Natural Language Descriptions of Visual Scenes: Corpus Generation and Analysis

Muhammad Usman Ghani Khan **Rao Muhammad Adeel Nawab** **Yoshihiko Gotoh**
University of Sheffield, United Kingdom
{ughani, r.nawab, y.gotoh}@dcs.shef.ac.uk

Abstract

As video contents continue to expand, it is increasingly important to properly annotate videos for effective search, mining and retrieval purposes. While the idea of annotating images with keywords is relatively well explored, work is still needed for annotating videos with natural language to improve the quality of video search. The focus of this work is to present a video dataset with natural language descriptions which is a step ahead of keywords based tagging. We describe our initial experiences with a corpus consisting of descriptions for video segments crafted from TREC video data. Analysis of the descriptions created by 13 annotators presents insights into humans' interests and thoughts on videos. Such resource can also be used to evaluate automatic natural language generation systems for video.

1 Introduction

This paper presents our experiences in manually constructing a corpus, consisting of natural language descriptions of video segments crafted from a small subset of TREC video[1] data. In a broad sense the task can be considered one form of machine translation as it translates video streams into textual descriptions. To date the number of studies in this field is relatively small partially because of lack of appropriate dataset for such task. Another obstacle may be inherently larger variation for descriptions that can be produced for videos than a conventional translation from one language to another. Indeed humans are very subjective while annotating video streams, *e.g.*, two humans may produce quite different descriptions for the same video. Based on these descriptions we are interested to identify the most important and frequent high level features (HLFs); they may be 'keywords', such as a particular object and its position/moves, used for a semantic indexing task in video retrieval. Mostly HLFs are related to humans, objects, their moves and properties (*e.g.*, gender, emotion and action) (Smeaton et al., 2009).

In this paper we present these HLFs in the form of ontologies and provides two hierarchical structures of important concepts — one most relevant for humans and their actions, and another for non human objects. The similarity of video descriptions is quantified using a bag of word model. The notion of sequence of events in a video was quantified using the order preserving sequence alignment algorithm (longest common subsequence). This corpus may also be used for evaluation of automatic natural language description systems.

1.1 Background

The TREC video evaluation consists of on-going series of annual workshops focusing on a list of information retrieval (IR) tasks. The TREC video promotes research activities by providing a large test collection, uniform scoring procedures, and a forum for research teams interested in presenting their results. The high level feature extraction task aims to identify presence or absence of high level semantic features in a given video sequence (Smeaton et al., 2009). Approaches to video summarisation have been explored using rushes video[2] (Over et al., 2007).

[1] www-nlpir.nist.gov/projects/trecvid/

[2] Rushes are the unedited video footage, sometimes referred to as a pre-production video.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 38–47,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

TREC video also provides a variety of meta data annotations for video datasets. For the HLF task, speech recognition transcripts, a list of master shot references, and shot IDs having HLFs in them are provided. Annotations are created for shots (*i.e.*, one camera take) for the summarisation task. Multiple humans performing multiple actions in different backgrounds can be shown in one shot. Annotations typically consist of a few phrases with several words per phrase. Human related features (*e.g.*, their presence, gender, age, action) are often described. Additionally, camera motion and camera angle, ethnicity information and human's dressing are often stated. On the other hand, details relating to events and objects are usually missing. Human emotion is another missing information in many of such annotations.

2 Corpus Creation

We are exploring approaches to natural language descriptions of video data. The step one of the study is to create a dataset that can be used for development and evaluation. Textual annotations are manually generated in three different flavours, *i.e.*, selection of HLFs (keywords), title assignment (a single phrase) and full description (multiple phrases). Keywords are useful for identification of objects and actions in videos. A title, in a sense, is a summary in the most compact form; it captures the most important content, or the theme, of the video in a short phrase. On the other hand, a full description is lengthy, comprising of several sentences with details of objects, activities and their interactions. Combination of keywords, a title, and a full descriptions will create a valuable resource for text based video retrieval and summarisation tasks. Finally, analysis of this dataset provides an insight into how humans generate natural language description for video.

Most of previous datasets are related to specific tasks; PETS (Young and Ferryman, 2005), CAVIAR (Fisher et al., 2005) and Terrascope (Jaynes et al., 2005) are for surveillance videos. KTH (Schuldt et al., 2004) and the Hollywood action dataset (Marszalek et al., 2009) are for human action recognition. MIT car dataset is for identification of cars (Papageorgiou and Poggio, 1999). Caltech 101 and Caltech 256 are image datasets with 101 and 256 object categories respectively (Griffin et al., 2007) but there is no information about human actions or emotions. There are some datasets specially generated for scene settings such as MIT outdoor scene dataset (Oliva and Torralba, 2009). Quattoni and Torralba (2009) created indoor dataset with 67 different scenes categories. For most of these datasets annotations are available in the form of keywords (*e.g.*, actions such as sit, stand, walk). They were developed for keyword search, object recognition or event identification tasks. Rashtchian et al. (2010) provided an interesting dataset of 1000 images which contain natural language descriptions of those images.

In this study we select video clips from TREC video benchmark for creating annotations. They include categories such as news, meeting, crowd, grouping, indoor/outdoor scene settings, traffic, costume, documentary, identity, music, sports and animals videos. The most important and probably the most frequent content in these videos appears to be a human (or humans), showing their activities, emotions and interactions with other objects. We do not intend to derive a dataset with a full scope of video categories, which is beyond our work. Instead, to keep the task manageable, we aim to create a compact dataset that can be used for developing approaches to translating video contents to natural language description.

Annotations were manually created for a small subset of data prepared form the rushes video summarisation task and the HLF extraction task for the 2007 and 2008 TREC video evaluations. It consisted of 140 segments of videos — 20 segments for each of the following seven categories:

Action videos: Human posture is visible and human can be seen performing some action such as 'sitting', 'standing', 'walking' and 'running'.

Close-up: Human face is visible. Facial expressions and emotions usually define mood of the video (*e.g.*, happy, sad).

News: Presence of an anchor or reporters. Characterised by scene settings such as weather boards at the background.

Meeting: Multiple humans are sitting and communicating. Presence of objects such as chairs and a table.

Grouping: Multiple humans interaction scenes that do not belong to a meeting scenario. A

table or chairs may not be present.

Traffic: Presence of vehicles such as cars, buses and trucks. Traffic signals.

Indoor/Outdoor: Scene settings are more obvious than human activities. Examples may be park scenes and office scenes (where computers and files are visible).

Each segment contained a single camera shot, spanning between 10 and 30 seconds in length. Two categories, 'Close-up' and 'Action', are mainly related to humans' activities, expressions and emotions. 'Grouping' and 'Meeting' depict relation and interaction between multiple humans. 'News' videos explain human activities in a constrained environment such as a broadcast studio. Last two categories, 'Indoor/Outdoor' and 'Traffic', are often observed in surveillance videos. They often shows for humans' interaction with other objects in indoor and outdoor settings. TREC video annotated most video segments with a brief description, comprising of multiple phrases and sentences. Further, 13 human subjects prepared additional annotation for these video segments, consisting of keywords, a title and a full description with multiple sentences. They are referred to as **hand annotations** in the rest of this paper.

2.1 Annotation Tool

There exist several freely available video annotation tools. One of the popular video annotation tool is *Simple Video Annotation tool*[3]. It allows to place a simple tag or annotation on a specified part of the screen at a particular time. The approach is similar to the one used by *YouTube*[4]. Another well-known video annotation tool is *Video Annotation Tool*[5]. A video can be scrolled for a certain time period and place annotations for that part of the video. In addition, an annotator can view a video clip, mark a time segment, attach a note to the time segment on a video timeline, or play back the segment. '*Elan*' annotation tool allows to create annotations for both audio and visual data using temporal information (Wittenburg et al., 2006). During that annotation process, a user selects a section of video using the timeline capability and writes annotation for the specific time.

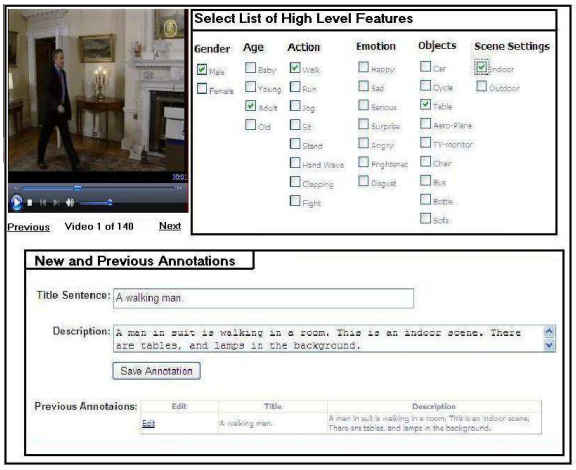

Figure 1: *Video Description Tool* (VDT). An annotator watches one video at one time, selects all HLFs present in the video, describes a theme of the video as a title and creates a full description for important contents in the video.

We have developed our own annotation tool because of a few reasons. None of existing annotation tools provided the functionality of generating a description and/or a title for a video segment. Some tools allows selection of keywords in a free format, which is not suitable for our purpose of creating a list of HLFs. Figure 1 shows a screen shot of the video annotation tool developed, which is referred to as *Video Description Tool* (VDT). VDT is simple to operate and assist annotators in creating quality annotations. There are three main items to be annotated. An annotator is shown one video segment at one time. Firstly a restricted list of HLFs is provided for each segment and an annotator is required to select all HLFs occurring in the segment. Second, a title should be typed in. A title may be a theme of the video, typically a phrase or a sentence with several words. Lastly, a full description of video contents is created, consisting of several phrases and sentences. During the annotation, it is possible to stop, forward, reverse or play again the same video if required. Links are provided for navigation to the next and the previous videos. An annotator can delete or update earlier annotations if required.

[3] videoannotation.codeplex.com/

[4] www.youtube.com/t/annotations_about

[5] dewey.at.northwestern.edu/ppad2/documents/help/video.html

2.2 Annotation Process

A total of 13 annotators were recruited to create texts for the video corpus. They were undergraduate or postgraduate students and fluent in English. It was expected that they could produce descriptions of good quality without detailed instructions or further training. A simple instruction set was given, leaving a wide room for individual interpretation about what might be included in the description. For quality reasons each annotator was given one week to complete the full set of videos.

Each annotator was presented with a complete set of 140 video segments on the annotation tool VDT. For each video annotators were instructed to provide

- a title of one sentence long, indicating the main theme of the video;
- description of four to six sentences, related to what are shown in the video;
- selection of high level features (*e.g.*, male, female, walk, smile, table).

The annotations are made with open vocabulary — that is, they can use any English words as long as they contain only standard (ASCII) characters. They should avoid using any symbols or computer codes. Annotators were further guided not to use proper nouns (*e.g.*, do not state the person name) and information obtained from audio. They were also instructed to select all HLFs appeared in the video.

3 Corpus Analysis

13 annotators created descriptions for 140 videos (seven categories with 20 videos per category), resulting in 1820 documents in the corpus. The total number of words is 30954, hence the average length of one document is 17 words. We counted 1823 unique words and 1643 keywords (nouns and verbs).

Figure 2 shows a video segment for a meeting scene, sampled at 1 fps (frame per second), and three examples for hand annotations. They typically contain two to five phrases or sentences. Most sentences are short, ranging between two to six words. Descriptions for human, gender, emotion and action are commonly observed. Occasionally minor details for objects and events are also stated. Descriptions for the background are often associated with objects rather than humans. It is interesting to observe the subjectivity with the task; the variety of words were selected by individual annotators to express the same video contents. Figure 3 shows another example of a video segment for a human activity and hand annotations.

Hand annotation 1
(title) interview in the studio;
(description) three people are sitting on a red table; a tv presenter is interviewing his guests; he is talking to the guests; he is reading from papers in front of him; they are wearing a formal suit;

Hand annotation 2
(title) tv presenter and guests
(description) there are three persons; the one is host; others are guests; they are all men;

Hand annotation 3
(title) three men are talking
(description) three people are sitting around the table and talking each other;

Figure 2: A montage showing a meeting scene in a news video and three sets of hand annotations. In this video segment, three persons are shown sitting on chairs around a table — extracted from TREC video '*20041116_150100_CCTV4_DAILY_NEWS_CHN33050028*'.

3.1 Human Related Features

After removing function words, the frequency for each word was counted in hand annotations. Two classes are manually defined; one class is related directly to humans, their body structure, identity, action and interaction with other humans. (Another class represents artificial and natural objects and scene settings, *i.e.*, all the words not directly related to humans, although they are important for semantic understanding of the visual scene — described further in the next section.) Note that some related words (*e.g.*, '*woman*' and '*lady*') were replaced with a single concept ('*female*'); concepts were then built up into a hierarchical structure for each class.

Figure 4 presents human related information observed in hand annotations. Annotators paid full attention to human gender information as the number of occurrences for '*female*' and '*male*' is

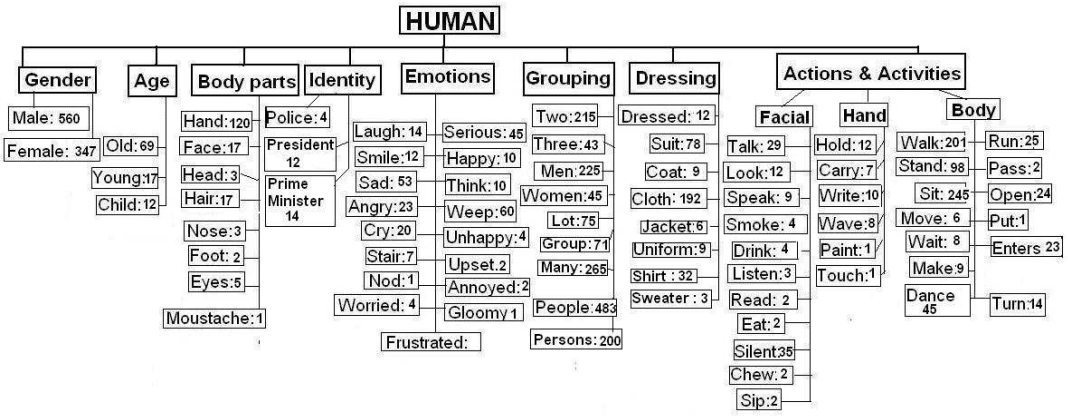

Figure 4: Human related information found in 13 hand annotations. Information is divided into structures (gender, age, identity, emotion, dressing, grouping and body parts) and activities (facial, hand and body). Each box contains a high level concept (*e.g.*, '*woman*' and '*lady*' are both merged into '*female*') and the number of its occurrences.

Hand annotation 1
(title) outdoor talking scene;
(description) young woman is sitting on chair in park and talking to man who is standing next to her;

Hand annotation 2
(title) A couple is talking;
(description) two person are talking; a lady is sitting and a man is standing; a man is wearing a black formal suit; a red bus is moving in the street; people are walking in the street; a yellow taxi is moving in the street;

Hand annotation 3
(title) talk of two persons;
(description) a man is wearing dark clothes; he is standing there; a woman is sitting in front of him; they are saying to each other;

Figure 3: A montage of video showing a human activity in an outdoor scene and three sets of hand annotations. In this video segment, a man is standing while a woman is sitting in outdoor — from TREC video '*20041101_160000_CCTV4_DAILY_NEWS_CHN_41504210*'.

the highest among HLFs. This highlights our conclusion that most interesting and important HLF is humans when they appear in a video. On the other hand age information (*e.g.*, '*old*', '*young*', '*child*') was not identified very often. Names for human body parts have mixed occurrences ranging from high ('*hand*') to low ('*moustache*'). Six basic emotions — anger, disgust, fear, happiness, sadness, and surprise as discussed by Paul Ekman[6] — covered most of facial expressions.

Dressing became an interesting feature when a human was in a unique dress such as a formal suit, a coloured jacket, an army or police uniform. Videos with multiple humans were common, and thus human grouping information was frequently recognised. Human body parts were involved in identification of human activities; they included actions such as standing, sitting, walking, moving, holding and carrying. Actions related to human body and posture were frequently identified. It was rare that unique human identities, such as police, president and prime minister, were described. This may indicate that a viewer might want to know a specific type of an object to describe a particular situation instead of generalised concepts.

3.2 Objects and Scene Settings

Figure 5 shows the hierarchy created for HLFs that did not appear in Figure 4. Most of the words are related to artificial objects. Humans interact with these objects to complete an activity —

[6] en.wikipedia.org/wiki/Paul_Ekman

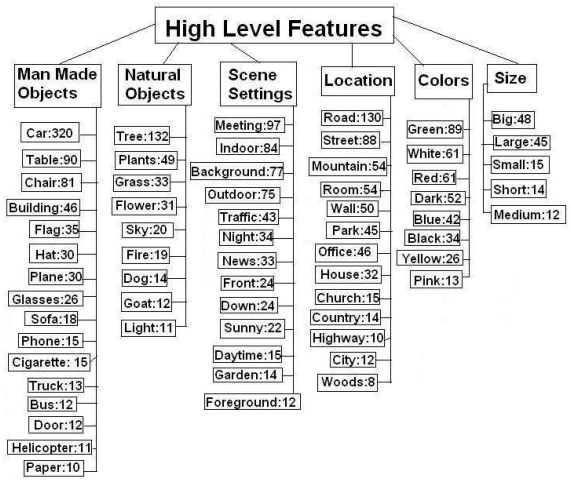

Figure 5: Artificial and natural objects and scene settings were summarised into six groups.

e.g., '*man is sitting on a chair*', '*she is talking on the phone*', '*he is wearing a hat*'. Natural objects were usually in the background, providing the additional context of a visual scene — *e.g.*, '*human is standing in the jungle*, '*sky is clear today*'. Place and location information (*e.g.*, room, office, hospital, cafeteria) were important as they show the position of humans or other objects in the scene — *e.g.*, '*there is a car on the road*, '*people are walking in the park*'.

Colour information often plays an important part in identifying separate HLFs — *e.g.*, '*a man in black shirt is walking with a woman with green jacket*', '*she is wearing a white uniform*'. The large number of occurrences for colours indicates human's interest in observing not only objects but also their colour scheme in a visual scene. Some hand descriptions reflected annotator's interest in scene settings shown in the foreground or in the background. Indoor/outdoor scene settings were also interested in by some annotators. These observations demonstrate that a viewer is interested in high level details of a video and relationships between different prominent objects in a visual scene.

3.3 Spatial Relations

Figure 6 presents a list of the most frequent words and phrases related to spatial relations found in hand annotations. Spatial relations between HLFs are important when explaining the semantics of visual scenes. Their effective use leads to the smooth description. Spatial relations can be categorised into

in (404); with (120); on (329); near (68); around (63); at (55); on the left (35); in front of (24); down (24); together (24); along (16); beside (16); on the right (16); into (14); far (11); between (10); in the middle (10); outside (8); off (8); over (8); pass-by (8); across (7); inside (7); middle (7); under (7); away (6); after (7)

Figure 6: List of frequent spatial relations with their counts found in hand annotations.

static: relations between stationary objects;

dynamic: direction and path of moving objects;

inter-static and dynamic: relations between moving and not moving objects.

Static relations can establish the scene settings (*e.g.*, '*chairs around a table*' may imply an indoor scene). Dynamic relations are used for finding activities present in the video (*e.g.*, '*a man is running with a dog*'). Inter-static and dynamic relations are a mixture of stationary and non stationary objects; they explain semantics of the complete scene (*e.g.*, '*persons are sitting on the chairs around the table*' indicates a meeting scene).

3.4 Temporal Relations

Video is a class of time series data formed with highly complex multi dimensional contents. Let video X be a uniformly sampled frame sequence of length n, denoted by $X = \{x_1, \ldots, x_n\}$, and each frame x_i gives a chronological position of the sequence (Figure 7). To generate full description of video contents, annotators use temporal information to join descriptions of individual frames. For example,

> *A man is walking. After sometime he enters the room. Later on he is sitting on the chair.*

Based on the analysis of the corpus, we describe temporal information in two flavors:

1. temporal information extracted from activities of a single human;

2. interactions between multiple humans.

Most common relations in video sequences are '*before*', '*after*', '*start*' and '*finish*' for single humans, and '*overlap*', '*during*' and '*meeting*' for multiple humans.

Figure 8 presents a list of the most frequent words in the corpus related to temporal relations. It can be observed that annotators put much focus

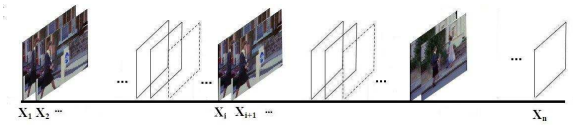

Figure 7: Illustration of a video as a uniformly sampled sequence of length n. A video frame is denoted by x_i, whose spatial context can be represented in the d dimensional feature space.

single human:
then (25); end (24); before (22); after (16); next (12); later on (12); start (11); previous (11); throughout (10); finish (8); afterwards (6); prior to (4); since (4)
multiple humans:
meet (114); while (37); during (27); at the same time (19); overlap (12); meanwhile (12); throughout (7); equals (4)

Figure 8: List of frequent temporal relations with their counts found in hand annotations.

on keywords related to activities of multiple humans as compared to single human cases. '*Meet*' keyword has the highest frequency, as annotators usually consider most of the scenes involving multiple humans as the meeting scene. '*While*' keyword is mostly used for showing separate activities of multiple humans such as '*a man is walking while a woman is sitting*'.

3.5 Similarity between Descriptions

A well-established approach to calculating human inter-annotator agreement is kappa statistics (Eugenio and Glass, 2004). However in the current task it is not possible to compute inter-annotator agreement using this approach because no category was defined for video descriptions. Further the description length for one video can vary among annotators. Alternatively the similarity between natural language descriptions can be calculated; an effective and commonly used measure to find the similarity between a pair of documents is the overlap similarity coefficient (Manning and Schütze, 1999):

$$Sim_{overlap}(X,Y) = \frac{|S(X,n) \cap S(Y,n)|}{\min(|S(X,n)|, |S(Y,n)|)}$$

where $S(X,n)$ and $S(Y,n)$ are the set of distinct n-grams in documents X and Y respectively. It is a similarity measure related to the Jaccard index (Tan et al., 2006). Note that when a set X is a subset of Y or the converse, the overlap coefficient is equal to one. Values for the overlap coefficient range between 0 and 1, where '0' presents the situation where documents are completely different and '1' describes the case where two documents are exactly the same.

Table 1 shows the average overlap similarity scores for seven scene categories within 13 hand annotations. The average was calculated from scores for individual description, that was compared with the rest of descriptions in the same category. The outcome demonstrate the fact that humans have different observations and interests while watching videos. Calculation were repeated with two conditions; one with stop words removed and Porter stemmer (Porter, 1993) applied, but synonyms NOT replaced, and the other with stop words NOT removed, but Porter stemmer applied and synonyms replaced. It was found the latter combination of preprocessing techniques resulted in better scores. Not surprisingly synonym replacement led to increased performance, indicating that humans do express the same concept using different terms.

The average overlap similarity score was higher for 'Traffic' videos than for the rest of categories. Because vehicles were the major entity in 'Traffic' videos, rather than humans and their actions, contributing for annotators to create more uniform descriptions. Scores for some other categories were lower. It probably means that there are more aspects to pay attention when watching videos in, *e.g.*, 'Grouping' category, hence resulting in the wider range of natural language expressions produced.

3.6 Sequence of Events Matching

Video is a class of time series data which can be partitioned into time aligned frames (images). These frames are tied together sequentially and temporally. Therefore, it will be useful to know how a person captures the temporal information present in a video. As the order is preserved in a sequence of events, a suitable measure to quantify sequential and temporal information of a description is the longest common subsequence (LCS). This approach computes the similarity between a pair of token (*i.e.*, word) sequences by simply counting the number of edit operations (insertions and deletions) required to transform one sequence into the other. The output is a sequence of common elements such that no other longer string is

	Action	Close-up	Indoor	Grouping	Meeting	News	Traffic
unigram (A)	0.3827	0.3913	0.4217	0.3809	0.3968	0.4378	0.4687
(B)	0.4135	0.4269	0.4544	0.4067	0.4271	0.4635	0.5174
bigram (A)	0.1483	0.1572	0.1870	0.1605	0.1649	0.1872	0.1765
(B)	0.2490	0.2616	0.2877	0.2619	0.2651	0.2890	0.2825
trigram (A)	0.0136	0.0153	0.0301	0.0227	0.0219	0.0279	0.0261
(B)	0.1138	0.1163	0.1302	0.1229	0.1214	0.1279	0.1298

Table 1: Average overlapping similarity scores within 13 hand annotations. For each of unigram, bigram and trigram, scores are calculated for seven categories in two conditions: (A) stop words removed and Porter stemmer applied, but synonyms NOT replaced; (B) stop words NOT removed, but Porter stemmer applied and synonyms replaced.

	raw	synonym	keyword
Action	0.3782	0.3934	0.3955
Close-up	0.4181	0.4332	0.4257
Indoor	0.4248	0.4386	0.4338
Grouping	0.3941	0.4104	0.3832
Meeting	0.3939	0.4107	0.4124
News	0.4382	0.4587	0.4531
Traffic	0.4036	0.4222	0.4093

Table 2: Similarity scores based on the longest common subsequence (LCS) in three conditions: scores without any preprocessing (raw), scores after synonym replacement (synonym), and scores by keyword comparison (keyword). For keyword comparison, verbs and nouns were presented as keywords after stemming and removing stop words.

available. In the experiments, the LCS score between word sequences is normalised by the length of the shorter sequence.

Table 2 presents results for identifying sequences of events in hand descriptions using the LCS similarity score. Individual descriptions were compared with the rest of descriptions in the same category and the average score was calculated. Relatively low scores in the table indicate the great variation in annotators' attention on the sequence of events, or temporal information, in a video. Events described by one annotator may not have been listed by another annotator. The News videos category resulted in the highest similarity score, confirming the fact that videos in this category are highly structured.

3.7 Video Classification

To demonstrate the application of this corpus with natural language descriptions, a supervised document classification task is outlined. *Tf-idf* score can express textual document features (Dumais et al., 1998). Traditional *tf-idf* represents the relation between term t and document d. It provides a measure of the importance of a term within a particular document, calculated as

$$tfidf(t,d) = tf(t,d) \cdot idf(d) \quad (1)$$

where the term frequency $tf(t,d)$ is given by

$$tf(t,d) = \frac{N_{t,d}}{\sum_k N_{k,d}} \quad (2)$$

In the above equation $N_{t,d}$ is the number of occurrences of term t in document d, and the denominator is the sum of the number of occurrences for all terms in document d, that is, the size of the document $|d|$. Further the inverse document frequency $idf(d)$ is

$$idf(d) = \log \frac{N}{W(t)} \quad (3)$$

where N is the total number of documents in the corpus and $W(t)$ is the total number of document containing term t.

A term-document matrix X is presented by $T \times D$ matrix $tfidf(t,d)$. In the experiment Naive Bayes probabilistic supervised learning algorithm was applied for classification using *Weka* machine learning library (Hall et al., 2009). Ten-fold cross validation was applied. The performance was measured using precision, recall and F1-measure (Table 3). F1-measure was low for 'Grouping' and 'Action' videos, indicating the difficulty in classifying these types of natural language descriptions. Best classification results were achieved for 'Traffic' and 'Indoor/Outdoor' scenes. Absence of humans and their actions might have contributed obtaining the high classification scores. Human actions and activities were present in most videos in various categories, hence the 'Action' category resulted in the lowest results. 'Grouping' category also showed

	precision	recall	F1-measure
Action	0.701	0.417	0.523
Close-up	0.861	0.703	0.774
Grouping	0.453	0.696	0.549
Indoor	0.846	0.915	0.879
Meeting	0.723	0.732	0.727
News	0.679	0.823	0.744
Traffic	0.866	0.869	0.868
average	0.753	0.739	0.736

Table 3: Results for supervised classification using the *tf-idf* features.

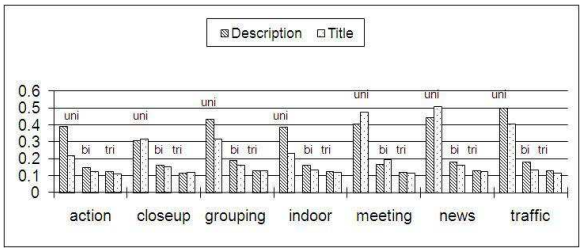

Figure 9: The average overlap similarity scores for titles and for descriptions. 'uni', 'bi', and 'tri' indicate the unigram, bigram, and trigram based similarity scores, respectively. They were calculated without any preprocessing such as stop word removal or synonym replacement.

weaker result; it was probably because processing for interaction between multiple people, with their overlapped actions, had not been fully developed. Overall classification results are encouraging which demonstrates that this dataset is a good resource for evaluating natural language description systems of short videos.

3.8 Analysis of Title and Description

A title may be considered a very short form of summary. We carried out further experiments to calculate the similarity between a title and a description manually created for a video. The length of a title varied between two to five words. Figure 9 shows the average overlapping similarity scores between titles and descriptions. It can be observed that, in general, scores for titles were lower than those for descriptions, apart from 'News' and 'Meeting' videos. It was probably caused by the short length of titles; by inspection we found phrases such as '*news video*' and '*meeting scene*' for these categories.

Another experiment was performed for classification of videos based on title information only. Figure 10 shows comparison of classification per-

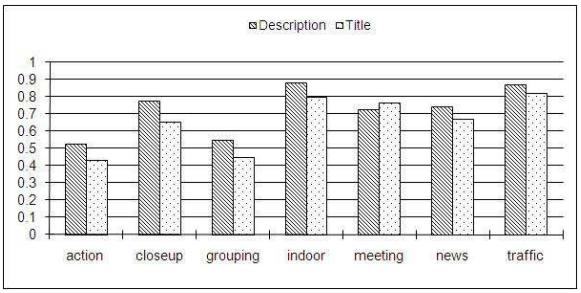

Figure 10: Video classification by titles, and by descriptions.

formance with titles and with descriptions. We were able to make correct classification in many videos with titles alone, although the performance was slightly less for titles only than for descriptions.

4 Conclusion and Future Work

This paper presented our experiments using a corpus created for natural language description of videos. For a small subset of TREC video data in seven categories, annotators produced titles, descriptions and selected high level features. This paper aimed to characterise the corpus based on analysis of hand annotations and a series of experiments for description similarity and video classification. In the future we plan to develop automatic machine annotations for video sequences and compare them against human authored annotations. Further, we aim to annotate this corpus in multiple languages such as Arabic and Urdu to generate a multilingual resource for video processing community.

Acknowledgements

M U G Khan thanks University of Engineering & Technology, Lahore, Pakistan and R M A Nawab thanks COMSATS Institute of Information Technology, Lahore, Pakistan for funding their work under the Faculty Development Program.

References

S. Dumais, J. Platt, D. Heckerman, and M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM.

B.D. Eugenio and M. Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

R. Fisher, J. Santos-Victor, and J. Crowley. 2005. Caviar: Context aware vision using image-based active recognition.

G. Griffin, A. Holub, and P. Perona. 2007. Caltech-256 object category dataset.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

C. Jaynes, A. Kale, N. Sanders, and E. Grossmann. 2005. The terrascope dataset: A scripted multi-camera indoor video surveillance dataset with ground-truth. In *Proceedings of the IEEE Workshop on VS PETS*, volume 4. Citeseer.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

M. Marszalek, I. Laptev, and C. Schmid. 2009. Actions in context.

A. Oliva and A. Torralba. 2009. Mit outdoor scene dataset.

P. Over, A.F. Smeaton, and P. Kelly. 2007. The trecvid 2007 bbc rushes summarization evaluation pilot. In *Proceedings of the international workshop on TRECVID video summarization*, pages 1–15. ACM.

C. Papageorgiou and T. Poggio. 1999. A trainable object detection system: Car detection in static images. Technical Report 1673, October. (CBCL Memo 180).

M.F. Porter. 1993. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

A. Quattoni and A. Torralba. 2009. Recognizing indoor scenes.

C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

C. Schuldt, I. Laptev, and B. Caputo. 2004. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE.

A.F. Smeaton, P. Over, and W. Kraaij. 2009. High-level feature detection from video in trecvid: a 5-year retrospective of achievements. *Multimedia Content Analysis*, pages 1–24.

P.N. Tan, M. Steinbach, V. Kumar, et al. 2006. *Introduction to data mining*. Pearson Addison Wesley Boston.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006. Citeseer.

D.P. Young and J.M. Ferryman. 2005. Pets metrics: On-line performance evaluation service. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 317–324.