noisy samples for the classes with smaller numbers of training samples. Thus, having different numbers of samples for each class affects the performance of the RELIEF-F algorithm negatively.

(v) *Late fusion inputs:* In regular use of RELIEF based algorithms, the distances between instances are calculated by using the feature values. However, the late fusion approaches usually rely on prediction scores and the feature values may

not be available at the time of fusion. Thus, a procedure that enables using the prediction scores is necessary.

In this chapter, we propose a new RELIEF extension for multimedia data (RELIEF for Multimedia data: RELIEF-MM) to handle the above given research issues. First, we restate the RELIEF-F algorithm in a class-specific way and show that the weights produced by the original RELIEF-F are equal to the average of all class-specific weights. Thus, generating class-specific weights does not have a negative effect on the computational complexity of the algorithm. Secondly, we deal with the multi-label and noise issues, and extend the weight estimation function by including the representation and reliability characteristics of the features in addition to the currently used discrimination capabilities. These characteristics of features are calculated based on the statistics of distances between the training instances, by complying with the distance-space criteria discussed before. The mean distances between the samples of each class are employed as the representative characteristics, and the correctness ratios of features for each class are used as the reliability characteristics. For the discriminative property, we calculate the distance between the means of classes, as in the original RELIEF-F. Thirdly, we deal with the unbalanced data problem, and propose the use of dynamic k nearest neighbor selection. In dynamic k selection, a different k value is calculated for each class, instead of the same k value for all classes. The dynamic k value is used as a predefined ratio of the number of samples in each class. This modification makes the algorithm deal with approximately the same ratio of noisy instances for all classes and give more regularized weight assessments. Lastly, we enable RELIEF-F algorithm for use with classifier predictions by converting the prediction scores into distances between instances.

We evaluate the RELIEF-MM algorithm with the TRECVID 2007 [91], TRECVID 2008 [92] and Columbia Consumer Video (CCV) Database [55] datasets. For each of the issues discussed above, we perform comparative tests against the RELIEF-F algorithm. In addition, we compare the multimedia retrieval accuracies of the RELIEF-MM based linear weighted fusion approach with single modalities, simple averaging and exhaustive search. As a general overview, we can state that the proposed RELIEF-MM algorithm generates better feature weights than the RELIEF-F algorithm and the computational complexity is still asymptotically the same as the original algorithm.

It has been observed that the fusion methods empowered by RELIEF-MM guarantee higher accuracies than any single modality. RELIEF-MM also demonstrates much better performance than simple averaging and RELIEF-F based methods. Moreover, RELIEF-MM gives nearly the same performance as the exhaustive-search based approach, yet it is computationally much more efficient than the exhaustive one.

The remainder of this chapter is organized as follows: In 6.2, an overview of modality selection in information fusion, feature selection methods and a detailed description of the RELIEF algorithms are given. In Section 6.3, the RELIEF-MM algorithm is presented in detail. In this section, first of all, the RELIEF algorithm is restated in a class-specific way, then the extensions for multi-label, noisy and unbalanced data problems are described. After introducing the extensions in detail, the combined algorithm is presented, along with a computational complexity analysis. Lastly, the strategy for using the RELIEF-MM with late fusion inputs (i.e. prediction scores) is described. In Section 6.4, the empirical results and the evaluations of our proposed solutions are given. In Section 6.5, an evaluation of the proposed fusion architecture is done based on the general fusion framework for fusion (Section 3.1). In the last section, some conclusions are drawn and some possible future studies are discussed.

6.2 Related Work

In multimedia retrieval, the most popular strategies for combining multimodal information are early fusion and late fusion. Early fusion is the concatenation of all available modalities into a single feature vector, whereas late fusion is the linear combination of classifier outputs after processing each modality by a separate classifier [51]. The studies in the literature do not present a clear winner between these two approaches, in terms of accuracy. Yet, early fusion usually leads to the "curse of dimensionality problem" because of concatenation of the modalities. On the other hand, late fusion is simple in calculation and has a reasonable performance despite its simplicity. Thus, late fusion has attracted much more attention than early fusion in recent studies [4, 51]. However, the selection of modalities (i.e. assigning weighs for each modality) is an important issue in late fusion, and affects the retrieval accuracy in fusion results. In this study, we focus on efficiently determining the effectiveness of modalities. Be-

low, we first present recent studies on modality selection for multimedia data. Then, with a machine learning point of view, the modality selection problem is compared with the feature selection problem in machine learning literature, and the well-known approaches for feature selection are presented. Lastly, we discuss the family of the RELIEF algorithms.

6.2.1 Modality Selection / Weighting

In the multimedia domain, the majority of the fusion studies prefer simplistic solutions for combining all available modalities by performing an empirical weighting scheme or a simple averaging [4, 46, 121]. An empirical weighting method is based on empirical observations and manual selection of the features. Besides, a simple averaging approach assumes that all of the modalities are equally effective although any of the features can be non-informative or redundant. Some successful utilizations of simple averaging can be found in [55, 56], where they obtain higher retrieval accuracies than any single modality. Yet there are several studies that perform the selection / weighting by evaluating the effectiveness of each modality, and some of the recent ones are summarized below.

One popular approach for modality selection is the use of the accuracy values as the weight estimations. In [37], Fumera et al. provide a theoretical analysis of this idea. Some recent utilizations of this idea can be found in [44, 85, 103]. Another approach applied in the literature is to find the independent feature subsets, considering that the result of the fusion process is improved if complementary (independent) inputs are combined [64]. Towards this direction, Wu et al. [143] redefine 'modality' as an 'independent component' among the available features and find statistically independent modalities from raw features by employing principle component analysis (PCA), independent component analysis (ICA) and independent modality grouping (IMG) techniques. Kludas et al. [63] apply the independency idea and use correlation coefficients to measure the dependency between features. Besides these, Atrey et al. [5], Kankanhalli et al. [58] and Snidaro et al. [120] study the problem in another perspective, and try to combine multiple data streams (e.g. data obtained from several different sensors like video camera, microphone, etc.), where each data stream can be

accepted as a different modality. Atrey et al. [5] use a dynamic programming approach to find the optimal subset of media streams based on several criteria which maximizes the information gain obtained. Kankanhalli et al. [58] propose an experiential sampling based solution for selecting the most informative subset of data streams. Snidaro et al. [120] define a quality metric for the data streams and dynamically regulate the fusion process. Further recent studies on the topic is as follows: Kalamaras et al. [57] takes the advantage of user feedback and learns the modality weights via an interactive user feedback scheme. Huang et al. [42] tailors the genetic algorithm to learn modality weights and applies it to alleviate the local minima problem during the process of finding an optimal solution. Moulin et al. [82] reformulate the modality weighting problem as a dimensionality reduction problem in a binary classification context and find the linear combination that best separate relevant and non-relevant documents for all queries by using a Fisher Linear Discriminant Analysis based approach. Chen et al. [22] calculate the modality weights by measuring the discriminative capability of each visual feature by a voting scheme, where the voting scheme is applied by processing all triples of the training samples (candidate, positive and negative) and assigning a vote for the candidate according to whether the candidate is closer to the positive or the negative. Wu et al. [142] consider the interactions among the multimodal classifier outputs and employ a fuzzy integral based approach in order to find modality weights. The fuzzy integral approach provides an importance measure for each subset of available information sources [129].

However, each of these methods has their own limitations and drawbacks. First of all, they are either computationally complex or their weight estimation capabilities are limited. Furthermore, the selection process is usually class-common, which means, the same set of features are used for all classes. In addition, they usually evaluate the features individually, which may cause loss of the information that is obtained from the correlation between features. In this study, we propose a timely efficient and effective way for modality weighting, which exploits the class-specific information for modalities and enables the use of correlation between modalities.

6.2.2 Feature Selection / Weighting Approaches

In addition to the above given methodologies, the *feature selection / weighting* studies in machine learning literature provide many different approaches for feature selection. Existing methods in the literature are categorized as filter or wrapper methods. Filter methods assess the relevance of features by looking only at the intrinsic properties of the data, whereas in wrapper methods the performance of a learning algorithm is used to evaluate the fitness of the feature subsets in the feature space. Filter methods are usually computationally much more efficient than wrapper methods; however, wrapper methods usually provide solutions closer to the optimal solution. Another weakness of the filter methods is that they usually evaluate the features individually. Thus, the quality of combined feature subsets is not analyzed and the correlation information between features cannot be exploited. Some well-known filter methods are Information Gain [43], Gain Ratio [101], Correlation based feature selection (CFS) [40], Chi-squared selection and RELIEF [59]. Some well-known wrapper methods are as follows: Exhaustive Search [47], Sequential Forward selection (SFS) [60], Sequential Backward elimination (SBE) [60], Plus q take-away r [36], Simulated Annealing and Genetic Algorithms. For more detailed discussions, interested readers can refer to [39, 47, 110] and the references therein.

With a machine learning point of view, the modality weighting problem is similar in nature to the feature weighting problem and, thus, efficient and effective feature weighting solutions can be applied for the modality weighting problem. However, it is not trivial to apply the available methods to the modality weighting problem due to several differences between the problems. The most crucial difference is the intrinsic dimensionality of modalities. In feature weighting, the input is a feature vector, which is a multi-dimensional vector of numerical/nominal values representing some pattern. Besides, in modality weighting, the input is multiple feature vectors. Feature weighting methods rank the dimensions of the input feature vector by assigning a weight for each dimension, whereas in modality weighting, the intrinsic dimensions of each modality are not the main concern. Modality weighting methods rank the modalities by assigning weights to each modality as a black-box. Still, an early combination (i.e. concatenation) of available modalities corresponds to a single multi-dimensional

feature, which makes any feature weighting method applicable. However, a big majority of the multimodal fusion studies employ late fusion approaches, in which each modality is processed separately. Thus, ranking the available modalities, instead of the intrinsic high-dimensional features, is still a crucial need for the multimodal information fusion. In addition, another concern may be performing some feature selection operations for each of the modalities. However, it can be assumed as a preprocessing step before modality selection / weighting. The second difference between feature and modality weighting is the values of the inputs. The inputs of a multimodal fusion system are not necessarily feature values; the most frequently utilized inputs in state-of-the-art fusion studies are the prediction scores. Thus, the modality selection approach should work under any of these inputs. One more issue related with the input values is that most of the frequently applied methods (e.g. Information Gain, Chi-squared) require the feature values to be binary or discretized. However, discretization of the modalities makes the process computationally complex, since each modality is represented by a multi-dimensional feature.

An applicable idea to deal with these problems is to work in a distance based metric space, instead of in a feature space. Utilizing a distance space solves the intrinsic dimensionality problem of multiple modalities by converting the multi-dimensional feature values of a modality to a uni-dimensional distance value. Furthermore, it enables handling the scores, ranks and decisions after converting them into applicable dissimilarity values with appropriate conversion functions. Thus, we focus on a RELIEF based algorithm, which generates the weights based on the distances between training samples. Being a filter approach, RELIEF avoids an exhaustive search and provides computationally a more efficient solution than the wrapper methods. Besides, it takes the context into account, exploits correlation information between features and thus usually performs better than the filter approaches. Details of the family of RELIEF algorithms are given below.

6.2.3 RELIEF Algorithms

Among the available feature selection and weighting methods, the RELIEF algorithm [59] is among the most successful. It is a simple and effective way for feature

Algorithm 1: Basic RELIEF

Input: list of features $\mathcal{F} = \langle f_i \rangle_{i=1}^{n}$, number of iterations m, set of training instances $\mathcal{D} = \{d_j\}_{j=1}^{t}$

Output: the weight vector W of estimations for the qualities of features

```
1  begin
2      for i ← 1 to n do                                   //for each feature in F
3          W[i] ← 0;
4      end
5      for j ← 1 to m do
6          r ← randomInstance(D);
7          ⟨H, M⟩ ← findNearestHitMiss(r, D);
8          for i ← 1 to n do                               //for each feature in F
9              W[i] ← W[i] − diff(f_i, r, H)/m + diff(f_i, r, M)/m;
10         end
11     end
12 end
```

selection [28]. In addition, RELIEF does not make a conditional independence assumption for features, as many other feature selection methods do, and can correctly estimate the quality of features with dependencies [105]. The key idea of RELIEF is to estimate weights for each feature according to their ability to discriminate between neighboring training samples by iterating through randomly selected instances in the training space. In [126], Sun presents the discrimination based approximation of RELIEF with a novel mathematical interpretation from the optimization perspective, and shows that RELIEF utilizes a margin based nonlinear classifier for searching useful features.

The basic RELIEF algorithm is given in Algorithm 1. The weight estimation function in Line 9 exploits the discrimination capability. The algorithm selects a random sample r, one Near-Hit H (nearest neighbor with the same class with the random sample) and one Near-Miss M (nearest neighbor with a different class with the random sample) and distances between them are calculated. In this calculation, the distance between instances in different classes indicates a discrimination between classes, so $diff(f_i, r, M)$ increases the weight. Inversely, distance between instances with the same class inhibits discrimination, so $diff(f_i, r, H)$ decreases the weight.

Considering several deficiencies of the basic RELIEF algorithm, Kononenko [66]

proposes several extensions for RELIEF: RELIEF-A uses k nearest neighbors instead of one and averages the contribution of k nearest instances in order to eliminate the effect of noisy instances; RELIEF-B, RELIEF-C and RELIEF-D extend the use of $diff$ function in order to handle incomplete datasets; RELIEF-E and RELIEF-F improve the weight update function for multi-class problems. Other well-known extensions for RELIEF are as follows: Sikonja et al. [104] propose RRELIEF-F for handling regression problems. In [116], Sikonja proposes using k-d trees for the selection of nearest neighbors in order to decrease the computation complexity of the RELIEF algorithm. In [126], Sun introduces Iterative RELIEF (I-RELIEF), which uses an Expectation Maximization algorithm in order to eliminate outlier data. Also, Liu et al. [76] try to eliminate outlier data and propose using selective sampling by means of a modified kd-tree instead of random sampling (at Line 6 in Algorithm 1).

Among the available extensions of the RELIEF algorithm, RELIEF-F is the most widely utilized. RELIEF-F enables working with multi-class problems, by selecting k nearest misses for each class. Thus, the RELIEF-F algorithm updates Line 7 of Algorithm 1 with the following;

$$\langle \mathcal{H}, \mathcal{M} \rangle \leftarrow findNearestHitsMisses(r, \mathcal{D}, k, \mathcal{C});$$

where k is the number of nearest neighbors, and $\mathcal{C} = \langle c_u \rangle_{u=1}^{s}$ is the list of classes. $\mathcal{H}$ is the k-sized list of hit instances, where $\mathcal{H}_v$ denotes the vth nearest hit instance. Besides, $\mathcal{M}$ is the $s \times k$ sized matrix, where $\mathcal{M}_v^u$ represents the vth miss instance for class $c_u \in \mathcal{C}$. In addition, the weight estimation function in Line 9 is also updated as;

$$\begin{aligned} W[i] \leftarrow W[i] &- \sum_{v=1}^{k} \frac{diff(f_i, r, \mathcal{H}_v)}{m \cdot k} \\ &+ \sum_{\substack{u=1 \\ c_u \neq C(r)}}^{s} \left(\frac{P(c_u)}{1 - P(C(r))} \sum_{v=1}^{k} \frac{diff(f_i, r, \mathcal{M}_v^u)}{m \cdot k} \right), \end{aligned} \tag{6.1}$$

where $P(c_u)$ represents the prior probability of class c_u, and $C(r)$ indicates the class of sample r.

In this study, we utilize RELIEF-F for multimodal feature selection in multimedia retrieval, which has not been done before, to the best to our knowledge. Using the RELIEF-F algorithm for multimodal feature selection on multimedia data enables

us to identify some weaknesses of the RELIEF-F algorithm. Thus, we extend the RELIEF-F algorithm due to the aspects discussed in Section 6.1.

6.2.4 Complexity Analysis

The feature selection / weighting problem is known as NP-hard, in terms of the number of features $n = |\mathcal{F}|$. An exhaustive search for generating all possible subsets requires $O(p^n)$ actions, where p is the number of assignable weights ($p = 2$ for binary selection). Considering that an exhaustive search is a wrapper method, it requires an evaluation for each of these subsets. Assuming a simple evaluation similar to RELIEF, based on the similarities / distances between m randomly selected instances to all t training instances, the total complexity of the exhaustive search becomes $O(m \cdot t \cdot n \cdot p^n)$. Moreover, if a class-specific approach is applied, the total complexity becomes $O(m \cdot t \cdot s \cdot n \cdot p^n)$, where s is the number of classes ($s = |\mathcal{C}|$).

On the other hand, the RELIEF algorithms provide solutions in polynomial time. The complexity of the basic RELIEF algorithm is $O(m \cdot t \cdot n)$, considering that the most complex operation is the selection of the nearest hit and miss instances since the distances between r and the other training instances should be calculated for each feature, which requires $O(t \cdot n)$ comparisons. Different from the basic RELIEF algorithm, the complexity of RELIEF-F depends on the number of nearest neighbors (k). If we use a priority queue, which is implemented with a heap structure, for the selection of k nearest neighbors, where the construction of the heap is $O(t)$ and the retrieval of k neighbors from each class is $O(k \cdot s \cdot \log t)$; the total complexity of selecting k nearest hits/misses becomes $O(m \cdot t \cdot n + m \cdot k \cdot s \cdot \log t + m \cdot k \cdot s \cdot n)$. In this equation, the first term is for the distance calculation, the second is for selecting nearest instances from the heap and the last is for the weight calculation (Eq. (6.1)). If the dataset is a balanced one and the value of k is considerably small with respect to t, then the computational complexity of the RELIEF-F algorithm becomes the same as the basic RELIEF algorithm ($O(m \cdot t \cdot n)$). A computationally better solution can be obtained by utilizing k-d trees for improving the nearest hit and miss selection process ($O(n \cdot t \cdot \log t)$).

If the space complexity is considered, both the basic RELIEF and RELIEF-F is in

linear time in terms of the number of features. The biggest space required for these algorithms is for the feature values for training dataset. The required space for the dataset is bounded by $O(t \cdot n \cdot A)$, where A is assumed as the average size for a single feature. The basic RELIEF has additional space requirements for features ($O(n)$), weights ($O(n)$), nearest neighbor selection ($O(n \cdot A)$). For RELIEF-F calculation, additional space requirements are; features ($O(n)$), weights ($O(n)$), classes ($O(s)$) and nearest neighbor selection ($O(k \cdot s + n \cdot A)$). Considering that $t \geq k \cdot s$, both basic RELIEF and RELIEF-F is bounded by $O(t \cdot n \cdot A)$, for space complexity.

6.3 RELIEF-MM: Modality Weighting Approach for Multimedia Data

In order to benefit from the simplicity and effectiveness of RELIEF algorithms, we propose a RELIEF based multimodal feature selection solution, by extending the RELIEF-F algorithm. Below, each of our extensions is presented in a separate subsection.

6.3.1 Class Specific Feature Weighting

Multimedia retrieval requires dealing with a high number of different queries, where each query usually denotes a different concept occurring in videos. Thus, multimedia retrieval is accepted as a multi-class classification problem with a high number of classes, where each class is a concept occurring in videos. In addition, the variety of such concepts is so wide that they can be associated with different sets of features / modalities. In other words, each concept can be represented better with different features specific to the concept [137, 151]. For instance, an $explosion$ concept can be represented relatively more accurately by the audio modality, whereas it is better to utilize visual modality for detecting a $mountain$ concept. Similarly, it can be easier to recognize a $meeting$ concept by using both the visual and the audio modalities. Hence, a class-specific modality weighting approach is inevitable to be used in the multimedia retrieval systems, in order to handle the high number of classes / concepts. However, the traditional feature selection methods, including the RELIEF-F algorithm, propose class-common solutions in which the selection is performed independently from the classes.

Based on the motivation above, we propose a substantial extension on RELIEF-F, which is converting it to a class-specific solution. Since the RELIEF-F algorithm iterates over available training samples to obtain the final value of the modality weights, grouping the training samples according to their classes and processing samples of each class separately can achieve a class-specific solution.

Assuming that we iterate over m training samples $\mathcal{R} = \{r_i\}_{i=1}^{m}$, which are randomly selected from the set of all training samples $\mathcal{D} = \{d_i\}_{i=1}^{t}$, the final weight of f_i can be formalized as;

$$W(f_i) = \sum_{j=1}^{m} \left[-\sum_{v=1}^{k} \frac{diff(f_i, r_i, \mathcal{H}_v)}{m \cdot k} + \sum_{\substack{u=1 \\ c_u \neq C(r_j)}}^{s} \left(\frac{P(c_u)}{1 - P(C(r_j))} \sum_{v=1}^{k} \frac{diff(f_i, r_j, \mathcal{M}_v^u)}{m \cdot k} \right) \right]. \tag{6.2}$$

Here, we can rewrite Eq. (6.2) as in Eq. (6.4), by assigning the effect of one training sample r_j on the final weight calculation of modality f_i into ΔW_j^i (Eq. (6.3)).

$$\Delta W(f_i, r_j) = -\sum_{v=1}^{k} \frac{diff(f_i, r_j, \mathcal{H}_v)}{k} + \sum_{\substack{u=1 \\ c_u \neq C(r_j)}}^{s} \left(\frac{P(c_u)}{1 - P(C(r_j))} \sum_{v=1}^{k} \frac{diff(f_i, r_j, \mathcal{M}_v^u)}{k} \right), \tag{6.3}$$

$$W(f_i) = \frac{1}{m} \sum_{j=1}^{m} \Delta W(f_i, r_j) \, . \tag{6.4}$$

If the samples in $\mathcal{R}$ are grouped according to the class they belong to, we can represent the final weight of f_i as in Eq. (6.5). Here, each group is represented by $\mathcal{R}_u = \{r \mid r \in \mathcal{R} \wedge C(r) = c_u\}$, where $\mathcal{R} = \bigcup \{\mathcal{R}_u\}_{u=1}^{s}$ and $C(r)$ represents the class of r.

$$W(f_i) = \sum_{u=1}^{s} \left(P(c_u) \frac{1}{|\mathcal{R}_u|} \sum_{r \in R_u} \Delta W(f_i, r) \right). \tag{6.5}$$

Here, we can define a class-specific weight $\omega(c_u, f_i)$ as in Eq. (6.6).

$$\omega(c_u, f_i) = \frac{1}{|R_u|} \sum_{r \in R_u} \Delta W(f_i, r) \, . \tag{6.6}$$

Algorithm 2: Class-Specific Adapt. of RELIEF-F

Input: list of features $\mathcal{F} = \langle f_i \rangle_{i=1}^{n}$, number of iterations m, set training instances $\mathcal{D} = \{d_j\}_{j=1}^{t}$, list of classes $\mathcal{C} = \langle c_u \rangle_{u=1}^{s}$, number of nearest neighbors k

Output: the weight matrix ω of estimations for the qualities of features

begin
  for $u \leftarrow 1$ **to** s **do** //for each class in $\mathcal{C}$
    for $i \leftarrow 1$ **to** n **do** //for each feature in $\mathcal{F}$
      $\omega[u][i] \leftarrow 0$;
    end
  end
  for $u \leftarrow 1$ **to** s **do** //for each class in $\mathcal{C}$
    $\mathcal{D}_u \leftarrow getClassInstances(\mathcal{D}, c_u)$;
    $m' \leftarrow m \cdot P(c_u)$; // P(c_u)=size($\mathcal{D}_u$)/size($\mathcal{D}$)
    for $j \leftarrow 1$ **to** m' **do**
      $r \leftarrow randomInstance(\mathcal{D}_u)$;
      $\langle \mathcal{H}, \mathcal{M} \rangle \leftarrow findNearestHitsMisses(r, \mathcal{D}, k, \mathcal{C})$;
      for $i \leftarrow 1$ **to** n **do** //for each feature in $\mathcal{F}$, apply Equation 6.6
        $\omega[u][i] \leftarrow \omega[u][i] - \sum_{v=1}^{k} \frac{diff(f_i, r, \mathcal{H}_v)}{m' \cdot k} + \sum_{\substack{u'=1 \\ u' \neq u}}^{s} \left(\frac{P(c_{u'})}{1 - P(c_u))} \sum_{v=1}^{k} \frac{diff(f_i, r, \mathcal{M}_v^{u'})}{m' \cdot k} \right)$
      end
    end
  end
end

The original class-common weight estimation function of RELIEF-F can also be rewritten as in Eq. (6.7), in terms of class-specific weights.

$$W(f_i) = \sum_{u=1}^{s} P(c_u) \omega(c_u, f_i) \; . \tag{6.7}$$

As seen in Eq. (6.7), RELIEF-F estimates the weights of the features by taking a weighted average of all class-specific weights and, thus, cannot reflect the characteristics of each class separately. Instead, we here propose to use weight estimations of each class separately. Consequently, this class-specific adaptation of RELIEF-F is presented with Algorithm 2.

We should also note that converting the original RELIEF-F algorithm into a class-specific version does not change computational complexity, since k hit / miss selection procedures and the number of processed samples do not change. As a result of having the same computational complexity, the approach can be accepted as scalable in terms

of the number of class, since the computational complexity of the algorithm is linearly proportional to the number of classes (as given in Section 6.2.4).

6.3.2 Multi-labeled / Noisy Datasets

In a typical multimedia retrieval task, each multimedia document (i.e. shot or video) is usually associated with a number of different semantic concepts. This situation reveals the problem of the multi-label feature selection, in which each sample is associated with multiple labels. In multimedia data, the multi-labeled characteristic of the data can be originated from either having more than one concept for each multimedia document in any single modality contained (e.g. having both an $airplane$ and a $mountain$ in a visual scene, as given in Figure 6.1), or containing different concepts in different modalities of the same document (e.g. having an $explosion$ sound in the audio modality and $military$ related vehicles in the visual modality at the same moment of the video).

In multi-label datasets, the samples are not mutually exclusive in terms of assigned labels, thus the discrimination of the samples between class labels becomes complicated. The discrimination of the samples between the retrieval classes is crucial to an effective feature selection. However, state-of-the-art studies accept the problem as a structural one, deal with converting the multi-labeled dataset into a single-labeled one for use with traditional feature selection methods [29, 65], and leave aside the cognitive aspect of the problem, which is also an important part of the problem. Here, the 'structural' side of the problem refers to the impossibility of using traditional learning / selection methods with the multi-labeled dataset due to the structure of the dataset, whereas the 'cognitive' side denotes the loss of the discrimination capability for learning. In this study, we regard both issues depicted and propose a two-step solution.

As the first step, we consider that it is not possible to use the RELIEF-F algorithm directly for a multi-label dataset, since having multi-labeled samples makes the nearest hit / miss selection procedure ambiguous. For instance, we need a solution to select the nearest hits/misses of a random instance with two different class labels, or a nearest item is labeled with two different classes. Thus, we first look into the state-of-the-art transformation methods. The most popular transformation methods in the literature

(a) (b)

Figure 6.1: Examples for multi-labeled shots. (a) $airplane$ and $mountain$ (b) car, $accident$, $people$ and $street$

are random assignment (RA), binary relevance (BR), label power set (LP) and pruned problem transformation (PPT) [132]. In the RA approach, a multi-labeled sample is randomly assigned to one of its classes. In BR, the dataset is transformed into $|\mathcal{C}|$ single-label datasets, where $\mathcal{C} = \langle c_u \rangle_{u=1}^{s}$ is the list of available classes. In any $\mathcal{D}_u = \{d \mid d \in \mathcal{D} \wedge C(d) = c_u\}$ of these datasets, the samples are labeled in a binary form, depending on whether a sample d is associated with class c_u or not. In LP, the basic idea is to convert the set of classes $\mathcal{C}$ into $\mathcal{C}'$ such that $\mathcal{C}'$ is the power set of $\mathcal{C}$ ($\mathcal{C}' = \mathcal{P}(\mathcal{C})$). PPT is an improvement on LP, where unused subsets are removed from $\mathcal{C}'$. However, using any of these approaches causes loss of either the effectiveness or the efficiency of the algorithm. Using RA makes the process nondeterministic and also loses a large amount of valuable information due to the random class selection; thus results in an ineffective solution. On the other hand, although BR, LP and PPT are potentially good solutions to prevent information loss, the process becomes computationally complex. Hence we focus on an alternative solution that enables use of the RELIEF-F algorithm for multimedia data and does not increase the computational complexity.

Assuming that c_i, c_j and c_k are three classes different from each other, we decompose the multi-label problem for RELIEF-F into three cases:

- Case-1: A random sample x is associated with both classes c_i and c_j. In this case, it is not clear which class will be accepted for hits and misses.

- Case-2: Random sample x is labeled with c_i, and y is one of the nearest neighbors of x. If y is labeled with both c_i and c_j, it is unclear whether such a neighbor instance is a hit or a miss.

- Case-3: Random sample x is labeled with c_i, and y is one of the nearest neighbors of x. The neighbor instance y is labeled with both c_j and c_k. In this case, it is clear that y is a miss. However, it is not clear which class of miss it is.

We first start with a BR-like method, which is very compatible with the class-specific extension of RELIEF-F discussed in Section 6.3.1. Different from BR, we do not generate $|\mathcal{C}|$ number of binary valued datasets. In accordance with the class-specific extension, an intuitive way to deal with these cases is to transform each multi-labeled sample into multiple single-labeled samples with the same feature values but different classes (as illustrated in Figure 6.2), and group the samples according to the class that they belong to. Thus we divide the training dataset into $|\mathcal{C}|$ number of subsets, each having the samples of a different class. During the execution of the algorithm, the random samples are selected among each subset iteratively, and finding the associated class of a sample is not problematic anymore, even if it is a multi-labeled sample originally. Thus, Case-1 is discarded. Actually, the use of a class-specific extension helps to prevent Case-1. For handling Case-2 and Case-3, the same transformation as with Case-1 is applicable. For Case-2, any multi-labeled neighbor instance y is replicated and transformed into y_{c_i} and y_{c_j}. Then, y_{c_i} is used as a hit instance and y_{c_j} is used as a miss instance, which actually means y is used both as a hit and a miss instance. Similarly, for Case-3, y is transformed into y_{c_j} and y_{c_k}, then y_{c_j} is used as a miss instance for class j, whereas y_{c_k} is used for class k.

Although this solution is an efficient approach to deal with the multi-labeled structure of training data and does not cause information loss as in BR transformation, it is still possible to lose some information due to the use of the same neighboring instances as both hits and misses (i.e. Case-2). Considering the weight estimation function of RELIEF-F given in Eq. (6.2), while calculating the weight of modality f by using random sample x, the effect of a neighbor hit instance y is as follows;

$$\delta_{hit} = -\frac{diff(f, x, y)}{m \cdot k} \ . \tag{6.8}$$

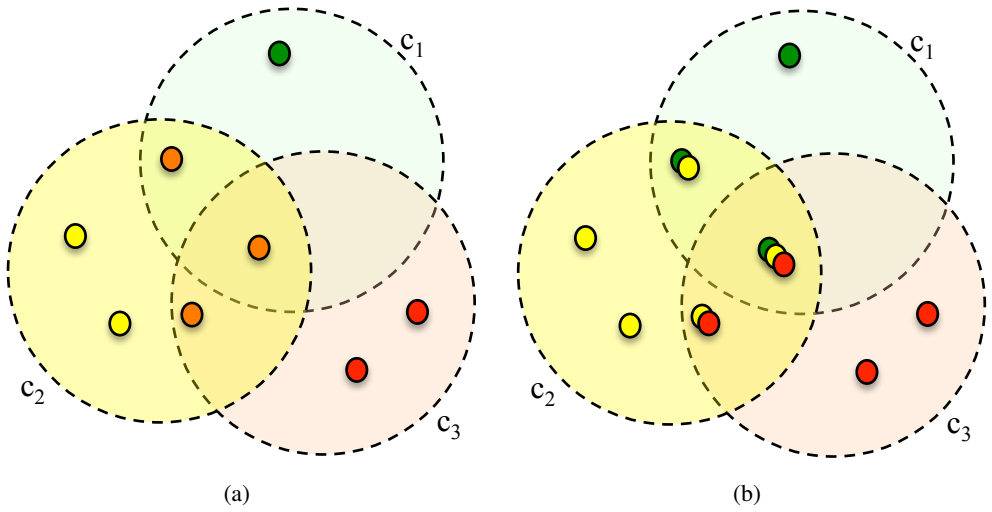


Figure 6.2: Transforming multi-labeled samples into multiple single-labeled samples. Small green, yellow and red circles denote c_1,c_2 and c_3 instances, respectively. Orange circles in (a) are multi-labeled instances, each of which is transformed into multiple single-labeled instances in (b).

However, if the neighbor instance y is a multi-labeled one as in Case-2, the same instance is used both as a hit and a miss instance. Thus, the net effect of the neighboring hit instance becomes:

$$\delta'_{hit} = \left(\frac{P(c_j)}{1 - P(c_i)} - 1 \right) \frac{diff(f, x, y)}{m \cdot k} \,. \tag{6.9}$$

In other words, the effect of the hit instance is decreased because of being a multi-labeled instance. The worst case of this situation, although practically impossible, occurs when the instance is labeled with all available classes. In such a situation, the effect of the instance equals to zero. In [65], Kong et al. propose to ignore the instances of Case-2, which is practically the same as assuming the situation is always the worst case. In our approach, we do not ignore such instances, since they may still provide some valuable information as long as the situation is not the worst case. We accept the decrease in the effect of the hit instances as a sort of noise and loss in the discrimination capability of the features.

In this aspect, we also consider the effect of noise in multimedia data. In addition to the fact that the multimedia data have an expected internal noise, the way we model the multimedia data can create an artificial noise. Since the multimedia data is usually large –even huge–, some sub-sampling (i.e. using shots and keyframes instead of

each particular frame) is done before processing it. The extracted features represent only subsamples from the video, whereas the ground truth labels are based on the full content of the video. Such a situation makes the evaluation of features complicated and eventually some of the ground truth instances appear as noisy instances. Similar to the multi-label issue, having noise in the samples prevents a correct discrimination between classes. In addition, depending directly on the distances between training instances affects the performance of the algorithm negatively, considering the noisy instances.

Consequently, the second step of our approach is based on strengthening the feature weighting mechanism of RELIEF-F. Thus, we introduce two new factors for the calculation of the weights, in addition to the discrimination capability: the representation and reliability characteristics. Having additional components in the weight calculation makes the algorithm less dependent on the discrimination capability, and provides better estimations. Hence, the class-specific weight of a feature, which was previously defined in Eq. (6.6), is updated as the following;

$$\varpi(c_u, f_i) = \begin{cases} \big(\omega(c_u, f_i)\big)^{\alpha} \cdot \gamma(c_u, f_i) \cdot \eta(c_u, f_i), & \text{if } \omega_f^c > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{6.10}$$

where ω, γ and η functions provide the discrimination, representation and reliability based weights, respectively. In addition, α is an experimental constant for tuning. Considering that RELIEF-F based weights are in $[-1, 1]$, and weights smaller than zero denote irrelevant features, we discard these by assigning zero. The proposed functions are discussed in detail below.

6.3.2.1 Discrimination Based Weight

The discrimination based weight ($\omega(c_u, f_i)$) refers to the weight calculated by using the data from all available classes with an aim to discriminate between those classes. The calculation of $\omega(c_u, f_i)$ is basically accepted as the way to calculate class-specific RELIEF-F (Eq. (6.6)).

6.3.2.2 Representation Based Weight

The representation based weight ($\gamma(c_u, f_i)$) refers to the weight calculated by using the data only from any single class, with an aim to represent that class independent of other classes. In order to measure its effectiveness by using only its characteristics and calculate such a weight, we assume that we can isolate the samples of a particular class from other classes. Here, isolation means that any sample labeled with other classes is always at a farthest location. Applying this idea to the class-specific RELIEF-F weight calculation gives the following: The distance of a random sample to any of its nearest misses always equals to 1 (note that $diff(f, x, y) \in [0, 1]$). Hence, the representation based weight becomes the following:

$$\gamma(c_u, f_i) = \frac{1}{|\mathcal{R}_u|} \sum_{r \in \mathcal{R}_u} \left[1 - \sum_{v=1}^{k} \frac{diff(f_i, r, \mathcal{H}_v)}{m \cdot k} \right] . \tag{6.11}$$

Eq. (6.11) can also be interpreted as the complement of the mean distance of a class to itself, so the weight of a feature is inversely proportional to the mean distance of the class to itself. Here, the mean distance of a class to itself is the average of all distances from each sample of a class to its k neighbor hits. It is expected for a particular class that the features with lower mean distance values represent the class better. Thus, $\gamma(c_u, f_i)$ is a sound metric to estimate the representation capability of a feature.

6.3.2.3 Reliability Based Weight

Reliability based weight ($\eta(c_u, f_i)$) refers to the weight calculated by using accuracy with respect to a feature for a particular class, with an aim to see whether it is reliable for that class or not. The idea of using the accuracies of features is based on the theoretical analysis of Fumera et al. [37]. Fumera et al. work on a late fusion scheme and show that the weight of a classifier for feature f_i should be inversely proportional to the error of the classifier.

Considering that RELIEF-MM is a filter method, and classification results are not available during the feature weighting, we propose to estimate the accuracy of each feature by comparing the intra-class distance of each class with the inter-class distances to other classes. The intra-class distance is defined as the mean distance of the samples

in c_u to their nearest k hits, whereas the inter-class distance is the mean distance of the samples in c_u to their nearest k misses from each different class $c_{u'} \neq c_u$. It is important for a feature to give the lowest distance values for the instances in a class which is the same as the class of the query instances. Thus, $\eta(c_u, f_i)$ provides an estimation for reliability by finding the number of inter-class distances (by means of different classes) that has a larger value than the intra-class distance. The formal representation of $\eta(c_u, f_i)$ is given in Eq. (6.12);

$$\eta(c_u, f_i) = \frac{\left|\{\mu(c_u, c_{u'}, f_i) | \begin{smallmatrix} c_{u'} \in \mathcal{C} - \{c_u\} \\ \wedge\ \mu(c_u, c_{u'}, f_i) > \mu(c_u, c_u, f_i) \end{smallmatrix}\}\right|}{s-1}\ , \tag{6.12}$$

$$\mu(c_u, c_{u'}, f_i) = \frac{1}{|\mathcal{R}_u|} \sum_{r \in \mathcal{R}_u} \left[\sum_{v=1}^{k} \left(diff(f_i, r, \mathcal{N}_v^{c_{u'}}) \right) \right]\ . \tag{6.13}$$

where $\mathcal{N}_v^{c_{u'}}$ is the vth $c_{u'}$-labeled nearest instance of sample r. Thus, $\mu(c_u, c_u, f_i)$ refers to the mean distance to k hits (intra-class distance), whereas $\mu(c_u, c_{u'}, f_i)$ with $c_{u'} \in \mathcal{C} - \{c_u\}$ is the mean distance to the k misses of any other class (inter-class distance).

6.3.3 Unbalanced Datasets

In multimedia datasets, some of the concepts occur less frequently than others, which causes the annotated training data to be unbalanced among different classes. We can consider the occurrences of *flag* vs. *car* objects through a random video as an example; a *flag* object usually occurs less often than a *car* object. Thus, the number of *car* samples is usually larger than the number of *flag* samples. One important consequence of frequent occurrence is having more representative and descriptive data than the infrequent concepts, e.g. it is possible to find several different models and colors of *car* samples, but it is hard to find the variations of *flag* samples. Hence, having an unbalanced dataset may prevent an adequate learning process.

Although the unbalanced dataset problem is usually discussed in the scope of classification and learning [19], the RELIEF-F algorithm, as a feature selection method, is also negatively affected by unbalanced data. The reason why RELIEF-F is affected by the imbalance in the data is the use of k nearest neighbors during the weight calculation.

As discussed in Section 6.2.3, RELIEF-F uses average distance to k nearest neighbors while calculating the weights, in order to eliminate the effect of outlier data. However, the placement of training samples in the multi-dimensional space and the amount of outliers are highly data-dependent, and can be very different for different domains and classes. Thus, we point out that selecting k number of nearest neighbors for every class is not a fair preference, when each class has a different number of samples. Using the same k number of neighbors hinders the use of an equal amount of information from all classes. For instance, a certain value of k may provide for the acquiring of all available patterns of a particular class. However, for another class, the same k value may provide for the acquisition of only a small ratio of the available patterns. The situation is not different if we consider the outlier data. Selecting the same number of neighbor instances from different classes (each of which has a different number of samples) may result in different ratios of outlier data for each class.

Considering the above given issues, we propose to select the value of k dynamically, i.e. a class-specific k value. However, enabling a class-specific k selection makes the process more complicated, despite the potential improvement in the estimation of feature weights. Thus, we propose another promising idea; using the k value as a certain ratio of sample count in a class. By employing such an idea, the k value of class c_u can be calculated by;

$$k_u = k_R \cdot |\mathcal{D}_u| \,, \tag{6.14}$$

where $\mathcal{D}_u = \{d \mid d \in \mathcal{D} \wedge C(d) = c_u\}$ is the set of training instances with class c_u, and $k_R \in [0, 1]$ is the nearest neighbor selection ratio, which is defined independently of the classes.

A weak point of this idea is that it requires us to assume approximately the same ratio of noise for all classes. Yet, this assumption can be practically applicable, considering that the datasets mostly do not suffer from the outliers because of mislabeling, but because of complexities related with the internal characteristics of video data, such as lighting variations, camera motion, occlusion, and noise in the sensed data. Mislabeling is a human-oriented noise, in which we cannot assume that the ratio of outliers are equal for different classes (e.g. it may be harder to annotate the samples with less frequently occurring classes). However, we can assume that the complexities in the video occur approximately in the same ratio for any class, especially when we have a broad range

of videos.

6.3.4 The Final Algorithm

The finalized RELIEF-MM algorithm including all of the extensions that we describe above is given in Algorithm 3. In order not to make the presentation of the algorithm more complex, some of the calculation including loops are represented by some mathematical functions (e.g. sum operations). The $\varpi(c_u, f_i)$ in Eq. (6.10) is represented with W matrix in the algorithm. The other parameters for calculating $\varpi(c_u, f_i)$ are given as they are given in Eq. (6.10).

The RELIEF-MM algorithm consists of three parts. Firstly, the parameters of the weight estimation function (ω, γ and η) are initialized. Secondly, these parameters are updated iteratively by encountering random training samples in the total of m. This process is performed separately for each class, thus some percentage of m (proportional to the prior probability of each class) is used for each class. Lastly, the calculated parameters are used to find the final values of weight estimations for each feature and class.

Here, it should be noted that the original RELIEF-F algorithm is an online algorithm, which means that the algorithm processes training instances one-by-one in a serial fashion, and can give an output after processing each instance. However, the RELIEF-MM algorithm presented in Algorithm 3 is offline, since the final weights are calculated as a batch instruction. Our choice eliminates further complexity in the algorithm. Yet, it is fairly straightforward to convert Algorithm 3 into an online version by moving the block between lines 23–31 into the for loop between lines 13–22, as the last instruction.

6.3.4.1 Complexity Analysis

We assume that n denotes number of features ($n = |\mathcal{F}|$), m denotes number of iterations, k_R denotes nearest neighbor selection ratio, s denotes number of classes ($s = |\mathcal{C}|$) and t denotes number of training instances ($t = |\mathcal{D}|$). Considering the

Algorithm 3: RELIEF-MM

Input: list of features $\mathcal{F} = \langle f_i \rangle_{i=1}^{n}$, number of iterations m, set of training instances $\mathcal{D} = \{d_j\}_{j=1}^{t}$, list of classes $\mathcal{C} = \langle c_u \rangle_{u=1}^{s}$, nearest neighbor selection ratio k_R, tuning constant α

Output: the weight matrix W

begin

`// Initialization`

for $u \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

  for $i \leftarrow 1$ **to** n **do** `//for each feature in` $\mathcal{F}$

    $\omega[u][i] \leftarrow 0$;

    $\gamma[u][i] \leftarrow 1$;

    $\eta[u][i] \leftarrow 0$;

    for $u' \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

      $\mu[u][u'][f] \leftarrow 0$;

`// Calculations`

for $u \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

  $\mathcal{D}_u \leftarrow getClassInstances(\mathcal{D}, c_u)$;

  $k_u \leftarrow k_R \cdot size(\mathcal{D}_u)$;

  $m' \leftarrow m \cdot P(c_u)$; `// P(`c_u`)=size(`$\mathcal{D}_u$`)/size(`$\mathcal{D}$`)`

  for $j \leftarrow 1$ **to** m' **do**

    $r \leftarrow randomInstance(\mathcal{D}_u)$;

    $\langle \mathcal{H}, \mathcal{M} \rangle \leftarrow findNearestHitsMisses(r, \mathcal{D}, k_u, \mathcal{C})$;

    for $i \leftarrow 1$ **to** n **do** `//for each feature in` $\mathcal{F}$

      $\omega[u][i] \leftarrow \omega[u][i] - \sum_{v=1}^{k_u} \frac{diff(f_i, r, \mathcal{H}_v)}{m' \cdot k_u} + \sum_{\substack{u'=1 \\ u' \neq u}}^{s} \left(\frac{P(c_{u'})}{1 - P(c_u))} \sum_{v=1}^{k_u} \frac{diff(f_i, r, \mathcal{M}_v^u)}{m' \cdot k_u} \right)$;

      $\gamma[u][i] \leftarrow \gamma[u][i] - \sum_{v=1}^{k_u} \frac{diff(f_i, r, \mathcal{H}_v)}{m' \cdot k_u}$;

      $\mu[u][u][i] \leftarrow \mu[u][u][i] + \sum_{v=1}^{k_u} \frac{diff(f_i, r, \mathcal{H}_v)}{m' \cdot k_u}$;

      for $u' \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

        if $u' = u$ **then** continue;

        $\mu[u][u'][i] \leftarrow \mu[u][u'][i] + \sum_{v=1}^{k_u} \frac{diff(f_i, r, \mathcal{M}_v^u)}{m' \cdot k_u}$;

`// Finalization`

for $u \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

  for $i \leftarrow 1$ **to** n **do** `//for each feature in` $\mathcal{F}$

    for $u' \leftarrow 1$ **to** s **do** `//for each class in` $\mathcal{C}$

      if $u' \neq u \wedge \mu[u][u'][i] > \mu[u][u][i]$ **then**

        $\eta[u][i] \leftarrow \eta[u][i] + \frac{1}{s-1}$

    if $\omega[u][i] > 0$ **then**

      $W[u][i] \leftarrow (\omega[u][i])^{\alpha} \cdot \gamma[u][i] \cdot \eta[u][i]$;

    else

      $W[u][i] \leftarrow 0$;

Algorithm 3, RELIEF-MM includes three main loops for initialization, calculation and finalization.

The first main loop (lines 2–8) initializes the parameter matrices and takes;

$$L_{init} = O(s^2 \cdot n) \; . \tag{6.15}$$

The second main loop (lines 9–22) is basically used for iterating over m instances from any class in C. Inside the loop, there are three operations, which are not $O(1)$; (1) filtering c_u-labeled instanced in $\mathcal{D}$, in line 10, (2) selection of hits and misses, in line 15, (2) weight parameter calculations, between lines 16–22. The first operation is performed once for each class in $\mathcal{C}$. The operation checks whether each instance in $\mathcal{D}$ is labeled with c_u or not, and takes $O(t)$ time. The second operation includes the distance calculation from random instance to all instances in $\mathcal{D}$, heap construction using the distances and neighbor selection from the heap. This process is similar to the case for RELIEF-F in Section 6.2.4, and takes $O(t \cdot n + t + k_u \cdot s \cdot \log t)$ steps. The third operation contains four instructions and is repeated for each feature. It takes $O(2 \cdot k_u \cdot s \cdot n + 2 \cdot k_u \cdot n)$ steps in total. The bounds for the second and third operations include a k_u term which is dependent on a class c_u. In other words, for each class $c_u \in \mathcal{C}$, the k_u value gets a different value based on Eq. (6.14). Considering that these operations are repeated for m' instances of s number of classes, the total complexity of these three operations becomes;

$$\begin{aligned}
&= \sum_{u=1}^{s} \left[O\Big(mP(c_u)(tn + k_R|\mathcal{D}_u|s \log t + k_R|\mathcal{D}_u|sn) \Big) \right] \; , \\
&= O\Big(m\, t\, n \; + \; m\, k_R\, s \; \log t \; \sum_{u=1}^{s} \big(P(c_u)|\mathcal{D}_u| \big) \\
&\qquad + \; m\, k_R\, s\, n \; \sum_{u=1}^{s} \big(P(c_u)|\mathcal{D}_u| \big) \Big) \; .
\end{aligned} \tag{6.16}$$

Considering that $P(c_u)$ is the prior probability of the classes and can be calculated by using the instance counts in each class, the summation term $\sum_{u=1}^{s}(P(c_u)|\mathcal{D}_u|)$ in Eq. (6.16) can be rewritten as $\sum_{u=1}^{s}(|\mathcal{D}_u|^2/t)$. The minimum value of this term is obtained when the dataset is balanced. For such a case, the term equals to t/s. The maximum value of the term is obtained with an unbalanced dataset, where one of the classes contains all t instances and the other classes contain no instances, although

this is practically impossible. In this case, the term equals to t. Thus the complexity bounds for the term is $\Omega(t/s)$ and $O(t)$. By applying this result in Eq. (6.16), the total complexity of the second main loop becomes

$$L_{calc} = O\big(m \cdot t \cdot n + m \cdot (k_R \cdot t) \cdot s \cdot \log t + m \cdot (k_R \cdot t) \cdot s \cdot n\big) \; . \tag{6.17}$$

The third main loop (lines 23–31) calculates the final weights by looping over all features and classes. It also includes a s-sized loop for finding the final value of η. The total complexity of the third main loop is;

$$L_{fin} = O(s^2 \cdot n) \; . \tag{6.18}$$

The total complexity of the RELIEF-MM algorithm can be obtained by adding the values in Eqs. (6.15),(6.17) and (6.18). Here, we consider that $t \geq s$ and $m \geq s$ should be true, since the algorithm implicitly makes an assumption that there should be at least one instance of each class, and also at least one instance should be selected from each class, in order to calculate the feature weights of each class. Hence, $m \cdot t \cdot n \geq s^2 \cdot n$. Consequently, the terms coming from the Eqs. (6.15) and (6.18) can be omitted for the calculation of the asymptotic upper bound. Then, the complexity of the RELIEF-MM algorithm equals

$$O(MM) = O\big(m \cdot t \cdot n + m \cdot (k_R \cdot t) \cdot s \cdot \log t + m \cdot (k_R \cdot t) \cdot s \cdot n\big) \; . \tag{6.19}$$

If the complexity of RELIEF-MM is compared with the complexity of RELIEF-F given in Section 6.2.4, it can be seen that the only difference lies in the terms related with the nearest neighbor selection. RELIEF-MM includes $(k_R \cdot t)$, whereas RELIEF-F has k. Essentially, these two terms are asymptotically equal in terms of complexity since both reside in the same range. Furthermore, if we consider using RELIEF-MM with balanced datasets, the $(k_R \cdot t)$ term turns into $(k_R \cdot \frac{t}{s})$, as described above. Thus, for small values of k_R, the complexity of RELIEF-MM for balanced datasets is $O(m \cdot t \cdot n)$, as it is for RELIEF-F. Here, the $m \cdot t \cdot n$ term is an asymptotic upper bound on $m \cdot (k_R \cdot t) \cdot \log t$ and $m \cdot (k_R \cdot t) \cdot n$, for small values of k_R. In conclusion, it can be said that the complexity of the RELIEF-MM algorithm is asymptotically the same as the original RELIEF-F algorithm.

The space complexity of RELIEF-MM also does not differ dramatically since all the work is on the same resources (inputs), and the biggest size requirement comes from the inputs. As discussed before, the space complexity of the RELIEF-F is bounded by $O(t \cdot n \cdot A)$, where A is assumed as the average size for a single feature. The space requirements for RELIEF-MM are; features ($O(n)$), weights ($O(n)$), classes ($O(s)$), nearest neighbor selection ($O(k \cdot s + n \cdot A)$) and parameters of estimation function ($O(3 \cdot s \cdot n)$). Considering that $t \geq s$, the space complexity of RELIEF-MM is bounded by $O(t \cdot n \cdot A)$.

6.3.5 Using RELIEF-MM with Prediction Scores

As mentioned before, in late fusion, the fusion is performed after a classification step. Thus, the inputs for the fusion process are the prediction scores obtained from the classifiers. In other words, the feature values of the samples may not available during the fusion process, in many cases. However, the original RELIEF algorithm uses the feature values of the samples in order to calculate the distances between them. Thus, in late fusion scenarios, where the feature values are not available, it is not possible to utilize the RELIEF algorithm. It is necessary to extend the weight calculation process of the RELIEF algorithm so that it can be used with the prediction score inputs.

Given that the classes are $\mathcal{C} = \langle c_u \rangle_{u=1}^{s}$, the modalities are $\mathcal{F} = \langle f_i \rangle_{i=1}^{n}$ and the training samples are $\mathcal{D} = \langle d_j \rangle_{j=1}^{t}$; the list of prediction probabilities for class c_u and modality f_i is $S_{c_u,f_i} = \{s_j^{c_u,f_i}\}_{j=1}^{t}$, where $0 \leq s_j^{c_u,f_i} \leq 1$. Note that the order of the samples in $\mathcal{D}$ and the score values in S_{c_u,f_i} are given correspondingly.

While using RELIEF-MM with prediction score inputs, the algorithm remains the same, but the $diff$ function calculation should be rewritten, since we do not have feature values anymore. Considering that an $s_j^{c_u,f_i}$ value of a sample d_j corresponds to the similarity of the sample to a predefined class (c_u), we utilize the following idea: The difference between similarities of two samples to the same pattern corresponds to a reasonable distance metric of these samples. Thus the $diff$ function in the RELIEF-MM algorithm can be updated as the differences of the score values of the samples. However, for each sample, there exist s number of scores of each modality, where each score is the similarity value for a different class. Thus, we consider that the

RELIEF-MM algorithm iterates over the training samples, and we use the score list which corresponds to the class of the randomly selected sample, on each turn. Thus, the $diff$ function becomes;

$$diff(f_i, d_x, d_y) = |s_x^{C(d_x),f_i} - s_y^{C(d_x),f_i}| \,, \tag{6.20}$$

where d_x is the randomly selected sample, d_y is one of the hit/miss instances for d_x, and $C(d_x)$ function corresponds to the ground truth class value of the sample d_x given as parameter.

6.4 Empirical Study

In this section, we evaluate the proposed modality weighting approach for semantic retrieval of multimedia data. For the retrieval task, the multimedia data is queried based on the semantic concepts. First, retrieval for each single modality is performed, then a multimodal retrieval is done. During the multimodal retrieval, the modalities are combined with a linear (weighted averaging) combiner based late-fusion approach, where the weights of the modalities are generated via different approaches.

In order to perform a detailed comparison, we carry out our empirical study in two major steps:

- Comparison with Other Approaches: We compare the retrieval accuracies of the RELIEF-MM based linear weighted fusion approach with a RELIEF-F based one, as well as the single modalities, basic approaches (simple averaging and maximum) and exhaustive search. Also, we compare the modality selection performance of RELIEF-MM with RELIEF-F in terms of the accuracies for each different number of feature selections.

- Tests for Each Extension Idea: After a comparison with alternative approaches, we focus on the issues that motivated us to develop RELIEF-MM, and perform tests comparing (i) class-common and class-specific selection, (ii) performances with multi-label, uni-label data and noisy cases (iv) using a dynamic vs. static nearest neighbor selection (k_R vs. k).

Table 6.1: Datasets

		TRECVID 2007	TRECVID 2008	CCV
Dataset length (hours)	Train	∼50	∼100	∼105
	Test	∼50	∼100	∼105
Number of videos	Train	110	219	4659
	Test	109	215	4658
Number of shots	Train	21,532	39,674	N/A
	Test	18,142	33,726	N/A

Considering that one of the important contributions of this study is the use of prediction scores with the RELIEF algorithm, the experiments are conducted with both of the following scenarios:

- We assume that the feature values are available, and use them to calculate the feature weights.
- We apply a pure late-fusion scenario by assuming that the feature values are not available. Thus, the prediction scores are used for weight calculation.

6.4.1 Experimental Setup

6.4.1.1 Datasets

Experiments are carried out on three frequently utilized benchmark datasets: TRECVID 2007 [91], TRECVID 2008 [92] and the Columbia Consumer Video (CCV) Database [55]. The dataset characteristics are summarized in Table 6.1. Further details and a performance comparison of TRECVID participants can be found in the corresponding references.

While using the TRECVID 2007 and 2008 dataset, we prefer using the outputs of common shot reference, for shot segmentation. For these datasets, the shots are used as the retrieval documents. Besides, for the CCV dataset, each video is accepted as a retrieval document. During the tests, the shots (for TRECVID 2007 / 2008) and the videos (for CCV) are considered as individual and independent documents, which

(a) TRECVID 2007 dataset

(b) TRECVID 2008 dataset

(c) CCV dataset

Figure 6.3: Query concepts for each dataset and sample shot images from query concepts.

means no contextual information or interaction is taken into account between shots / videos.

Each of the utilized datasets provides different sets of concept annotations. The annotations on all three datasets are provided in a multi-label manner, which means each shot can contain more than one label. A complete list of these concepts is given in Figure 6.3 with sample images. The semantic queries performed during the tests are based on these semantic concepts.

6.4.1.2 Modalities

For all datasets, we consider a multimodal setting, and use features from different modalities. However, we prefer a relaxed definition for 'modality' [143]. The modalities of multimedia data are usually accepted as audio, visual and text modalities, but each of these modalities can be expanded. For instance, visual data can be defined with several modalities like color, shape, texture and face. Here, each of these modal-

Table 6.2: Modalities Utilized for each Dataset

Dataset	Modalities
TRECVID 2007	MPEG-7 Color Layout (CL) MPEG-7 Region Shape (RS) MPEG-7 Edge Histogram (EH) Zero Crossing Rate and Energy (ZCRE) Mel-freq. Cepstrum Coefficients (MFCC) Term Freq.–Inverse Doc. Freq. (TF-IDF)
TRECVID 2008	Gabor Texture (GT) Edge Direction Histogram (EDH) Scale Inv. Feature Transform (SIFT) Grid-based Color Moment (GCM) Grid-based Wavelet Texture (GWT)
CCV	Scale Inv. Feature Transform (SIFT) Spatial-Temporal Interest Points (STIP) Mel-freq. Cepstrum Coefficients (MFCC)

ities is a different type of information source, and contains a significant amount of complementary information. Thus, we accept each different type of information (i.e. each complementary feature) as a different modality. The multimodal features utilized during the test are listed in Table 6.2.

As presented on the table, visual, audio and textual features are extracted from the videos of the TRECVID 2007 dataset. For visual features, one key frame per shot is adopted and the middle frame for each shot is selected as the key frame. The feature extraction and distance calculation tasks of the visual features are performed by using the MPEG-7 reference software (eXperimentation Model, XM) [83]. For audio features, the entire audio of each shot is processed and Yaafe toolbox [12] is utilized for feature extraction. For the textual features, the Automatic Speech Recognition and Machine Translation texts, which are provided by TRECVID, are employed. During the calculations, no stop-word filtering or preprocessing is done.

For TRECVID 2008, the features are not extracted; instead, the prediction score values of each shot for the concept queries are obtained from the CU-VIREO374 [53] dataset. In the CCV dataset some well-known features are already provided, as well as the videos and annotations. For more detailed explanations, interested readers can refer

to [53] and [55].

Considering that we combine the modalities with a late fusion process, features from each modality should be processed with a classifier and the prediction scores should be obtained before the combination (for TRECVID 2007 and CCV datasets). For the classification task, a Support Vector Machine (SVM) classifier with appropriate Radial Basis Function (RBF) based kernels is preferred, and LibSVM [18] is utilized.

6.4.1.3 Metrics

To measure the retrieval accuracy, *Precision*, *Recall*, *Average Precision (AP)* and *Mean Average Precision (MAP)* are used. *Precision* is the fraction of retrieved documents that are relevant to the query concept, while *Recall* is the fraction of relevant documents that are retrieved. The *AP* is the sum of the precision at each relevant hit in the retrieved list, divided by the minimum of the number of relevant documents in the collection and the length of the list. Regarding the evaluation rules of TRECVID, AP is measured at 2000. *MAP* is the *AP* averaged over several query concepts. In other words, the *AP* of each concept is calculated separately and then the *MAP* is found by averaging them. Beyond the measurements of accuracy, we also present the statistical significance of the obtained results. To do so, we perform a *student's t-test* with paired samples, where the pairs are the accuracy results for different concept queries. A paired t-test gives a p-value which denotes the significance of the improvement between two tests. The smaller the p-value, the more significant the difference of the two average values. We assume a confidence level at 0.95 and accept the results with p-value$<$0.05 as significant.

We define another metric, named Fusion Gain (FG), to perceive the effect of the fusion process. Fusion gain gives the relative performance increase between to two different configurations:

$$FG(x,y) = \frac{MAP(x) - MAP(y)}{MAP(y)} \ , \tag{6.21}$$

where x and y denote different configurations (i.e. different feature selections). In our experiments we calculate two *FG*s:

- FG_{BS}: The fusion gain is calculated by comparison with the best single modal-

ity.

- FG_{AVG}: The fusion gain is calculated by comparison with the simple averaging approach.

6.4.2 Comparison with Other Approaches

In order to see the effectiveness of RELIEF-MM, we first compare its retrieval accuracy with the following alternative methods,

- Each single modality,
- Basic approaches like maximum (MAX) and averaging (AVG),
- Class-common exhaustive search (Exh-CC), Class-specific exhaustive search (Exh-CS)
- Original RELIEF-F algorithm,

Using each single modality and basic approaches represents the lower accuracy bounds for the fusion system. A fusion system is accepted as successful if it provides better accuracy than any of the single modalities. We also consider the MAX and AVG approaches as lower bounds, since these are the most frequently utilized fusion approaches due to their simplicity in calculation. In the MAX approach, the decision in the fusion process is calculated by taking the maximum score value of the available modalities. In the AVG approach, the mean of the score values of all available modalities is accepted as the final decision. On the other hand, we also present the accuracies of the exhaustive search for finding optimal modality weights, which provides an upper bound for the retrieval accuracies. For the exhaustive search approach, we perform both class-common and class-specific weighting processes. Exh-CC evaluates every different weight set in order to find the optimal weight of each modality. In Exh-CS, the same process is repeated for each class, separately. Lastly, we compare our proposed approach with the original RELIEF-F algorithm, which exhibits the major contribution of this study. During these comparisons, for RELIEF-F and RELIEF-MM, the performances at the optimal k_R values are presented. The v value for RELIEF-MM is used as 2.

The use of an exhaustive search usually causes infeasible test situations. In our tests, the feasibility of the weight selection process via an exhaustive search depends on the precision of the weights, as well as the number of modalities. For instance, if we want to have a precision of 0.01 between weights with 6 modalities, we should check 100^6 cases. Assuming that we already have the prediction scores of each modality beforehand, such a process for TRECVID 2007 dataset would take so long that even parallelization of the process would not be a solution. Thus, we follow a computationally simpler search process without damaging the fairness of the comparisons. We perform the following two different near-exhaustive search process[3], and then select the best one: (i) We first perform an exhaustive binary selection among available modalities, and select the best 4 modalities. Then, we perform a weight search on the selected 4 modalities with 0.01 precision ($w \in \{0, 0.01, 0.02, ..., 0.99, 1\}$). After finding the optimal weights and fixing them, we perform a weight search on the remaining 2 modalities. (ii) We first perform an exhaustive weight search on all available modalities with 0.1 precision and find the optimal weights for each feature. Then, as a second step, we tune up the weights by performing a selection between $[w - 0.05, w + 0.05]$ with 0.01 precision.

In order to evaluate the proposed approach, one may argue that there should be comparisons with other available feature selection / weighting methods. However, as described in Section 6.1 and Section 6.2.2, currently available filter based feature selection / weighting methods in the literature are not easily applicable to the modality weighting problem, due to the issues of the intrinsic multi-dimensionality of modalities and the multivariate inputs of fusion systems. Thus, adapting other approaches to modality weighting problem is beyond the focus of this study. Besides, we do not consider comparing our method with several different wrapper approaches since we perform a comparison with the exhaustive search, which gives the best possible accuracy. It is also known that any wrapper approach is much more computationally complex than our approach. Consequently, we think that the comparisons included in this study are enough to evaluate the effectiveness and efficiency of our proposed approach.

[3] This two step process is applied for the TRECVID 2007 and 2008 datasets, where the number of modalities lead to inefficient situations. For the CCV dataset, an exhaustive weight search process is performed with 0.01 precision.

In Table 6.3, the MAP values of the above listed approaches are presented for the TRECVID 2007, TRECVID 2008 and CCV datasets. For a better understanding of which weighting approach provides more effective fusion, the Fusion Gains of these approaches are calculated and presented in Table 6.4. In addition to the accuracy results included here, a statistical significance analysis of the results is presented in Table 6.5. In the tables, (F) denotes the use of feature values as inputs to the RELIEF based algorithms, whereas (P) represents the cases where the predictions scores are used as the inputs.

From these experimental results, we arrive at the following observations:

- Combinations of different modalities give more accurate results than the single modalities. However, selection of modalities is a critical issue. A wrong selection can lead to worse results than the best of the single modalities. For instance, AVG cannot provide a positive gain in the TRECVID 2007 and 2008 datasets, compared to the best single modality. Similarly, a MAX approach is not successful in any of the three datasets. This is because of the fact that these simple methods do not perform an effective evaluation on the modalities, and thus they cannot discard the unfavorable modalities. Although these approaches provide the most efficient solutions, they cannot always provide an effective solution and they are not robust against different datasets. Consequently, a more robust and effective approach is highly recommended despite the risk of some decrease in efficiency.

- RELIEF-F is significantly better than the best single modality in one of two datasets where feature values are used as input, and one of three datasets where predictions scores are used. If compared with the AVG approach, RELIEF-F has a significant improvement in only one case out of all five. Hence, RELIEF-F is not a robust solution against different datasets. Still, it can be accepted as an applicable modality selection approach, since it does not provide retrieval accuracies worse than the best single modality, and usually performs slightly better.

- RELIEF-MM provides a significant improvement over the best single modality for all datasets when the feature values are used as input, and two of three datasets

Table 6.3: Comparison of Retrieval Accuracies. The first column denotes the configuration: (S) Single modalities, (B) Basic approaches, (E) Exhaustive Search, (F) RELIEF methods using feature values, (P) RELIEF methods using prediction scores

	TRECVID 2007		TRECVID 2008		CCV	
		MAP (%)		MAP (%)		MAP (%)
(S)	CL	8.711	EDH	10.479	SIFT	49.676
	EH	9.032	GT	10.802	STIP	39.959
	RS	6.762	SIFT	19.032	MFCC	27.585
	ZCRE	6.385	GCM	13.027		
	MFCC	6.884	GWT	9.094		
	TFIDF	6.286				
(B)	MAX	6.639	MAX	17.126	MAX	52.071
	AVG	8.270	AVG	18.969	AVG	57.340
(E)	Exh-CC	10.322	Exh-CC	20.034	Exh-CC	57.403
	Exh-CS	12.988	Exh-CS	22.183	Exh-CS	57.783
(F)	RELIEF-F	9.847			RELIEF-F	56.027
	RELIEF-MM	10.563			RELIEF-MM	57.511
(P)	RELIEF-F	9.076	RELIEF-F	19.760	RELIEF-F	55.380
	RELIEF-MM	9.454	RELIEF-MM	20.559	RELIEF-MM	57.562

Table 6.4: Fusion Gains wrt. Best single modality and AVG approach.

		TRECVID 2007		TRECVID 2008		CCV	
		FG_{BS} (%)	FG_{AVG} (%)	FG_{BS} (%)	FG_{AVG} (%)	FG_{BS} (%)	FG_{AVG} (%)
(B)	MAX	-26.500	-19.726	-10.013	-9.718	4.822	-9.188
	AVG	-8.439		-0.327		15.428	
(E)	Exh-CC	14.283	24.816	5.266	5.612	15.556	0.110
	Exh-CS	43.792	57.045	16.558	16.941	16.321	0.773
(F)	RELIEF-F	9.016	19.064			12.786	-2.289
	RELIEF-MM	16.944	27.723			15.773	0.298
(P)	RELIEF-F	0.483	9.744	3.829	4.170	11.483	-3.418
	RELIEF-MM	4.669	14.316	8.023	8.378	15.877	0.388

Table 6.5: Statistical Significance Analysis using Paired T-Test. Pairs are based on query concepts. Statistically significant results according to the confidence level 0.95 (p-value$<$0.05) are given with an asterisk. p-value (BEST), (AVG) and (RELIEF-F) denote the p-values with respect to the best single modality, simple averaging and RELIEF-F approaches, respectively.

			p-value (BEST)	p-value (AVG)	p-value (RELIEF-F)
TRECVID 2007	(F)	RELIEF-F	7.39E-02	*4.67E-02	
		RELIEF-MM	*1.96E-02	*2.90E-02	*1.95E-02
	(P)	RELIEF-F	4.75E-01	2.45E-01	
		RELIEF-MM	2.65E-01	1.51E-01	*1.89E-02
TRECVID 2008	(P)	RELIEF-F	4.04E-01	5.60E-02	
		RELIEF-MM	*1.87E-02	*2.30E-03	*4.02E-04
CCV	(F)	RELIEF-F	*6.74E-06	*2.38E-03	
		RELIEF-MM	*2.02E-06	1.24E-01	*4.43E-05
	(P)	RELIEF-F	*2.08E-04	*4.72E-13	
		RELIEF-MM	*1.92E-06	*4.29E-02	*6.99E-10

when the prediction scores are used. If compared with AVG results, RELIEF-MM is significantly better in one of the two datasets with the feature value inputs, and in two out of three datasets with the prediction scores. When RELIEF-MM is compared with RELIEF-F, it is observed that RELIEF-MM obtains higher retrieval accuracies than RELIEF-F in all cases, each having a p-value$<$0.05. In addition, it should be noted that RELIEF-MM achieves higher accuracy results than the best single modality, AVG and MAX approaches, and even slightly better results than the Exh-CC approach. Thus, there is strong evidence that the RELIEF-MM approach introduces a significant improvement and can be accepted as a robust and effective solution as a modality weighting approach for multimedia data. Therefore, RELIEF-MM can be regarded as a practical enhancement for the multimedia retrieval studies using simple averaging for fusion.

- An exhaustive search finds the optimal feature selection since it evaluates all possible combinations. The accuracy results show that the use of a class-specific approach in the exhaustive search (Exh-CS) helps to improve retrieval accuracy in all three datasets. Besides, being a class-specific approach, RELIEF-MM is not upper-bounded with Exh-CC, whereas the accuracies of RELIEF-F are always less than Exh-CC.

- The performance of using prediction scores instead of feature values for calculating the modality weights depends on the characteristics of the dataset. In our experiments, results with the TRECVID 2008 and CCV datasets are reasonably good. However, for TRECVID 2007 dataset, there exists a considerable decrease in accuracy according to the results of using feature values. Thus, it may be hard to give a conclusive decision about the effectiveness of using prediction scores, with the current evidence. Nevertheless, the accuracies with the prediction scores outperform best single modality, MAX and AVG approaches. Consequently, we observe that the results of using prediction scores is promising and they are applicable when the feature values are not available during the fusion process.

The efficiency of the proposed approach is another important concern. A running time comparison of RELIEF-F, RELIEF-MM, Exh-CC and Exh-CS is presented in

Table 6.6: Approximate Execution Times of Exhaustive and RELIEF based methods, on three different datasets. The column with an asterisk denotes estimated values for a real exhaustive search scenario.

	RELIEF-F	RELIEF-MM	Exh-CC	Exh-CS	*Exh-CC	*Exh-CS
TRECVID 2007	3 sec	6 sec	17 hours	340 hours	19 years	380 years
TRECVID 2008	10 sec	11 sec	22 hours	440 hours	100 days	5.5 years
CCV	2 sec	2 sec	0.14 hours	2.7 hours	0.14 hours	2.7 hours

Table 6.6. The measurements are taken on a machine with "Intel(R) Xeon(R) CPU E5530 @2.40GHz". The values on the graph and table are obtained without a parallel programming approach. The values given in the table correspond to the cases presented in Table 6.3. The table includes both the near-exhaustive search running times and the estimated real exhaustive search times. The basic approaches (AVG and MAX) are not given in the table since they are done at no cost. Furthermore, a detailed comparison of RELIEF-MM and RELIEF-F, for different k_R nearest neighbor selections, is presented in Figure 6.4. According to the given experimental results, execution of RELIEF-MM results in a small increase in time, which is in parallel with the complexity analysis given in Section 6.3.4.1. Besides, the exhaustive search methods, even the near-exhaustive search, require a high time cost, as expected. Hence, RELIEF-MM can be accepted as an efficient modality weighting approach, considering that the

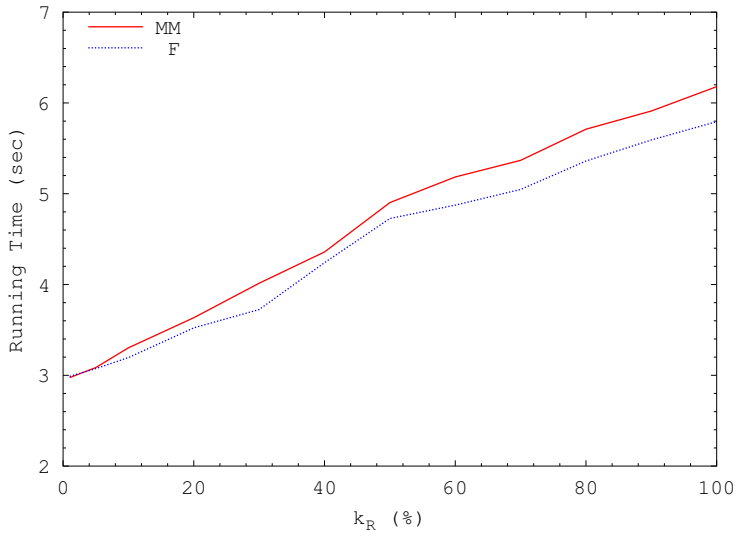


Figure 6.4: Running Time Comparison of RELIEF-F and RELIEF-MM on TRECVID 2007 dataset for different k_R values.

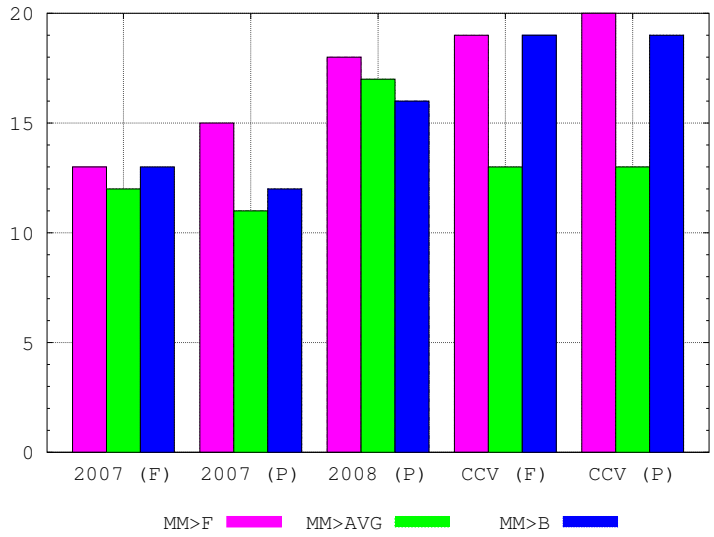


Figure 6.5: Concept-based Accuracy Comparison of RELIEF-MM with other approaches. Columns indicate the number concepts that RELIEF-MM provides higher accuracy than the compared approach. Each group of columns denote a different dataset with a specific input type. MM>F: RELIEF-MM vs. RELIEF-F, MM>AVG: RELIEF-MM vs. Simple Averaging, MM>B: RELIEF-MM vs. Best Single Modality.

time cost is a polynomial function of the number of modalities, and thus much more efficient than the exhaustive search. When compared with basic approaches, the cost of RELIEF-MM is still acceptable, considering the improvement in the retrieval accuracy.

Up until now, the average query performances have been compared. In order to make a more detailed comparison, we also perform a concept-based analysis. Figure 6.5 illustrates a concept-based comparison and presents the number of concepts for which RELIEF-MM provides higher accuracy when compared with a particular approach. In addition, precision-recall graphs for some of the query concepts are given in Figure 6.6 and Figure 6.7. According to the given experimental results, RELIEF-MM achieves higher accuracies in a larger number of concepts than RELIEF-F, the best single modality and AVG approaches, regardless of the used dataset and the input type (feature values vs. prediction scores). Nonetheless, the success rate of RELIEF-MM compared to RELIEF-F is more pronounced than the best single modality and AVG approaches. Under this observation, we can infer that the improvement provided by RELIEF-MM is reasonably good, due to the extensions introduced in this study. However the RELIEF idea in general may lead to difficulties in some particular data

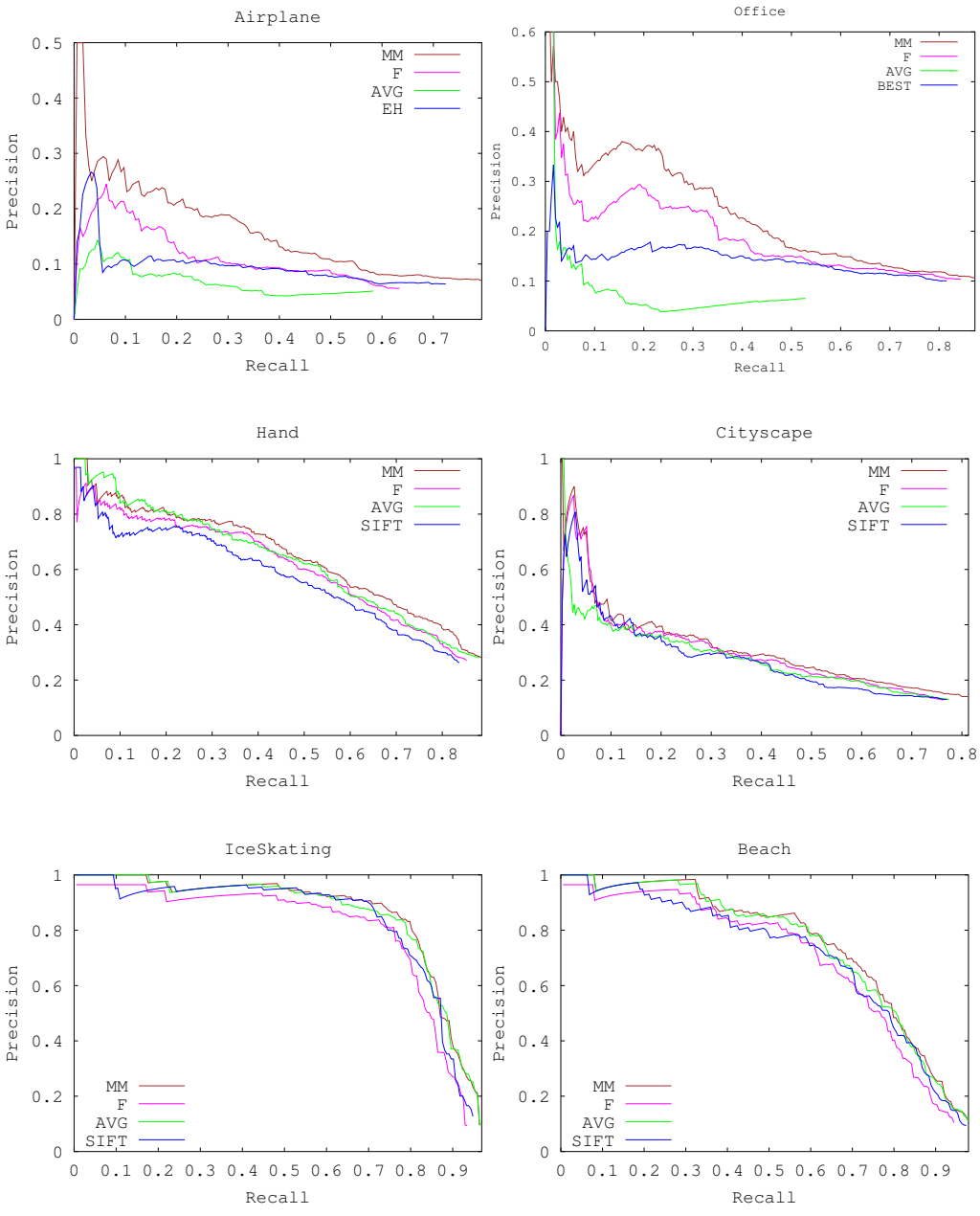


Figure 6.6: Precision-Recall graphs of some selected concepts, which are best-case examples for RELIEF-MM (in terms of accuracy). The rows contain concepts from TRECVID 2007, TRECVID 2008 and CCV datasets, respectively.

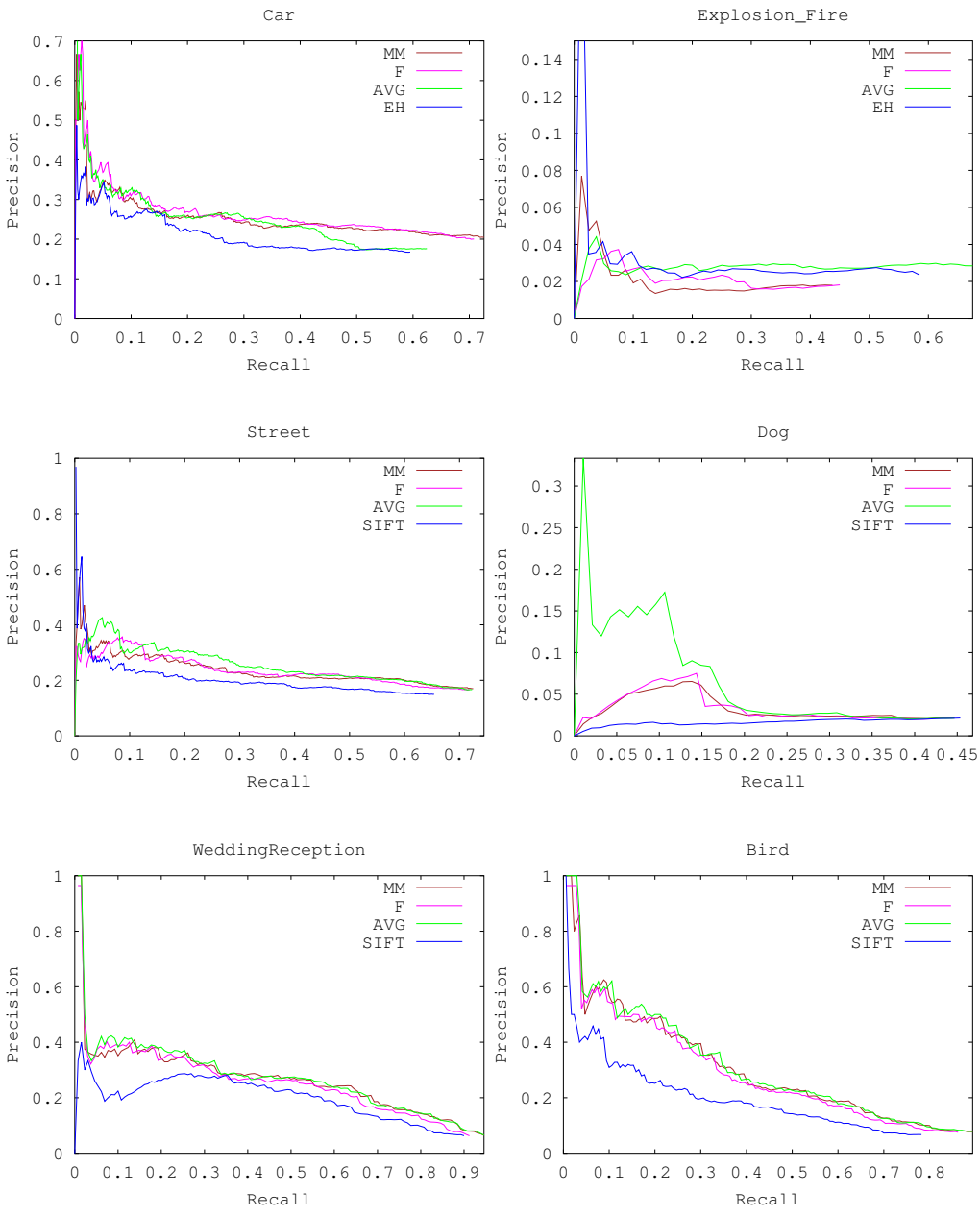


Figure 6.7: Precision-Recall graphs of some selected concepts, which are worst-case examples for RELIEF-MM (in terms of accuracy). The rows contain concepts from TRECVID 2007, TRECVID 2008 and CCV datasets, respectively.

distributions and is open to improvement. Even though the RELIEF algorithm utilizes a margin based nonlinear classifier [126] to evaluate the features and a margin based nonlinear classifier is known to be successful in general, the way RELIEF uses the input data is based on a standard procedure of employing the distances from each training sample to its neighbors, and does not benefit from any feature transformations in kernel space. This approach may be inadequate for some particular concepts that have unique data distributions. Just as employing various kernel types in SVM classifiers according to the characteristics of data and features leads to more effective classification results, so performing some appropriate kernel transformations on the RELIEF input data will help to make the RELIEF approach superior in a larger number of query concepts. However, such a problem is not included within the scope of this study, and has been left for future work.

Beyond the discussion on kernel transformation, one may focus on the comparison between RELIEF-MM and RELIEF-F, and expect that a class-specific approach, i.e. RELIEF-MM, should have an ability to optimize the weights for every query concept individually and thus achieve higher retrieval accuracies in any concept. Insofar as our observations have shown, we think that there exist two important factors that prevent RELIEF-MM from giving the best accuracies in some of the concepts.

The first reason for RELIEF-MM's less-than-optimal accuracy with some concepts is the small number of training samples for some particular concepts, which lead to incomplete representation of the concept. As explained in Section 6.3.1, RELIEF-MM takes the samples of each concept into account for the weight calculation, whereas RELIEF-F uses all training samples without considering the concept that they belong to. As a result, the weight calculation of the concepts with a small number of training samples may lead to ineffective results. On the other hand, RELIEF-F gains a general insight into the effectiveness of each modality, which usually provides better results than the estimations of RELIEF-MM which are based on inadequate data. *Explosion_-Fire*, *Desert*, *Flag* and *Truck* in the TRECVID 2007 dataset are some of the concepts for which RELIEF-F gives better accuracies. These concepts include 46, 67, 12 and 126 samples, respectively, whereas the dataset contains more than 350 samples per concept on average. A performance visualization for these kinds of concepts is given in the first column of Figure 6.7. It is also worth noting that TRECVID 2008 and

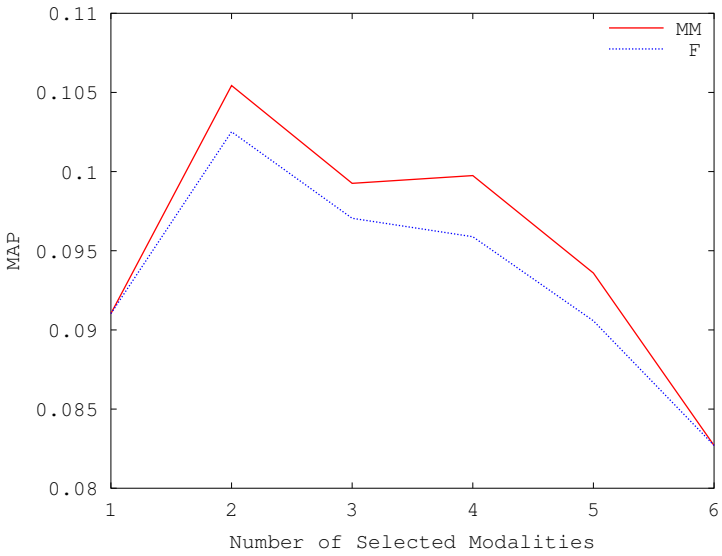


Figure 6.8: Modality Selection Performances

CCV include less concepts with a small number of samples, and thus the performance of RELIEF-MM is better in these two datasets than TRECVID 2007, as seen in Figure 6.5.

Second reason for RELIEF-MM's weakness for some concepts is the intra-concept sample variety, which can be accepted as a side effect of including γ_f^c and η_f^c into the weight estimation function. As mentioned above, the way in which the margin based classifier is utilized in RELIEF may be inadequate for some particular concepts that have unique data distributions. In RELIEF-MM we extend the weight estimation function and include γ_f^c and η_f^c into the formula. This preference increases the effect of margin based calculations in the function since both γ_f^c and η_f^c are calculated by using the intra-concept and inter-concept distances. Even though such preference makes the weight estimations better in most of the cases, increasing the effect of margin based calculations without a feature transformation in the kernel space may lead to worse weight estimations. As a solution, two alternatives can be considered, a kernel transformation or including a non-margin based variable into the weight estimation function, which can be considered for future work.

As a last comparison between RELIEF-MM and RELIEF-F, we try a different scenario from the previous tests, and combine the modalities with a simple averaging approach

after a hard selection of the modalities instead of weighting. In modality selection for fusion, the ultimate goal is to find which subset of the modalities is more effective for the retrieval task. It is therefore important to rank the modalities correctly, and this scenario helps us to do so. In Figure 6.8, the retrieval accuracies of RELIEF-F and RELIEF-MM are presented for those cases where a different number of modalities are selected and combined. During the test, firstly the weights of the modalities are obtained via the RELIEF-F and RELIEF-MM algorithms. Then for a particular number of modalities are selected according the assigned weights. The results show that RELIEF-MM is clearly superior to the original RELIEF-F algorithm in this task. Hence, it can be said that the ranking capability of RELIEF-MM is more effective than that of RELIEF-F.

6.4.3 Tests for Each Extension Idea

In order to further analyze the improvements that RELIEF-MM provides, we compare our proposed algorithm with the baseline RELIEF-F algorithm with respect to each idea presented in Section 6.3. Below, each idea is discussed in a separate sub-section. Through this evaluation, the TRECVID 2007 dataset is utilized.

The first improvement issue in RELIEF-MM is the conversion of the original RELIEF-F algorithm, which is a class-common approach, into a class-specific one. Thus, we compare the retrieval accuracies of the class-specific adaptation of RELIEF-F algorithm, which is introduced in Algorithm 2, with the original RELIEF-F. Moreover, we include the retrieval performances of RELIEF-MM algorithm in order to provide a more complete representation. In Figure 6.9, precision recall curves of these three methods are compared for optimized k selections. In addition, Figure 6.10 presents the retrieval performances of the given approaches with respect to different values of k nearest neighbors.

6.4.3.1 Class-Common vs. Class-Specific Feature Weighting

Figure 6.9 and Figure 6.10 show that RELIEF-MM provides higher retrieval accuracies than both of the original RELIEF-F algorithm and the class-specific adaptation of

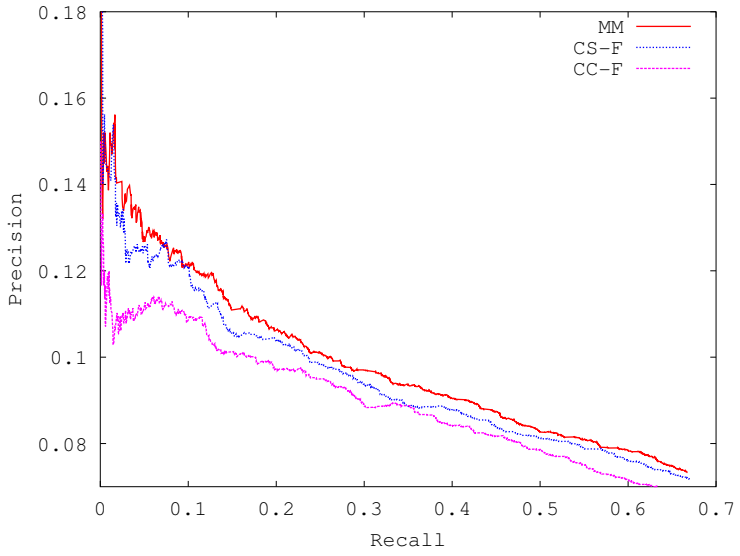


Figure 6.9: Precision-Recall Curves of the original RELIEF-F (CC-F), class-specific RELIEF-F (CS-F) and RELIEF-MM (MM) algorithms.

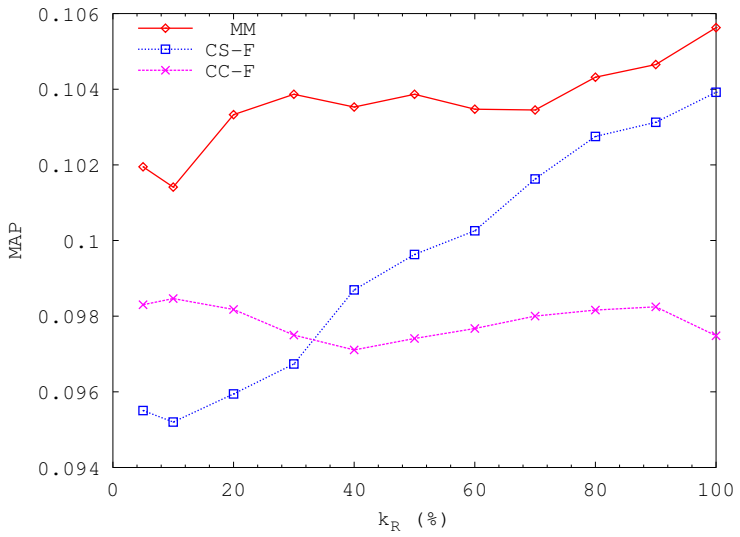


Figure 6.10: Retrieval Performances of original RELIEF-F (CC-F), class-specific RELIEF-F (CS-F) and RELIEF-MM (MM) algorithms.

RELIEF-F, for different values of nearest neighbors. Furthermore, Figure 6.9 presents the clear superiority of the class-specific approach over the original one, and Figure 6.10 shows that the accuracy of the original approach decreases, as the number of neighbors is increased. However, in the class-specific RELIEF-F, the accuracy is almost directly proportional to the number of neighbors. In addition, until some point around 33% of nearest neighbors selection, the original RELIEF-F performs better than the class-specific RELIEF-F, which means that the original algorithm is more powerful than the class-specific approach for a small number of neighbors. The reason for this situation is discussed below.

In discrimination based approaches, one of the most important factors that affects the success of the approach is the variety of the encountered samples. The original RELIEF-F algorithm estimates the weights by processing the randomly selected m samples and k neighbors of each sample from $s-1$ classes. Equivalently, class-specific RELIEF-F allocates the randomly selected m samples into s classes according to the prior probabilities of each class, and processes the samples of each class separately. Hence, the weights of each class are estimated by using a smaller number of samples according to m. If the number of nearest neighbors k is also small, the information obtained from the distances between samples becomes limited, which directly affects the success of class-specific RELIEF-F. If the number of nearest neighbors is increased, it is certain that the algorithm encounters some neighboring samples which have not been seen before, so that the algorithm obtains some adequate number of sample distances to estimate more effective weights. On the contrary, the original RELIEF-F usually does not encounter new samples when the number of nearest neighbors is increased, since many of the samples are seen through the m sample selection. If m is chosen as all training samples, there is no new instance that can provide new information while k is increased. Therefore, the only factor affecting the success of the original RELIEF-F algorithm becomes the noisy information obtained due to the increase in k. Consequently, it is more beneficial in this test to see which of the approaches can achieve higher accuracy in any configuration, since those upper bounds present how effectively they can use the available information. In Figure 6.9, it is apparent that class-specific RELIEF-F uses the available information more effectively.

6.4.3.2 Performances with Uni-label, Multi-label and Noisy Data

Another improvement of RELIEF-MM is its ability to handle multi-label data. Thus, we compare the retrieval accuracies of the fusion systems using RELIEF-F and RELIEF-MM for weight generation in uni-label and multi-label data. This comparison helps us to understand whether RELIEF-MM is more effective in multi-label data.

In order to obtain a uni-label data, we first process the training dataset and remove the multi-labeled instances from the dataset. We use the newly constructed uni-label dataset only for the weight generation step of the fusion process. The classifiers, which give the inputs to the fusion process, are always trained with the multi-label dataset. Thus, we manage to compare only the effect of different weight generation methods. Furthermore, it should be noted that constructing the uni-label dataset by removing the multi-labeled instances may cause the loss of some information (e.g. approximately 40% of the training instances is removed) and affect the performance of weight generation. Still, using a completely different uni-label dataset prevents us from comparing the accuracies of a weighting approach across datasets. Consequently, we find this setting fair enough to compare the effectiveness of RELIEF-MM and RELIEF-F.

The tests are conducted for several k_R nearest neighbor selections. Figure 6.11 presents the retrieval accuracies of RELIEF-MM and RELIEF-F by using uni-label and multi-label data for weight generation. In order to understand the effect of using multi-label data, the differences between the accuracies of RELIEF-F and RELIEF-MM can be compared for uni-label and multi-label datasets. Such differences can be best understood by the area between the curves of RELIEF-F and RELIEF-MM in the given graph. As seen on the graph, the area between RELIEF-F and RELIEF-MM curves is larger for multi-label data, which can be evaluated as RELIEF-MM working better in multi-label data.

In addition to the uni-label vs. multi-label data comparison, we also consider the performances of the algorithms for noisy data. In Section 6.3 it is proposed that RELIEF-MM should perform better than RELIEF-F even in noisy data cases. Thus, we compare the performances of RELIEF and RELIEF-MM with noisy data-sets.

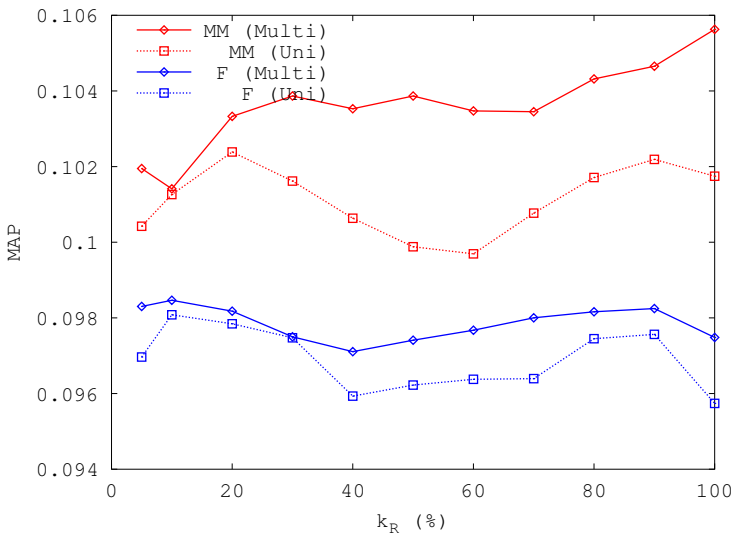


Figure 6.11: Retrieval Accuracies of RELIEF-F and RELIEF-MM for different k_R values, with Uni-label and Multi-label Training Data for Weight generation. MM and F denote RELIEF-MM and RELIEF-F. (Uni) and (Multi) denote the use of Uni-label and Multi-label datasets for weight generation.

For this purpose, we manually add mislabeled instances into the multi-label dataset, and construct 10%, 30% and 50% noisy datasets. Similar to the tests for uni-label data, these noisy datasets are used only for the weight generation step. The retrieval accuracies at given noise levels are presented in Figure 6.12, as well as the zero noise level.

Figure 6.12 demonstrates that the decrease in accuracy is usually larger for RELIEF-F, as the noise increases. Furthermore it is observed that RELIEF-MM is superior to RELIEF-F at any noise level. It can be stated that RELIEF-MM is more robust against noise.

6.4.3.3 Using k vs. k_R

One more improvement on the original RELIEF-F is the dynamic selection of k nearest neighbor as a ratio value of the class sample counts. The changes in retrieval accuracy change according to different k nearest neighbors are shown in Fig 6.13(a). The change according to different k_R nearest neighbors is shown in Fig. 6.13(b).

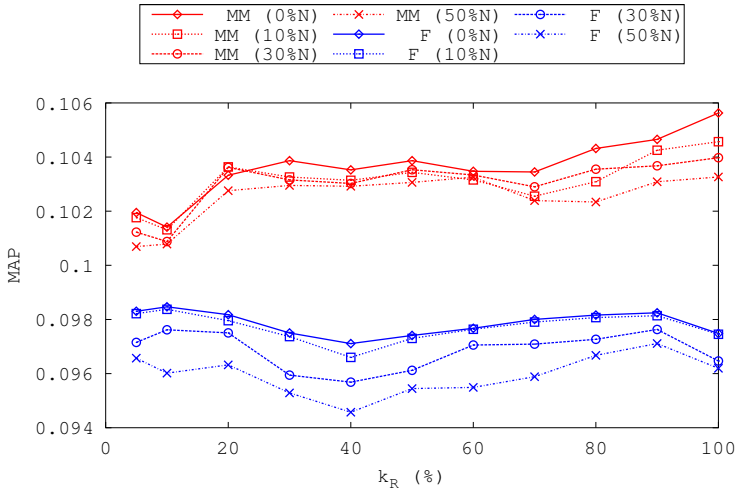
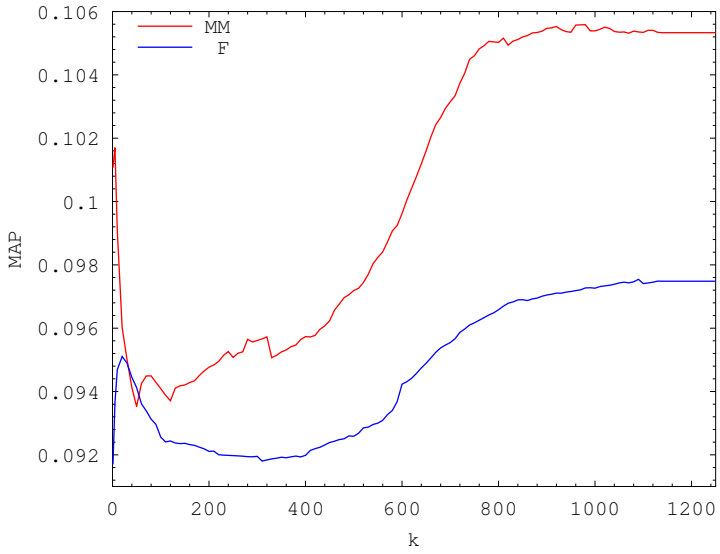
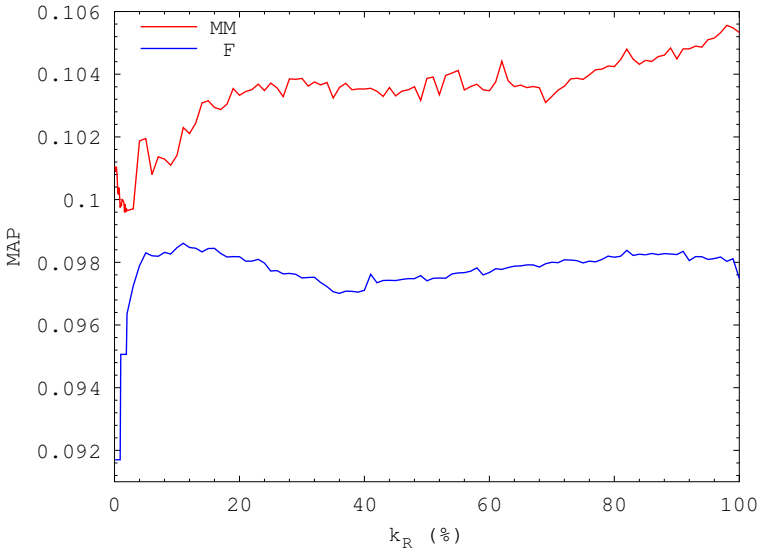


Figure 6.12: Retrieval Performances with Different Levels of Noisy Training Data for Weight generation.

For each of the methods, it is expected that accuracy values will converge into the same value when k reaches the number of all training instances and k_R reaches 100%. The improvement that k_R provides is more apparent in lower numbers of training samples. Figure 6.13 shows that both approaches exhibit a decrease in performance when 100-500 nearest neighbors are used. The main reason for the decrease is the use of imbalanced hit and miss instances for concepts that have a smaller number of samples, e.g. using $k = 400$ for a concept with only 200 samples causes the algorithm to use 200 hit instances, but 400 miss instances. Considering that RELIEF-MM works with class specific preference, the decrease in accuracy becomes more dramatic for RELIEF-MM. On the other hand, the use of a dynamic selection with ratios (k_R) prevents such a decrease for both methods and enables more robust accuracy results against a different number of nearest neighbor selections. k_R is bounded by the number of samples in the class, thus the decrease caused by imbalanced hits/misses does not occur any more.



(a) k Nearest Neighbors



(b) k_R Nearest Neighbors

Figure 6.13: Retrieval Performances According to k vs. k_R Nearest Neighbors

6.5 Evaluation of Fusion System Design

Considering the general fusion framework proposed in Section 3.1, an evaluation of the fusion architecture described in this chapter is given below. The approach is based on a 'multi-modal, multi-classifier' fusion scenario and focuses on the 'What to Fuse' problem. Below, how each affecting factor is handled through the proposed solution is described.

- Fusion Setting: The approach combines multiple modalities, each of the modality being a different feature. Before combination, the data of each modality is classified with a separate classifier, and the results of the classifiers are combined.
- Selection of Sources: RELIEF-MM is a modality weighting approach, which has a capability to be used as an online algorithm. Thus, the approach can be accepted as a dynamic solution for feature weighting. In addition, the weights have a context relation, since the approach is based on the class-specific feature selection idea.
- Fusion Strategy: The approach focuses on the use of complementary information for fusion.
- Content Representation: For content representation, both feature-based and score-based representation is applicable. In score-based representation, the classification scores of the samples are stored and processed.
- Normalization of Sources: The fusion inputs are classifier outputs, where each of them lays in between $[0, 1]$. Thus, a normalization process is not applied on the fusion inputs.
- Fusion Level: The approach is a late fusion approach.
- Fusion Methodology: Considering that the focus of the study is the feature / modality weighting, linear weighted averaging approach is utilized as the fusion methodology.
- Operation Modes: The mode for operation is a parallel scheme.
- Synchronization: A simple shot or video based synchronization is applied.

- Adaptation: Considering that proposed algorithm can be used as an online algorithm, the approach can be accepted as an adaptive solution.

6.6 Remarks

In this chapter, the problem of modality weighting for multimodal information fusion is studied. As an effective and efficient modality weighting solution, a RELIEF based approach is proposed. Considering the problems with RELIEF-F when using it with multimedia data for multimodal fusion, we focus on five crucial issues and extend the original RELIEF-F algorithm in these aspects. We first convert the original algorithm into a class-specific representation. Then we extend the algorithm and weight estimation function so that they estimate the modality weights better with multi-label and noisy data. For better estimations, we include the representation and reliability characteristics of modalities into the weight estimation function, in addition to the currently available discrimination capability. We also make an extension in order to make the algorithm more effectively with unbalanced datasets. Lastly, we introduce a conversion procedure that enables the use of classifier predictions in RELIEF, considering that feature values may not be available during a fusion process.

Our approach is extensively tested on TRECVID 2007, TRECVID 2008 and CCV datasets with several modalities in a multimodal information fusion scenario. The results show that using RELIEF-MM guarantees higher accuracies than any single modality, and shows much better performance than simple averaging and RELIEF-F based methods. In addition, RELIEF-MM provides slightly better performance than the class-common exhaustive-search based approach, although it is computationally much more efficient. We also perform several comparative tests against the RELIEF-F approach, aiming to examine each extension idea, and confirm that the proposed extensions lead to improvements on RELIEF-F. Consequently, we argue that our proposed approach is a timely efficient, accurate and robust way of modality selection.

The experiments carried out also exhibit some situations for future work. In order to further improve RELIEF-MM, we put forward the following ideas for future study:

- The RELIEF-MM algorithm utilizes a margin based discrimination approach,

like the original RELIEF, while evaluating features. Performing some appropriate feature transformations on the kernel space may improve the quality of the weight estimations, especially for those particular concepts that have unique data distributions. Performing such transformations separately for each concept type, like a one-vs.-all approach, may yield better results.

- Another improvement idea for RELIEF-MM is to include a non-margin based evaluation metric into the weight estimation function (e.g. mutual information, information gain, correlation, etc). Any considerable metric may have its own complications when being used with modalities instead of features, deficiencies for multimedia data and extra computational complexity, however. All these factors should be analyzed in detail.

- It is possible to further increase the efficiency of the RELIEF-MM algorithm by employing some caching mechanisms (e.g. k-d trees, hashing).

CHAPTER 7

COMBINING BAGS-OF-WORDS: A NOVEL MINING AND GRAPH BASED APPROACH

In this chapter, the problem of finding a way to fuse the modalities effectively is taken into consideration. The approach focuses on the combination of the components used in the state-of-the-art studies. Thus, the proposed approach provides a novel mining and graph based combination method for combining the Bags of Words (BoW) obtained from different modalities. Considering the fact that most of the studies do not use intramodal and intermodal relations of the available words effectively, our approach combines the classification outputs of each single modality, intramodal relations and intermodal relations with a late fusion approach.

7.1 Overview

The key to perform a successful multimedia retrieval operation is to analyze the semantic content of the multimedia data adequately. For an adequate analysis, the multimodal nature of the data should be analyzed carefully and the information contained in the data should be used completely. In this respect, combining the information gathered from multiple modalities is an empirically validated approach to increase the retrieval accuracy [4]. Yet, two major issues are pointed out by many researchers as attractive research areas [4, 98, 143]: (i) How to determine the best modalities? (ii) How to fuse them the best way? This study focuses on the second problem and presents a modality combination approach in order to use the multiple modalities effectively.

The previous chapters of this thesis has already presented solutions on the above given problems. However, proposed solutions are mostly focused on the first problem and there is the need for an effective solution on the 'How-to-fuse' problem. Through this direction, we focus on the following needs;

- Use of correlated information: A big majority of the available approaches accept the fusion as a complementary process, and assume that the fusion inputs are independent. Thus, the dependency / correlation between different modalities is usually ignored and each modality is processed separately. Considering that any object or event occurring in multimedia data is also multimodal (e.g. A 'car' has a visual appearance, a characteristic sound and has some parts including text on it), it can be argued that there exist a strong dependency between the different modalities of multimedia data. Hence, the fusion solution should benefit from such information and support working with both complementary and dependent (correlated) inputs.

- Working with contemporary approaches: The fusion literature contains a huge amount of studies that are usually grouped under two broad levels, according to when the fusion process is applied: early and late fusion. As described in Chapter 3 in detail, early fusion approaches combine the information on sensory-level or feature-level, whereas late fusion approaches deal with the outputs of the classifiers (scores, ranks and decisions). However, contemporary learning approaches introduce a 'new' level into the learning process (i.e. the use of bag-of-words). Considering that such bag-of-words based learning approaches are highly popular on the multimedia retrieval domain, we can advocate the need for combining the information at this level.

- Combining multi-level inputs: Almost all of the fusion approaches assume the homogeneity of the fusion inputs. However, the fusion inputs may be in different levels (e.g. combining two systems, one providing features and another providing scores of classification), or may be in different class-spaces (e.g. combining the results of two classifiers, one performing a classification into classes $C_1, C_2, ..., C_n$, but the other one into $S_1, S_2, ..., S_m$). Although it may be possible to convert all inputs into the the same level performing a

classification on the low-level inputs (features), such an operation may lead to loss of information. Thus, we need a general fusion framework that enables combining multi-level inputs.

Regarding the above given need, we focus on a solution that combines bags of words that are generated from different modalities. Through this direction, we propose a general fusion framework based on BoWs by converting any type of information into BoW format. After converting all types of inputs into BoW representation, we incorporate both the complementary and the correlated information into fusion process. For exploiting the correlated information, we analyze the intramodal relations within each modality, and the intermodal relations between modalities. Such correlation information and the provided BoW based features of different modalities exhibit a complementary behavior, thus they are combined with late fusion approach. Hence, our proposed approach is composed of four steps: (i) classification of information in each modality, (ii) intramodal correlation analysis for each modality and classification of the obtained information, (iii) intermodal correlation analysis between modalities and classification of the obtained information, (iv) late fusion of all classification results. For the late fusion, a linear weighted averaging approach is utilized with the weights generated by using the RELIEF-MM algorithm.

For the intramodal and intermodal correlation analysis problem, we propose a novel mining and graph based solution. Throughout the intramodal process, the *word*s of each modality and the correlation between these *words* are converted into a graph representation, and then the meaningful *phrase*s are extracted by using these *word*s. For calculating the correlations between the words, a frequent itemset mining (FIM) procedure is executed. In order to extract the *phrase*s, the most together occurring words are extracted from the constructed *word* graph. For the intermodal process, first, the correlation between the extracted phrases of different modalities are calculated based on the Pearson's correlation coefficients, and then obtained information is converted into a graph representation. After that, the *multimodal phrase*s are extracted from the graph. Both of these processes end up with using the extracted *phrase*s for classification. For the evaluation of the proposed approach, an experimental study is conducted on TRECVID 2011 dataset with visual, audio, text modalities. The test results show that the proposed approach is an effective way for fusing BoW-

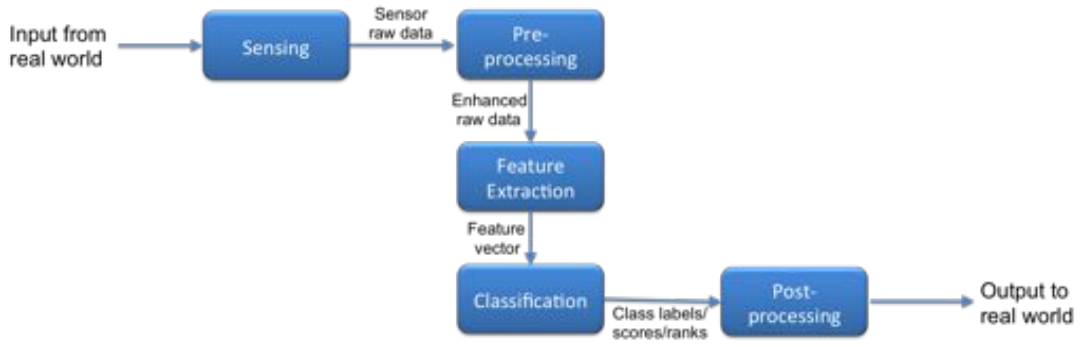


Figure 7.1: A Typical Multimedia Analysis Process

based feature vectors of different modalities. In addition, the use of intramodal and intermodal correlation information helps to improve retrieval performance and the fusion gain.

The remainder of this chapter is organized as follows: In Section 7.2, an overview of the contemporary learning approaches and the some descriptions on the use of bags-of-words are given. In Section 7.3, some related work on combining BoWs and an analysis of the state-of-the-art approaches are presented. Then, in Section 7.4, the proposed approach for combining BoWs is given in detail. In Section 7.5, the empirical results and the evaluations of our proposed solution are given. In Section 7.6, an evaluation of the proposed fusion architecture is done based on the general fusion framework for fusion (Section 3.1). In the last section, some conclusions are drawn and some possible future studies are discussed.

7.2 Background Knowledge

Pattern recognition and computer vision literatures contain a huge amount of multimedia analysis (especially image analysis) studies [26, 118]. In a traditional analysis system (as given in Figure 7.1), the process starts with the perception of some input from the real world via some hardware called sensors. After converting physical inputs (i.e. sounds or images) into signal data via some sensors and preprocessing such signals (i.e. enhancement and segmentation operations), a feature extraction step is employed and several important properties of the real world input that are useful for

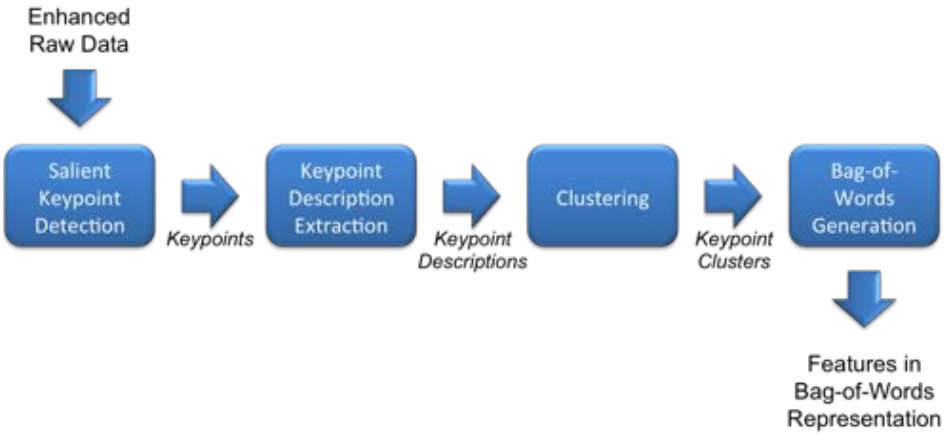


Figure 7.2: BoW Generation Process

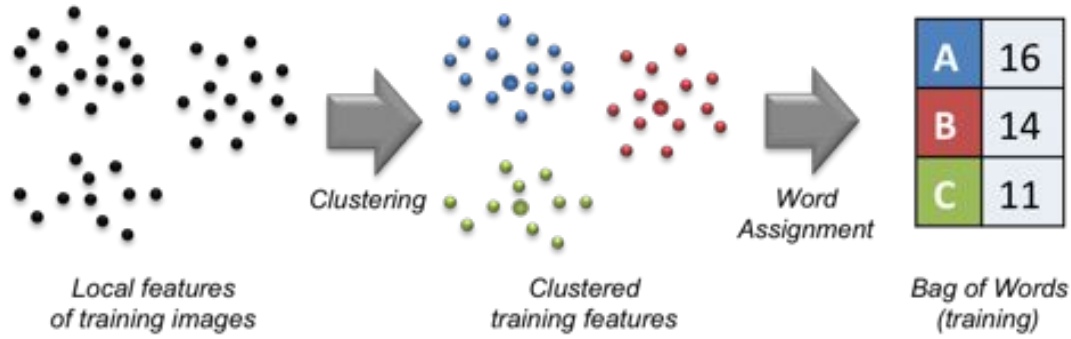
classification are extracted by using sensor data. Afterwards, features are used for classification and class label, score or a ranked list is obtained as the classification result. Lastly it is possible to have an enhancing post-processing mechanism on the results (i.e. fusion of several classifiers, features, etc.). In such systems, extracted features from the input signals are usually global features, which represent the overall characteristics of given multimedia frame, segment, shot or video.

Different from the traditional approaches, contemporary approaches enhance the feature extraction process. The contemporary approaches bases on two solid ideas:

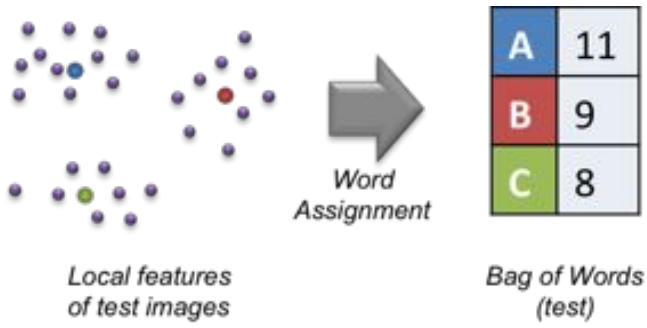
- Use of local parts and features
- Employing "Bag of Words (BoW)" approach

Brief descriptions on these ideas are given in the following subsections. As a summary, Figure 7.2 illustrates such feature extraction process. First, salient local keypoints are selected from the given multimedia frame or segment and representative local parts of the given frame/segment are found. Then feature descriptions of these points are generated. Lastly, the bag-of-words approach is applied on these keypoint feature descriptions and keypoint features of each multimedia frame/segment are converted into vector space format by clustering them. A complete example of how BoWs are generated is presented in Figure 7.3. A detailed description of these ideas can be found in [117].

(a) Example ‘Car’ images and local salient points on the images



(b) BoW Generation during training phase



(c) BoW Generation during test phase

Figure 7.3: BoW Generation Example

7.2.1 Using Local Parts and Features

The idea of using local parts (keypoints) and features is based on the need for identifying objects in images . Although the pattern recognition literature present many mature solutions for object detection, the problems due to viewpoint changes, lighting conditions or partial occlusion make the problem still challenging [117]. Considering these problems, trying to find objects by some segmentation approaches is not adequate. A popular solution on this problem is the use of representative local parts (keypoints). By using an effective local feature, the objects are represented by a set of local regions (keypoints) each of which is modeled with a feature vector (descriptor) computed from the region. The keypoints and their descriptors are generated with a controlled degree of invariance to viewpoint and illumination conditions so that similar descriptors are computed for all images in the database [117]. Consequently, the idea of using local parts and features contains two major steps: Salient Keypoint Detection and Keypoint Description Extraction. These steps are given in detail below.

7.2.1.1 Salient Keypoint Detection

Given an image (or a multimedia frame / segment), salient keypoints are extracted by employing a saliency detection algorithm. The methods for keypoint detection can be grouped in two: Dense-sampling and sparse-sampling. In dense-sampling, the given frame is partitioned into $m \times n$ grid and each cell is used as a keypoint. In sparse-sampling, a keypoint detection algorithm is used and salient points are decided by sampling a sparse set of locally stable points [52]. The sampled keypoints are expected to be invariant to geometric and photometric changes. Some sparse-sampling algorithms are as follows: Laplacian of Gaussian (LoG) [75], Difference of Gaussian (DoG) [77], Harris Laplace [80], Hessian Laplace [79], Harris Affine [80], Hessian Affine [79].

7.2.1.2 Keypoint Description Extraction

Keypoint descriptors are used to describe the regions around the keypoints. The descriptors usually provide a description for each keypoint which is invariant to location,

scale and rotation, and robust to affine transformations (changes in scale, rotation, shear, and position) as well as changes in illumination. The most famous keypoint descriptor is SIFT (Scale Invariant Feature Transform) [77]. SIFT is a 128 dimensional feature vector that captures the spatial structure and the local orientation distribution of a region surrounding a keypoint. Recent studies have shown that SIFT is one of the best descriptors for keypoints [52]. Some other well-known keypoint descriptors are as follows: SURF (Speeded-Up Robust Features) [8], OpponentSIFT [135], RGB-SIFT [122], HSV-SIFT [14], Hue-SIFT [136], W-SIFT [38], Color Moment [135], Self Similarities (SSIM) [115], GIST [88], HOG [25].

Here, it should be noted that any global feature can be used as a local feature after performing keypoint detection by extracting global features for each local patch. However, SIFT-like invariant approaches perform superior.

7.2.2 Bag of Words (BoW)

Respecting the huge amount of work done in text retrieval, Bag-of-Words approach is adapted into multimedia domain from the text retrieval literature. Text retrieval systems employ a number of steps for text retrieval purposes. Firstly, the words in the document are extracted and their stems are found. Then, a stop word list is used to prune very common words (e.g. ‘a’, ‘an’, ‘the’, etc.), which are not discriminative for any document. After that, the remaining words are processed and the frequency of each unique word is calculated. The documents are represented with a BoW format, where each document is a feature vector containing the frequencies of the words the document contains. The word frequency calculation can be done in various ways (Term Frequency (TF), Term Frequency - Inverse Document Frequency (TF-IDF), etc.) [117]. The retrieval is based on the constructed vectors for each document.

With a multimedia aspect, it is possible to make an analogy for each modality with the textual bag of words. For instance, images including some number of salient keypoints resemble the documents having some number of words parsed. In other words, visual bags of words can be constructed by using the visual keypoints in images / videos. In this respect, the BoW idea is to represent each image / video as an orderless collection of local keypoint features. In order make the representation more

compact, the keypoints are first clustered into a visual vocabulary with a predefined size (Figure 7.3), and each keypoint cluster is accepted as a "visual word" in the visual vocabulary. Then each image / video is represented with a vector containing the presence of each word in the vocabulary [52]. Not only the presence of the words, any other weighting scheme can be adapted (i.e. counts, TF, TF-IDF, etc.). This process can be applied for any other modality of the multimedia data.

7.3 Related Work and Analysis of the State-of-the-Art Approaches

As discussed in detail in previous chapters, combining the information gathered from multiple modalities is an empirically validated approach to increase the retrieval accuracy. Considering that the most popular and effective methods in multimedia analysis studies in the last decade are based on the use of local parts / features in multimedia documents and employing Bag-of-Words (BoW) approaches, we would like fuse all available information obtained via BoWs of different modalities. In this section, we first analyze BoW approach with regard to the aspects given in Section 3.1. Then the approaches for combining bags of words are presented with a brief literature analysis and prototype implementation of these approaches.

An important aspect for information fusion is the content representation method (Section 3.1.4). Bag-of-words representation is a new type of content representation considering the feature-based, similarity-based and preference-based representations, which are discussed in Section 3.1.4. Although its usage is very similar with feature based representation, the information contained in BoWs has a crucial difference. The low level features in feature-based representation do not have direct semantic meaning, and cause the well-known semantic gap problem when used. However, BoWs contain information on the representative parts of object/concepts occurring in the multimedia data, without assigning a label to the represented parts. Thus, a BoW-level processing is a new semantic level between the low level features and the high level concepts.

Another important characteristic of BoWs, which provide simplicity during fusion is that it is self-normalized or very easy to normalize, which is related to the normalization issue given in Section 3.1.5. Considering that BoWs are generated according to some

pre-decided vocabulary sizes, it is possible to set the size of the BoW vector. In addition, the values in the vector are usually frequencies which are in a predefined range. Since we generate the BoWs of each modality, we use the same metric and scales for the values in the vector. Therefore, the issues related to the normalization of sources, that are discussed in Section 3.1.5, are no a longer problem, when the combination inputs are in BoW format.

In addition to the representation and normalization issues, the fusion strategy to be used is also a crucial issue (Section 3.1.3). As mentioned in Chapter 3, the correlation between the inputs is an important source of information for fusion. Such a fusion strategy can be applied for the fusion of BoWs. Essentially, although employing a BoW approach and combining multiple BoWs have a big potential to provide a high accuracy for multimedia analysis, there is still a crucial source of information that is not used effectively, which is the relations between words. Standard use of BoWs do not include any relation information between the keypoints (neither a spatial relation, nor temporal). Thus, we propose to exploit the spatial relationships between keypoints in each multimedia frame and temporal relationships between the keypoints occurring in the successive multimedia frames. The relations between keypoints can be either intramodal or intermodal. An intramodal relation refers to the correlation between keypoints with respect to a single modality, whereas an intramodal relation is the correlation between different modalities. Exploiting such relationships can also be named as extracting the co-occurrences of interesting patterns or mining multimedia data for frequent patterns.

Another important affecting factor for fusion is the fusion level (Section 3.1.6). Considering the common fusion approaches and the time to apply the fusion process, which are discussed in Section 3.1.6, the combination of BoWs can be done basically by assuming BoW vectors are usual feature vectors. Thus, basic approaches for combining several BoWs is to combine the feature vectors (early fusion) or the classification results (late fusion). However, it is possible to reformulate the fusion levels for combining BoWs as follows, due to the the additional step of clustering in BoW generation:

- Pre-Early Fusion: First the keypoints of the multimedia document and the

corresponding descriptions for each keypoint are extracted. Such extraction is done for each different feature, and the descriptions of different features for each keypoint are concatenated. After concatenation, vector quantization (clustering) is performed and the words are generated. These words can be used in any type of learning architecture. Here, it should be noted that pre-early fusion is not applicable for multimodal fusion, but it can be used with a single-modality multi-feature scheme.

- Early Fusion: First, the keypoints and the corresponding keypoint descriptions are extracted for each modality. Then, keypoints of each modality are clustered separately, and the BoWs of each modality is obtained. Lastly, the vector representations of bags are concatenated.

- Late Fusion: For each modality, keypoint selection, keypoint description extraction, BoW generation and classification using the BoWs are performed separately. Then, the classification results are combined.

Originality of the proposed approach is that it performs mining operation on the BoW vectors, which means it uses the representative parts of objects/concepts to mine. Thus, it prevents one from dealing with the unnecessary details in low level features and performs the mining in an effective way. In addition, it enables more information than working with high level semantic concepts for mining. Using high level semantic concepts for mining depends on the success of the high level semantic extraction, which includes an important problem, namely the semantic gap between low level and high level features. Also it is highly affected from the viewpoint, lighting, occlusion problems. Using parts of objects/concepts eliminates such inefficiencies.

Regarding the above given analysis, the solution framework consists of two parts. First part combines the available BoWs in Early and Late Fusion schemes. These approaches are widely applied in the literature [50, 52, 81, 122]. Second part performs a mining operation in the multimedia data. Actually, video data mining is an attractive topic in recent years and the literature contains a considerable amount of studies. However, most of these studies perform the mining operation directly on the high level semantic concepts, by ignoring the semantic gap problem [11, 33, 154]. The rest of them, work with low level features for mining and do not use spatial and temporal

relationships between patterns [111]. In addition, most of the studies do not use intramodal and intermodal relations effectively and designed in a domain-dependent way [10]. Below, some applicable approaches for multimodal mining are discussed in detail.

7.3.1 N-grams

Considering that we would like to exploit the interaction between words, the use of the spatial and temporal proximity between the words can provide us some valuable information. In [54], Jiang et al. make an analogy between the spatial co-occurrence of visual words and the bi-grams or N-grams in text categorization and try to obtain the geometrical structure of an image by using the spatial proximity of words. The use of N-gram offers a perspective of modeling the spatial and temporal co-occurrence of multimodal words. Some of the recent similar ideas include the studies of [54, 71, 147]. In [54], Jiang et al. construct a two-dimensional co-occurrence histogram to represent the images based on the visual bi-grams. After eliminating the word couples having an euclidean distance smaller than a predefined threshold, the resting couples used for a learning process. With a similar idea, Lazebnik et al. [71] group the neighboring keypoints for object recognition, and Ye et al. propose a joint audio-visual bi-modal representation by using the temporal co-occurrence of the audio and visual words.

We can analyze some examples in order to understand the given ideas more clearly. Let's consider two BoWs, one of which is visual words, the other one is audio words. It is mentioned that the spatial proximity of different visual words is important for classification because it captures the geometrical structure of the image. For example, visual words depicting 'tire' may frequently co-occur with visual words characterizing 'headlight'. In addition, we can consider the temporal proximity of visual words depicting 'tire' with audio words 'car voice', considering that 'tire' may frequently co-occur with the sound of the car which the tire belongs to.

Although N-gram is originally used for predicting a word from a number of consecutive previous words (in text retrieval domain), the approach provides the probabilities of occurrence for a set of words, as used in [54]. In order to be used with a multimodal mining, the definition of N-gram should be extended. Originally, N-gram is

one-dimensional and neighborhood based. So we should modify N-gram selection algorithm from one-dimensional neighborhood based selection to a spatial and temporal proximity based algorithm.

7.3.1.1 Prototype Implementation

We implement a preliminary test to analyze the usability of N-grams approach. For the test, we use the TRECVID 2011 dataset with visual SIFT feature. We construct bag-of-words representations with a vocabulary size of 1000, for both. For the weighting, we prefer a binary approach (1 if word exists, 0 otherwise), for simplicity. We select the 'Car' concept for learning and construct bi-grams for visual BoWs and audio BoWs. As a choice for the spatial proximity between words, we assume that the words occurring in the same frame are close enough. However, considering that the size of all 2-grams for such configuration is

$$\binom{1000}{1} \cdot \binom{1000}{1} = 10^6 \,, \tag{7.1}$$

it is necessary to reduce number of bi-gram outputs. Thus, a similar approach with [147] is preferred for pruning the set of word couples and top 1000 results having the largest $SupportDifference$ value are selected;

$$SupportDifference = support(`Car') - support(`Non - Car') \,, \tag{7.2}$$

where $support(X)$ is the count of a bi-gram selection among the all training samples. After finding the bi-grams, we perform an SVM-based classification with the SIFT and bi-gram features separately and combine the results with a Linear Weighted Fusion approach with the following formula

$$D_{FUSION} = (1 - w) \cdot D_{SIFT} + w \cdot D_{2-GRAM} \,, \tag{7.3}$$

where w is the weight of 2-gram decisions. Figure 7.4 presents the resulting Average Precision values.

As seen on the graph, using 2-grams of SIFT increases the accuracy from 64.63% to 65.24%, which is not statistically significant. The reason why the increase is not statistically significant is the very simple configuration of the test. We have made two crucial assumptions for simplicity, that possibly affected the fusion performance:

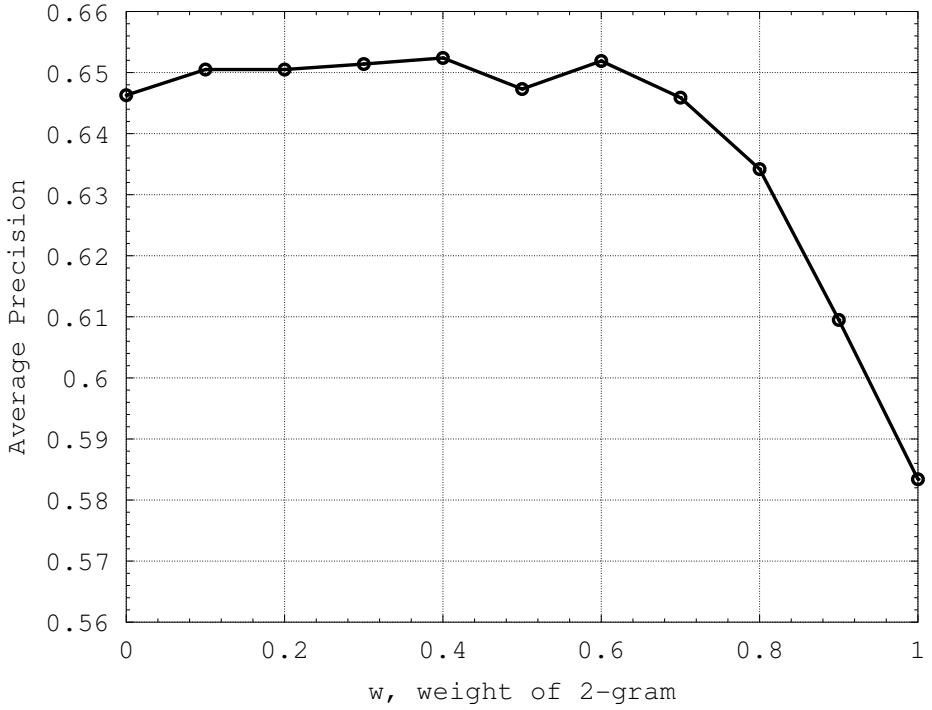


Figure 7.4: Average Precision for the fusion of SIFT and bi-gram based retrieval

- The number of 2-gram components are fixed at 1000, which means we use 0.1% $(1,000/1,000,000)$ of all 2-gram components.
- We used N-grams for only n=2. It is possible to extend the test n=2 to m, where m is tractable.

The above given preliminary test and the analysis show us that obtaining a reasonable increase in the performance is possible, but the process may be computationally complex. Consequently, N-gram approach has a potential to provide extra intramodal and intermodal correlation information, however it suffers from the combinatorial explosion problem.

7.3.2 Frequent Itemset Mining

Considering the efficiency problem with N-gram approach, a practical solution for the extended N-gram idea can be obtained by finding the co-occurrences of interesting patterns directly instead of an exhaustive search by N-gram approach. Finding interesting

patterns is also called as mining multimedia data for frequent patterns.

In the literature, there are several studies using data mining, more specifically association rule mining (ARM) or frequent itemset mining (FIM), techniques in order to perform semantic indexing (mostly on image data). In [10], Bhatt et al. perform a recent survey on multimedia data mining. Pioneers of the multimedia mining studies mostly deal with high-level features occurring in multimedia data and try to mine frequent patterns on these high-level features. Actually, this approach is very suitable for finding the correlations between modalities. However, a big majority of the currently available studies perform the mining only in an intramodal way and do not deal with multiple modalities. In addition, building the solution on top of the high-level features causes ignoring the semantic gap problem. Although the BoW representation is a mid-level representation and BoWs may be applicable instead of high-level features, still the solutions should be revisited for such an extension. Another deficiency with these approaches is that they use the same itemsets for learning any of the classes. However, it is highly probable that each class has a different list of frequent itemsets.

Some of the recent studies [11, 21, 49, 154] discuss various solutions for multimedia data mining. For instance, in [154], Zhu et al. work on the frequent high-level features. Firstly, they employ several video processing techniques to find some audio/visual cues such as court field, camera motion activity, applause in basketball videos; then perform association mining among these cues and assign each association a high-level class label. In [11], Bhatt et al. perform the probabilistic mining process, considering the accuracies of the detected events may change over a time interval. In [21], the authors focus on the temporal information in the video and propose a hierarchical temporal association mining approach by extending the traditional association mining method. Not only the study of Chen et al., but the former two studies are also based on traditional association mining approaches. Yet, considering that the traditional association mining approaches may lead to combinatorial explosion problem, in [49], Jiang et al. propose a non-traditional association mining approach, and use a neural network to learn direct mapping between the visual and textual features by automatically and incrementally summarizing the associated features.

7.3.2.1 Prototype Implementation

Considering above ideas, we carry out a prototype implementation by using the TRECVID 2011 dataset. We have used 71,502 training shots and 34,179 test shots with 50 concepts, which is the configuration for the Lite Run of the Semantic Indexing task of TRECVID 2011. We extract visual SIFT and audio MFCC features and construct bag-of-words representations with a vocabulary size of 1000 for both. For the weighting, we prefer the TF approaches (TF: number of word occurrences in shot).

During the evaluation, we first calculate the retrieval accuracies for SIFT and MFCC features, with an SVM based approach. Then we calculate the intramodal correlations for each multimodal feature. For correlation analysis, we employ frequent itemset mining and utilize a FP-Growth implementation [13] to calculate the frequent itemsets. We prefer selecting the maximal frequent itemsets during the FIM. After finding frequent itemsets, we accept each itemset as an attribute of our new feature vector. We perform an SVM based classification approach in order to find retrieval accuracy of the correlation based features.

Before the classification step, considering that the number of itemsets obtained from the FIM is really huge, we perform a filtering step on the attributes and try different alternatives for how the attributes are selected. For the filtering of attributes, we prefer support (Sup) based filtering, for which we select the itemsets with top-k support values, where k is a predefined value. Here, the support value denotes the occurrence percentage of the itemsets among the samples of each class. For the attribute selection step, we compare three alternatives:

- Common: A single feature vector is constructed for using with any of the classes, which also means the same itemset list is used for learning any of the classes. This is the way how the current studies construct attributes from itemsets.

- Combined: Top-*k* attributes are selected for each class, separately. Then, all attributes are combined into a single vector.

- Separate: Top-*k* attributes are selected for each class, separately. Then, only the corresponding attributes of each class is used as the feature vector.

For the above given configurations, we perform several tests. The tests are based on the semantic retrieval of the video shots according to the semantic concepts. The Mean Average Precision MAP) values are presented in Table 7.1. In addition, Average Precision results of some example concepts are also presented.

According to the empirical results presented in Table 7.1, following evaluations can be done:

- The retrieval performances seems to be very low and insufficient, in general. However, we should note that we do not spend much time on optimizing the SIFT and MFCC features (specifically, for salient keypoint detection), finding the optimal number of vocabulary count and kernel optimization of SVM classifier, since this is only a prototype testing. In addition, it is should be remarked that best MAP accuracy obtained in TRECVID 2007 Semantic Indexing task is around 15% and the median of all participants is about 5.6%. Thus, we do not stick to the low performance and continue with the tests.

- It is clear that 'Combined' attribute selection performs better than a 'Common' selection, however a 'Separate' selection is the worst. It is expected that selecting top-k itemsets is better than using the same itemsets for all classes; however, it is interesting to see that it is necessary to combine itemsets of all classes into a single vector. This can be evaluated as a consequence of early fusion of attributes.

- Considering the results of SF_2, SF_3; we can state that the number of attributes incorporated affects the performance, however, using more attributes increases the training time for SVM classifiers. Thus, it is important to find an optimum value of attributes by filtering the attributes.

- Considering that accuracy of $S+SF_1$ is better than S and SF_1, and $M+MF_1$ is better than M and MF_1; it can be resulted that using an intramodal FIM provide useful information for the fusion process.

- Considering the accuracies of the SMF and the configurations including SMF, it seems that intermodal FIM provides more useful information than the intramodal FIMs.

Table 7.1: Prototype Analysis for Frequent Itemset Mining

Test Name	Attribute Selection	Attribute Count	Text AP	Adult AP	Indoor AP	Car AP	Dancing AP	Explosion AP	MAP
S	N/A	1000	15.980%	12.758%	11.036%	7.051%	5.418%	1.898%	4.366%
SF_0	Common	1000	13.629%	11.821%	10.691%	4.513%	1.544%	1.127%	3.810%
SF_1	Combined	1024*	15.849%	11.454%	10.629%	5.033%	2.391%	1.529%	4.202%
SF_2	Seperate	40	17.060%	11.359%	5.384%	4.967%	1.404%	0.963%	3.730%
SF_3	Separate	1000	13.945%	11.119%	7.921%	4.574%	1.729%	0.918%	3.834%
SMF	Combined	2000*	14.519%	11.093%	11.229%	5.776%	1.912%	2.238%	4.116%
M	N/A	1000	10.134%	12.568%	11.572%	4.583%	1.550%	3.455%	3.958%
MF_1	Combined	1361*	9.675%	11.760%	10.676%	3.396%	0.956%	1.836%	3.448%
$S + M$	N/A	N/A	16.170%	12.415%	11.435%	7.361%	5.246%	2.838%	4.593%
$S + SF_1$	N/A	N/A	15.823%	11.957%	11.691%	7.155%	5.372%	1.806%	4.437%
$M + MF_1$	N/A	N/A	10.199%	12.523%	11.494%	4.579%	1.504%	3.386%	4.027%
$SF_1 + MF_1$	N/A	N/A	14.290%	11.004%	10.434%	5.145%	1.221%	1.876%	4.197%
$S + M + SMF$	N/A	N/A	15.852%	11.592%	11.044%	7.191%	2.719%	1.696%	4.459%
$S + SF_1 + M + MF_1$	N/A	N/A	16.616%	12.291%	11.882%	6.685%	4.907%	3.151%	4.695%
$S+SF_1+M+MF_1+SMF$	N/A	N/A	16.895%	12.312%	11.660%	6.844%	5.169%	3.336%	4.773%

S: SIFT, M:MFCC, SF: SIFT_FIM, MF: MFCC_FIM, SMF: SIFT_MFCC_FIM

* (40x50): Top-40 attributes of 50 classes are combined into a single vector.

$+$ given in the test configurations denotes the late fusion of the features.

- During the intermodal FIM, we combine the attributes of both modalities (SIFT and MFCC). When multiple modalities are combined with a concatenation approach, an important problem occurs since each modality have different support intervals. Actually, this is the well-known problem of 'rare itemsets' in data mining domain. If we apply a unique support threshold for both of the modalities, one of them (in our case, it is SIFT) becomes more dominant. Thus, it is required to solve such problem in an intelligent way which is not affected from the different support levels of each modality. In this test, we simply selected the itemsets having items from both of the modalities. Here, we should note that performing a combined frequent itemset mining as we do here is not studied even in the data mining domain before (e.g. finding frequent itemsets from two market-baskets collaboratively).

- The 'rare itemset problem' also occurs while working with different classes. This leads us to use class-specific support thresholds for each class.

7.3.3 Improving Frequent Itemset Mining with Locality and Graphs

In the previous subsection, it is stated that the multimedia mining studies usually focus on mining based on the high-level features extracted from multimedia data. Yet, in accordance with the popularity of local features, keypoints and BoWs, the frequent itemset mining on BoWs has also become popular in the last few years. By using the BoWs, the studies usually calculate most frequently occurred word-sets as the itemset and use these itemsets for classification. In addition to the use of BoWs, the video/image mining literature has more sophisticated approaches that makes use of frequent itemset mining. Majority of these studies benefit from the occurrence of similar patterns in the local parts of the images and aims to find frequently occurring objects or scenes in the images. The rationale behind the use of locality with FIM is the fact that parts of a particular object or a scene usually occurs together in different samples of that object/scene. Thus, obtaining the salient keypoints, constructing transactions as the salient keypoint and the neighbors of it and then performing a FIM operation can give a valuable information about the particular object / scene.

Some of the recent studies, which employs the locality idea, are summarized as follows:

In [100] Quack et al. introduce a novel method for mining frequently occurring objects and scenes from videos. They incorporate the keypoints in the frames of the videos and spatial information of the keypoints, and generate transactions for mining by listing the neighboring keypoints of each keypoint with their positional relation (top-left, top-right, etc.) as the items. After generating transactions, they apply the Apriori algorithm to select frequent patterns. Such selection gives the possible frequent objects /scenes. In [99], they extend their approach to automatically find spatial configurations of local features occurring frequently on foreground objects of target types, and rarely on the background objects. In [153] Yuan et al. carry out a similar study and name it as obtaining visual phrases from visual words. They use the neighborhood of the visual words (keypoints) and try to group them as visual phrases, by using the frequent itemset mining techniques. In [35] Fernando et al. state that most of the studies using FIM for image classification were not able demonstrate competitive results, and propose an improved approach. In their approach, they improve the way the transactions are generated. They propose to find frequent local histograms (FLH) which is based on the neighboring keypoints. However they employ a TF similar approach instead of a binary calculation. Furthermore, they propose a bag-of-FLH approach.

Considering the above given studies, a typical BoW-based mining-oriented classification approach can be summarized as given in Algorithm 4. The algorithm identifies the training phase of the classification and it is assumed that the algorithm is executed separately for each different class / concept. First, each multimedia document (image, frame / shot of the video, etc.) of given collection is processed and salient keypoints in the document are extracted. Then, neighborhoods of the keypoints are calculated and a transaction (for mining operation) is generated for each keypoint by including the keypoint and the neighbors. After processing all keypoints and constructing the transaction set, an elimination on the transactions is performed, according to some predefined rules (e.g. removing the transactions not occurring on some number of consecutive frames). Then, the frequent itemset mining operation is performed, frequent itemsets are obtained and a pruning on the frequent itemsets (e.g. removing some percentage of most and least occurring) is done if necessary. Lastly, each itemset is accepted as a representative pattern, and used for a learning process. For instance, a rule based approached can be used simply, and the representative pattern can be

Algorithm 4: A Typical BoW-based mining-oriented classification

Input: Multimedia documents $\mathcal{D} = \{d_i\}_{i=1}^{t}$, corresponding class c

Output: Learning schema L

```
1  begin
2      T ← {} ;                                   // Transaction list
3      for d_i ∈ D do
4          K ← extractSalientKeypoints(d_i);
5          for k ∈ K do
6              N ← calculateNeighbors(k);
7              T ← T + {k, N} ;                   // Add a new transaction
8          end
9      end
10     prune(T);
11     I ← findFreqItemset(T);
12     prune(I);
13     L ← performLearning(I, c);
14 end
```

associated with the class. Alternatively a classification approach can be used; the existence / frequencies of each pattern can be calculated for each training document as a feature value, and a training feature vector can be generated. Querying phase of the classification is not different from the training phase. First the keypoints are extracted, then the training frequent itemsets (representative patterns) are used to convert the query keypoints to query feature vector and passed to the classifier.

An important deficiency with the current studies is that they are usually limited to single modality and reflect only the intermodal correlations. The use of spatial relations is still limited, and temporal relations are not used at all. Since we work on the video data, it is very probable to apply the mining process both with intra and inter modal ways. However, if multiple modalities are combined with a simple approach like concatenation, similar problems with the use of FIM occurs. So, performing FIM with locality does not solve the problems with the FIM approach, and a better way of dealing with multimodality and employing spatial/temporal relations is necessary.

A promising way of employing spatial/temporal relationships between the salient keypoints is to represent the document as a graph. With a basic definition, in this approach, the salient keypoints are assumed as the nodes of the graph, and the edges are the connections between the nodes. In order to draw an edge between two nodes, different considerations can be applied such as neighborhood, a threshold based distance, or a correlation value. Some of the recent studies are given below.

In [93], Ozdemir et al. apply the idea on satellite images. They extract the interesting regions/points in the image and accept the points as the nodes of the graph. For the edges, in order to provide a scale invariant solution, they construct voronoi diagrams on the image by drawing a voronoi cell for each point. Then, they connect each two node if they are neighbors in the voronoi diagram. After constructing the graph, they perform a Frequent Subgraph Mining operation by extending the $gSpan$ [146] algorithm, and try to find frequent patterns for each class. In [87], Nowozin et al. work on the images with objects in them, extract salient keypoints in the images (with SIFT) and accept each keypoint as a vertex on the graph. Then, they connect each two vertex with an undirected edge and obtain a completely connected graph. For each edge, they assign a label including the following information; ratio of scales of the two vertices connected, normalized distance between vertices and a horizontal orientation measure. They use the $gSpan$ [146] algorithm for graph mining. Although there exists several recent studies about graph mining, given two studies are enough to understand the solution technique. However, we include one more interesting study which offers a novel technique for the edge labels. In [23], Chen et al. work on predicting the relationships between people in group photos by converting the arrangement of the faces in the photo to a graph. Then, they assign each face as a vertex on the graph, and calculate the label of each edge as the order distance between the vertices. In order to find the order distances, the Minimum Spanning Tree (MST) is calculated on the graph and the order distance is calculated as the length of the shortest path between vertices.

Performing graph mining is a sound alternative for finding the intramodal and intermodal correlations. However, during the graph mining process, checking the similarities of graphs (isomorphism tests) is a required operation and executed several times. Thus, the biggest problem of this approach is the high computational complexity of the isomorphism tests. Another problem with the graph construction is the unreliability

of spatial distance measures inside the visual frames, due to the scale changes. For this problem, similar approaches with the studies presented above can be applied (i.e. MST based orders, neighboring based approaches, etc.). Still, the problem of finding the intramodal and intermodal correlations may be highly computationally-complex, considering that using all the salient keypoints of all modalities makes the graph a giant one.

Regarding the inefficiencies of using all salient keypoints in a graph, the ideas of using FIM and subgraph can be used together. In such a scheme, firstly a frequent itemset mining is applied on the salient keypoints, and then each itemset is converted into a graph with the keypoints as the vertices and the spatial / temporal distances between keypoints as the edges. Then a frequent subgraph mining can be applied in a more efficient way. This procedure enables to work on much smaller graphs than the previous, thus reduces the computational complexity. However, following such algorithm may lead to lose the correlations between different modalities since different modalities usually have different support intervals. Furthermore, many of the previously discussed problems remain unsolved. Thus, a better approach is proposed in the next section.

7.4 Combining Bags of Words: A Novel Mining and Graph Based Approach

Considering the above discussed motivation and analysis, we focus on a solution that combines bags of words that are generated from different modalities. As mentioned before, BoW model is a middle level information between low level features and high level concepts. Thus, we can still use valuable information contained in the low level features, but behave the features like high level concepts, since words represent the parts or regions in the concepts.

The proposed approach assumes that all inputs are in the BoW form. However, converting any type of information into the BoW form is not complicated. Fusion inputs can be analyzed under two types as discussed below. Up until now, only the first type is mentioned and focused due to the effectiveness of the type. Yet, we can propose a general fusion framework based on BoWs by converting any type of information

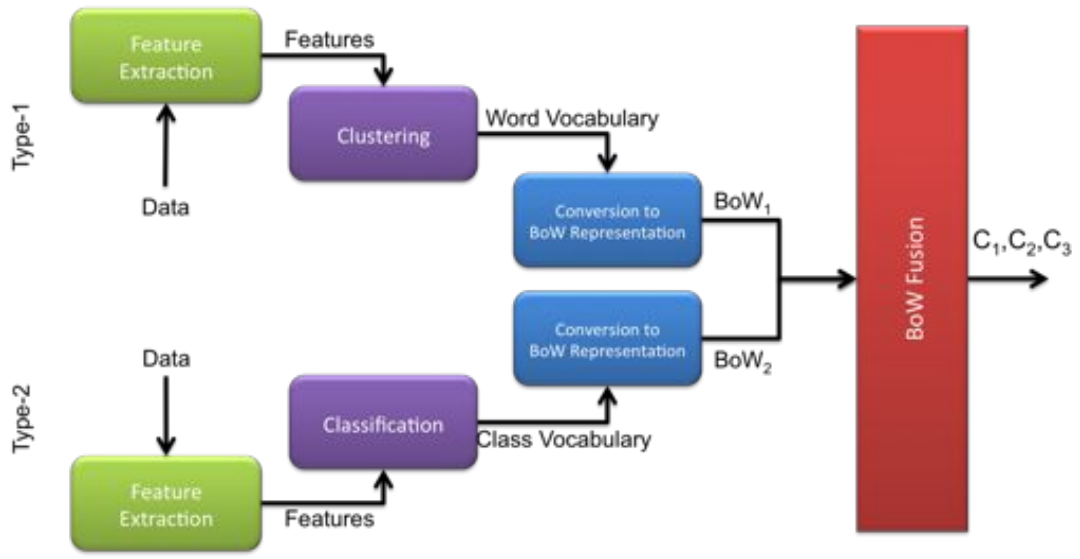


Figure 7.5: BoW-based General Framework for Fusion

into the BoW format. The idea is summarized in Figure 7.5.

- Type-1 (Working with low-level features): Regardless of being invariant or not, or generated after a sparse or dense sampling, the extracted low level features can be used as the inputs for fusion. However, the requirement is to have some keypoints, or local parts. After obtaining the keypoints, they are clustered into words to construct a vocabulary. Then the training data is converted into the BoW format by using the vocabulary, as given in Section 7.2 in detail. Any other type of information, that cannot be represented via local parts, can be processed as the second type given below.

- Type-2 (Working high-level features): High-level features (concepts) are the second input type for fusion. The complete set of high-level features occurred in the training set of multimedia data is assumed to be the vocabulary, and each high-level concept to be a word. Then the data can easily be converted into the BoW format. Through this way, depending on the type of classifier, or the target class types of classifier, we can generate and combine many types of bags: Bag of Objects, Bag of Events, Bag of Actions, Bag of Faces, Bag of Silhouettes, Bag of Activities, etc. In addition to the use of high-level concepts directly, any type of data that cannot be represented via local parts can also be

accepted in this type. This type of data is firstly classified into some defined high-level concepts, and then converted into the BoW format. Here, it should be remarked that having a more number of high-level concepts, and having a successful classification step are important issues for the fusion performance.

After converting all types of input into the BoW representation, we would like to use both complementary and correlated information during the fusion process. Thus we follow the procedure given in Figure 7.6. The figure represents the fusion of two different modalities. The procedure basically includes four steps, as listed below. These steps are presented in details, in the following subsections.

- Classification of information in each modality,
- Intramodal correlation analysis for each modality and classification of the obtained information,
- Intermodal correlation analysis between modalities and classification of the obtained information,
- Late fusion of all classification results

Here, it should be noted that, this study mostly focuses on the combination of correlated information within the BoW inputs, considering that the late fusion step of the given procedure is rather simple and attacked during the What-to-Fuse part of the thesis work. Combination of correlated information is usually accepted as a type of early fusion. Hence, the proposed fusion mechanism is a multi-level approach, which includes both early and late fusion stages. The early fusion focuses on the correlated information, whereas the late fusion mostly deals with the complementary information.

As a solution to the intramodal and intermodal correlation discovery problem, we prefer a novel mining, graph and correlation based solution, which is constructed on the ideas discussed in Section 7.3. It has been previously stated that the N-gram approach has a potential to provide extra intramodal and intermodal correlation information, however it suffers from the combinatorial explosion problem. A practical solution for this problem can be obtained by finding the co-occurrences of interesting patterns by using the association rule mining (ARM) or the frequent itemset mining (FIM)

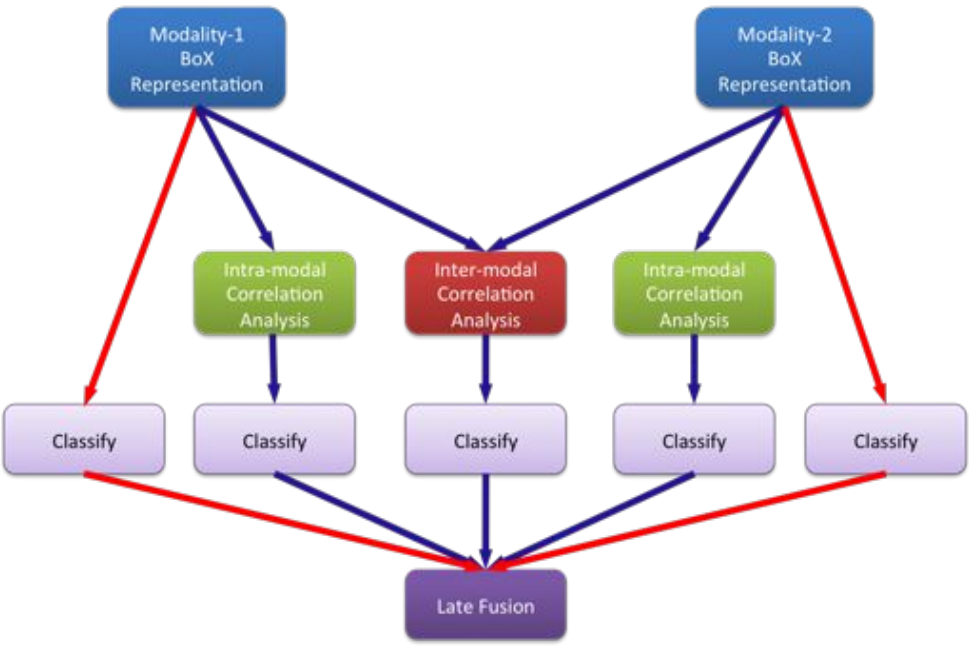


Figure 7.6: Workflow for Combining Bags of X

techniques instead of an exhaustive search of the N-gram approach. While using these techniques, using the local information (keypoints and scale/rotation invariant local features) and representing the extracted information with graphs are promising techniques. After the analysis of pros / cons of each idea discussed, we come up with a novel solution, as detailed in the following subsections.

7.4.1 Learning from Single Modalities

Even though we conduct a study that focuses on the fusion of multimodal inputs, it should be noticed that fusion enhances the recognition performance, where the performance without fusion is still important. Thus, the learning process by using each single modality should be as effective as possible. In this direction, Jiang et al. [54] perform a detailed analysis on the optimal use of bag-of-words based approaches. In this study, we follow the findings of Jiang et al. and propose to use the following learning scheme:

- Classifier & Kernel Choice: Support Vector Machine (SVM) is one of the most popular classifiers for the BoW-based classification. For SVMs, the choice of an

appropriate kernel function is a critical issue for the classification performance. In literature, existing studies usually prefer Linear Kernel, Histogram Intersection Kernel, Gaussian Radial Basis Function (RBF) Kernel and χ^2-RBF Kernel. In [54], Jiang et al. experimentally show that χ^2-RBF Kernel is superior to the others when BoW-based feature vectors are used for classification. Thus, we prefer SVM with χ^2-RBF kernels for the classification tasks. Another issue with the classification procedure is how to handle the multi-class classification. Considering that multimedia data is usually multi-labeled, the classification procedure should be correspondingly multi-class and multi-labeled. For such a purpose, we prefer a 1-against-all approach, where we train k number of classifiers for k different class labels.

- Weighting scheme: Weighting in BoWs is statistical information about the occurrences of words in the multimedia documents. The most basic scheme is the binary weighting, which indicates the presence / absence of a word in each document. More sophisticated schemes include term frequency (TF) and/or term frequency - inverse document frequency (TF-IDF), which perform superior than binary weighting. Thus, we prefer TF weighting.

- Vocabulary size: For BoW modeling, vocabulary is the set of keypoint clusters in the clustering process. Having a small sized vocabulary may cause a loss of discriminative power since two keypoints may be assigned into the same cluster although they are not similar to each other. On the contrary, a large vocabulary is less generalizable, less tolerant to noises, and causes extra processing. The studies in the literature work with a large range of vocabulary sizes from 100 to 10,000. In [54], it is shown that the impact of vocabulary size is less significant when sophisticated weighting schemes are employed. Hence, we prefer a mid-level size (4096).

7.4.2 Intramodel Correlation Analysis

For the intramodal correlation discovery part of the problem, each modality is processed separately. Considering that parts of a particular object or a scene usually occurs together in different samples of that object/scene, grouping the together oc-

curring words is a promising idea. Thus, we propose to use *phrase*s as the groups of frequent word occurred together. In order to find the *phrase*s, a mining and graph based algorithm (Unimodal Phrasing Algorithm, given in Algorithm 5) is used on the training dataset. Through the algorithm, we first try to find some meaningful *phrase*s from the *word*s in each modality, with the help of FIM and graph representation. Thus we manage to exploit the intramodal relations within each modality.

Here, it should be noted that samples of each class should be processed separately, which means that the algorithm is executed separately for each class, in order not to cause the 'rare itemset' problem. While performing frequent itemset mining, support thresholds for frequent patterns highly differs in different classes. Thus, class-specific support thresholds should be applied for each class. In addition, processing samples of different classes separately is beneficial since the samples of multimedia data is multi-labeled (multi-label issue is discussed in Section 6.3.2 in detail).

The algorithm is composed of two parts; mining and phrase extraction. The first part requires performing a frequent itemset mining (FIM) operation and calculating support values of frequent 2-itemsets. For this purpose, all BoW feature vectors of all multimedia documents (i.e. videos or video shots) are processed and a transaction for each document is generated. Each transaction includes the words existing in the corresponding feature vector. Here, an 'existing word' refers to the words having weights larger than zero. After constructing the transaction list, a FIM operation is performed (i.e. via FP-Growth or Apriori algorithms) for calculating the frequent 2-itemsets from the obtained transaction list. The algorithm requires the $freqThr$ parameter as the minimum support threshold for calculating the frequent 2-itemsets at this step. Each 2-itemset is a set consisting of two words. The support value of 2-itemset is the co-occurrence probability of the corresponding two words.

The second part of the algorithm is based on a graph construction. In order to extract phrases and find the words included by each phrase, a graph representation is constructed by using the support values of 2-itemsets and the vocabulary of words. In the graph, the words are assumed as the vertices, and the edges are valued by using the support values of 2-itemsets. As mentioned before, previous studies on this issue usually prefer the neighborhood information or a threshold based spatial / temporal

Algorithm 5: Unimodal Phrasing Algorithm

Input: Multimedia documents $\mathcal{D} = \{d_i\}_{i=1}^{t}$, vocabulary $\mathcal{W} = \{w_i\}_{i=1}^{m}$, minimum support for itemsets $freqThr$, threshold for graph pruning $neiThr$, depth threshold for graph pruning h

Output: Phrases list P

begin
  // Mining
  $T \leftarrow \langle\rangle$; //Initialize transactions list
  for $d_i \in \mathcal{D}$ **do**
    $b_i \leftarrow extractBoW(d_i, \mathcal{W})$;
    $t_i \leftarrow generateTransaction(b_i)$;
    $T \leftarrow add(T, t_i)$; //Add a new transaction
  end
  $I \leftarrow findFreqItemsets(T, freqThr, 2)$; //2-itemsets, sup>*freqThr*
  // Phrase extraction
  $G\langle V, E\rangle \leftarrow constructGraph(\mathcal{W}, I)$;
  $P \leftarrow \langle\rangle$; //Initialize phrases list
  for $v_i \in V$ **do**
    $N \leftarrow selectNN(v_i, neiThr, h)$; //NN within h-depth, sup$(v_i, v_j) \geq$*neiThr*
    $phr_i \leftarrow \{v\} + N$;
    $add(P, phr_i)$;
  end
end

distance as the edge values. However, we find it promising and helpful to use the support values of the 2-itemsets as the edge values. This preference is based on two assumptions; (i) different words occurring together in several shots/videos can be accepted in the same phrase; since it is actually not required for these words to be the parts of the same object / concept, (ii) support value is a better metric than the distances between words for phrase construction, since co-occurrence is more important than closeness for the words.

After constructing the graph, the phrases are extracted by processing the graph. Alternative approaches may be applicable for the efficient handling of the graph. Considering that the task is to extract group of words from the word graph, finding the maximum cliques in the graph is an applicable idea. However, we take the following restrictions

into account; (i) the clique decision problem is NP-complete, (ii) the graph constructed by using the words and the 2-itemset supports is a large-sized complete graph (i.e. we use a 4096-word vocabulary), (iii) pruning of the edges is possible, however, a low threshold for pruning cannot prevent the graph to be densely connected, whereas a high threshold may cause the loss of valuable information, (iv) finding a maximal clique is not enough; for fairness among the words, a maximal clique should be calculated for each word (or for the most frequent ones, at least). Thus, even the heuristics for maximal clique calculation are not timely-efficient for our problem. Therefore, we propose an alternative heuristic for the phrase extraction task, which is based on the k-nearest neighbors approach.

In the above mentioned phrase extraction task, nearest neighbors of each word v_i in the vocabulary are selected considering the following requirements; (i) the neighbor (v_j) should be within h-depth with respect to the word, (ii) the support value $sup(v_i, v_j)$ should be larger than the $neiThr$ threshold value. Thus, the most co-occurring words for each word in the vocabulary are found. Although the results are not the cliques for each word, they are still meaningful for using as phrases.

After obtaining phrases with the above given procedure, phrase-based feature vectors should be extracted from the training and test data. Considering that each phrase contains multiple words, we need an aggregation method to assign some numerical values for each phrase. For this purpose, we prefer a simple averaging approach. In this approach, average of TF values of the words in each phrase is calculated as the phrase value. After performing the aggregation task, phrase-based feature vectors are obtained for each training and test document.

The learning procedure by using the extracted phrase-based feature vectors is similar to the procedure given in Section 7.4.1. The data is classified with a χ^2-RBF kernel based SVM classifier.

The time complexity of the algorithm depends on the most complex operation, which is the frequent itemset mining task. If the FP-Growth implementation is preferred for this task, the task performs two passes on the whole training documents, which makes our algorithm bounded by linear time space complexity, in terms of number of multimedia documents. In fact, the number of modalities is also an important concern for this

study. However, the number of modalities is very small according to the number of multimedia documents. Yet, if the number of modalities is concerned, the algorithm is still in linear time since operation is performed in a single modality.

Besides, the space complexity of the algorithm is highly dependent on the implementation. The largest space requirements are caused by the multimedia documents, transaction list, word graph, phrase list and frequent itemset mining task. Actually, it is not reasonable to hold features for all documents and the constructed transaction list in the memory. Instead, one document at a time is read, processed and the corresponding transaction is written to disk for further processing. So, the space requirement for reading features and transaction construction is only the size of the feature vector for a single training sample, which is bounded by the vocabulary size. On the other hand, the space complexity of the frequent itemset mining task is bounded by the number of items (or the vocabulary size, for our case), if an FP-Growth implementation is utilized. The space necessary for the word graph can be calculated by adding vocabulary size (number of nodes) and number of 2-itemsets (edges), which is in quadratic time in terms of vocabulary size. Lastly, the space required for phrase is also bounded by squared vocabulary size. Consequently, the space complexity of the algorithm is $O(m^2)$, where m is the vocabulary size .

7.4.3 Intermodel Correlation Analysis

For the intermodal correlation discovery part of the problem, all modalities are processed together and the correlation between the words and phrases in different modalities is exploited. The idea given for intramodal correlation analysis is also applicable for intermodal processing. The sensed data of a particular scene, which is collected though different channels (modalities) usually has some parts occurring together in different samples of that scene. For instance, if different samples of 'car' videos are processed, it is highly probable that several visual words depicting that parts of a car will occur together with some sound signals belonging to a car.

Considering that we have already grouped the frequently co-occurred *word*s into *phrase*s during the intramodal correlation analysis, in this section we propose to extract *multimodal phrases* by grouping phrases of different modalities. In order

to find the *multimodal phrase*s, a correlation and graph based grouping algorithm (Multimodal Phrasing Algorithm, given in Algorithm 6) is applied on the training dataset. Through the algorithm, firstly the correlation between the pairs of phrases, where each of the phrases in a pair belongs to different modalities, is calculated. Then, multimodal phrase groups are formed, by selecting only one phrase from each modality and generating a multimodal phrase for each phrase of each modality. Similar with the intramodal analysis, the given algorithm is executed separately for each class. Thus, we manage to exploit the intermodal relations within each modality.

The proposed intermodal correlation analysis approach differs from the intramodal analysis approach for the following aspects:

- The intramodal analysis is based on the words of a single modality and outputs phrases as the group of words. In each phrase, there is not a limit for the included number of words. On the other hand, intermodal analysis is not based on the words, it exploits correlation between phrases, and outputs groups of phrases. In addition, the number phrases in each group (multimodal phrase) is limited and equals to the number of modalities. The reason for such a preference is two-fold; (i) to prevent the domination of a particular modality, since strong intramodal correlation of a particular modality may cause adding the phrases of that modality densely, (ii) not to include any intramodal correlation information into the multimodal phrases, since adding more than one phrase from any modality will contain an intramodal correlation information of that modality.

- As mentioned in Section 7.3.2.1, when the words / phrases of multiple modalities are combined for a FIM operation, the well-known problem of 'rare itemsets' occurs. The problem is caused by applying a single support threshold for both of the modalities. For instance, we have experimentally experienced that SIFT-based words usually have higher support values than the MFCC-based words. Considering that the actual problem is not the selection of co-occurring items in a set, but items of different sets, and combining the different sets into a single set for FIM causes inefficiencies, we prefer a correlation based selection approach. In our approach, the correlation between all pairs of phrases from different modalities are calculated, although it can be argued that the approach

Algorithm 6: Multimodal Phrasing Algorithm

Input: Modalities $\mathcal{M} = \{m_i\}_{i=1}^{n}$, multimedia documents $\mathcal{D} = \{d_i\}_{i=1}^{t}$, phrase vocabulary list of all modalities $\mathcal{PW} = \langle PW^i \rangle_{i=1}^{n}$ s.t. each phrase vocabulary $PW^i = \{phr_j\}_{j=1}^{r}$

Output: Multimodal phrases list MMP

begin
  // Correlation calculation
  for $d_k \in \mathcal{D}$ **do**
    for $m_i \in \mathcal{M}$ **do**
      $P^i \leftarrow getPhraseVector(d_k, m_i)$;
      for $p_a \in P^i$ **do**
        $mean[m_i][p_a] \leftarrow mean[m_i][p_a] + value(p_a)/size(\mathcal{D})$
      end
    end
  end
  for $d_k \in \mathcal{D}$ **do**
    foreach $\{m_i, m_j\} \in \mathcal{M} \times \mathcal{M},\ i \neq j$ **do**
      $P^i \leftarrow getPhraseVector(d_k, m_i)$;
      $P^j \leftarrow getPhraseVector(d_k, m_j)$;
      for $\{p_a, p_b\} \in P^i \times P^j,\ p_a \in P^i \wedge p_b \in P^j$ **do** //Pearson's corr.coeff.calculation
        $partX \leftarrow value(p_a) - mean[m_i][p_a]$;
        $partY \leftarrow value(p_b) - mean[m_i][p_b]$;
        $partCov \leftarrow partX \times partY$;
        $cov[m_i][m_j][p_a][p_b] \leftarrow cov[m_i][m_j][p_a][p_b] + partCov$;
        $stdDev[m_i][p_a] \leftarrow sdtDev[m_i][p_a] + partX^2$;
        $stdDev[m_j][p_b] \leftarrow sdtDev[m_j][p_b] + partY^2$;
      end
    end
  end
  foreach $\{m_i, m_j\} \in \mathcal{M} \times \mathcal{M},\ i \neq j$ **do**
    foreach $\langle phr_k, phr_l \rangle \in PW^i \times PW^j, phr_k \in PW^i \wedge phr_l \in PW^j$ **do**
      $r[m_i][m_j][phr_k][phr_l] \leftarrow cov[m_i][m_j][phr_k][phr_l]/(stdDev[m_i][phr_k] \cdot stdDev[m_j][phr_l])^{1/2}$
    end
  end
  // Phrase extraction
  $MMP \leftarrow \langle\rangle$; //Initialize multimodal phrases list
  for $m_i \in \mathcal{M}$ **do**
    for $phr_k \in PW^i$ **do**
      $mmPhr_i \leftarrow \{phr_j\}$;
      for $m_j \in \mathcal{M} - m_i$ **do**
        $phr_l \leftarrow argMax(r[m_i][m_j][phr_k])$; //Get max correlated phrase
        $mmPhr_i \leftarrow mmPhr_i + \{phr_l\}$;
      end
      $add(MMP, mmPhr_i)$;
    end
  end
end

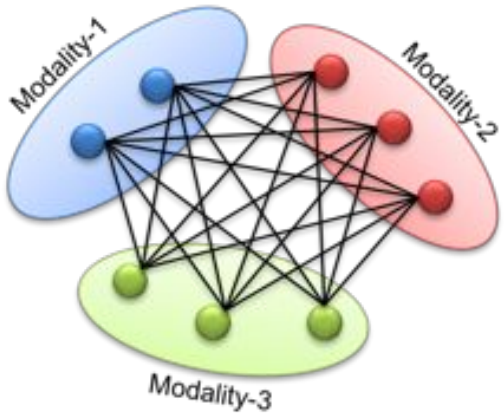


Figure 7.7: A sample graph representation for intermodal analysis

is computationally complex. Yet, the approach is just in quadratic (or sub-quadratic) time in terms of the number of modalities. However, for a mining task, the number of modalities is not the main concern since the number of modalities is very small according to the number of multimedia documents. Considering that the documents are traversed twice, as the FP-Growth algorithm does, the complexity is not a problem.

- Intramodal analysis constructs a complete graph, from a single modality, by adding the words as the nodes and support values of the word pairs as the edge weights. In intermodal analysis, not only a single modality but all modalities are included into the graph, by adding the phrases of all modalities as the nodes of the graph. Actually, this graph is composed of several sub-graphs, each sub-graph is composed of the nodes a different modality. There does not exist any edge between the nodes within a sub-graph; the edges are between the nodes from different sub-graphs. Edge weights are based on the calculated correlation values between the phrases. Such a graph is illustrated in Figure 7.7. During the multimodal phrase extraction, for each node (phrase) in the graph, a single nearest neighbor (having the largest edge weight) from each sub-graph is selected.

The algorithm is composed of two parts; correlation calculation and phrase extraction. The first part requires calculating the correlation between the phrases of different

modalities. The correlation is calculated based on the Pearson's correlation coefficient.

Definition 1. *Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. For samples X and Y, the sample Pearson's correlation coefficient is calculated as the following:*

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} . \tag{7.4}$$

Considering that the number of multimedia documents used for training may be so large that it may not be possible to load the features all documents into memory to process, the algorithm should process the documents sequentially. Thus the algorithm pass over the whole documents twice. In the first turn, the mean values for all phrases in each modality are calculated and stored in an array. In the second turn, the covariance (numerator) and the standard deviation (denominator) calculations are performed, which are necessary to calculate the correlation coefficient. After these calculations, the third step is performed by passing over all phrases of all modalities. In this step, the final Pearson's correlation coefficients for all phrase pairs are obtained.

The second part of the algorithm enables finding the most correlated phrases in each modality for all phrases and construct the list of multimodal phrases. The number phrases in each multimodal phrase is limited and equals to the number of modalities. As mentioned above, it is possible to construct a graph by including the phrases of all modalities as the nodes of the graph, in order to perform the phrase extraction, as it is done in intramodal correlation analysis. Such a graph should contain several sub-graphs, each sub-graph is composed of the nodes a different modality. For multimodal phrase extraction, for each phrase in the graph, a single nearest neighbor from each sub-graph is selected. Considering that we are interested in selecting 1-nearest neighbor, the algorithm is given with a simpler representation without mentioning the graph construction. Actually, 1-nearest neighbor selection, or the $argMax$ operation given in the algorithm can be easily implemented by a heap data structure. In short, the constructed correlation coefficients array is used to find the maximum correlated phrase pairs. Thus, the list of multimodal phrases is generated.

After obtaining the multimodal phrases with the above given procedure, feature vectors

for multimodal phrases should be extracted from the training and test data. Similar with the intramodal analysis, a simple averaging approach is used for aggregation the multiple phrases into a single multimodal phrase and assign some numerical values for each multimodal phrase. After averaging the TF values of the phrases and assigning these values as the multimodal phrase values, multimodal phrase-based feature vectors are obtained for each training and test document.

The learning procedure for intermodal analysis is also is similar to the procedure given in Section 7.4.1. For learning and querying, the extracted multimodal phrase-based feature vectors are used. The data is classified with a χ^2-RBF kernel based SVM classifier.

The time complexity of the given algorithm is bounded by the correlation calculation operation, which is the most complex operation. As mentioned above, the number of modalities is very small according to the number of multimedia documents. Thus, the given algorithm in linear time in terms of the number of multimedia documents, since the algorithm requires two passes on the whole dataset. If the number of modalities is concerned, the complexity of the algorithm is in quadratic time (or sub-quadratic time, depending on the implementation) in terms of the number of modalities, since all pairs of phrases from different modalities are calculated.

As discussed in the unimodal phrasing algorithm, the space complexity of the algorithm is highly dependent on the implementation. The space requirements are based on the following components; multimedia documents, correlation calculation and multimodal phrases. Similar with the previous algorithm, it is not reasonable to hold features for all documents, thus one document at a time is read, and included into the correlation calculation. The space requirement for reading features is only the size of feature vector for a single training sample, which is bounded by the phrase vocabulary size (r). The correlation calculation requires holding three parameters; $mean$ matrix ($O(n \cdot r)$), cov matrix ($O(n^2 \cdot r^2)$) and $stdDev$ matrix ($O(n \cdot r)$), where n is the number of modalities. The extracted multimodal phrases costs an $O(n^2 \cdot r)$ complexity space, for having r number of multimodal phrases for each modality, and having n number of items in each multimodal phrase. Consequently, the space complexity of the algorithm is $O(n^2 \cdot r^2)$.

7.4.4 Late Fusion of All Inputs

As mentioned in Section 7.4.1, fusion helps to improve the retrieval performance, yet the major component for retrieval is the learning procedure by using each single modality. Learning schemes provided by different modalities and the intramodal / intermodal analyses abstract videos from different aspects. Each of these learning schemes most likely complement each other, and the sets of patterns misclassified by different learning schemes do not necessarily overlap. Thus, all these schemes should be combined to improve the recognition capability.

Figure 7.6 illustrates the workflow of our proposed combination approach. After performing the classification procedures of each modality and also the intramodal / intermodal analyses, the results of classifications are combined with a late fusion scheme. The results are fused by applying a Linear Weighted Averaging approach. As discussed previously, Linear Weighted Averaging is the most frequently utilized approach in the information fusion literature [37, 133, 145], due to its simplicity and reasonable performance despite its simplicity. The approach requires good selection of the weights for successful results, thus it is supported with the RELIEF-MM algorithm for modality / feature weighting.

7.5 Empirical Study

In this section, we evaluate the proposed modality fusion approach for semantic retrieval of multimedia data. For the retrieval task, the multimedia data is queried based on the semantic concepts. First, retrieval for each single modality is performed, then a multimodal retrieval is done.

7.5.1 Experimental Setup

Experiments are carried out on the TRECVID 2011 dataset [90], which is a frequently utilized benchmark datasets for multimedia retrieval. The dataset characteristics are summarized in Table 7.2. Further details and a performance comparison of TRECVID participants can be found in the corresponding references.

Table 7.2: TRECVID 2011 dataset characteristics

	Train	Test
Dataset length (hours)	$\sim$ 400	$\sim$ 200
Number of videos	$\sim$ 16,000	$\sim$ 8,000
Number of shots	71,502	34,179
Concepts	Adult, Anchorperson, Beach, Car, Charts, Cheering, Dancing, Demonstration_Or_Protest, Door_Opening, Doorway, Event, Explosion_Fire, Face, Female_Human_Face, Female_Person, Female-Human-Face-Closeup, Flags, Flowers, Hand, Head_And_Shoulder, Indoor, Male_Human_Face, Male_Person, Mountain, News, News_Studio, Nighttime, Old_People, Overlaid_Text, People_Marching, Quadruped, Reporters, Running, Scene_Text, Singing, Sitting_Down, Skating, Sky, Speaking, Speaking_To_Camera, Sports, Streets, Studio_With_Anchorperson, Table, Text, Traffic, Two_People, Urban_Scenes, Walking, Walking_Running	

While using the TRECVID 2011 dataset, we prefer using the outputs of common shot reference provided with the dataset, for shot segmentation. For these datasets, the shots are used as the retrieval documents. The dataset also provides concept annotations for each shot. The annotations are provided in a multi-label manner, which means each shot can contain more than one label. In this experimental evaluation, we prefer working with the configuration for the Lite Run of the Semantic Indexing task of TRECVID 2011. Thus, 50 concepts are used as the shot annotations. A complete list of these concepts is given in Table 7.2. The semantic queries performed during the tests are based on these semantic concepts.

For a multimodal setting, we use three features from different modalities. The modalities of multimedia data are usually accepted as audio, visual and text, thus we employ one BoW-based feature for each of these modalities. Detailed description of how these modalities have been obtained and utilized are as follows:

- SIFT (Visual): The BoW-based SIFT features are not extracted from scratch. We prefer using the 4096-bin histograms of the SIFT BoW features extracted by INRIA from IRIM consortium [7] for the TRECVID 2012 evaluation. Having 4096-bin histogram means that we use 4096 visual words in the vocabulary of visual modality. During the feature extraction, INRIA prefers dense-sampling

for keypoint extraction, and keypoints are extracted from the frames provided by the common shot reference. The weighting scheme for the features is TF based.

- MFCC (Audio): Similar with the SIFT features, we use the 4096-bin histograms of the MFCC BoW features extracted by LIRIS from IRIM consortium [7] for the TRECVID 2012 evaluation. In the provided dataset, the MFCC features are extracted by using the audio waves of 2 seconds around the keyframes of each video shot, with parameters of 20 ms window length and 10 ms window shift. After extracting the MFCC features, the extracted audio keypoints are clustered into a 4096 word vocabulary. The weighting scheme for the features is TF based.

- TF-IDF (Text): The textual TF-IDF features are calculated by using the Automatic Speech Recognition and Machine Translation texts, which are provided by TRECVID. Before the calculations, a stop-word filtering procedure is applied and all the remaining words are used as textual vocabulary, without feature selection or any further processing.

As mentioned in the previous section, we combine the modalities, intramodal / intermodal phrases with a late fusion process, thus each of these sources are first processed with a Support Vector Machine (SVM) classifier, with χ^2-RBF kernel. For the SVM implementation, LibSVM [18] is utilized. After classification, the classifier outputs are combined with a Linear Weighted Averaging approach. The weights are calculated using the RELIEF-MM algorithm.

During the intramodal correlation analysis, first the dataset is converted into transactions for frequent itemset mining. For this purpose, the BoW based feature vectors are converted into binary transactions. If the TF value of any word is larger than zero, it is accepted as 1 in the transaction, and 0 otherwise. For the parameters $freqThr$ (minimum support for itemsets) and $neiThr$ (threshold for graph pruning), the average of the 1-itemset supports is used, since such preference is both practical to calculate and halves the number of words used during mining. For the frequent itemset mining implementation, FP-Growth implementation of Borgelt [13] is utilized.

To measure the retrieval accuracy, *Average Precision (AP)* and *Mean Average Precision*

(MAP) are used. The *AP* is the sum of the precision at each relevant hit in the retrieved list, divided by the minimum of the number of relevant documents in the collection and the length of the list. Here, *precision* is the fraction of retrieved documents that are relevant to the query concept. Regarding the evaluation rules of TRECVID, AP is measured at 2000. *MAP* is the *AP* averaged over several query concepts. In other words, the *AP* of each concept is calculated separately and then the *MAP* is found by averaging them.

To perceive the effect of the fusion process, we also measure the Fusion Gain (FG). Fusion gain gives the relative performance increase between to two different configurations:

$$FG(x,y) = \frac{MAP(x) - MAP(y)}{MAP(y)} \ , \tag{7.5}$$

where x and y denote different configurations (i.e. different feature selections). In our experiments we calculate three *FG*s:

- FG_{BestSM}: The fusion gain is calculated by comparing with the best single modality.
- FG_{Comb_S}: The fusion gain is calculated by comparing with the combination of all single modalities.
- $FG_{Comb_{SI}}$: The fusion gain is calculated by comparing with the combination of all single modalities and the results of the intramodal analyses of all modalities.

7.5.2 Test Results and Evaluations

In order to see the effectiveness of the proposed fusion approach, and also the effect of intramodal and intermodal correlation analysis procedures, retrieval accuracies of the following configurations are measured.

- *Each single modality ($Visual$, $Audio$, $Text$):* Retrieval accuracy of each single modality is measured.
- *Intramodal analysis of each single modality (Phr_V,Phr_A,Phr_T):* Intramodal analysis outputs phrases for each modality. Phr_V,Phr_A,Phr_T denotes the use of extracted phrases from each modality; Visual, Audio and Text, respectively.

- *Best single modality ($BestSM$):* This is not an actual configuration, but the results for using each single modality is used select the best single modality for each concept.
- *Combination of all modalities ($Comb_S$):* All modalities ($Visual$, $Audio$, $Text$) are combined with a late fusion scheme.
- *Combination of all intramodal analyses outputs ($Comb_I$):* Modality phrases (Phr_V,Phr_A,Phr_T) are combined with a late fusion approach.
- *Intermodal analysis ($MMPhr$):* Intermodal analysis outputs multimodal phrases. $MMPhr$ denotes the use of extracted multimodal phrases for retrieval.
- *Combination of modalities and intramodal analyses outputs ($Comb_{SI}$):* $Comb_{SI}$ is the combination $Comb_S$ and $Comb_I$
- *Combination of all inputs ($Comb_{All}$):* $Comb_{All}$ is the combination $Comb_S$, $Comb_I$ and $MMPhr$.

Using each single modality and reporting the best single modality helps to represent the lower accuracy bound for the fusion system. A fusion system can be accepted successful if it provides better accuracy than any of the single modalities. Moreover, the retrieval accuracy of intramodal phrases for each modality and the intermodal (multimodal) phrases are measured to see how much successful they are before fusion. In addition to those representing the accuracies without fusion, several fusion results are also reported; combination of single modalities ($Comb_S$), combination of all intramodal phrase based classification results ($Comb_I$), combination of single modalities and intramodal outputs ($Comb_{SI}$), and finally combination of all inputs ($Comb_{All}$). $Comb_S$ is important to see how much increase provided by the intramodal and intermodal analysis. Similarly, $Comb_{SI}$ is necessary to see the improvement obtained by the intermodal process.

In order to evaluate the proposed approach, one may argue that there should be comparisons with other studies using the TRECVID 2011 dataset. However, considering below given restrictions, we do not find it fair to compare our experimental results with other studies. Yet, we present a brief information about the retrieval results of TRECVID participants in Table 7.3. In the table, the best and the median retrieval

Table 7.3: Retrieval Performances of TRECVID Participants. 'Best MInfAP' refers to the best retrieval performance achieved in terms of Mean Inferred Average Precision. 'Median MInfAP' is the median of the retrieval performances reported by the all participants.

	Best	Median
TRECVID 2010	10.3%	2.1%
TRECVID 2011	14.9%	5.6%
TRECVID 2012	35.8%	21.2%
TRECVID 2013	32.1%	12.8%

performances performed by the TRECVID participants are given. It is worth noting that the retrieval performances highly differ in different years. This is probably caused by the amount of annotated information, the quality of annotations and the semantic indexing concepts used for retrieval. The restrictions that prevents a fair comparison is as follows:

(i) As described in the TRECVID dataset descriptions, the training and test samples of an evaluation at year t is used as training samples at year $t+1$. In order to obtain annotated training and test dataset, we use the training samples of the TRECVID 2012 dataset and the corresponding annotations. The training set of the TRECVID 2012 dataset is composed of the training and the test samples of TRECVID 2011. Thus, we state that our tests are performed with TRECVID 2011 dataset. In principle, the samples of these two years are the same. However, the annotations are improved every year, so it is not fair to compare the results of TRECVID 2011 participants with a system using TRECVID 2012 annotations.

(ii) In TRECVID 2011 evaluation, there are 137.327 shots in the test dataset. However, we consider only 34.179 shots, which are the ones annotated as including at least one concept in it. Thus, it is not fair to compare two systems that has different sampling scheme for testing.

(iii) In TRECVID evaluation, the basic accuracy metric is *Inferred Average Precision* [3], which estimates average precision very well using a small sample of test samples. However, we prefer measuring with *Average Precision*. Although the quality of *Inferred Average Precision* has been experimentally confirmed during

previous TRECVID evaluations, they are still different metrics and there is a possibility of deviation.

(iv) The fine-tuning of the classifiers has an important effect on the retrieval performance. In the TRECVID evaluations, participants try to do their best performance in retrieval. However, our tests are primarily focused on investigating the effect of fusion, and thus, the SVM classifiers are usually not fine-tuned.

Table 7.4 and Table 7.5 presents the AP values for the retrieval of 50 concepts. In the tables, retrieval accuracies for above given configurations are reported. The concepts are given in alphabetical order. Figure 7.8 illustrates the AP values of all concepts for $BestSM$, $Comb_S$ and $Comb_{All}$ configurations. In Figure 7.9, corresponding MAP values are illustrated. In addition, Table 7.6 shows the fusion gains of each fusion configuration with respect to other fusion configurations as described above.

From the given experimental results, we arrive at the following observations:

- Combining different modalities provide more accurate results than the single modalities. The AP values of $Comb_S$ is higher than the $BestSM$ for 48 of the 50 concepts. In addition, MAP for $Comb_S$ is also better than the $BestSM$, and $Comb_S$ provides a 8.92% fusion gain on the best single modality based on the MAP values. Here, it should be noted that the selection / weighting of modalities is a critical issue, and the use of RELIEF-MM enables such a successful result. A wrong selection can lead to worse results than the best of the single modalities.

- Our proposed approach, including both intramodal and intermodal analyses, performs superior than any single modality and the combination of the single modalities. The proposed approach achieves 40.64% retrieval rate, whereas the best single modality performs 36.15% success and the combination of single modalities obtains a rate of 38.94%. Considering these results, exploiting the intramodal and intermodal correlations in / between modalities enables doubling the fusion gain. The fusion gain of the proposed approach ($Comb_{All}$) is 17.12% with respect to best single modality, whereas the fusion gain for $Comb_S$ is 8.92%.

Table 7.4: Retrieval accuracies for the first half of concepts.

	$Visual$	$Audio$	$Text$	Phr_V	Phr_A	Phr_T	$MMPhr$	$BestSM$	$Comb_S$	$Comb_I$	$Comb_{SI}$	$Comb_{All}$
Adult	28.49%	15.87%	11.98%	21.14%	14.19%	12.69%	21.22%	28.49%	29.25%	21.72%	29.52%	30.86%
Anchorperson	72.42%	52.68%	8.66%	34.87%	32.65%	1.84%	21.73%	72.42%	80.04%	45.32%	80.61%	81.43%
Beach	33.37%	16.46%	1.09%	20.73%	7.44%	1.33%	22.17%	33.37%	37.12%	20.94%	37.12%	38.25%
Car	52.35%	7.05%	4.90%	13.21%	6.93%	8.06%	13.81%	52.35%	53.68%	13.84%	54.01%	54.93%
Charts	9.23%	0.46%	0.51%	10.08%	0.85%	0.00%	9.39%	9.23%	10.71%	8.83%	14.71%	19.93%
Cheering	21.59%	10.96%	1.90%	7.72%	2.83%	2.09%	12.89%	21.59%	25.78%	8.02%	25.98%	27.59%
Dancing	11.81%	3.58%	0.70%	6.27%	3.17%	0.79%	6.48%	11.81%	11.81%	6.42%	12.05%	13.29%
Demonstration_Or_Protest	33.70%	36.44%	1.91%	2.81%	23.18%	2.13%	8.33%	36.44%	45.61%	23.74%	47.04%	48.27%
Door_Opening	13.34%	0.43%	0.32%	4.56%	0.29%	0.18%	3.55%	13.34%	13.54%	4.60%	13.54%	13.54%
Doorway	40.06%	7.70%	2.95%	20.72%	1.56%	1.88%	18.08%	40.06%	41.86%	21.16%	42.69%	43.37%
Event	8.96%	7.15%	5.85%	5.85%	6.75%	6.01%	7.40%	8.96%	9.76%	7.21%	9.81%	10.29%
Explosion_Fire	36.86%	17.51%	3.14%	7.54%	12.61%	5.23%	9.65%	36.86%	40.27%	13.23%	41.01%	41.85%
Face	30.11%	11.77%	9.24%	19.42%	12.04%	9.07%	18.01%	30.11%	31.09%	20.01%	32.01%	32.68%
Female_Human_Face	30.41%	15.36%	3.73%	14.81%	9.96%	1.74%	18.12%	30.41%	33.10%	15.05%	33.12%	34.60%
Female_Person	33.64%	20.23%	6.20%	14.97%	17.57%	5.86%	22.30%	33.64%	37.71%	19.19%	38.41%	39.84%
Female-Human-Face-Closeup	38.89%	6.61%	3.01%	21.41%	2.26%	1.59%	14.74%	38.89%	41.51%	21.68%	44.13%	44.89%
Flags	34.53%	4.07%	1.44%	5.09%	1.47%	3.25%	6.69%	34.53%	35.36%	5.55%	35.36%	35.36%
Flowers	45.57%	15.36%	1.92%	26.99%	4.99%	1.04%	23.99%	45.57%	50.27%	27.61%	50.27%	51.05%
Hand	33.12%	2.56%	1.67%	24.21%	2.21%	0.85%	21.46%	33.12%	33.69%	25.13%	35.17%	36.04%
Head_And_Shoulder	31.37%	12.97%	5.44%	13.70%	6.88%	5.51%	15.01%	31.37%	32.50%	14.39%	32.88%	33.42%
Indoor	33.25%	21.77%	10.07%	12.98%	18.23%	10.73%	17.62%	33.25%	36.02%	18.89%	36.02%	37.02%
Male_Human_Face	40.23%	17.35%	5.78%	13.08%	12.84%	6.11%	18.99%	40.23%	43.47%	14.97%	43.74%	44.91%
Male_Person	27.19%	23.58%	10.16%	13.72%	21.12%	10.42%	20.64%	27.19%	33.07%	21.70%	34.07%	35.39%
Mountain	70.13%	5.68%	7.55%	41.31%	6.04%	1.83%	40.85%	70.13%	71.32%	42.51%	71.32%	71.32%
News	55.39%	41.04%	0.95%	37.77%	29.06%	2.63%	18.35%	55.39%	66.52%	44.54%	71.88%	72.78%

Table 7.5: Retrieval accuracies for the second half of concepts.

	$Visual$	$Audio$	$Text$	Phr_V	Phr_A	Phr_T	$MMPhr$	$BestSM$	$Comb_S$	$Comb_I$	$Comb_{SI}$	$Comb_{All}$
News_Studio	70.92%	47.59%	1.27%	38.96%	30.19%	1.62%	28.88%	70.92%	77.74%	43.67%	78.45%	79.48%
Nighttime	45.98%	2.60%	0.79%	18.82%	3.41%	2.26%	13.72%	45.98%	46.15%	19.23%	46.39%	47.24%
Old_People	24.33%	12.52%	4.46%	4.55%	13.27%	5.80%	11.66%	24.33%	27.93%	13.98%	28.87%	29.59%
Overlaid_Text	68.17%	11.43%	9.06%	66.54%	8.68%	8.58%	67.70%	68.17%	70.25%	67.51%	71.78%	72.51%
People_Marching	3.48%	3.67%	1.47%	2.37%	2.43%	0.85%	1.29%	3.67%	5.89%	2.75%	8.92%	9.15%
Quadruped	24.55%	9.64%	3.56%	4.91%	1.43%	3.61%	5.03%	24.55%	26.20%	5.70%	26.50%	27.40%
Reporters	76.91%	43.83%	1.33%	43.46%	29.35%	1.31%	31.39%	76.91%	80.55%	49.66%	80.97%	81.82%
Running	24.57%	3.50%	1.41%	6.62%	1.52%	7.31%	7.69%	24.57%	26.58%	7.39%	27.11%	29.32%
Scene_Text	22.27%	6.09%	5.87%	7.65%	4.94%	7.87%	10.74%	22.27%	24.03%	8.13%	24.17%	24.74%
Singing	17.75%	5.92%	1.13%	7.13%	2.65%	0.62%	7.29%	17.75%	19.93%	7.23%	20.13%	21.00%
Sitting_Down	0.87%	0.32%	0.27%	0.33%	0.33%	0.80%	0.42%	0.87%	0.87%	0.80%	0.87%	0.87%
Skating	57.39%	68.35%	0.95%	7.09%	40.28%	0.18%	39.83%	68.35%	75.02%	43.03%	75.02%	75.02%
Sky	45.62%	12.05%	10.92%	32.63%	8.72%	3.73%	35.90%	45.62%	46.63%	33.40%	48.43%	50.41%
Speaking	24.57%	11.63%	4.48%	12.22%	10.48%	2.58%	13.45%	24.57%	27.47%	12.75%	28.51%	29.85%
Speaking_To_Camera	40.13%	15.75%	3.47%	17.87%	13.14%	4.06%	15.01%	40.13%	43.99%	20.19%	46.02%	46.75%
Sports	39.96%	38.49%	5.28%	7.47%	13.49%	6.72%	9.68%	39.96%	49.36%	14.19%	49.69%	50.56%
Streets	27.64%	5.35%	4.53%	5.10%	4.09%	5.21%	7.99%	27.64%	28.71%	6.04%	29.35%	30.10%
Studio_With_Anchorperson	81.93%	51.92%	1.62%	52.68%	31.91%	2.91%	38.11%	81.93%	82.65%	56.92%	83.44%	84.25%
Table	19.20%	3.97%	1.93%	5.91%	1.94%	2.46%	5.18%	19.20%	19.66%	6.25%	19.66%	19.66%
Text	56.89%	13.29%	13.10%	43.37%	9.12%	15.21%	40.18%	56.89%	57.86%	44.69%	57.95%	58.29%
Traffic	35.69%	8.72%	2.62%	5.06%	4.71%	1.82%	6.47%	35.69%	37.92%	5.25%	38.03%	39.20%
Two_People	14.18%	7.24%	1.53%	6.96%	3.26%	1.75%	6.02%	14.18%	14.62%	7.21%	14.91%	15.73%
Urban_Scenes	50.69%	12.77%	5.94%	20.35%	13.56%	6.72%	21.80%	50.69%	52.82%	21.46%	53.34%	54.42%
Walking	28.85%	9.09%	6.04%	8.39%	4.42%	4.31%	9.39%	28.85%	31.97%	8.91%	32.44%	33.17%
Walking_Running	25.06%	7.35%	5.79%	5.28%	4.12%	8.63%	6.39%	25.06%	26.91%	7.96%	27.62%	28.44%

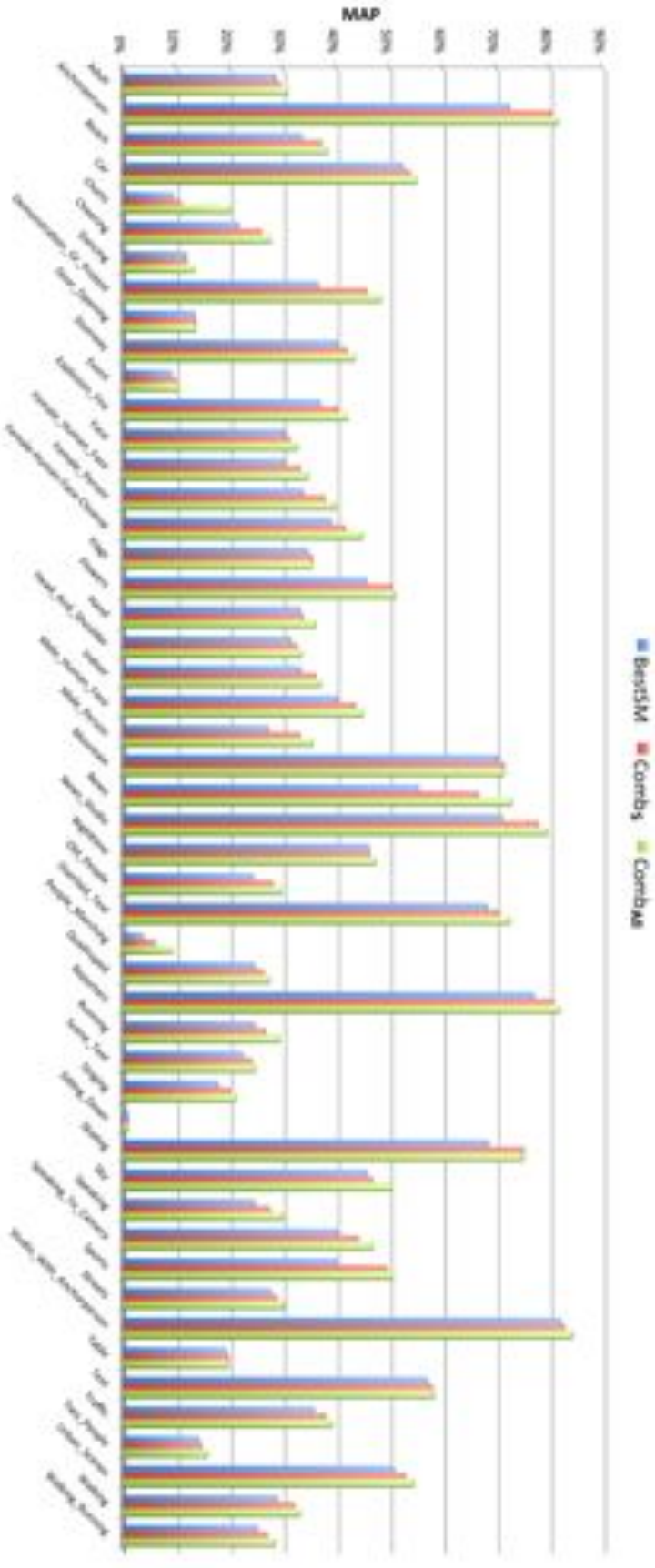


Figure 7.8: AP comparisons for all concepts

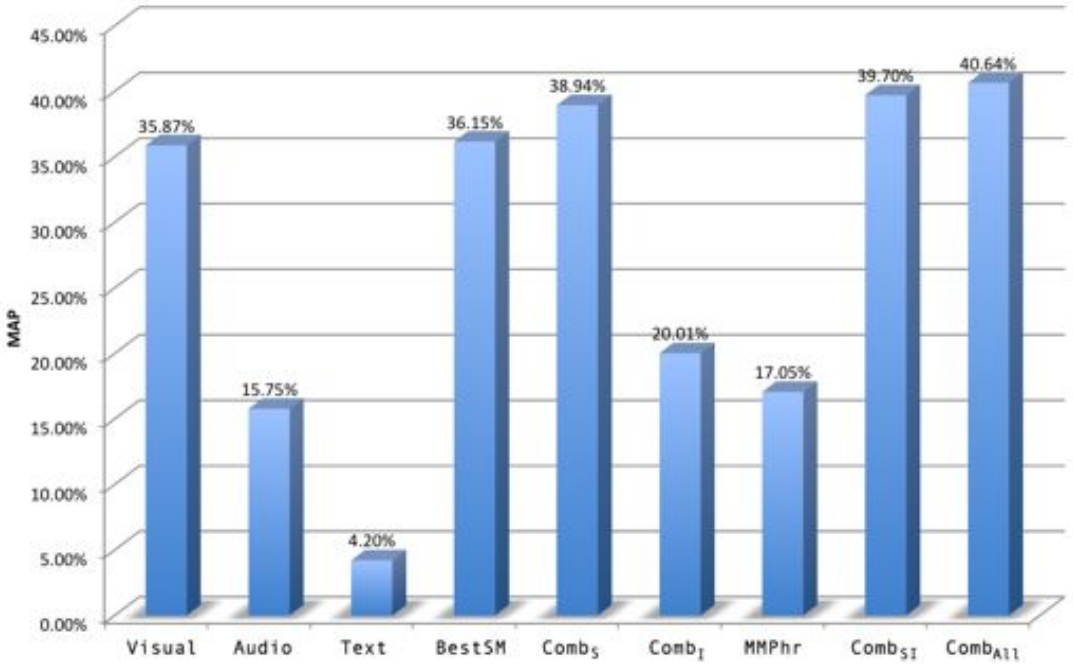


Figure 7.9: MAP comparison for different fusion configurations

Table 7.6: Fusion Gains for combination approaches

	$FG(BestSM)$	$FG(Comb_S)$	$FG(Comb_{SI})$
$Comb_S$	8.92%		
$Comb_{SI}$	13.12%	3.31%	
$Comb_{All}$	17.12%	6.89%	3.24%

- The phrases constructed by the intramodal process (Phr_V, Phr_A and Phr_T) do not provide an adequate level of retrieval accuracy. Moreover, the combination of the intramodal phrase based classifications ($Comb_I$) is not enough to be evaluated as successful. Considering that the phrases include only the information about the co-occurrences of the words in each modality, but not a definitive information about the occurrences of the words, such a result is expected. Thus, the intramodal phrases are not intended to be used for a retrieval operation on their own, but, as described in previous chapters, the actual idea is to use the classification results of intramodal phrases as an input to the final late fusion process. Here, one may argue that the intramodal phrases obtained from each modality can be early-fused with that modality by using a concatenation

approach. However, using such concatenation causes the number of features used for classification to be doubled in size, which has a similar effect with increasing the vocabulary size. Increasing the number of features or using a large vocabulary may lead to a 'curse of dimensionality' effect and over-fitting to the training samples. Hence, the solution may be less generalizable and less tolerant to noises. In addition, performing a classification with a double sized feature vector causes extra processing. Consequently, we do not include the given idea and the problems into the scope of this study, and consider that the idea is straight-forward to apply but time-consuming, leave it as a future work.

- The multimodal phrases constructed by the intermodal process ($MMPhr$) is also not adequate for a retrieval operation on their own. The issues discussed for intramodal phrases is also valid for the multimodal phrases.

- In order to evaluate the effectiveness of intramodal and intermodal phrases, we can compare the accuracies of $MMPhr$ with Phr_V, Phr_A and Phr_T. Considering the presented experimental results, following observations are done.

  - for 31 concepts $MMPhr$ performs superior than the Phr_V,
  - for 39 concepts $MMPhr$ performs superior than the Phr_A,
  - for 48 concepts $MMPhr$ performs superior than the Phr_T,
  - for 20 concepts $MMPhr$ performs superior than the combination of Phr_V, Phr_V and Phr_V, which is $Comb_I$.

  Thus, based on the observations of our experiments, we can state that the combination of the intramodal phrase based classifications is better than multimodal phrases although using multimodal phrases provides better retrieval than the phrases of any single modality. The MAP values for these configurations also support such an evaluation. However, this results may be dependent on the characteristics of the utilized concepts due to the multimodal capability of each concept, so this conclusion should be evaluated in the scope of utilized concepts.

- It may be also attractive to understand if there exists a correlation between the retrieval performances of phrases and multimodal phrases. However, it is hard to infer such a correlation based on the available experimental results. As

summarized in the previous evaluation, the performance of multimodal phrasing may be better than all the phrasing results for some of the concepts. For some others, multimodal phrasing may be worse than all phrasing results, or better than audio and text but worse than visual phrasing. In short, current evidences show that there is not a correlation between the performances of phrases and multimodal phrases, and thus a successful phrasing cannot guarantee a successful multimodal phrasing, or vice versa. The reason for this situation is also the same with the previous evaluation; the accuracies of phrasing and multimodal phrasing is dependent on the nature of the concepts.

- The late fusion of all modalities and the intramodal phrases of all modalities ($Comb_{SI}$) provides a fusion gain of 13.12% on the best single modality ($BestSM$), and a gain of 3.31% on the combination of all modalities ($Comb_S$). In addition, $Comb_{SI}$ performs better retrieval than $Comb_S$ for 41 of the available concepts, whereas for 9 concepts, they are equal in accuracy. These observations show that including the intramodal correlation information into the fusion process improves the retrieval performance. Hence, intramodal correlations within the modalities are important sources of information. Although it is hard to concretize which parts of the concepts are correlated, the experimental results prove that they are correlated, and such information helps to increase the retrieval accuracy.

- The late fusion of all components ($Comb_{All}$) provides a fusion gain of 17.12% on $BestSM$, a gain of 6.89% on $Comb_S$, and a gain of 3.24% on $CombSI$. In addition, $Comb_{All}$ performs better retrieval than $Comb_S$ for 44 of the available concepts, whereas for 6 concepts, they are equal in accuracy. Considering that $Comb_{All}$ includes intermodal correlation information on top of the configuration in $Comb_{SI}$, these observations show that using the intermodal correlation information improves the retrieval performance. Thus, we can argue that the correlations between different modalities is also an important source of information.

- In the above given two evaluations, it is seen that $Comb_{SI}$ never performs worse than $Comb_S$ and also $Comb_{All}$ does not perform worse than $Comb_{SI}$ for any concepts. Such a situation is the result of using RELIEF-MM for

feature weighting, instead of a simple averaging. RELIEF-MM provides a weighting scheme for maximizing the retrieval performance, and eliminating the unsuccessful components by assigning zero weight.

In order to make a clearer visualization on the fusion gains of each component, the fusion gains for all concepts are illustrated in a stacked bar chart, in Figure 7.10. The information sources in our experiments are the modalities, the intramodal analyses of all modalities and the intramodal analysis. Thus, the components for fusion gain comparison are these three information sources ($Comb_S$, $Comb_{SI}$, $Comb_{All}$). The given chart helps to perceive the contribution of each component. Based on Figure 7.10, the following evaluations can be done:

- It is observed that the contribution of modality fusion is more than the intramodal and intermodal analysis for 39 concepts, where the intramodal analysis has more contribution in 7 concepts and the intermodal analysis has more effect in 10 concepts. For 5 concepts, the only contribution comes from the modality fusion.
- If the intramodal and intermodal analysis is compared, the intermodal analysis has more contribution in 35 concepts, whereas the intramodal analysis has more effect for 10 concepts. For 5 concepts, they are equal.
- It is clear that the modality data is an indispensable source of information. Yet, the correlation data provided by the intramodal and intermodal analysis has an important contribution on the fusion result. For one third of the concepts, only correlation data has more contribution than the modality fusion. Moreover, including the phrases and multimodal phrases into final late fusion step increases the fusion gain from 8.92% to 17.12%.
- The contribution of the intermodal process is more than the intramodal process. This is somehow expected, since phrases of each modality is extracted from that single modality. Having the same source of information may limit the contribution. On the other hand, the multimodal phrases of the intermodal process is extracted by using the multiple modalities and captures the correlation between different modalities, which is not an available information in any single modality.

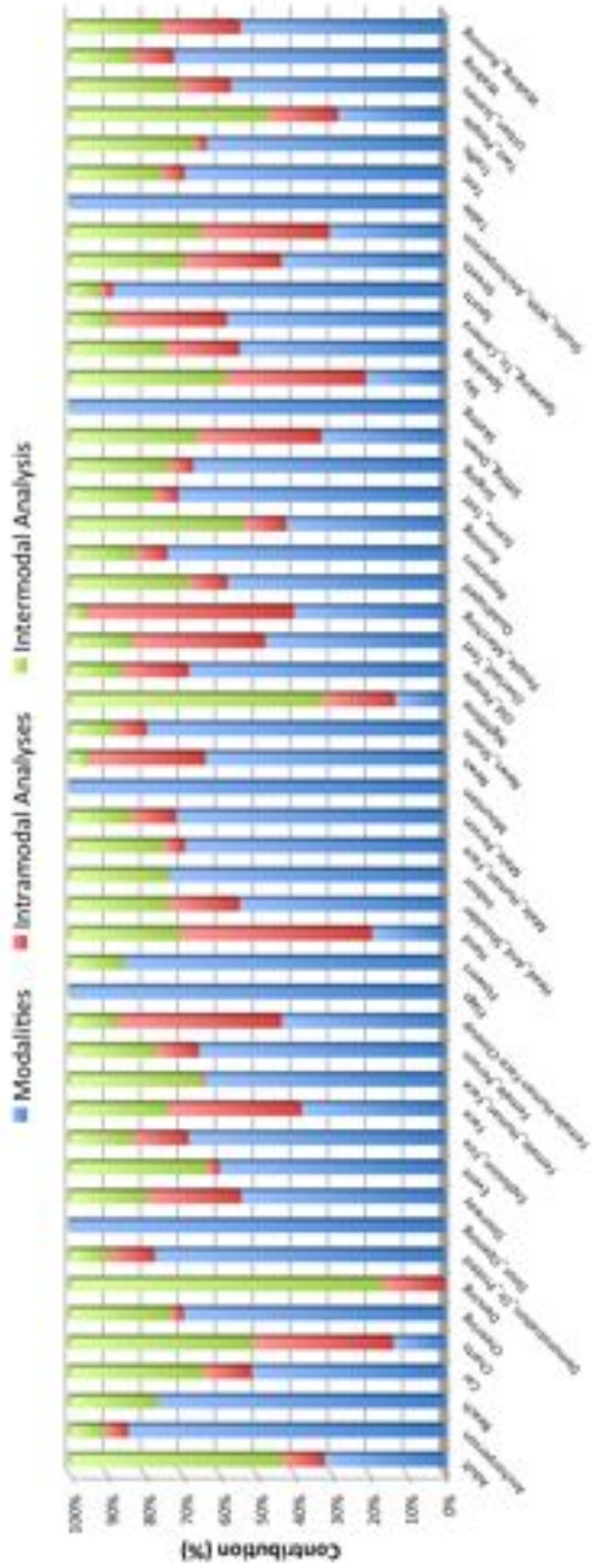


Figure 7.10: Fusion Contribution Analysis

In addition to the above given discussions, the following general evaluations can also be done:

- The proposed approach provides an effective way for fusing BoW-based feature vectors of different modalities. The approach put the intramodal and intermodal correlation analysis forward, however it is actually based on a late fusion of several information sources. Yet, the correlation analysis part enhances the fusion process, and doubles the fusion gain.

- It has been stated that including intramodal and intermodal correlations into the fusion process enhances the fusion performance. The correlation based phrasing and multimodal phrasing procedures can be accepted as a type of early fusion since the inputs are processed before the fusion process and combined in a systematical way. Considering that we apply a late fusion step to combine all modalities and the intramodal / intermodal phrases, the whole process is a multi-level fusion approach including both the early and late fusion steps.

- The approach is based on the extraction of the intramodal and intermodal correlations, and provide promising results. Thus, it can be argued that the correlated information, or in other words the redundant information, has a potential to increase the performance.

7.6 Evaluation of Fusion System Design

Considering the general fusion framework proposed in Section 3.1, an evaluation of the fusion architecture described in this chapter is given below. The approach is based on a 'multi-modal, multi-classifier' fusion scenario and focuses on the 'How to Fuse' problem. Below, how each affecting factor is handled through the proposed solution is described.

- Fusion Setting: The approach combines multiple modalities, each of the modality being a different feature. Before combination, the data of each modality is classified with a separate classifier, and the results of the classifiers are combined.

- Selection of Sources: The RELIEF-MM algorithm proposed in the previous chapter is utilized for the late fusion part of the proposed fusion mechanism.
- Fusion Strategy: The approach focuses on the use of both the complementary and correlated information.
- Content Representation: A feature-based representation is preferred. Considering that the proposed solution enables combining BoWs, the representation can be also be accepted as BoW-based.
- Normalization of Sources: A normalization process is not applied on the fusion inputs.
- Fusion Level: The approach is a multi-level fusion approach, where both an early and a late fusion approaches are included.
- Fusion Methodology: In this study, a new fusion methodology is proposed. The approach is a novel mining and graph based solution. The approach first performs the intramodal and intermodal correlation analyses on the modalities. Then, the results of the correlation analyses and the modalities are combined with a linear weighted averaging approach.
- Operation Modes: The mode for operation is a parallel scheme.
- Synchronization: Since BoW features are extracted in a shot-based manner, the synchronization is shot-based.
- Adaptation: The weighing part of the solution is adaptive, since RELIEF-MM is utilized. However, early fusion part is not adaptive.

7.7 Remarks

In this chapter, the 'how to fuse' problem of information fusion is studied, and a novel mining and graph based combination method for combining the Bags of Words (BoW) obtained from different modalities. The preference of combining BoWs is based on learning schemes used by the state-of-the-art studies. The approach is basically based on the late fusion of modalities, yet powered by the use of the

intramodal and intermodal correlations. Considering the fact that most of the studies do not use the intramodal and intermodal relations of the available words in the BoW model effectively, the proposed approach combines the classification outputs of each single modality, intramodal relations within each modality and intermodal relations between modalities. In addition, the approach is presented in a way that any type of input can be converted into the BoW format, and included in the proposed fusion approach. Consequently, the approach fulfills three important needs: (i) using correlated (redundant) information for fusion, (ii) a novel approach compatible with the contemporary learning methods, (iii) combining multi-level inputs.

Empirical study conducted on TRECVID 2011 dataset with visual, audio and text modalities validate the usability and effectiveness of the proposed approach. The use of the intramodal and intermodal relations improves the retrieval performance and enhances the fusion gain.

The experiments carried out exhibit several ideas for future work. In order to further improve the proposed approach, we put forward the following ideas for future study:

- As discussed before, the phrases obtained from intramodal analyses are accepted as a new source of information and included in the fusion process with a late fusion approach. However, it may be reasonable to early-fuse the intramodal phrases obtained from each modality with that modality by using a concatenation approach. Despite the risk of 'curse of dimensionality' and extra processing cost, the idea may provide an improvement on the retrieval performance.

- The procedure for applying frequent itemset mining (i.e. FP-Growth) requires to convert BoW-based feature vectors of a training sample into transactions for mining, where the transactions are composed of binary values stating occurrence of each word in the corresponding sample. Although we prefer a TF-based weighting for the words in our BoW models, converting them to binary values for mining may cause an information loss. Yet, it is not possible to utilize FP-Growth algorithm with real values. Thus, alternative approaches can be designed to perform the mining procedure with the real word values instead of converting them into binary values.

- In the scope of phrasing algorithms, a pruning step is required while selecting the nearest neighbors. Instead, a traditional feature selection can be applied before construction of the transaction list, accepting the risk of losing some information. Such an action may facilitate the procedures during / after graph construction, but causes extra processing before mining.
- Implementation of a multi-level input schema, including local features, global features and classification results is also left as a future work item.

CHAPTER 8

A DEMO APPLICATION FOR MULTIMODAL INFORMATION RETRIEVAL

In this chapter, a demo application for multimedia information retrieval is presented. The application is based on the studies presented in Section 6.4 and helps to visualize how a multimodal fusion enhances the retrieval performance.

8.1 Brief Description

This thesis study advocates that fusing multimodal information in multimedia data improves the retrieval performance. Though this assertion, new approaches for fusion are proposed and extensive experimental analyses are conducted. Although an experimental study is enough to prove the benefits of fusion, having an application that demonstrates the comparison between the retrieval performances of single modalities and multimodal fusion makes the idea clearer and more visually comprehensible.

For this purpose, we have prepared a demo application for multimodal information retrieval. The demo application enables;

- querying and retrieval of previously trained concepts,
- visualizing the top retrieval results with marks for true/false retrieval,
- comparison of retrieval results for single modalities and multimodal fusion,
- watching retrieved videos.

8.1.1 The Dataset & Modalities

The demo is based on the Columbia Consumer Video (CCV) Database [55] dataset. The dataset is composed of 9,317 YouTube videos, which takes approximately 210 hours in total, and equally partitioned into training and test sets.

The videos in the dataset is manually labeled into 20 target concept classes, as presented in Figure 8.1. The trained concepts are; Basketball, Baseball, Soccer, Ice Skating, Skiing, Swimming, Biking, Cat, Dog, Bird, Graduation, Birthday, Wedding Reception, Wedding Ceremony, Wedding Dance, Music Performance, Non Music Performance, Parade, Beach, Playground.

In the demo application, three different modalities are visualized: visual, audio and motion. For these modalities, SIFT, MFCC and STIP features are used from the video dataset, respectively. The features are provided by the CCV dataset.

Figure 8.1: Samples for each concept in CCV dataset

8.1.2 Multimodal Fusion Approach

The demo application visualizes the benefits of using multimodal fusion by comparing the retrieval results of multimodal fusion with the results of single modalities. The approach used for multimodal fusion is a late fusion approach, specifically Linear Weighted Averaging. As mentioned before, in this approach, the data for each modality is classified separately and then the results are combined by a weighted averaging. The weights are determined with the RELIEF-MM approach, which is presented in Chapter 6.

8.1.3 Implementation Details

The demo application is actually a user interface for visualizing the previously indexed data. Although retrieval operation is an online process, it should be backed by an off-line indexing process, which is usually provided by some learning procedures. For our demo application, the test videos in the dataset are first classified into the target concepts and the classification results are stored in particular index files. The classification results and index data are based on the results of experimental studies presented in Section 6.4. When a query is performed, the queried concept is found in the previously stored index files and the results are visualized by showing the keyframes of each resulting video, decision score and accuracy of the resulting video (true/false). The application supports the visualization the results for each modality separately, and also the multimodal fusion result.

Considering that the dataset is an online resource (YouTube), the demo application is implemented to be an online application. Thus, the implementation of the demo application is performed by using HTML, CSS, and Javascript with the help of Bootstrap framework [89], and served online[1]. The application is lightweight, and fully integrated with YouTube. No images or videos are stored; all of them are accessed via YouTube on runtime.

[1] The demo application can be accessible online via http://www.turgayyilmaz.net/mvs

8.2 The Demo Application

A web-based graphical user interface (GUI), which is served online, is prepared for all operations mentioned above. A basic scenario for using the application is described below, with sample screenshots.

Figure 8.2 presents the home page of the application. In this page, brief information on the dataset, modalities used, query concepts and the multimodal fusion approach are given, as well as the search input form.

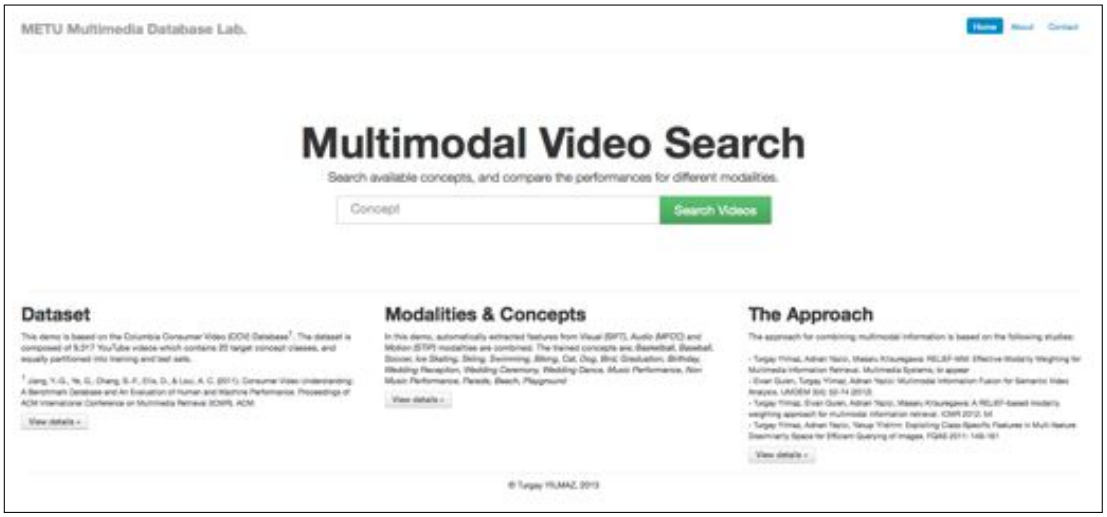


Figure 8.2: Homepage for the demo application

In order perform a query, the query concept is written into the search input area. When a few letters are entered, a pop up list for the alternative query concepts is shown (Figure 8.3). After selecting a query concept, the "Search Videos" button is pressed and the search operation is started.

In Figure 8.4, the query result page is given. In the result page, the retrieval results are presented in four columns. The first column presents the retrieval results for using only visual modality, the second is for audio modality, the third is for motion modality, and finally the last column show the results for the multimodal (combined) retrieval. Each column includes small boxes, each of which represents a video object and contains a frame from the video. Each box is colored with green or red, green representing a correct retrieval, and red means a false retrieval. This accuracy information is

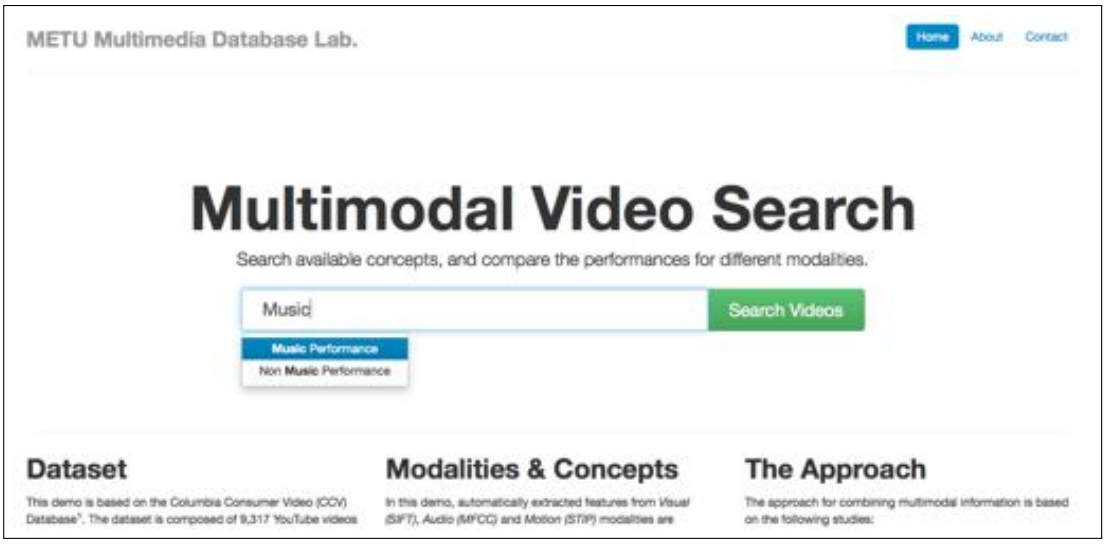


Figure 8.3: Performing a query

based on the CCV dataset ground truths. In addition to the accuracy information, the probabilities of being a queried concept type is also given in the left bottom parts of each box.

The result page lists four videos in a line, for each modality type, as default (Figure 8.5). However, it can be customized to show two, three or six videos, in order to see more or less retrieval results in a single view. In Figure 8.4, query results for the *dog* concept is given in a 3-in-a-line format. In Figure 8.5, the results of the *baseball* query is presented as four videos in each line. In Figure 8.6, the results for query *cat* is given with a 6-in-a-line format.

After obtaining a general view on the query results, each of the retrieved videos can be watched by clicking on the video frames. When clicked, the video is loaded from YouTube and can be played (Figure 8.7).

8.3 Evaluation

An evaluation on the implemented demo application is presented below.

- The implemented demo application presents the retrieval results of single modalities and multimodal fusion side by side on the same view. Such view makes

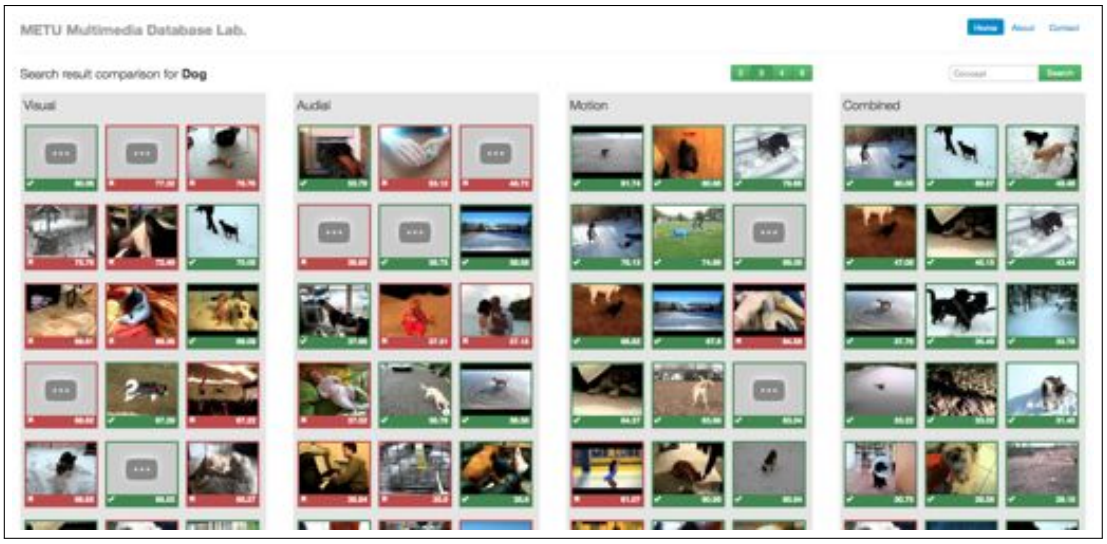


Figure 8.4: Retrieval results page for *dog* query concept

Figure 8.5: Retrieval results page for *baseball* query concept

it easy to compare the retrieval performances. Hence, the advantages of fusion becomes more clear and visually comprehensible. For example, Figure 8.5 shows the retrieval results for *baseball* query concept. In the top-24 results, visual modality returns 19 correct results, audio modality gives 4 correct results and motion modality returns 10 correct results. Besides, multimodal fusion returns 24 correct results.

- In retrieval operations, average precision (AP) or mean average precision (MAP)

Figure 8.6: Retrieval results page for *cat* query concept

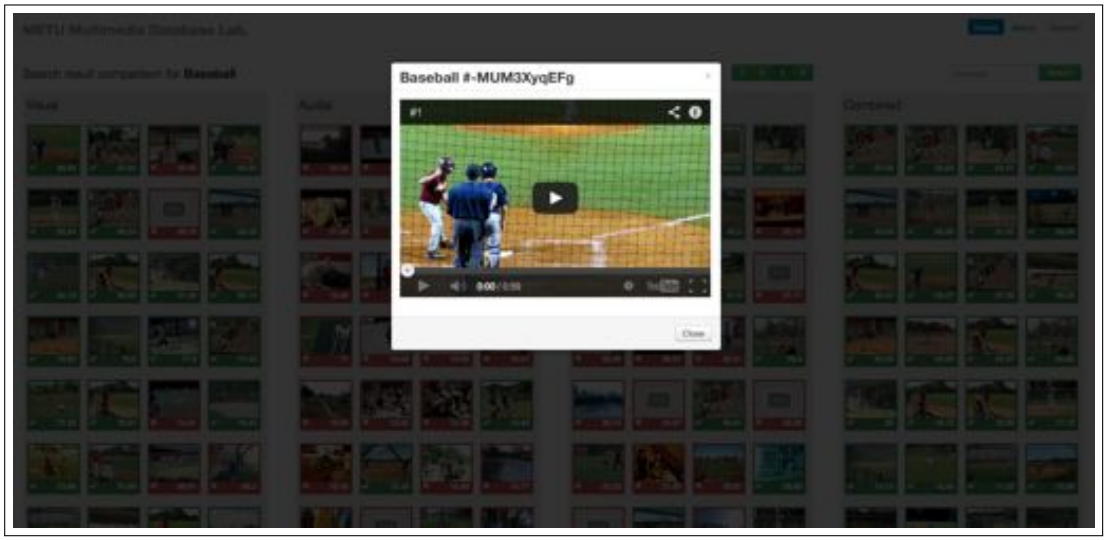


Figure 8.7: Watching a retrieved video

metrics are more important than precision and recall metrics. The user interface of the demo application shows top-K retrieval results with an information about the accuracy (green for correct, red for false retrievals), which helps to understand which method is superior in terms of average precision.

- The demo applications supports querying only the trained 20 concepts. However, the application can be extended by increasing training dataset size and concepts. Also, any keyword search can be handled by including a text retrieval mechanism

and mapping the keywords to trained concepts. Yet, increasing the number of concepts trained depends on the dataset characteristics. For example, CCV dataset is annotated for 20 concepts, whereas TRECVID dataset is annotated for 500 concepts.

CHAPTER 9

CONCLUSION

The increasing usage of digital capture devices has shown the need for multimedia retrieval systems. In order to provide more relevant results, searching multimedia data requires content analysis in the data. Although the content of the multimedia data can be modeled with some low level features, the variety of the features and how they are utilized have an importance on the performance of the constructed retrieval system. The use of just one low-level feature is unsatisfactory, because of the fact that multimedia data usually has a complex structure containing multimodal information in it. The information contained by different modalities usually complements each other, and thus, fusing multimodal information in multimedia data improves the retrieval performance. This thesis has been concerned with an investigation of the problems of multimodal information fusion in the context of multimedia information retrieval.

The studies conducted throughout the thesis include the construction of a general fusion framework by performing a literature survey, and the following two core challenges of information fusion in the context of multimedia information retrieval: (i) How to determine the best modalities? (ii) How best to fuse them? For the former problem, firstly, a class-specific feature/modality selection approach is proposed. Then, this approach is extended into RELIEF-based feature/modality weighting algorithm. For the latter problem, a novel mining and graph based combination approach, which enables an effective combination of the modalities represented with Bag-Of-Words models, is proposed. In addition, a non-linear weighted averaging approach, which attacks the both problems together, is proposed.

The thesis first focuses on identifying the design aspects of a general information

fusion system and make a contribution for the construction of a general fusion framework. To this end, a literature survey on information fusion is conducted, and a general framework, which helps to represent a big picture for designing information fusion systems, is proposed. The proposed general framework describes the affecting variables of fusion systems, helps to design a fusion system by presenting the alternative approaches and pros/cons, and simplifies the construction and evaluation of new fusion systems.

After defining a general framework for fusion, as the first step, two major problems of fusion are considered together, and an efficient fusion architecture is investigated by regarding the most frequently utilized approaches in the literature. Thus, an ANP-based non-linear weighted averaging method, which is a non-linear extension on the linear weighted fusion, is proposed. The linear weighted fusion suffers from the performance upper-bound of linearity and dependency on the selection of weights. The method extends linear weighted fusion with two crucial ideas; interdependency between classes and dependency of classes on the features. The approach is tested on the Columbia Consumer Video Database by using multimodal features of SIFT, MFCC and STIP. The results demonstrate that the proposed non-linear weighting approach is superior than linear weighting, and is less-dependent on the selection of weights. Hence, we argue that the proposed non-linear weighted averaging approach is a sound alternative for linear weighting.

As the second step, the study is focused on the problem of finding the best features / modalities for fusion. For this purpose, a class-specific feature selection (CSF) approach for the fusion of multiple features is proposed. In order to eliminate the high-dimensionality of multiple features and provide efficient querying over the multimedia documents, a dissimilarity based approach is utilized. The class-specific features are determined by using the representativeness and discriminativeness of features for each class / concept. The calculations of representativeness and discriminativeness are based on the statistics on the dissimilarity values of training data. In order to evaluate the proposed approach, experiments with multi-feature and multimodal settings are conducted. In addition, utilization of the approach in a Wireless Video Sensor Networks application is performed. The multi-feature experiments are performed by using the CalTech 101 dataset with 8 MPEG-7 visual features. For the multimodal

experiments, TRECVID 2007 dataset is used with 3 visual, 2 audio and 1 textual modalities. For all experiments, the retrieval performance of the proposed approach is compared with the performances of single features, simple combination approaches and exhaustive search approach. The results obtained from these tests show that the proposed class-specific feature selection approach is an effective and efficient feature selection method.

Considering that the proposed CSF approach is a promising idea and the problem of modality weighting / selection is a major one, a more sophisticated algorithm, named as the RELIEF-MM algorithm, is also introduced. RELIEF-MM is a RELIEF algorithm extension, and utilizes the ideas proposed for the CSF approach. The original RELIEF-F algorithm is employed for multimodal feature selection on multimedia data, and several weaknesses of the algorithm are identified. Considering these weaknesses, the following issues are focused for extending the original RELIEF-F algorithm: class-specific representation, multi-labeled data, noisy data, unbalanced datasets, using classifier predictions instead of feature values for weighting. The proposed approach is extensively tested on TRECVID 2007, TRECVID 2008 and CCV datasets with several modalities in a multimodal information fusion scenario. In these tests, RELIEF-MM has achieved higher accuracies than any single modality, and showed much better performance than simple averaging and RELIEF-F based methods. In addition, RELIEF-MM has provided slightly better performance than the class-common exhaustive-search based approach, although it is computationally much more efficient than class-common exhaustive-search. In addition, several comparative tests are performed against the RELIEF-F approach, aiming to examine each extension idea, and it is confirmed that the proposed extensions lead to improvements on RELIEF-F. Consequently, we argue that our proposed approach is a timely efficient, accurate and robust way of modality selection.

As the third step, the problem of finding a way to fuse the modalities effectively is taken into consideration. For this purpose, the most popular and effective methods in multimedia analysis studies in the last decade are considered. Considering that these studies are usually based on the use of local parts / features in multimedia documents and employing one of the Bag-of-Words (BoW) approaches, the last part of the thesis is focused on combining the BoWs obtained from different modalities. Hence, a novel

mining and graph based combination approach, which exploits the intramodal and intermodal relations, is proposed. The intramodal process extracts *phrase*s from *word*s of each modality by using the correlation between the words within each modality. Besides, the intermodal process extracts *multimodal phrase*s by using the correlation between phrases of different modalities. The proposed approach is tested on TRECVID 2011 dataset by using visual, audio and text modalities. The results show that the use of the intramodal and intermodal analysis doubles the fusion gain, and thus improves the retrieval performance.

In addition to the proposed algorithms, an application is implemented to demonstrate the comparison between the retrieval performances of single modalities and multimodal fusion. The demo application makes the fusion idea more clear and visually comprehensible.

The studies conducted throughout this thesis study have shown that multimodal fusion is a beneficial approach for improving the retrieval performance for multimedia information retrieval. It has been experienced that different modalities abstracts videos from a different aspect, thus different modalities in multimedia data complement each other. Although fusion is usually beneficial, an inadequate configuration for fusion may lead to inefficiencies. Thus the components to fuse, and the fusion approach should be selected carefully. The general fusion framework presented in this thesis helps to evaluate alternative techniques and construct a new fusion system. In addition, the proposed algorithms for modality weighting, non-linear weighted averaging and combining BoW-based features are promising approaches for performing successful fusion operations.

One important issue to mention is the applicability of the proposed approaches in different domains and research areas. In this study, the domain focused for information fusion is multimedia retrieval. The proposed approaches are designed by considering the needs for multimedia data, and the experimental study has been conducted only on the multimedia data. However, the approaches are beneficial for other research areas utilizing fusion for improving performance, including pattern recognition, other information retrieval systems, geospatial information systems, cheminformatics, bioinformatics, wireless sensor networks, biometrics systems. For instance, the non-linear

weighted averaging approach (Chapter 4), focuses on a simple alternative approach for linear weighting, which is not limited to the linear boundaries and less-dependent on the selection of weights. Such an improvement will be useful for any study, regardless of the domain and the research area, which performs a late fusion approach and combines the results of several classifiers. Similarly, Chapter 6 proposed a modality weighting approach which handles multi-label, noisy and unbalanced data issues, as well as the use class-specific feature selection. These issues are not only the problems for multimedia data, but for any other domain or research area, which may face similar problems depending on the dataset. Any retrieval or classification study utilizing a fusion method can benefit from the CSF (Chapter 5) and RELIEF-MM Chapter 6 algorithms for feature weighting / selection. Not only the fusion based studies, but many other studies can also benefit from the ideas applied for the proposed algorithms. For instance, in pattern recognition, feature selection is an important research topic, the CSF and RELIEF-MM algorithms will be useful for feature selection before performing a classification. Applying the proposed approaches for other domains and research areas, and further modifications and improvements are left as future work. In addition to such a utilization study, some other future work items are discussed below.

9.1 Future Work

Considering that potential future work of the studies in each chapter are presented specifically within each chapter, here a list of general future work items are given. Actually, the variables depicted in the general framework given in Section 3.1 and the open issues given in Section 3.2 can be used to point out the future work for information fusion. Below, some future work items are listed:

- The fusion studies in the literature usually describe the fusion method in their context of application. Although we provide a general framework in this thesis to identify all affecting variables, a theoretical background as well as the theoretical performance boundaries and experimental evaluation of such theoretical boundaries are still missing. However, determining general theoretical performance boundary may not be feasible due to the variety of fusion approaches and high dependency of the performance on the fusion inputs.

- Using multimodal information requires the synchronization of modalities according to each other as mentioned in Section 3.1.9. In addition, the occurrences of concepts in different modalities does not necessarily overlap in terms of timing, and some delay between the modalities may be required. However, this is an issue that has not yet been explored exhaustively.

- In this thesis we provide an important contribution to the development of an effective modality weighting for fusion. Although the algorithm is an online procedure and support adaptiveness, no further analysis has been performed on the adaptivity issue. Potential improvements and an adaptive way of determining best sources can be accepted as a future study.

- The available fusion studies usually assume that all of the sources are ready at the fusion time and fused at once, in a parallel operation mode. However, the serial and hybrid architectures are not studied adequately. We think that a serial operation mode has a potential to improve the efficiency of the algorithms and decrease the required execution time for fusion.

- We show that correlated information has a potential to improve the fusion gain. However, most of the recent studies focus on the complementary components for improving accuracy. Thus, alternative approaches for analyzing correlations in dataset can be considered as an early fusion approach, which can also be evaluated as a promising future work.

REFERENCES

[1] Ian F. Akyildiz, Tommaso Melodia, and Kaushik R. Chowdhury. A survey on wireless multimedia sensor networks. *Computer Networks*, 51(4):921–960, March 2007.

[2] Miguel Arevalillo-Herráez, Juan Domingo, and Francesc J. Ferri. Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, 29:2174–2181, Dec 2008.

[3] Javed A. Aslam, Virgil Pavlu, and Emine Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 541–548, New York, NY, USA, 2006. ACM.

[4] Pradeep Atrey, M. Hossain, Abdulmotaleb El Saddik, and Mohan Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379, 2010. 10.1007/s00530-010-0182-0.

[5] Pradeep K. Atrey, Mohan S. Kankanhalli, and John B. Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications and Applications*, 3, February 2007.

[6] Noboru Babaguchi, Yoshihiko Kawai, and Tadahiro Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, 2002.

[7] Nicolas Ballas, Benjamin Labbe, Aymen Shabou, Herve Le Borgne, Miriam Redi Philippe Gosselin, Bernard Merialdo, Herve Jegou, Jonathan Delhumeau, Remi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stephane Ayache, Abdelkader Haadi, Bahjat Safadi, Franck Thollard, Nadia Derbas, Georges Quenot, Herve Bredin, Matthieu Cord, Boyang Gao, Chao Zhu, Yuxing tang, Emmanuel Dellandreav, Liming Chen Charles-Edmond Bichot, Alexandre Benot, Patrick Lambert, Tiberius Strat, Joseph Razik, Sebastion Paris, Herve Glotin, Tran Ngoc Trung, Dijana Petrovska, Geerard Chollet, Andrei Stoian, and Michel Crucianu. IRIM at TRECVID 2012: Semantic Indexing and Instance Search. In *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, 2012.

[8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understing*, 110(3):346–359, June 2008.

[9] Rachid Benmokhtar and Benoit Huet. Classifier fusion: Combination methods for semantic indexing in video content. In Stefanos Kollias, Andreas Stafylopatis, Wlodzislaw Duch, and Erkki Oja, editors, *Artificial Neural Networks - ICANN 2006*, volume 4132 of *Lecture Notes in Computer Science*, pages 65–74. Springer Berlin Heidelberg, 2006.

[10] Chidansh Amitkumar Bhatt and Mohan S. Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011.

[11] Chidansh Amitkumar Bhatt and Mohan S. Kankanhalli. Probabilistic temporal multimedia data mining. *ACM Transactions on Intelligent Systems*, 2(2):17, 2011.

[12] B.Mathieu, S.Essid, T.Fillon, J.Prado, and G.Richard. Yaafe, an easy to use and efficient audio feature extraction software, 2010. Proceedings of the 11th ISMIR conference, Utrecht, Netherlands.

[13] Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 1–5, New York, NY, USA, 2005. ACM. Software available at `http://www.borgelt.net/fpgrowth.html`.

[14] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, April 2008.

[15] Eric Bruno and Stephane Marchand-Maillet. Multimodal preference aggregation for multimedia information retrieval. *Journal of Multimedia*, 4(5):321–329, 2009.

[16] Eric Bruno, Nicolas Moënne-Loccoz, and Stéphane Marchand-Maillet. Design of multimodal dissimilarity spaces for retrieval of multimedia documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1520–1533, 2008.

[17] J. Callan, F. Crestani, H. Nottelmann, P. Pala, and X. M. Shou. Resource selection and data fusion in multimedia distributed digital libraries (poster). In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[18] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Exploration Newsletter*, 6(1):1–6, June 2004.

[20] Haifeng Chen and P. Meer. Robust fusion of uncertain information. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(3):578 –586, 2005.

[21] Min Chen, Shu-Ching Chen, and Mei-Ling Shyu. Hierarchical temporal association mining for video event detection in video databases. In *IEEE 23rd International Conference on Data Engineering Workshop, 2007*, pages 137–145, April 2007.

[22] Yan-Ying Chen, W.H. Hsu, and H.-Y.M. Liao. Automatic training image acquisition and effective feature selection from community-contributed photos for facial attribute detection. *IEEE Transactions on Multimedia*, 15(6):1388–1399, 2013.

[23] Yan-Ying Chen, Winston H. Hsu, and Hong-Yuan Mark Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 669–678, New York, NY, USA, 2012. ACM.

[24] C.C. Chibelushi, F. Deravi, and J.S.D. Mason. Adaptive classifier integration for robust pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(6):902 –907, December 1999.

[25] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, volume 1, pages 886–893 vol. 1, June 2005.

[26] Ritendra Datta, Jia Li, and James Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proceedings of the 7th ACM SIGMM workshop on Multimedia Information Retrieval*, pages 253–262, New York, NY, USA, 2005. ACM Press.

[27] Ritendra Datta, Jia Li, and James Z. Wang. Content-based image retrieval: Approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '05, pages 253–262, New York, NY, USA, 2005. ACM.

[28] Thomas G. Dietterich. Machine-learning research: Four current directions. *The AI Magazine*, 18(4):97–136, 1998.

[29] Gauthier Doquire and Michel Verleysen. Feature selection for multi-label classification problems. In *Proceedings of the 11th international conference on Artificial neural networks conference on Advances in computational intelligence - Volume Part I*, IWANN'11, pages 9–16, Berlin, Heidelberg, 2011. Springer-Verlag.

[30] Robert P. W. Duin. The combining classifier: To train or not to train? In *Proceedings of the 16th International Conference on Pattern Recognition, 2002*, volume 2, pages 765–770, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[31] Robert P. W. Duin and David M. J. Tax. Experiments with classifier combining rules. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 16–29, London, UK, 2000. Springer-Verlag.

[32] Robert P.W. Duin, Marco Loog, Elżbieta P ekalska, and David M.J. Tax. Feature-based dissimilarity space classification. In Devrim Ünay, Zehra Çataltepe, and Selim Aksoy, editors, *Recognizing Patterns in Signals, Speech, Images and Videos*, volume 6388 of *Lecture Notes in Computer Science*, pages 46–55. Springer Berlin Heidelberg, 2010.

[33] Nastaran Fatemi, Florian Poulin, Laura Elena Raileanu, and Alan F. Smeaton. Using association rule mining to enrich semantic concepts for video retrieval. In Ana L. N. Fred, editor, *KDIR 2009, Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 119–126. INSTICC Press, 2009.

[34] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594, 2006.

[35] Basura Fernando, Élisa Fromont, and Tinne Tuytelaars. Effective use of frequent itemset mining for image classification. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV 2012 - Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Part I*, volume 7572 of *Lecture Notes in Computer Science*, pages 214–227. Springer, 2012.

[36] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler. Comparative study of techniques for large-scale feature selection, 1994. Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems.

[37] Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27(6):942–956, June 2005.

[38] Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W.M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, December 2001.

[39] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal Machine Learning Research*, 3:1157–1182, March 2003.

[40] Mark A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, April 1999.

[41] S. Hengstler, D. Prashanth, Sufen Fong, and H. Aghajan. Mesheye: A hybrid-resolution smart camera mote for applications in distributed intelligent surveillance. In *6th International Symposium on Information Processing in Sensor Networks, 2007. IPSN 2007.*, pages 360–369, April 2007.

[42] Kuan-Chieh Huang, Hsueh-Yi Sean Lin, Jyh-Chian Chan, and Yau-Hwang Kuo. Learning collaborative decision-making parameters for multimodal emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME), 2013*, pages 1–6, 2013.

[43] Earl B. Hunt, Philip J. Stone, and Janet. Marin. *Experiments In Induction*. Academic Press, New York :, 1966.

[44] Nakamasa Inoue, Yusuke Kamishima, Toshiya Wada, Koichi Shinoda, and Shunsuke Sato. Tokyotech+canon at trecvid 2011. In *NIST TRECVID Workshop*, Gaithersburg, MD, December 2011.

[45] A.K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Proceedings of the International Conference on Image Processing 2002*, 2002.

[46] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270 – 2285, 2005.

[47] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.

[48] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.

[49] Tao Jiang and Ah-Hwee Tan. Learning image-text associations. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):161–177, 2009.

[50] Yu-Gang Jiang. Super: Towards real-time event recognition in internet videos. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 7:1–7:8, New York, NY, USA, 2012. ACM.

[51] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, pages 1–29, 2012.

[52] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 494–501, New York, NY, USA, 2007. ACM.

[53] Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University ADVENT #223-2008-1, August 2008.

[54] Yu-Gang Jiang, J. Yang, Chong-Wah Ngo, and AG. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, Jan 2010.

[55] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 29:1–29:8, New York, NY, USA, 2011. ACM.

[56] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In Paul Over, George Awad, Jonathan G. Fiscus, Brian Antonishek, Martial Michel, Wessel Kraaij, Alan F. Smeaton, and Georges Quénot, editors, *TRECVID*. National Institute of Standards and Technology (NIST), 2010.

[57] Ilias Kalamaras, Athanasios Mademlis, Sotiris Malassiotis, and Dimitrios Tzovaras. A novel framework for retrieval and interactive visualization of multimodal data. *Electronic Letters on Computer Vision and Image Analysis*, 12(2), 2013.

[58] M.S. Kankanhalli, Jun Wang, and R. Jain. Experiential sampling on multiple data streams. *IEEE Transactions on Multimedia*, 8(5):947 –955, 2006.

[59] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the 9th International Workshop on Machine Learning*, ML '92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[60] J. Kittler. Feature Set Search Algorithms. *Pattern Recognition and Signal Processing*, pages 41–60, 1978.

[61] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, March 1998.

[62] Jana Kludas. Multimedia retrieval and classification for web content. In *Proceedings of BCS IRSG Symposium: Future Directions in Information Access (FDIA)*, 2007.

[63] Jana Kludas, Eric Bruno, and Stéphane Marchand-Maillet. Information fusion in multimedia information retrieval. In *Proceedings of 5th International Workshop on Adaptive Multimedia Retrieval (AMR)*, Paris, France, July 5-6 2007.

[64] Jana Kludas, Eric Bruno, and Stéphane Marchand-Maillet. Can feature information interaction help for information fusion in multimedia problems? *Multimedia Tools and Applications*, 42(1):57–71, 2009.

[65] Deguang Kong, C. Ding, Heng Huang, and Haifeng Zhao. Multi-label relieff and f-statistic feature selections for image annotation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2352 –2359, june 2012.

[66] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *Proceedings of the European Conference on Machine Learning*, pages 171–182, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc.

[67] L.I. Kuncheva. Switching between selection and fusion in combining classifiers: an experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 32(2):146 –156, April 2002.

[68] L.I. Kuncheva. "fuzzy" versus "nonfuzzy" in combining classifiers designed by boosting. *IEEE Transactions on Fuzzy Systems*, 11(6):729 – 741, 2003.

[69] Ludmila Kuncheva, James C. Bezdek, and Robert P. W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001.

[70] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[71] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *10th IEEE International Conference on Computer Vision, 2005. ICCV 2005*, volume 1, pages 832–838 Vol. 1, Oct 2005.

[72] K. LeBlanc and A. Saffiotti. Multirobot object localization: A fuzzy fusion approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(5):1259 –1276, 2009.

[73] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.

[74] Martin E Liggins, David L Hall, and James Llinas. *Handbook of Multisensor Data Fusion: Theory and Practice; 2nd ed.* Electrical Engineering and Applied Signal Processing Series. Taylor & Francis Ltd, Hoboken, NJ, 2008.

[75] Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[76] Huan Liu, Hiroshi Motoda, and Lei Yu. A selective sampling approach to active feature selection. *Artificial Intelligence*, 159:49–74, November 2004.

[77] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[78] José Martínez. Mpeg-7 overview (version 10). Requirements ISO/IEC JTC1 /SC29 /WG11 N6828, International Organisation For Standardisation, Oct 2003.

[79] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005.

[80] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, October 2004.

[81] C. Moulin, C. Barat, and C. Ducottet. Fusion of tf.idf weighted bag of visual features for image classification. In *2010 International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1 –6, june 2010.

[82] Christophe Moulin, Christine Largeron, Christophe Ducottet, Mathias Gèry, and Cècile Barat. Fisher linear discriminant analysis for text-image combination in multimedia information retrieval. *Pattern Recognition*, 47(1):260 – 269, 2014.

[83] MPEG. Mpeg-7 reference software experimentation model, 2003. http://standards.iso.org/ittf/PubliclyAvailableStandards/c035364_ISO_-IEC_15938 -6(E)_Reference_Software.zip [Online; accessed 01-Feb-2011].

[84] Milind R. Naphade and Thomas S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[85] Pradeep Natarajan, Prem Natarajan, Vasant Manohar, Shuang Wu, Stavros Tsakalidis, Shiv N. Vitaladevuni, Xiaodan Zhuang, Rohit Prasad, Guangnan Ye, Dong Liu, I-Hong Jhuo, Shih-Fu Chang, Hamid Izadinia, Imran

Saleemi, Mubarak Shah, Brandyn White, Tom Yeh, and Larry Davis. Bbn viser trecvid 2011 multimedia event detection system. In *NIST TRECVID Workshop*, Gaithersburg, MD, December 2011.

[86] Giang P. Nguyen, Marcel Worring, and Arnold W. M. Smeulders. Similarity learning via dissimilarity space in cbir. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, pages 107–116, New York, NY, USA, 2006. ACM.

[87] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. Bakir. Weighted substructure mining for image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, pages 1–8, June 2007.

[88] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[89] Mark Otto, Jacob Thornton, and Bootstrap contributors. Bootstrap, 2014. http://getbootstrap.com [Online; accessed 29-August-2014].

[90] Paul Over, George Awad, Jonathan Fiscus, Brian Antonishek, Martial Michel, Alan Smeaton, Wessel Kraaij, and Georges Quénot. Trecvid 2011 - goals, tasks, data, evaluation mechanisms and metrics. In Paul Over, George Awad, Jonathan Fiscus, Brian Antonishek, Martial Michel, Alan Smeaton, Wessel Kraaij, and Georges Quénot, editors, *TRECVID*. National Institute of Standards and Technology (NIST), 2011.

[91] Paul Over, George Awad, Wessel Kraaij, and Alan F. Smeaton. Trecvid 2007–overview. In Paul Over, George Awad, Wessel Kraaij, and Alan F. Smeaton, editors, *TRECVID*. National Institute of Standards and Technology (NIST), 2007.

[92] Paul Over, George Awad, R. Travis Rose, Jonathan G. Fiscus, Wessel Kraaij, and Alan F. Smeaton. Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics. In Paul Over, George Awad, R. Travis Rose, Jonathan G. Fiscus, Wessel Kraaij, and Alan F. Smeaton, editors, *TRECVID*. National Institute of Standards and Technology (NIST), 2008.

[93] Bahadir Ozdemir and Selim Aksoy. Image classification using subgraph histogram representation. In *Proceedings of the 20th International Conference on Pattern Recognition, 2010*, ICPR '10, pages 1112–1115, Washington, DC, USA, 2010. IEEE Computer Society.

[94] H. Oztarak, T. Yilmaz, K. Akkaya, and A. Yazici. Efficient and accurate object classification in wireless multimedia sensor networks. In *21st International Conference on Computer Communications and Networks (ICCCN), 2012*, pages 1–7, 2012.

[95] Hakan Oztarak, Kemal Akkaya, and Adnan Yazici. Providing automated actions in wireless multimedia sensor networks via active rules. In Erol Gelenbe, Ricardo Lent, and Georgia Sakellari, editors, *Computer and Information Sciences II*, pages 185–190. Springer London, 2012. 10.1007/978-1-4471-2155-8_23.

[96] D. Parikh and R. Polikar. An ensemble-based incremental learning approach to data fusion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(2):437 –450, 2007.

[97] Elzbieta Pekalska, Pavel Paclík, and Robert P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[98] N. Poh and J. Kittler. *Multimodal Information Fusion: Theory and Applications for Human-Computer Interaction*, chapter 8, pages 153–169. Academic Press, 2010.

[99] Till Quack, Vittorio Ferrari, Bastian Leibe, and Luc J. Van Gool. Efficient mining of frequent and distinctive feature configurations. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20*, pages 1–8. IEEE, 2007.

[100] Till Quack, Vittorio Ferrari, and Luc Van Gool. Video mining with frequent itemset configurations. In Hari Sundaram, Milind Naphade, JohnR. Smith, and Yong Rui, editors, *Image and Video Retrieval*, volume 4071 of *Lecture Notes in Computer Science*, pages 360–369. Springer Berlin Heidelberg, 2006.

[101] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, March 1986.

[102] M. Rahimi, R. Baer, O. Iroezi, J. Garcia, J. Warrior, and M. Srivastava. Cyclops: in situ image sensing and interpretation in wireless sensor networks. *Proceedings of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys 2005)*, pages 193–204, 2006.

[103] Md Mahmudur Rahman, Daekeun You, MatthewS. Simpson, SameerK. Antani, Dina Demner-Fushman, and GeorgeR. Thoma. Multimodal biomedical image retrieval using hierarchical classification and modality fusion. *International Journal of Multimedia Information Retrieval*, 2(3):159–173, 2013.

[104] Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression. In *Proceedings of the 14th International Conference on Machine Learning*, ICML '97, pages 296–304, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[105] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53:23–69, October 2003.

[106] A. Ross and A. K. Jain. Multimodal biometrics: An overview. In *Proceedings of 12th European Signal Processing Conference (EUSIPCO)*, pages 1221–1224, 2004.

[107] A. Rowe, A. Goode, D. Goel, and I. Nourbakhsh. Cmucam3: an open programmable embedded vision sensor. Technical report, ri-tr-07-13, Carnegie Mellon Robotics Institute, 2007.

[108] T.L. Saaty. How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, 48:9–26, 1990.

[109] T.L. Saaty. *Decision Making with Dependence and Feedback: The Analytic Network Process*. RWS Publications, Pittsburgh, 1996.

[110] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, September 2007.

[111] Yousuf Aboobaker Sait and Balaraman Ravindran. Visual object detection using frequent pattern mining. In Hans W. Guesgen and R. Charles Murray, editors, *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference, May 19-21, 2010, Daytona Beach, Florida*. AAAI Press, 2010.

[112] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *ACM Communications*, 18:613–620, November 1975.

[113] Conrad Sanderson and Kuldip K. Paliwal. On the Use of Speech and Face information for identity Verification. Technical Report, Idiap-RR Idiap-RR-10-2004, IDIAP, 3 2004.

[114] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[115] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, pages 1–8, June 2007.

[116] Marko Robnik Sikonja. Speeding up relief algorithm with k-d trees. In *Proceedings of Electrotehnical and Computer Science Conference*, pages 137–140, 1998.

[117] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–, Washington, DC, USA, 2003. IEEE Computer Society.

[118] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380, 2000.

[119] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machince Intelligence*, 22(12):1349–1380, 2000.

[120] L. Snidaro, Ruixin Niu, G.L. Foresti, and P.K. Varshney. Quality-based fusion of multiple video sensors for video surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(4):1044 –1051, 2007.

[121] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[122] Cees G. M. Snoek, Koen E. A. van de Sande, X. Li, M. Mazloom, Y. G. Jiang, Dennis C. Koelma, and Arnold W. M. Smeulders. The MediaMill TRECVID 2011 semantic video search engine. In *Proceedings of the 9th TRECVID Workshop*, Gaithersburg, USA, November 2011.

[123] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA, 2005. ACM.

[124] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Multi-modal classifier fusion for video shot content retrieval. In *Proceedings of 6th International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.

[125] Rohini K. Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28:49–56, 1995.

[126] Yijun Sun. Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1035–1051, 2007.

[127] Daniel L. Swets and John J. Weng. Shoslif-o: Shoslif for object recognition and image retrieval (phase ii). Technical Report CPS 95-39, Michigan State University, Department of Computer Science, 1995.

[128] David M. J. Tax, Martijn van Breukelen, Robert P. W. Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, 2000.

[129] Andrey Temko, Dusan Macho, and Climent Nadeu. Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. *Pattern Recognition*, 41(5):1814 – 1823, 2008.

[130] Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, May 1999.

[131] B.L. Titzer, D.K. Lee, and J. Palsberg. Avrora: scalable sensor network simulation with precise timing. In *Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on*, pages 477–482, April 2005. Software available at `http://compilers.cs.ucla.edu/avrora/`.

[132] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook, 2nd ed.*, pages 667–685. Springer, 2010.

[133] Kagan Tumer and Joydeep Ghosh. Linear and order statistics combiners for pattern classification. *CoRR Computing Research Repository*, cs.NE/9905012, 1999.

[134] Mutlu Uysal and FatoŞ T. Yarman-Vural. Selection of the best representative feature and membership assignment for content-based fuzzy image database. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*, CIVR'03, pages 141–151, Berlin, Heidelberg, 2003. Springer-Verlag.

[135] K. E A Van de Sande, T. Gevers, and C. G M Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept 2010.

[136] J. van de Weijer, T. Gevers, and AD. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150–156, Jan 2006.

[137] Lipo Wang, Nina Zhou, and Feng Chu. A general wrapper approach to selection of class-dependent features. *IEEE Transactions on Neural Networks*, 19(7):1267–1278, 2008.

[138] Wikipedia. Classification (machine learning) – Wikipedia, the free encyclopedia, 2008.
http://en.wikipedia.org/wiki/Classification_(machine_learning) [Online; accessed 21-December-2013].

[139] Wikipedia. Modality (semiotics) – Wikipedia, the free encyclopedia, 2008.
http://en.wikipedia.org/wiki/Modality_(semiotics) [Online; accessed 21-December-2013].

[140] Wikipedia. Pattern recognition – Wikipedia, the free encyclopedia, 2008.
http://en.wikipedia.org/wiki/Pattern_recognition [Online; accessed 21-December-2013].

[141] Kevin Woods, W. Philip Kegelmeyer Jr., and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.

[142] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and D.D. Feng. Realistic human action recognition with multimodal feature selection and fusion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(4):875–885, 2013.

[143] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 572–579, New York, NY, USA, 2004. ACM.

[144] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418 –435, May/Jun 1992.

[145] Rong Yan and Alexander G. Hauptmann. The combination limit in multimedia retrieval. In *Proceedings of the 11th ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 339–342, New York, NY, USA, 2003. ACM.

[146] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 721–, Washington, DC, USA, 2002. IEEE Computer Society.

[147] Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, D. T. Lee, and Shih-Fu Chang. Joint audio-visual bi-modal codewords for video event detection. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 39:1–39:8, New York, NY, USA, 2012. ACM.

[148] Turgay Yilmaz, Elvan Gulen, Adnan Yazici, and Masaru Kitsuregawa. A relief-based modality weighting approach for multimodal information retrieval. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 54:1–54:8, New York, NY, USA, 2012. ACM.

[149] Turgay Yilmaz, Adnan Yazici, and Masaru Kitsuregawa. Non-linear weighted averaging for multimodal information fusion by employing analytical network process. In *ICPR 2012 - 21st International Conference on Pattern Recognition*, pages 234–237, 2012.

[150] Turgay Yilmaz, Adnan Yazici, and Masaru Kitsuregawa. Relief-mm: effective modality weighting for multimedia information retrieval. *Multimedia Systems*, 20(4):389–413, 2014.

[151] Turgay Yilmaz, Adnan Yazici, and Yakup Yildirim. Exploiting class-specific features in multi-feature dissimilarity space for efficient querying of images. In Henning Christiansen, Guy De Tré, Adnan Yazici, Slawomir Zadrozny, Troels Andreasen, and Henrik Legind Larsen, editors, *Flexible Query Answering Systems - Proceedings of the 9th International Conference, FQAS 2011, Ghent,*

Belgium, October 26-28, 2011, volume 7022 of *Lecture Notes in Computer Science*, pages 149–161. Springer, 2011.

[152] Turgay Yilmaz, Yakup Yildirim, and Adnan Yazici. A genetic algorithms based classifier for object classification in images. In Erol Gelenbe, Ricardo Lent, and Georgia Sakellari, editors, *Computer and Information Sciences II*, pages 519–525. Springer London, 2012. 10.1007/978-1-4471-2155-8_66.

[153] Junsong Yuan, Ying Wu, and Ming Yang. Discovery of collocation patterns: from visual words to visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, pages 1–8, June 2007.

[154] Xingquan Zhu, Xindong Wu, Ahmed K. Elmagarmid, Zhe Feng, and Lide Wu. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):665–677, 2005.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: YILMAZ, Turgay
Date and Place of Birth: 1982, Bursa
Homepage: http://www.turgayyilmaz.net
E-mail: turgay@ceng.metu.edu.tr

EDUCATION

Degree	Institution	Date of Graduation
M.S.	Computer Engineering, METU	Jan 2008
B.S.	Computer Engineering, Bilkent University	May 2004

ACADEMIC EXPERIENCE

Date	Place	Enrollment
Nov 2011 - Jun 2013	Institute of Industrial Science, The Univ. of Tokyo, Tokyo, Japan	Visiting Research Associate
Mar 2008 - Dec 2012	Dept. of Computer Engineering, METU, Ankara	Project Research Assistant

PROFESSIONAL EXPERIENCE

Date	Place	Enrollment
Jun 2013 - Present	TÜRKSAT A.Ş., Ankara	System Architect
Feb 2013 - Jun 2013	Rakuten Inc., Tokyo, Japan	Engineer/Researcher
Sep 2006 - Nov 2011	HAVELSAN A.S., Ankara	Expert Engineer
Feb 2005 - Aug 2006	HAVELSAN A.S., Ankara	Engineer
Jul 2004 - Jan 2005	EES Ltd. Sti., Ankara	Software Engineer

RESEARCH INTERESTS

Multimedia databases, information fusion, multimedia retrieval, intelligent systems, fuzzy systems, machine learning, soft computing

JOURNAL PUBLICATIONS

Turgay Yilmaz, Adnan Yazici, Masaru Kitsuregawa. *RELIEF-MM: Effective Modality Weighting for Multimedia Information Retrieval*. Multimedia Systems, Vol. 20, Issue 4, pp 389-413, July 2014.

Yakup Yildirim, Adnan Yazici, **Turgay Yilmaz**. *Automatic Semantic Content Extraction in Videos using a Fuzzy Ontology and Rule-based Model*. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 47-61, Jan. 2013.

Elvan Gulen, **Turgay Yilmaz**, Adnan Yazici. *Multimodal Information Fusion for Semantic Video Analysis*. International Journal of Multimedia Data Engineering and Management (IJMDEM), 3(4), 52-74, 2012

CONFERENCE PUBLICATIONS

Fatih Senel, Kemal Akkaya, **Turgay Yilmaz**. *Autonomous Deployment of Sensors for Maximized Coverage and Guranteed Connectivity in Underwater Acoustic Sensor Networks*, 2013 IEEE 38th Conference on Local Computer Networks (LCN), 2013, 211-218, Oct 2013.

Elvan Gulen, **Turgay Yilmaz**, Adnan Yazici. *A Multimodal Fusion Approach by Exploiting Concept Interactions for Efficient Multimedia Analysis*. VLDB 2012 Workshop on Multimedia Databases and Data Engineering (MDDE), 2012.

Turgay Yilmaz, Adnan Yazici, Masaru Kitsuregawa. *Non-Linear Weighted Averaging for Multimodal Information Fusion by Employing Analytical Network Process*. 21th International Conference on Pattern Recognition (ICPR), 2012.

Turgay Yilmaz, Elvan Gulen, Adnan Yazici, Masaru Kitsuregawa. *A RELIEF-Based Modality Weighting Approach for Multimodal Information Retrieval*. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12, New York, NY, USA, 2012. ACM.

Hakan Oztarak, **Turgay Yilmaz**, Kemal Akkaya, Adnan Yazici. *Efficient and Accurate Object Classification in Wireless Multimedia Sensor Networks*. In Proceedings of 21st IEEE International Conference on Computer Communications and Networks (ICCCN'12), Munich, Germany, July 2012.

Turgay Yilmaz, Adnan Yazici, Yakup Yildirim. *Exploiting class-specific features in multi-feature dissimilarity space for efficient querying of images*. In Proceedings of the 9th international conference on Flexible Query Answering Systems (FQAS'11), Hen-

ning Christiansen, Guy Tré, Adnan Yazici, Slawomir Zadrozny, and Troels Andreasen (Eds.). Springer-Verlag, Berlin, Heidelberg, 149-161, 2011.

Utku Demir, Murat Koyuncu, Adnan Yazici, **Turgay Yilmaz**, Mustafa Sert. *Flexible content extraction and querying for videos*. In Proceedings of the 9th international conference on Flexible Query Answering Systems (FQAS'11), Henning Christiansen, Guy Tré, Adnan Yazici, Slawomir Zadrozny, and Troels Andreasen (Eds.). Springer-Verlag, Berlin, Heidelberg, 460-471, 2011.

Turgay Yilmaz, Yakup Yildirim, Adnan Yazici. *A Genetic Algorithms Based Classifier For Object Classification In Images*. 26th International Symposium on Computer and Information Sciences, ISCIS 2011, London, UK, September 26-28, 2011, Proceedings. Lecture Notes in Electrical Engineering, Springer, 2011.

Murat Koyuncu, **Turgay Yilmaz**, Yakup Yildirim, Adnan Yazici. *A Framework for Fuzzy Video Content Extraction, Storage and Retrieval*. In FUZZ-IEEE'10: Proceedings of the IEEE International Conference on Fuzzy Systems 2010, Barcelona, SPAIN, 2010.

Yakup Yildirim, **Turgay Yilmaz**, Adnan Yazici. *Ontology-supported object and event extraction with a genetic algorithms approach for object classification*. In CIVR'07: Proceedings of the 6th ACM international conference on Image and video retrieval, pages 202-209, New York, NY, USA, 2007. ACM.