HAL
open science

Using Topic Concepts for Semantic Video Shots Classification

Stéphane Ayache, Georges Quénot, Jérôme Gensel, Shin'Ichi Satoh

To cite this version:

Stéphane Ayache, Georges Quénot, Jérôme Gensel, Shin'Ichi Satoh. Using Topic Concepts for Semantic Video Shots Classification. International Conference on Image and Video Retrieval CIVR'06, 2006, Tempe, United States. hal-00953902

HAL Id: hal-00953902

https://inria.hal.science/hal-00953902

Submitted on 3 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using Topic Concepts For Semantic Video Shots Classification

Stéphane Ayache, Georges Quénot, Jérôme Gensel and Shin'ichi Satoh

CLIPS-IMAG, LSR-IMAG, NII

Abstract. Automatic semantic classification of video databases is very useful for users searching and browsing but it is a very challenging research problem as well. Combination of visual and text modalities is one of the key issues to bridge the semantic gap between signal and semantic. In this paper, we propose to enhance the classification of high-level concepts using intermediate topic concepts and study various fusion strategies to combine topic concepts with visual features in order to outperform unimodal classifiers. We have conducted several experiments on the TRECVID'05 collection and show here that several intermediate topic classifiers can bridge parts of the semantic gap and help to detect high-level concepts.

1 Introduction

In order to retrieve and browse videos into huge databases, needs for indexing understandable concepts are rapidly growing. The extraction of such concepts is one of the main objectives of the semantic video indexing community. Although using and combining visual and text modalities are expected to improve the performance of high-level concepts classification, new issues arise. Usual approaches, merge directly visual and text features into a single flat classifier. However, even with a clever choice of relevant features, the correlation between such low-level features and high-level concepts is still weak. Such approaches assume that there exists a correlation between uttered speech and high-level concepts to classify high-level features [3, 15, 18]. But, recent TRECVID evaluations [1] have shown the limitations of such approaches: a single classifier cannot bridge this large semantic gap. Furthermore, concerning visual modality, promising results have been obtained by integrating context information based on the merging of intermediate visual concepts [17, 7, 14, 1]. Such a stacked classifier [19] learns implicit relations between intermediate concepts to derive high-level concepts.

In this study, we extend the context based framework we proposed in [1] by exploiting intermediate concepts extracted from textual modality. In order to learn relations between uttered speech and visual content, we propose to classify video shots with several intermediate topic categories, then to combine them for high-level concepts detection. We show that such topic categories provide useful

[1] TREC Video Retrieval Evaluation: http://www-nlpir.nist.gov/projects/trecvid/

semantic context when combined with visual information. The main idea is that several intermediate classifiers can bridge small parts of the semantic gap in order to improve the detection of high-level concepts. Similarly to the early and late fusion schemes [18], we investigate the combination of intermediate concepts by means of one and two-level fusions. We show the improvement brought by the use of topic concepts for video shots classification through several experiments performed on TRECVID'05 corpora. This paper is structured as follows: in section 2, we argue for the use of intermediate concepts; in section 3, we describe our framework based on intermediate concepts classification; in section 4, we describe and comment some conducted experiments; we finally draw some conclusions in section 5.

2 The need for intermediate and understandable concepts

Video documents contain visual, textual and audio cues (where text is mainly extracted from speech transcription). Thus, high-level concepts can be extracted from various modalities and can, on the one hand, exploit contexts from other cues while, on the other hand, provide context to them. We have shown that combining intermediate visual concepts offers rich sources of contexts and increase the derivation of high-level concepts [1, 2].

Furthermore, the extraction of understandable concepts is extremely useful for video browsing, since users can use them for expressing non-trivial information needs. Many approaches extract mid-level features by using dimensionality reduction algorithms such as PCA or LSA [13, 5, 11], yielding discriminant features which fit especially well the learned data. However, such eigen-features are hard to interpret and, hence, unusable for other multimedia tasks. Thus, in order to enrich our basis of concepts, we focus on the use of supervised classifiers in order to extract usable intermediate concepts.

2.1 High-level concepts

High-level concepts are input devices which allow users to express their information need for video browsing or search tasks. They can be described in terms of other concepts and are independent of the modality in which they are naturally expressed [8]. In this paper, we focus on the 10 high-level concepts defined in TRECVID'05.

2.2 Visual concepts

We extract visual local concepts from image patches in each keyframe. Intermediate visual concepts provide spatial and semantic knowledge to higher level classifiers. We use a set of 15 visual concepts ('vegetation', 'sky', 'skin/face' ...) selected from the LSCOM ontology [12] which can be extracted from patches and are discriminative enough to help the classification of higher level concept.

2.3 Topic concepts

We propose to extract a set of topic concepts from speech transcription [6], then to classify shots according to these intermediate concepts. We use 25 categories of the TREC Reuters collection [10] to classify each speech segment. The advantages of extracting such concepts from the Reuters collection are that they cover a large panel of news topics like the TRECVID collection, they are obviously human understandable, and thus they can be used for video search tasks. Examples of such topics are 'Economics', 'Disasters', 'Sports' and 'Weather'. Reuters collection contains about 800000 text news items in the years 1996 and 1997. They are classified among 103 topics, hierarchically structured. We used the 25 top level categories as topic concepts.

3 Video indexing framework for high-level concepts classification

In our previous work, we have proposed a context-based approach for automatic image annotation based on intermediate local concepts [1]. Using stacking technique, a multi-layer SVM classifier learns topological and semantic contexts from image content. In the present work, topic concepts and visual global features are expected to enhance the discriminating power of semantic context. We also investigate the use of fusion classifiers in order to merge such intermediate concepts while exploiting the 2 following kinds of context:

- **Topologic context** learns the spatial distribution of a visual local concept and assign a score to the whole image. The idea behind the use of topologic context is that the confidence (or score) of an image performs better by taking into account the confidences obtained for each patch in the image for the same concept.
- **Semantic context** exploits the semantic relations between concepts based on co-occurrences learned by the classifier. The idea is that the confidence of a single concept is computed more accurately by taking into account the confidences obtained for other concepts co-occuring in the same image.

Additionally, we have shown on the TRECVID'05 evaluation that combining both contexts increases the accuracy of high-level concepts classification. By merging all the visual concept scores, a classifier learns **Topologic-Semantic** context associated with the local concepts and the high-level concepts. The classifier can activate or inhibit high-level concepts based on intermediate scores and learned relations [2].

Figure 1 shows a general view of our extended framework. At the intermediate layer, the Local classifier assigns visual local concept scores to each patch of keyframes, while the Topic classifier assigns topic concept scores to each speech segment. Then, the fusion function derives high-level concept scores at shot level, based on both the output of intermediate classifiers and the use of predefined contexts.

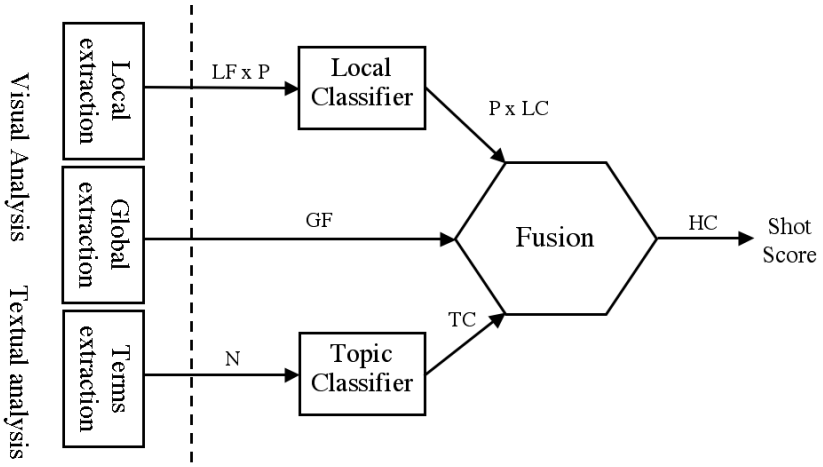

Fig. 1. Extended framework. LF and GF are numbers of Local and Global Low-level Features, P is the number of Patches, LC denotes the number of Visual Local concepts, HC denotes the number of high-level concepts, N is the number of inputs terms and TC the number of Topic Concepts.

3.1 Visual Concepts classification

We classify each keyframe using intermediate local concepts and global low-level features. One local concept corresponds to the score of one patch according to a visual concept. In order to derive high-level concepts, we merge local concepts using Topologic-Semantic context. Additionally, the use of global low-level features is expected to enhance local only classification.

Visual low-level extraction

As we want to handle the topologic context, we need to compute low-level features for parts of the image, as well as for the whole image. In order to compute local features, many approaches have been proposed. Since the automatic and a priori segmented regions are usually too far from the semantic meaning of image, we have decided to split images into patches. By doing so, we should be far from semantic, but with such granularity one patch is more likely to contain one single concept.

- **Local features:** We first split the image into overlapping patches. In our experiments, we use $P = 20 \times 13$ patches of 32×32 pixels. For each patch, we compute 9 color momentums (3 means + 6 co-variances), 24 Gabor wavelets for texture (3 scales x 8 orientations), and the 2 coordinates of the patches.
- **Global features:** More information can be extracted from a whole image than from a single patch. We extract $4 \times 4 \times 4$ three-dimensional histograms for color features based on RGB channels, Gabor wavelets for texture (5 scales $\times$ 8 orientations), and the first two momentums of motion vectors obtained from optical flow.

3.2 Topic Concepts classification

Based on speech transcription data, Topic classification relies on text retrieval techniques. The most common and widely used approach is the Vector Space Model [16], which has been successfully used in the traditional IR field. The model considers a vector space in which both documents and queries are represented by vectors of weighted terms calculated by the TF.IDF formula. For the classification task, a Rocchio classifier consists in first creating a prototype vector for each class, then assigning a test document to the nearest prototype using the Cosine similarity.

In TRECVID transcriptions, one speech segment corresponds to one speaking turn. We build the prototype vectors of each topic category on Reuters corpora and apply the Rocchio classification on each speech segment. Such granularity is expected to provide robustness in terms of covered concepts, as each speaking turn should be related to a single topic. Our assumption is that the statistical distributions of Reuters corpora and TRECVID transcriptions are similar enough to obtain relevant results. Finally, we derive high-level concepts by merging outputs of Topic concept classifiers.

Text analysis
We construct a vector representation for each speech segment by applying stoplist and stemming. Also, in order to avoid noisy classification, we reduce the number of input terms. While the whole collection contains more than 250000 terms, we have experimentally found that considering the top 2500 frequently occurring terms gives the better classification results on Reuters collection.

3.3 Combining intermediate features

Combining intermediate concepts aims at deriving high-level concepts from several unimodal intermediate concepts. Such strategy leads to a multimodal shot classifier. In order to unify outputs of intermediate classifiers, we report the topic concept scores of speech segments to each keyframe. A given keyframe, corresponding to the time point T, is associated with the speech-segment delimited by the time bounds TB and TE so that $TB \leq T \leq TE$.

We identify two possibilities to merge intermediate features, depending on the abstraction level considered. Similarly to the early and late fusion schemes defined in [18], we use either a one level or a two level fusion schemes, described as follows:

One-level fusion: In a one-level fusion process, intermediate features or concepts are concatenated into a single flat classifier, as in an early fusion scheme [18]. Such a scheme takes advantage of the use of the semantic-topologic context from visual local concepts, and semantic context from topic concepts and visual global features. However, it is constrained by the curse of dimensionality problem. Also, the small numbers of topic concepts and global features compared to the huge amount of local concepts can be problematic: the final score might strongly depend upon the local concepts.

Two-level fusion: In a two-level fusion scheme, we classify high-level concepts from each modalities separately at a first level of fusion. Then, we merge the obtained outputs into a second layer classifier. We investigate the following possible combinations. Classifying each high-level concept with intermediate classifiers, then merging outputs into a second level classifier is equivalent to the late fusion defined in [18]. Using more than two kinds of intermediate classifiers, we can also combine pairwise intermediate classifiers separately, then combine given scores in a higher classifier. For instance, we can first merge and classify global features with topic concepts, then combine the given score with outputs of local concept classifiers in a higher classifier. An other possibility is to merge separately local concepts with global features and local concepts with topic concepts, then to combine the given scores in a higher level classifier. Advantages of such schemes are numerous: the second layer fusion classifier avoids the problem of unbalanced inputs, and keep both topologic and semantic contexts at several abstraction levels.

We compute high-level concept scores for each keyframe using the predefined fusion classifiers. Then, in order to set a score to video shots, we keep the keyframe which has the maximum score, according to the idea that a concept occurs in a shot if one of the sub-shots contains this concept.

4 Experiments

We evaluate the use of visual and topic concepts and their combination for high-level concepts detection in the conditions of the TRECVID'05 evaluation. We show the 10 high-level concepts classification results evaluated with the trec_eval tool using the provided ground truth, and compare our results with the median over all participants. We have used a subset of the training set in order to exploit the speech transcription of the samples. As the quality of TRECVID'05 transcription is quite noisy due to both transcription and translation from Chinese and Arabic videos, some video shots do not have any corresponding speech transcription. In order to compare visual only runs with topic concept based runs, we have trained all classifiers using only keyframes whose transcript is not empty. In average, we have used about 300 positives samples and twice as many negative samples.

It has been shown in [10] that SVM outperforms a Rocchio classifier on text classification. In this experiment, we first show the improvement brought by the topic concepts based classification by comparing with a SVM text classifier based on the uttered speech occurring in a shot after same text analysis as topic classifiers. Then, we give some evidence of the relevance of using topic concepts, by showing the improvement of unimodal runs when combined with the topic concepts. In a second step, we compare one-level fusion with two-level fusion for combining intermediate concepts. We have implemented several two-level fusion schemes to merge the output of intermediate classifiers. Particularly, we show that pairwise combinations schemes can increase high-level concepts classification.

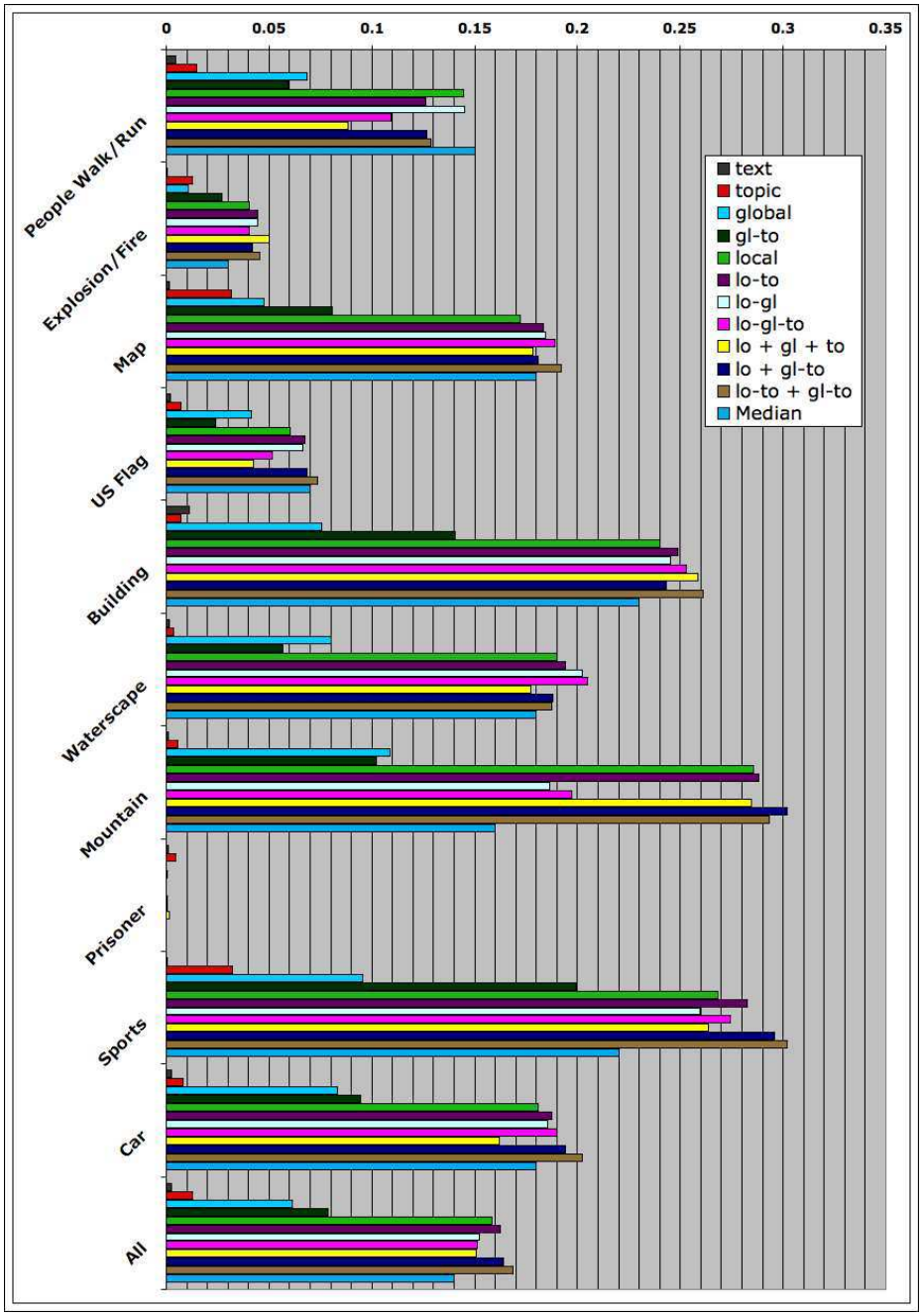

Fig. 2. Mean Average Precision of the 10 high-level concepts of TRECVID'05

We used a SVM classifier with RBF kernels as it has proved good performance in many fields, especially in multimedia classification. LibSVM [4] implementation is easy to use and provides probabilistic classification scores as well as efficient cross validation tool. We have selected the best combination of parameters C and Gamma out of 110, using the provided grid search tool.

Figure 2 shows the Mean Average Precision (MAP) results of the conducted experiments. We compare our results with the TRECVID'05 median result. 'Text', 'Topic' 'Global' and 'Local' experiments refers to the use of unimodal classifiers. On fusion experiments, the sign '-' refers to a one-level fusion and the sign '+' refers to the second-level fusion. For instance, in the run 'lo-to' we have concatenated inputs of the 'Local' and 'Topic' runs, and for the 'lo-to + gl-to' run, we have merged in a second level fusion scheme the classification results of 'lo-to' and 'gl-to' runs.

Topic concepts based classification performs much better than text based classifier, the gain obtained by topic concepts based classification is obvious. It means that despite the poor quality of speech transcription, intermediate topic concepts are useful to reduce the semantic gap between uttered speech and high-level concepts. Each intermediate topic classifier provides significant semantic information despite the differences between Reuters and TRECVID transcripts corpora. It is interesting to notice that the 'Sports' concept is also a Reuters category and has the best MAP value for the Topic concepts based classification.

For 'Global' run, we have directly classified high-level concepts using their corresponding global low level features. When combined with topic concepts, the average MAP increases by 30%, and up to 100% on Sports high-level concept. Also, some high-level concepts which have poor topic based classification MAP cannot benefit from the combination with topic concepts.

The use of the topologic-semantic context in local concepts based classification improves clearly the performance over the global based classifier. However, we observe a non significant gain when combined with topic concepts. This can be explained by the huge numbers of 'Local' inputs compared with the few numbers of 'Topic' inputs. Since we have used RBF kernel, the topic concepts inputs have a very small impact on the euclidian distance between two examples. A solution to avoid such unbalanced inputs could be to reduce the numbers of local concepts inputs using a feature selection algorithm before merging with the topic concepts. Despite this observation, we notice that we obtain better results by combining Local with Topic concepts than combining Local concepts with Global features.

We have conducted several experiments to combine 'Topic' concepts with 'Local' and 'Global' features. Where 'Local' only classification performs very well for some "visual" high-level concepts (Mountain, Waterscape), we can observe an improvement using fusion based runs for most of high-level concepts. The runs 'lo-go-to' and 'lo + go + to', which correspond respectively to the early and late fusion schemes, provide roughly similar results and do not outperform visual local classifier. This is probably due to the relative good performance of 'Local' run compared to other runs.

We have obtained the most significant results using two-level fusion when combining separately topic concepts with local and global features in the first fusion layer. In this case, the duplication of topic concepts at the first level fusion performs better by 10% than other fusion schemes. With such a scheme, topic concepts integrate useful context to visual features and achieve significant improvement, compared to unimodal classifiers, for most of high-level concepts.

5 Conclusion

In this paper, we investigate the use of topic concepts on a generic framework for high-level concepts video shots classification. We show that topic concepts based classification performs much better than a single text classifier to classify high-level concepts. In addition, we show that combined with visual cues, topic concepts can improve shots classification despite the poor quality of speech transcriptions. However, in some case, the 'Local' unimodal classifier does better than other fusion strategies. This could be due to the huge numbers of 'Local' inputs compared to the few numbers of 'Topic' inputs. The RBF kernel used in the presented experiments was not able to handle such unbalanced inputs.

Furthermore, regarding to the 'Sports' classification performance, the choice of topic categories seems to have a direct impact on the high-level concepts classification. Therefore, in future work, we intend to improve the topic based classification by carefully selecting the topic categories and also by appropriately normalizing the texts of Reuters and TRECVID collections. It should be also interesting to evaluate our approach on the TRECVID 2003 and 2004 collections which have better quality transcriptions.

Bibliography

[1] S. Ayache and G. Quénot and S. Satoh. Context-based conceptual image indexing. In ICASSP, 2006.

[2] S. Ayache and G. Quénot and J. Gensel and S. Satoh. CLIPS-LSR-NII experiments at TRECVID 2005. In TRECVID Workshop, 2005.

[3] S. Ayache and G. Quénot and M. Charhad. Video shot classification using lexical context. In European Conference on Information Retrieval, 2005.

[4] C. Chang and C Lin LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[5] A Garg, S Agarwal, T.S. Huang. Fusion of Global and Local Information for Object Detection. In 16th International Conference on Pattern Recognition (ICPR'02) - Volume 3, 2002.

[6] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. Speech Communication, 37(1-2):89-108, 2002.

[7] G. Iyengar and H.J. Nock. Discriminative model fusion for semantic concept detection and annotation in video. MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, 2003.

[8] G. Iyengar and H. Nock and C. Neti and M. Franz. Semantic indexing of multimediq using audio, text and visual cues. In IEEE Int. Conference on Multimedia and Expo, 2002.

[9] D. Lewis, F. Li, T. Rose, and Y. Yang. The reuters corpus volume I as a text categorization test collection. In Journal of Machine Learning Research, 2003.

[10] Lewis, D. D.; Yang, Y.; Rose, T.; and Li, F. RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 5:361-397, 2004.

[11] D. A. Lisin, M. A. Mattar, M B. BlMark C. Benfield and E. G. Learned-Miller. Combining Local and Global Image Features for Object Class Recognition. In CVPR, 2005.

[12] LSCOM Lexicon Definitions and Annotations Version 1.0. DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3 , March 2006.

[13] K. Murphy, A. Torralba, D. Eaton and W. Freeman. Object detection and localization using local and global features. Sicily Workshop on Object Recognition. Lecture Notes in Computer Science, 2005.

[14] M. Naphade. On supervision and statistical learning for semantic multimedia analysis. Journal of Visual Communication and Image Representation, 15(3):348369, 2004.

[15] H. J. Nock, G. Iyengar, C. Neti. Issues in speech-based retrieval of video. In ISCA Tutorial Workshop, 2003.

[16] G. Salton. Introduction to Modern Information Retrieval. McGraw-Hill, 1983

[17] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma and F.J. Seinstra. The MediaMill TRECVID 2004 Semantic Video Search Engine. InTRECVID Workshop, 2004.

[18] C.G.M. Snoek and M. Worring and A.W.M. Smeulders. Early versus Late Fusion in Semantic Video Analysis. Proceedings of ACM Multimedia, 2005.

[19] D.H. Wolpert Stacked Generalization. Neural Networks, Vol. 5, pp. 241-259, Pergamon Press.