The Stanford/Technicolor/Fraunhofer HHI Video Semantic Indexing System

A. F. de Araujo[1], F. Silveira[2], H. Lakshman[3], J. Zepeda[2], A. Sheth[2], P. Pérez[2] and B. Girod[1]
[1]Stanford University
{afaraujo,bgirod}@stanford.edu
[2]Technicolor
{fernando.silveira,joaquin.zepeda,anmol.sheth,patrick.perez}@technicolor.com
[3]Fraunhofer HHI
haricharan.lakshman@hhi.fraunhofer.de

***Abstract*—Video search has become a very important tool, with the ever-growing size of multimedia collections. This work introduces our Video Semantic Indexing system. Our experiments show that Residual Vectors provide an efficient way of aggregating local descriptors, with complementary gain with respect to BoVW. Also, we show that systems using a limited number of descriptors and machine learning techniques can still be quite effective. Our first participation at the TRECVID evaluation has been very fruitful: our team was ranked 6^{th} in the light version of the Semantic Indexing task.**

***Index Terms*—Video semantic indexing, video retrieval, video categorization, video search, concept detection**

I. Introduction

THE steep rise in the availability of video content, during the last decade, is hardly breaking news. Today, YouTube reports that 72 hours of videos are uploaded to its servers every minute [1]. Video is increasingly ubiquitous and video collections are increasingly large, throughout the spectrum ranging from public broadcasters to personal archives. The size of the databases is growing faster than the technology to handle it can be developed. Sooner or later, video retrieval will be a key technology of the new digital world [2].

Most existing commercial video search engines retrieve videos based on textual tags, descriptions and transcripts. This provides limited performance in important cases, such as when the searched visual content is not mentioned in the text, or when the accompanying text is in a different language.

In a study with broadcast news [3], Hauptmann et al. showed that text-based retrieval of videos performed poorly. This was based on 83 use cases developed by the LSCOM effort [4], a collaboration between several organizations that created a taxonomy of 1,000 concepts, along with realistic use cases and queries with a large annotated set of broadcast news videos.

Hauptmann's work [3] further shows that a system using 320 semantic concepts can improve performance significantly, even with quite low concept detection performance. Finally, their work also argues that, in realistic settings, a few thousand concepts would be sufficient to take video retrieval's performance to the level of existing web text search engines.

STRUCTURED ABSTRACT - SIN

- *Keywords for each run*
  L_A_stanford1_1. CENTRIST, SIFT, OppSIFT, dense extraction, HarLap keypoint detector, BoVW, SPM, Residual Vectors, late fusion by linear combination of validation-weighted scores.
  L_A_stanford2_2. CENTRIST, SIFT, OppSIFT, dense extraction, HarLap keypoint detector, BoVW, SPM, Residual Vectors, late fusion by average of scores.
  L_A_stanford3_3. CENTRIST, SIFT, OppSIFT, dense extraction, Residual Vectors, late fusion by linear combination of validation-weighted scores.
  L_A_stanford4_4. CENTRIST, SIFT, dense extraction, BoVW, SPM, Residual Vectors, late fusion by linear combination of validation-weighted scores.
- *Performance comparison among runs*
  The two first runs were meant to be the most competitive ones. They perform quite similarly, being the 19^{th} and 20^{th} top runs. Considering only one run per team, our best run achieves the 6^{th} rank.
  The third run achieves the 43^{rd} rank, and the fourth run achieves the 49^{th} rank. These naturally perform worse, since they utilize fewer elements. They are, however, much more computationally efficient, which demonstrate that, even with a limited set of descriptors/aggregation methods, reasonable Semantic Indexing is still possible.
- *Estimate of relative contribution of each component*
  A system using only the efficient CENTRIST descriptor, and SIFT descriptors, already achieves more than 0.2 MAP (Run 4). If we replace BoVW-aggregated SIFT by Residual Vectors-aggregated OppSIFT, performance improves further (Run 3). If all these elements are combined, performance improves by a significant 25% to our best runs (Runs 1 and 2).
  In our preliminary experiments, we verified that using both BoVW and Residual vectors improved overall system's performance, over the use of only one of them.
- *Learned lessons*
  - Systems using limited number of descriptors and machine learning techniques can still be quite effective;
  - Residual Vectors provide an efficient way of aggregating local descriptors for video indexing;
  - Residual Vectors provide complementary gain to BoVW.

Multimodal video indexing has been researched intensely recently – though mostly in an ad-hoc way, with groups of researchers tackling isolated problems. Snoek et al. [5] provide a coherent review, along with a unifying and multimodal framework containing the different modalities and granularities involved in video indexing.

The same authors provide a comprehensive overview of concept-based video retrieval [2], discussing the challenges not only of detecting the concepts, but also of conceiving a practical system for real users. A key element of such systems is the way users perform a query, which can be done by keyword, by example, or by concept. The system could then have an automatic selection of concepts based on query prediction techniques.

In concept-based indexing, the objective is to conceive detectors that, with a generic framework, can handle hundreds or even thousands of different concepts with reasonable performance. Such a technology should also be able to i) provide access to specific video segments, and not only to the entire video, ii) allow for searching videos in databases which do not have textual tags and iii) enhance video search in the cases text-based search is effective.

As this is our first participation in the TRECVID evaluation, we tackled only the light version of the Video Semantic Indexing task. The task definition is:

"Given the test collection, master shot reference, and concept definitions, return for each concept a list of at most 2000 shot IDs from the test collection ranked according to their likeliness of containing the concept."

In other words, the teams are given a number of shots and semantic concept definitions, and they have to submit a 2000-length ranked list with the most likely shots to contain the searched semantic concept. Training data are also provided, with which the systems are to be trained.

The task is characterized as a large scale one. There are more than 400k training shots (equivalent to 600 hours), and more than 100k testing shots (equivalent to 200 hours). The shot is the unit of analysis, based on which the annotations are provided.

The remaining of the paper is organized as follows. Section II gives an overview of our system's main components. Section III follows with the experimental setup and our results. Section IV concludes with a discussion of our main results.

II. Overview of the system

In this Section, we introduce our Video Semantic Indexing System. Figure 1 presents a high-level overview of the task, with the main system components.

A large set of videos, decomposed into shots, is provided, with partial labels regarding presence or absence of several concepts (more detail in the dataset will be provided in Subsection III-A). With these training data, features are extracted and used to train classifiers. At testing time, the same features are extracted from the testing videos, and the classifier predicts the presence or absence of each concept.

Figure 2 gives a more detailed view of each high-level block from Figure 1. Each color represents the use of a different feature type and classifier - we refer to each of those as 'feature channels'. In our system, the input to the feature channels is a selected representative frame from each shot, as will be detailed in Subsection II-A.

For example, blue might represent Dense SIFT + BoVW + HIK-kernel SVM, and purple might represent Spatial CENTRIST + RBF-kernel SVM. These technical details will be discussed in more detail in the following subsections.

Image-level features can be extracted either in the form of global descriptors, or by aggregating local descriptors. Classifier construction is composed by an initial validation step, which estimates the best classifier parameters, and then by classifier training.

The classifier can then be applied to the representation of the testing instance (the video shot). That gives a score for that particular feature channel, for each of the possible concepts. We then fuse the scores of the different feature channels, to obtain a final score for the specific concept with respect to the specific test instance.

We did not explore the use of label co-occurences to improve the performance over predicting the occurence of each concept individually. This was due to two reasons: i) not much co-occurence training data are available (as will be detailed in III-A) and ii) previous work shows that it takes 25 times longer to integrate classifier training if taking correlations among concepts into account [6].

In the following, we present in more detail the components of the system.

A. Extraction of descriptors

The TRECVID 2012 SIN dataset (presented in detail in subsection III-A) provides, together with the videos, outputs of a speech-to-text engine. For some videos, textual tags and short descriptions are also made available.

In total, four modalities are present and can potentially be used/combined in a Semantic Indexing system: i) visual, ii) audio, iii) textual tags and descriptions and iv) speech-to-text transcriptions.

Our system only makes use of the visual modality, for several reasons:

i) Textual tags and descriptions are very sparse, multilingual and not necessarily consistent among videos presenting similar concepts.

ii) Speech-to-text transcripts are provided for the English language, even if the videos are not in English. Also, even for English-spoken videos, the transcripts are quite unreliable.

iii) Previous editions of the TRECVID workshop show that audio-based descriptors can be useful for a handful of concepts, but that they are not useful for most concepts [7]–[9].

For those reasons, it becomes very difficult to build statistical models based on the textual elements. Due to limitation of time, our team decided to focus only on visual-based descriptors, which have been shown to be the most useful in previous TRECVID editions.

Visual descriptors can be extracted based on a keyframe, or based on a sequence of frames. According to previous reports,

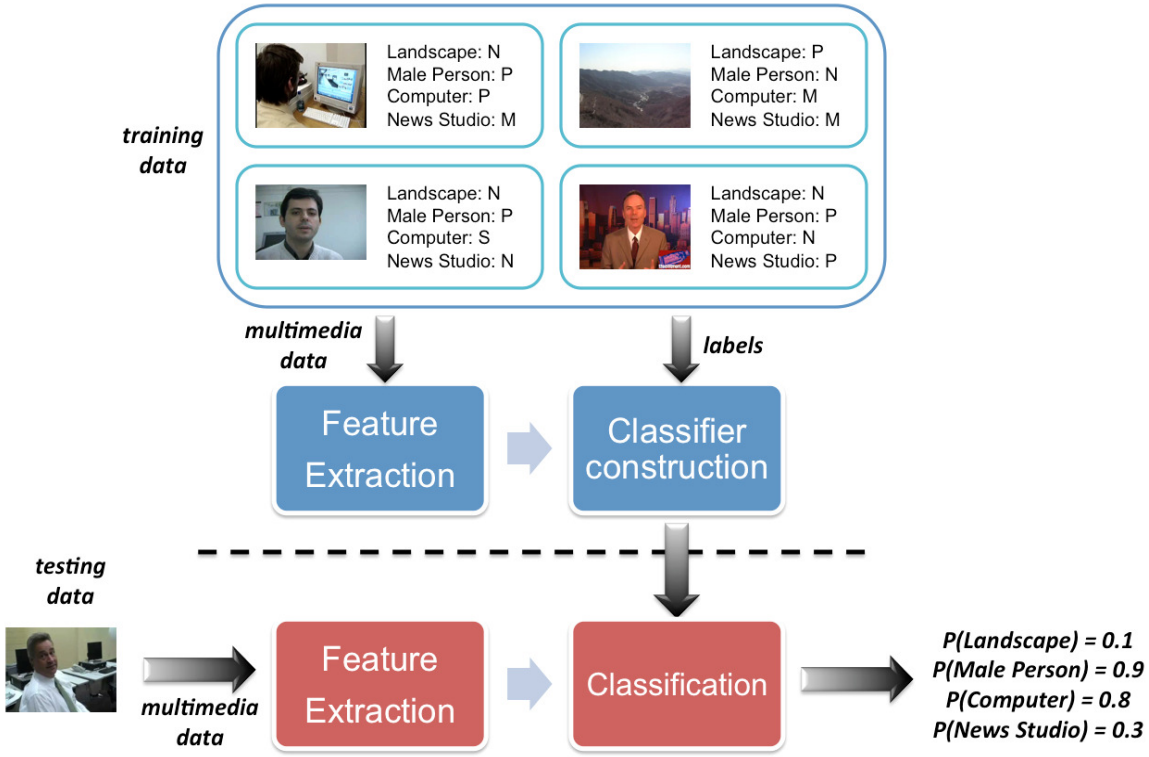

Fig. 1. High-level overview of the system, with example for the concepts 'Landscape', 'Male Person', 'Computer' and 'News Studio'. Each thumbnail represents a video shot. The annotations can be P, when the concept is clearly present, N when the concept is clearly not present, S when the annotator is not certain (skip), or M, when the annotation is missing.

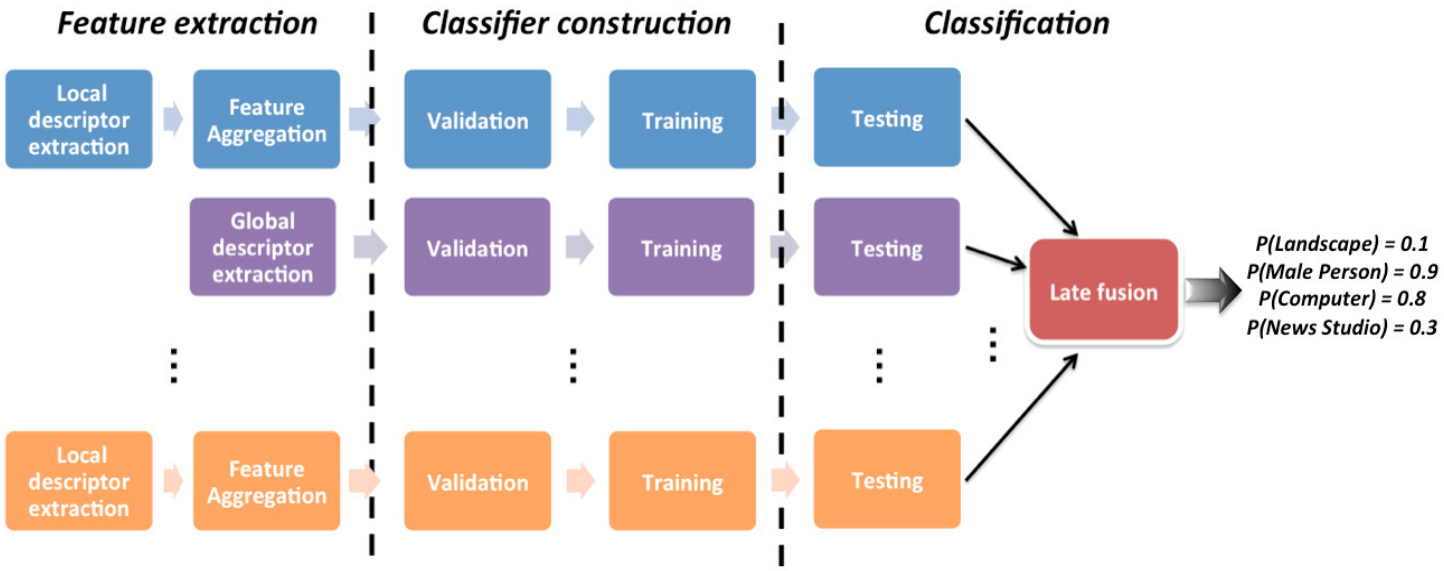

Fig. 2. The blocks from Figure 1 are presented in more detail. Each color represents a feature channel. Features are extracted from the shots by either extracting global descriptors or extracting local descriptors and then aggregating them.

the use of descriptors based on a sequence of frames can improve concept detection. However, keyframe-based descriptors are shown to account for most of the performance [7], [8]. Thus, due to limitation of time, we employ only keyframe-based visual descriptors.

The proposed system makes use of only one global descriptor and two types of local descriptors. In our system, local descriptors are extracted both in a dense grid and based on an interest-point detector. We mention them briefly in the following.

1) Local descriptors:

a) Extraction of interest points:

It has been reported that improvements in performance are obtained by combining the extraction of interest points in different ways, in the context of classification-based tasks [8]. We employ the Harris-Laplace [10] interest point detector and also extract patches on a dense grid.

The extraction of keypoints based on a dense grid has been shown to provide the best results for image classification-based tasks [11]. This is due to the fact that, in these scenarios, patterns other than the ones extracted by a common interest-point detector are statistically important.

b) Patch description:

We describe the extracted patches by using the well-known SIFT descriptors [12] and OppSIFT descriptors [13], which are a variant of SIFT calculated in each color component in the Opponent color space.

We verify experimentally that performance gains are obtained when combining these two.

2) Global descriptor: CENTRIST, a very efficient global descriptor for scene categorization, was introduced recently [14]. It has been shown that it works well for the Video Semantic Indexing [15] task.

It works by initially generating a binary pattern for each pixel, based on the comparison to its neighboring pixels. The final global descriptor (Spatial Principal component Analysis of Census Transform histograms - SPACT) is a histogram of appearances of each binary pattern, pooled over different spatial regions of the image.

B. Aggregation of local descriptors into feature vectors

In this subsection, we explain the methods by which we take the local descriptors and aggregate them in a fixed-size vector, suitable to common classification algorithms.

1) From variable-size sets to fixed-length feature vector: After the processing described in II-A1, we end up with a set of local descriptors representing each keyframe. The size of these sets can vary, due to variations on video resolution and number of detected keypoints.

In order to feed a fixed-size feature vector to a common machine learning tool, there is a need to transform the variable-size set of features into a fixed-size feature vector.

Many approaches have been used to perform such mapping, the most common being the Bag-of-Visual-Words (BoVW) method [16]. More recently, several approaches have been proposed, and a comparison of the most important ones was presented in [17].

We employ a simplified version of the Fisher Vector approach [18], inspired on Vector of Locally Aggregated Descriptors (VLAD) [19] and on Residual Enhanced Visual Vectors (REVV) [20]. Both are based on the aggregation of differences of local descriptors and the centroid of the Voronoi cell to which they are assigned.

In our system, these differences are averaged in each Voronoi cell. Our fixed-size feature vector is, then, a concatenation of these averaged differences. We will refer to these aggregated vectors as 'Residual Vectors' in the rest of the document.

We also employ a BoVW, and verify experimentally that the combination of both aggregation methods provide a significant performance gain.

Lastly, it is important to mention that we use only training data when training the parameters of the aggregation functions – in our case, the training of the visual codebook by a common k-means.

2) Using spatial information: The spatial arrangement of visual elements is clearly a source of important information for categorization tasks. The most common way of using it is by constructing Spatial Pyramids [21], which consist of aggregating descriptors over each image sub-region - using one of the methods described in II-B1.

More recently, an elegant formulation for dealing with spatial information was proposed in [22], but we did not have time to take advantage of it.

We use Spatial Pyramids when aggregating descriptors via BoVW, using a 3x1 grid (three horizontal stripes), and verify improvements in performance. Due to lack of time, we do not apply Spatial Pyramids to our residual vector-based aggregation approach.

C. Classification using Support Vector Machines

Having aggregated local descriptors into a fixed-size feature vector, and having the global descriptor, we can proceed to the training of classifiers that will distinguish between the presence and absence of a certain concept in a video.

Video Semantic Indexing is inherently a multi-label problem: i.e., labels can co-occur. Thus, the most common practice is to train one-vs-rest classifiers.

The most commonly used tools for classification are Support Vector Machines (SVM's), which were shown to be quite effective for Video Semantic Indexing [8], [9]. That is the technique we choose. It is usually recommended that the data instances be weighted according to how unbalanced the positive and negative classes are [23]. However, this has not proved useful in our preliminary experiments.

Depending on the type of features, different kernels might be used. For BoVW-based features, it is well-known that Histogram Intersection Kernels (HIK) are quite effective [8]. For residual vector features and for the SPACT global descriptor, we employ a Radial Basis Function (RBF) kernel, which is known to be a good general-purpose kernel. For RBF kernels (in our case defined as $\exp(-\gamma||u-v||^2)$), we set the parameter γ to an estimate of $\frac{1}{E(||x_i-x_j||^2)}$, where u and v are two feature vectors, as is commonly done in practice.

1) AP-based validation: The SVM regularization parameter C is chosen by performing validation experiments. For the Semantic Indexing task, since the evaluation is based on average precision (AP), we choose the C parameter that optimizes this measure.

We partition the training data in two sets. One partition is used for validation training, while the other is used for validation testing. We pick the parameter that achieves best estimated AP in these tests.

An interesting work studies the best way to perform cross-validation when optimizing AP, and in the context of Video Semantic Indexing [24]. It concludes that it is better to employ Balanced Average Precision (BAP) as a cross-validation measure. In our case, since we perform only one training and testing in validation experiments, the use of AP performs as well as the use of BAP.

Finally, we make sure that the validation classifier training is not influenced by validation testing data - so, codebook training and dimensionality reduction for validation experiments are performed based only on validation training data.

2) Final training and testing: After choosing the regularization parameter C, we train the final classifiers based on the entire training set. The final performance for a given concept is given by the estimated AP (see III-B) calculated on the results given by the application of the classifier on the testing set.

D. Late fusion

After obtaining the scores of the classifiers for each feature channel, a method is needed to be able to combine them. In the literature, this process is referred as 'late fusion', since the combination of the scores is done after the classification based on each feature channel.

This contrasts to what is known as 'early fusion', when the combination is done before classification. This is usually accomplished by summing up the kernel matrices of each feature channel, and is equivalent to concatenating the feature vectors in the space defined by the kernel.

In preliminary experiments, late fusion has proved to be more useful, so we decided to use it. Depending on the run, our late fusion strategy consists either of a simple average of scores from each feature channel, or of a linear combination of scores with coefficients given by each feature channel's validation performance.

III. Experiments and results

In this section, we present the experimental setup of the Semantic Indexing task and the results obtained by our team.

A. Dataset

The dataset consists of:

i) Training data: 19,701 videos, which correspond to 400,289 shots (roughly 20 shots/video) and 600 hours. These are drawn from the IACC.1.tv10.training, IACC.1.A and IACC.1.B collections.

ii) Testing data: 8,263 videos, which correspond to 145,634 shots and 200 hours. These are drawn from the IACC.1.C collection.

All these collections are provided by NIST, and are sampled from the Internet Archive [25]. All videos are provided in the H.264 format, with duration in the range from 10 seconds to 3.5 minutes. The videos are representative of user-generated content, with a large variety of topics and quality.

Some metadata, such as titles, textual tags and short descriptions might be provided with each video. Outputs of English speech-to-text software are also provided. For reasons explained in II-A, our system only makes use of visual information for retrieval.

All annotations are provided based on the shot, i.e., a shot might be tagged with the absence of the concept 'Ground vehicle', but with the presence of the concept 'Person'. A list of all concepts and their definitions can be found at http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.500.concepts_ann_v2.xls.

There are 500 concepts in total, of which 346 are chosen for the Full experiment type, and 50 are chosen for the Light experiment type. The 2,000 shots most likely to contain each concept must be submitted, in a ranked list. After submission, a subset of them are chosen for judgment, by the workshop organization: 15 in the Light task, 46 in the Full task. Our team decided to participate on the Light experiment only.

The annotation of the shots is done collaboratively, with the participation of members from the groups participating in TRECVID. It is important to mention that not all shots are annotated. Based on an active learning system, a subset of the shots are chosen to be annotated. At a high-level, the system estimates which shots can be more valuable for a Semantic Indexing system, and then suggests the most important ones to the human annotator. For more details, the reader is referred to [26].

Only the shots containing labels are used by our system. 9.75% of the shots in the training set do not contain labels, and thus are not used. It is known that unannotated shots can still help with a general classification task (usually referred as semi-supervised learning), but we did not exploit this strategy due to limited time constraints.

Not many shots contain enough annotations so as to leverage the correlations between concept appearances (for example, the presence of the concept 'News Studio' usually occurs together with the presence of the concept 'Person'). Only 47.2% of the shots contain more than 10 annotations, and only 16% of the shots contain more than 100 annotations. Also, as previously mentioned in II, previous work shows that the system becomes significantly more computationally demanding if using co-occurence information [6].

Lastly, two weeks before the submission deadline, colleagues at the Brno University of Technology released another 750k annotations on the training data [27]. Due to the closeness to the submission deadline, most teams (including us) did not make use of these annotations. For our team, another important issue was memory constraints: with the available annotations, there was just enough room in memory for the features of some concepts. Adding more data would probably require re-engineering some of the key algorithms of our system.

B. Evaluation measures

Since not all shots are annotated, it is not possible to calculate Average Precision (AP). A system that estimates this measure is then used [28]. The final measure for each concept is called infAP (or, in some cases, xinfAP), which stands for Inferred Average Precision (respectively, Extended Inferred Average Precision).

Whenever referring to the measure of performance taking into account all concepts, MinfAP (Mean of the Inferred Average Precisions) is used. This is simply the average of the infAP's over all concepts.

Similarly, a precise calculation of precision and recall is not possible. The same system is then used to estimate the Precision-Recall curve at different recall points.

C. Submitted concept detection runs

Each team could submit four different runs to the Semantic Indexing task. Our four runs are detailed as follows.

1) Run 1: This is our top-performing run, combining all of the techniques mentioned above, and performing late fusion with validation-based weights. It achieves the 19th best result among all runs in the light experiment, and the 6th best result considering only the best run per team. It achieves the 3rd best result for the concepts "Landscape" and "Male person".

The feature channels used in this run are composed of:

i) Dense keypoint extraction + OppSIFT. BoVW and Spatial Pyramid pooling in a 3x1 grid with a 4096-size visual dictionary;

ii) Dense keypoint extraction + SIFT. BoVW and Spatial Pyramid pooling in a 3x1 grid with a 4096-size visual dictionary;

iii) HarLap keypoint extraction + OppSIFT. BoVW pooling with a 4096-size visual dictionary;

iv) HarLap keypoint extraction + SIFT. BoVW pooling with a 4096-size visual dictionary;

v) Residual vectors on PCA-reduced (to 64 dimensions) densely extracted OppSIFT, using a 256-dimensional visual dictionary;

vi) Residual vectors on PCA-reduced (to 32 dimensions) densely extracted SIFT, using a 512-dimensional visual dictionary;

vii) SPACT: Spatial Principal component Analysis of Census Transform Histograms (CENTRIST).

2) Run 2: This run employs the exact same feature channels as Run 1, the sole difference being the way the scores are combined: in this case, we use a simple average of the scores. It achieves the 20th best result among all runs in the light experiment.

3) Run 3: This run evaluates the use of residual vector aggregation. We add the SPACT global descriptor, since it is very efficient and always helps, and perform late fusion with validation-based weights. It achieves the 43^{rd} highest result among all runs in the light experiment.

This run is composed of the following feature channels:

i) Residual vectors on PCA-reduced (to 64 dimensions) densely extracted OppSIFT, using a 256-dimensional visual dictionary;

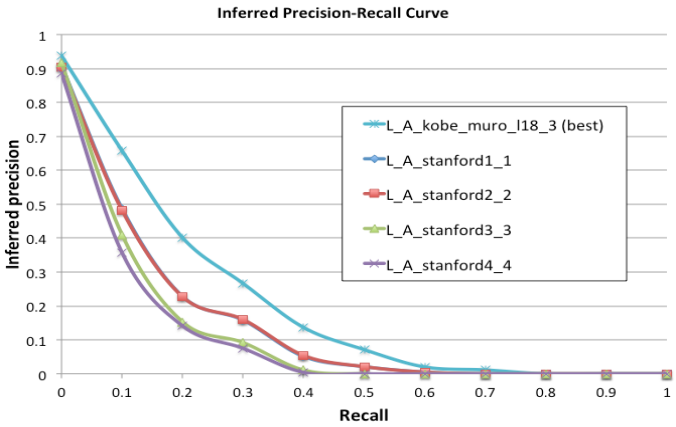

Fig. 3. Inferred precision-recall curve for each of our runs, and for the top-performing run. The runs "L_A_stanford1_1' and "L_A_stanford2_2"' overlap in the graph, since they are very similar.

ii) Residual vectors on PCA-reduced (to 32 dimensions) densely extracted SIFT, using a 512-dimensional visual dictionary;

iii) SPACT: Spatial Principal component Analysis of Census Transform Histograms (CENTRIST).

4) Run 4: One of the most computationally costly steps of the Semantic Indexing system is the extraction of local descriptors. This run evaluates the use of only one local descriptor – densely extracted SIFT – with different aggregation methods. We add the SPACT global descriptor, since it is very efficient and always helps, and perform late fusion with validation-based weights. It achieves the 49^{th} highest result among all runs in the light experiment.

This run is composed of the following feature channels:

i) Dense keypoint extraction + SIFT. BoVW and Spatial Pyramid pooling in a 3x1 grid with a 4096-size visual dictionary;

ii) Residual vectors on PCA-reduced (to 32 dimensions) densely extracted SIFT, using a 512-dimensional visual dictionary;

iii) SPACT: Spatial Principal component Analysis of Census Transform Histograms (CENTRIST).

D. Detailed results

Table I presents the detailed results for all our runs, together with the top-performing run, "L_A_kobe_muro_l18_3". Figures 3 and 4 show graphically Precision-Recall and Precision@n extracted from this Table.

Figure 5 compares the performance for each of our runs with respect to the best and median results for each concept – note: the comparison is done with respect to the best result for each concept, and not the result for each concept of the "L_A_kobe_muro_l18_3" run.

E. Timing

We report some timing numbers to give an idea of computational complexity. The following numbers are based on 24-core Intel Xeon 2.40GHz servers, with 64GB RAM.

TABLE I
DETAILED RESULTS FOR ALL OUR RUNS, AND FOR THE TOP-PERFORMING RUN.

Run	L_A_kobe_muro_l18_3 (best)	L_A_stanford1_1	L_A_stanford2_2	L_A_stanford3_3	L_A_stanford4_4
MinfAP	0.3578	0.265	0.263	0.212	0.206
Recall	**Inferred precision**				
0	0.9389	0.909	0.904	0.918	0.886
0.1	0.6577	0.487	0.482	0.41	0.357
0.2	0.4013	0.228	0.227	0.153	0.141
0.3	0.2663	0.158	0.16	0.092	0.075
0.4	0.1365	0.051	0.054	0.011	0.004
0.5	0.0711	0.021	0.021	0	0
0.6	0.0201	0.005	0.005	0	0
0.7	0.0113	0	0	0	0
0.8	0	0	0	0	0
0.9	0	0	0	0	0
1	0	0	0	0	0
Depth	**Inferred precision**				
10	0.82	0.74	0.747	0.747	0.693
100	0.678	0.587	0.589	0.527	0.515
1000	0.4424	0.359	0.36	0.323	0.315
2000	0.3633	0.3	0.298	0.267	0.258
Concept	**infAP per concept**				
Airplane_Flying	0.424	0.258	0.273	0.109	0.139
Bicycling	0.08	0.021	0.02	0.009	0.014
Boat_Ship	0.219	0.162	0.15	0.123	0.121
Computers	0.064	0.058	0.056	0.057	0.048
Female_Person	0.686	0.552	0.557	0.532	0.515
Instrumental_Musician	0.38	0.229	0.224	0.146	0.145
Landscape	0.55	0.607	0.605	0.571	0.532
Male_Person	0.918	0.936	0.933	0.716	0.846
Nighttime	0.297	0.18	0.175	0.16	0.145
Scene_Text	0.495	0.165	0.159	0.166	0.133
Singing	0.194	0.065	0.066	0.039	0.039
Sitting_Down	0.003	0.003	0.003	0.002	0.003
Stadium	0.203	0.215	0.208	0.138	0.075
Throwing	0.187	0.124	0.121	0.1	0.076
Walking_Running	0.667	0.395	0.393	0.308	0.256
All	0.358	0.265	0.263	0.212	0.206

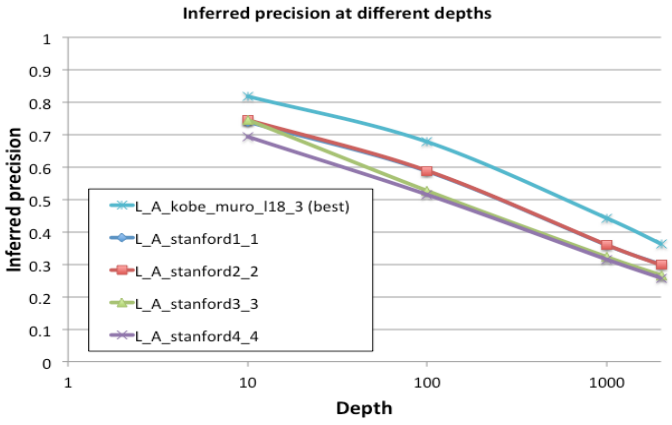

Fig. 4. Inferred precision@n for each of our runs, and for the top-performing run. The runs "L_A_stanford1_1' and "L_A_stanford2_2"' overlap in the graph, since they are very similar.

The computational cost breaks down into i) descriptor extraction, ii) local descriptor aggregation, iii) validation experiments, iv) classifier training and v) classifier testing. Late fusion does not take a significant amount of time.

Descriptor extraction and aggregation take from 1 to 2 days.

We precompute the SVM kernels for each pair of shots, which makes classifier training and testing very efficient. The major computational cost becomes, then, the precomputation of the kernel matrices. These depend basically on the feature dimension and the chosen kernel.

Precomputation of validation training matrix took from 2 hours to 2.5 days. Precomputation of validation testing matrix took roughly twice the time of precomputing validation training kernel matrix.

Precomputation of training kernel matrices could take from 10 hours to 5 days. Precomputation of the testing kernel matrix took roughly half of the time that it took for the the training kernel matrix precomputation.

F. Visualization of results

We present in Figure 6 top-10 results for our best run for nine different queries - 3 best-performing concepts, 3 worst-performing concepts and 3 average-performing concepts. The shots are represented by a keyframe, which usually contains most of the visual elements that appear along the shot.

Further visualization of results, including full submitted results for all concepts in the light experiment and precise

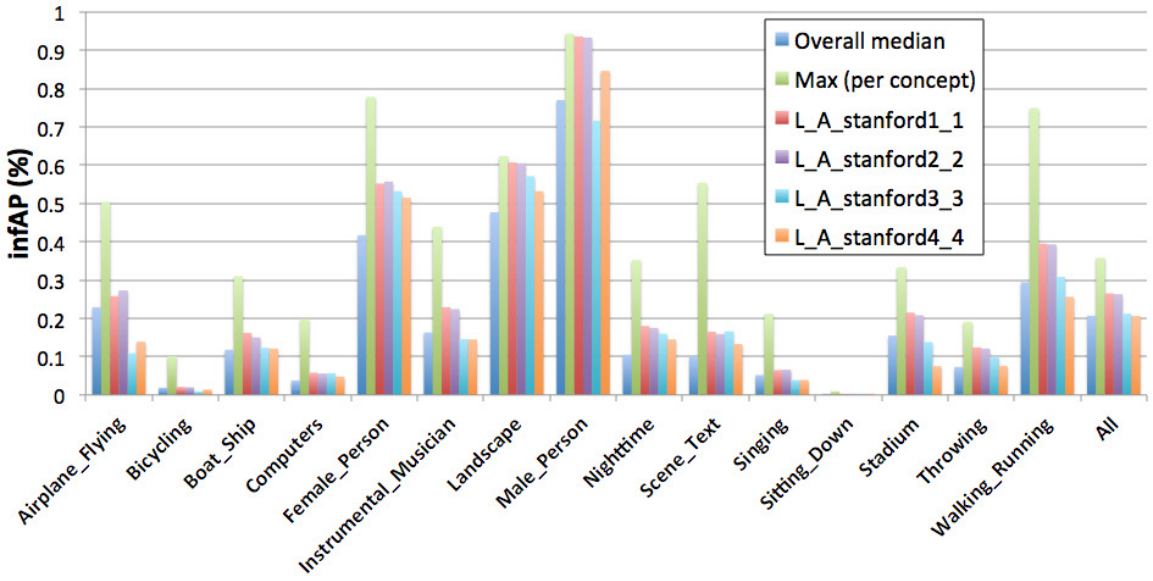

Fig. 5. infAP per concept, for each of our runs. As a reference, we also show the median infAP per concept, and the best infAP per concept.

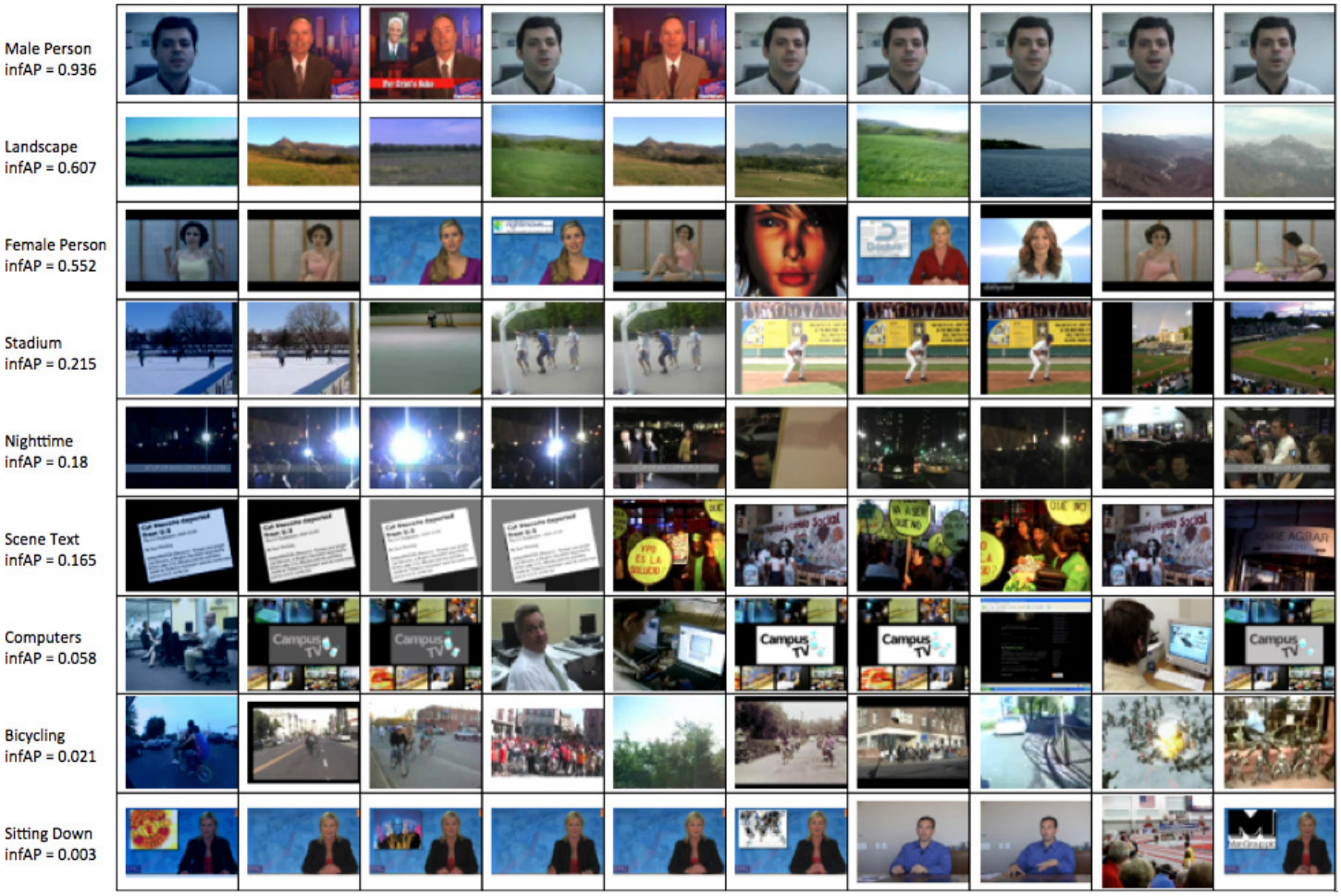

Fig. 6. Top-10 retrieved shots (only keyframe is shown) for selected concepts, for our best run (first retrieved shot on left). All results for all runs can be examined at http://stanford.edu/~afaraujo/trecvid.

definition of concepts, is available online, via the project webpage: http://stanford.edu/~afaraujo/trecvid.

IV. Conclusion

In this work, we have put together a system that automatically annotates video shots with semantic concepts. Tagging semantic concepts has been shown recently to be an effective tool for searching videos.

The system is based on state-of-the-art image processing and machine learning techniques. It uses SIFT, OppSIFT, CENTRIST descriptors, BoVW and Residual Vector aggregation, spatial pyramids and classification based on one-vs-rest SVM's.

The system has been evaluated on the large-scale TRECVID 2012 SIN task dataset, which contains more than 400k training shots and more than 100k testing shots.

In our team's first participation in the TRECVID workshop, the achieved performance was very satisfactory: our team ranked 6^{th} in the Semantic Indexing light experiment .

References

[1] "Youtube statistics," Oct. 2012. [Online]. Available: http://www.youtube.com/t/press_statistics

[2] C. G. M. Snoek and M. Worring, "Concept-Based Video Retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2007.

[3] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 958–966, 2007.

[4] M. Naphade, J. Smith, C. S. Tesic, Jelena, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[5] C. G. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, Jan. 2005.

[6] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *ACM Multimedia*, 2007.

[7] B. Delezoide, F. Precioso, P. Gosselin, M. Redi, B. Merialdo, L. Granjon, D. Pellerin, M. Rombaut, H. Jegou, R. Vieux, B. Mansecal, J. Benois-Pineau, S. Ayache, B. Safadi, F. Thollard, G. Quenot, H. Bredin, M. Cord, A. Benoit, P. Lambert, T. Strat, J. Razik, S. Paris, and H. Glotin, "Irim at TRECVID 2011: Semantic indexing and instance search," in *TRECVID Workshop*, 2011.

[8] C. Snoek, K. van de Sande, X. Li, M. Mazloom, Y.-G. Jiang, D. C. Koelma, and A. W. M. Smeulders, "The MediaMill TRECVID 2011 Semantic Video Search Engine," in *TRECVID Workshop*, 2011.

[9] N. Inoue, Y. Kamishima, T. Wada, K. Shinoda, and S. Sato, "TokyoTech + Canon at TRECVID 2011," in *TRECVID Workshop*, 2011.

[10] T. Tuytelaars and K. Mikolajczyk, "Local Invariant Feature Detectors: A Survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007.

[11] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference of Computer Vision (ECCV)*, 2006, pp. 490–503.

[12] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[13] 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*, vol. 32, no. 9, pp. 1582–96, Sep. 2010.

[14] J. Wu and J. M. Rehg, "CENTRIST: A Visual Descriptor for Scene Categorization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, Dec. 2010.

[15] M. Sjöberg, S. Ishikawa, M. Koskela, J. Laaksonen, and E. Oja, "PicSOM Experiments in TRECVID 2011," in *TRECVID Workshop*, 2011.

[16] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.

[17] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference 2011*, 2011.

[18] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[19] H. Jegou, M. Douze, C. Schmidt, and P. Perez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[20] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, H. Chen, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual Enhanced Visual Vectors for On-Device Image Matching," in *Asilomar Conference*, 2011.

[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *IEEE Conference on Computer Vision and Pattern Recognition*.

[22] J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with Fisher vectors for image categorization," in *International Conference on Computer Vision*, 2011.

[23] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid, "Towards good practice in large-scale learning for image classification," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[24] J. C. van Gemert, C. J. Veenman, and J.-M. Geusebroek, "Episode-Constrained Cross-Validation in Video Concept Retrieval," *IEEE Transactions on Multimedia*, 2009.

[25] "Internet archive," Oct. 2012. [Online]. Available: http://archive.org

[26] S. Ayache and G. Quénot, "Video Corpus Annotation using Active Learning," in *European Conference on Information Retrieval*, 2008.

[27] M. Hradiš, M. Kolár, A. Láník, J. Král, P. Zemcik, and P. Smrz, "Annotating images with suggestions - user study of a tagging system," in *Advanced Concepts for Intelligent Vision Systems*, 2012.

[28] E. Yilmaz, E. Kanoulas, and J. Aslam, "A simple and efficient sampling method for estimating AP and NDCG," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.