

# Using High-Level Semantic Features in Video Retrieval

Wujie Zheng, Jianmin Li, Zhangzhang Si, Fuzong Lin, and Bo Zhang

State Key Laboratory of Intelligent Technology and System  
Department of Computer Science and Technology  
Tsinghua University, Beijing, 100084, China

idiot00@mails.tsinghua.edu.cn, lijianmin@mail.tsinghua.edu.cn,  
scc02@mails.tsinghua.edu.cn, {linfz, dcszb}@mail.tsinghua.edu.cn

**Abstract.** Extraction and utilization of high-level semantic features are critical for more effective video retrieval. However, the performance of video retrieval hasn't benefited much despite of the advances in high-level feature extraction. To make good use of high-level semantic features in video retrieval, we present a method called pointwise mutual information weighted scheme (PMIWS). The method makes a good judgment of the relevance of all the semantic features to the queries, taking the characteristics of semantic features into account. The method can also be extended for the fusion of multi-modalities. Experiment results based on TRECVID2005 corpus demonstrate the effectiveness of the method.

## 1 Introduction

The wide availability of digital sensors, the high bandwidth Internet, and the falling price of storage devices have resulted in the increasing growth of unstructured digital media content. Therefore, developing effective information management technologies is a matter of great urgency. Since the last decade, the problem of content-based video retrieval has been actively researched by many communities.

Early research emphasizes on low-level image features such as color, texture and shape. However, "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [1], which is well known as semantic gap, makes the early retrieval systems disappointing. Moreover, the computation of high dimensional low-level features can lead to poor efficiency. On the other hand, using high-level semantic features for video retrieval allows people to perform search in the semantic level, which is more intuitive. Also, high-level semantic features can integrate additional knowledge of a specific domain as well as low-level features. What's more, they are compact enough so that the retrieval can be performed fast. Though the extraction of high-level semantic features is time consuming, it can be performed offline [11].

To take full advantage of the virtues of high-level semantic features for video retrieval, there are two issues to be addressed: 1) How to extract reliable high-level semantic features. 2) How to use high-level semantic features to describe

dataset and queries with relevant retrieval method. The advances in machine learning, and the availability of large annotated information sources, e.g., the TRECVID benchmark, have brought great advances in high-level feature extraction. The lexicon size of semantic features is believed to reach to 1000 in a few years, and the performances of high-level feature extraction in TRECVID 2005 are generally higher than before [5]. However, the performance of video retrieval hasn't benefited much from using high-level semantic features currently. High-level semantic features are considered not useful except a few topics which have well-performing correlated semantic features in [14], and semantic feature sets are used for only 5% of the interactions of interactive retrieval and contribute to negligible improvement [3].

One of the most important causes to the unsatisfying utilization of semantic features is that the issue of interaction between the semantic feature extraction and the search tasks has not been explored enough. Semantic features are mostly treated as complementary elements of other modalities. In most methods only the highly relevant semantic features are chosen, and then fused with other modalities based on complicated analysis of the whole retrieval system. In practice, it is hard to find highly relevant semantic features for most queries, and it is too rude to neglect any semantic feature which is not highly relevant since it may be also helpful for multi-modal fusion. Therefore, we present a method called pointwise mutual information weighted scheme (PMIWS). The method makes a good judgment of the relevance of all the semantic features to the queries, taking the characteristics of semantic features into account. The method can also be extended for the fusion of multi-modalities. Experiment results based on TRECVID2005 corpus demonstrate the effectiveness of the method.

The remainder of this paper is organized as follows. Section 2 gives a brief review of the related work. Then in Section 3, we present details of our approach. Experiments are presented in Section 4. And finally come the conclusions.

## 2 Related Work

High-level semantic features have been applied in different ways to improve the performance of video retrieval systems. Once semantic features are extracted, shots of dataset can be described by the relevance of them to semantic features, called feature scores. So are queries. With this representation, some machine learning approaches such as SVM are used to classify shots of dataset into two classes: related to the query, not related to the query [6]. The main challenge of this method is that there are a very small number of distinct positive examples and no negative examples. Currently the most popular method is simply the weighted-sum of shot feature scores. In this method, semantic features are assigned with proper weights, which represent the relevance of semantic features to queries. Thus the similarity measurement of shots and a query can be finished by the weighted-sum of shot feature scores.

The techniques for determining weights of features can be divided into three main categories:

1. *Manual methods.* For manual or interactive retrieval, the weights of semantic features can be assigned by human. But as pointed out in [3], it is too ambiguous for people to express a clear, consistent opinion about the relevance of the features to the queries. And it is not realistic while using a large set of semantic features.

2. *Text based methods.* Commonly, the text based methods calculate the feature weights by measuring the similarity between the query text and the concepts' description [14, 15, 16, 18], or using a preprocessed word-concept index [6, 7, 17]. In [15,16], only relevant concepts were utilized. Negative features were also considered in [17] while frequently-used features were added in [18]. More generally, all the semantic features are used in [6, 7, 14].

3. *Semantic feature based methods.* With semantic features already extracted from query as well as shots of dataset, it is natural to make use of them for determining weights of features, yet less work has been done this way [2, 7, 15]. In [2] it is said that the query feature scores are used to construct model vectors followed by appropriate normalization to remove bias and optionally by validity weighting to capture relative concept. But the detail of the normalization is not available. As this kind of methods can benefit more with the improved performances of high-level feature extraction, it needs to be further explored.

To distinguish expected results among hundreds of thousands of different shots, high-level features need to be integrated with other modalities. Though many methods are based on complicated analysis of the whole retrieval system, there are also some valuable works. Iyengar proposed a joint probability model for both the text and the visual components of multimedia documents [8]. Yan proposed to utilize query-class dependent weights within a hierarchical mixture-of-expert framework to combine multiple retrieval results in [10].

### 3 Our Approach

The task of video retrieval can be modelled as follows. Let  $D$  be a specific dataset of video shots,  $Q$  be the collection of queries which represent the user's information needs. For a given query  $q \in Q$ , let  $Y$  be the random variable which represents whether a shot  $d \in D$  meets the information needs described by  $q$ . There are two possible results: meet and not meet, which we can label as  $y_1$  and  $y_0$ . It is hard, if possible, to determine the value of  $Y$  for shots of  $D$ . Actually, the objective of video retrieval is to estimate the probability  $p(Y(d) = y_1)$ ,  $d \in D$ .

To realize this objective, features which describe different attributes of shots are firstly extracted, and then  $p(Y(d) = y_1)$  is estimated using the information provided by the features. In our approach, we model and extract 33 high-level features firstly, and then we present a method called pointwise mutual information weighted scheme (PMIWS) to utilize the information provided by these features. Text retrieval is also performed and the result is fused with semantic features by PMIWS, taking text of the query as another feature  $Q_{text}$ . The overview of our video retrieval system is illustrated in Figure 1.

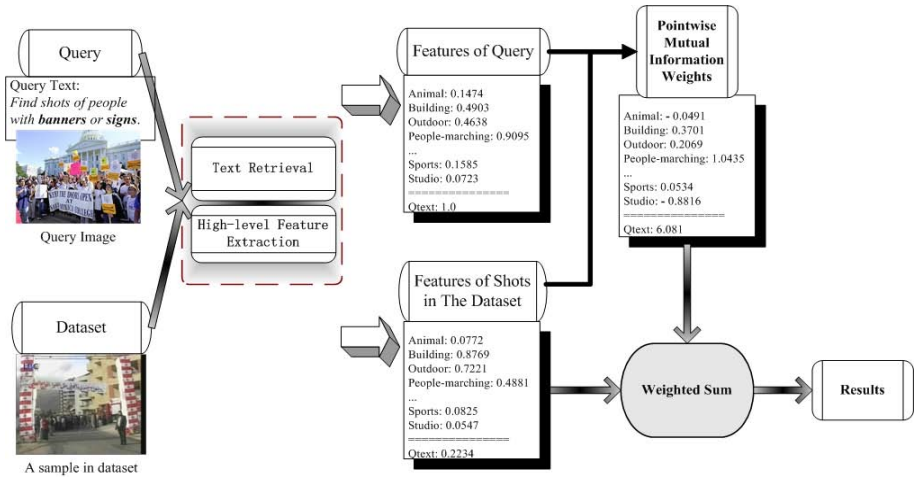


Fig. 1. Overview of the video retrieval system

### 3.1 High-Level Semantic Feature Extraction

Let  $C = \{C_1, C_2, \dots, C_n\}$  be the semantic feature set. For a given feature  $C_i \in C$ , let  $X_i$  be the random variable which represents the existence of  $C_i$  in a shot in  $D$  or a query  $q$ , where  $i \in \{1, 2, \dots, n\}$ . In this case, there are two possible results: exist and not exist, which we can label as  $x_{i1}$  and  $x_{i0}$ . Generally, in the semantic feature extraction, a detector for the semantic feature  $C_i$  is modelled and used to produce confidence scores, which represent the probability  $p(X_i(d) = x_{i1})$ , Where  $d \in D, i \in \{1, 2, \dots, n\}$ .

In our approach, we model and extract 33 high-level features which have support of more than 100 positive samples in the training set of TRECVID2005. The lexicon can refer to [12]. We use SVM as the base classifiers, and propose a Relay Boost approach to fuse the confidence scores produced by the base classifiers. The output of feature  $C_i$  for shot  $d \in D$  is

$$Confidence_i(d) = \sum_j \alpha_j^i \times conf_j^i(d), j \in \{1, 2, \dots, m\}$$

where  $m$  is the number of base classifiers of  $C_i$ ,  $\alpha_j^i$  is the weight of  $j^{th}$  base classifier of  $C_i$ , and  $conf_j^i(d)$  is the confidence score produced by  $j^{th}$  base classifier of  $C_i$ , which are mapped from SVM outputs by Platt's conversion method to represent the probabilities of  $C_i$  existing in  $d$  [13]. Details can be seen in [7]. Divided by the weights of base classifiers of  $C_i$ ,  $Confidence_i(d)$  can be normalized to give an estimation of  $p(X_i(d) = x_{i1})$ , which we refer to as *Feature Score* for  $C_i$ :

$$p(X_i(d) = x_{i1}) = \frac{Confidence_i(d)}{\sum_j \alpha_j^i} = \frac{\sum_j \alpha_j^i \times conf_j^i(d)}{\sum_j \alpha_j^i}, j \in \{1, 2, \dots, m\}$$

The semantic feature extraction is also done for  $q \in Q$ , which gives an estimation of  $p(X_i(q) = x_{i1})$  by the same formula.

### 3.2 Pointwise Mutual Information Weighted Scheme (PMIWS)

With semantic features extracted,  $p(Y(d) = y_1)$  can be estimated using the information provided by semantic features. We use pointwise mutual information to fulfill this work.

Originally, if we draw one sample of  $D$  at random, the entropy or uncertainty of  $Y$  is defined in terms of prior probabilities using Shannon’s definition:

$$H(Y) = - \sum_y p(y) \cdot \log(p(y)), y \in \{y_1, y_0\}.$$

After having observed  $X_i$ , the uncertainty of  $Y$  is the conditional entropy:

$$H(Y|X_i) = - \sum_{x_i} \sum_y p(y, x_i) \cdot \log(p(y|x_i)), y \in \{y_1, y_0\}, x_i \in \{x_{i1}, x_{i0}\}$$

Then the reduction in uncertainty of  $Y$  due to knowing about  $X_i$  is called mutual information:

$$I(Y, X_i) = H(Y) - H(Y|X_i) = \sum_{x_i, y} p(x_i, y) \cdot \log \frac{p(x_i, y)}{p(x_i)p(y)}, y \in \{y_1, y_0\}, x_i \in \{x_{i1}, x_{i0}\}$$

Furthermore, mutual information between two particular points is defined as pointwise mutual information [4]:

$$I(y, x_i) = \log \frac{p(x_i, y)}{p(x_i)p(y)} = \log \frac{p(x_i|y)}{p(x_i)}, y \in \{y_1, y_0\}, x_i \in \{x_{i1}, x_{i0}\}$$

The pointwise mutual information can be regarded as the amount of information  $x_i$  contains about  $y$ . The magnitude of  $I(y, x_i)$  indicates the power of influence of event  $\{X = x_i\}$  to event  $\{Y = y\}$ , while the sign of  $I(y, x_i)$  indicates the direction of influence of event  $\{X = x_i\}$  to event  $\{Y = y\}$ , i.e., increasing or decreasing the confidence of event  $\{Y = y\}$  happening. It equals to zero when  $\{X = x_i\}$  and  $\{Y = y\}$  are independent. Thus, we can use  $I(y_i, x_{i1})$  as the weight of semantic feature  $C_i$  with linear normalization:

$$weight_{C_i} = \text{Normalize}(I(y_1, x_{i1})) = \alpha \cdot \log \frac{p(x_{i1}|y_1)}{p(x_{i1})}$$

The normalization coefficient  $\alpha$  is the same for all the semantic features. And then we can estimate  $p(Y(d) = y_1)$  by the weight sum of  $p(X_i(d) = x_{i1})$ , where  $i \in \{1, 2, \dots, 33\}$ :

$$p(Y(d) = y_1) = \sum_i weight_{C_i} \cdot p(X_i(d) = x_{i1}) = \alpha \cdot \sum_i \log \frac{p(x_{i1}|y_1)}{p(x_{i1})} \cdot p(X_i(d) = x_{i1})$$

Here,  $p(X_i(d) = x_{i1})$  has been estimated by semantic feature extraction.  $p(x_{i1}) = \text{Mean}_{d \in D}(p(X_i(d) = x_{i1}))$ . And  $p(x_{i1}|y_1) \equiv p(X_i(d) = x_{i1}|Y(d) = y_1)$  can be approximated by  $p(X_i(q) = x_{i1})$ , since  $q$  certainly satisfies  $Y(q) = y_1$ .

The pointwise mutual information weight has grasped two principal issues of high-level semantic features in its expression, i.e., importance and reliability. Firstly, it brings in a term of  $p(x_{i1})$ , which expresses the underlying importance of different features compared with  $p(x_{i1}|y_1)$ . It is biased towards infrequent semantic features existing in  $q$ . For example, if there is a person and a ship shown in the example of  $q$ , then the feature of "ship" is believed to be more important as it is less frequent. However, if the person is George Bush, then the feature of "George Bush", if modelled, is believed to be more important. That is consistent with common usage. But there is also a risk that unreliable detectors of features with less frequency maybe mislead the retrieval. So secondly, it brings in the log factor, which makes it more robust to the unreliable detectors.

### 3.3 Fusion with Text Retrieval Result

The PMIWS method can not only handle the fusion of multiple semantic features, but also be easy to extend to the fusion of multi-modalities by treating other modalities as one or several kinds of high-level features. We will describe the fusion of high-level semantic features and text retrieval result below.

Our text retrieval system is based on an OKAPI-TF formula using the transcripts from the ASR/MT output provided by NIST. Pseudo feedback is also performed. For shot  $d$  the text retrieval system will give a score  $T(d)$ . Details can be seen in [7]. We treat text of the query as a high-level feature  $Q_{text}$ . Let  $X_{text}$  be the random variable which represents the existence of  $Q_{text}$  in a shot of  $D$ . There are two possible results: exist and not exist, which we can label as  $t_1$  and  $t_0$ . Similar to aforementioned analysis, we can use  $I(y_1, t_1)$  as the weight of  $Q_{text}$  by linear normalization:

$$\text{weight}_{Q_{text}} = \text{Normalize}(I(y_1, t_1)) = \beta \cdot \log \frac{p(t_1|y_1)}{p(t_1)}$$

The normalization coefficient  $\beta$  is the same for all the semantic features in place of  $\alpha$ . Then  $p(Y(d) = y_1)$  can be refined by adding one term of  $t_1$  ( $i \in \{1, 2, \dots, 33\}$ ):

$$p(Y(d) = y_1) = \beta \cdot \left\{ \sum_i \log \frac{p(x_{i1}|y_1)}{p(x_{i1})} \cdot p(X_i(d) = x_{i1}) + \log \frac{p(t_1|y_1)}{p(t_1)} \cdot p(X_{text}(d) = t_1) \right\}$$

Here,  $p(X_{text}(d) = t_1)$  is estimated by normalizing the text retrieval result  $T(d)$  to 0-1, according to the minimum and maximum value of  $T(d)$  in the dataset. And  $p(t_1) = \text{Mean}_{d \in D}(p(X_{text}(d) = t_1))$ ,  $d \in D$ . While  $p(t_1|y_1)$  is assumed to be 1.  $\beta$  can be omitted as it doesn't affect the result for final ranking.

## 4 Experiments

### 4.1 Dataset and Evaluation

The experiments are performed on TRECVID2005 dataset provided by NIST. The total amount of news video for the evaluated tasks is about 169 hours, in MPEG-1 format: 43 in Arabic, 52 in Chinese, 74 in English [5]. About 160 hours of them is used for Search Benchmark, the earlier half as development data, and the later half as test data. The data is divided into shots, which are the basic units of video retrieval. For each shot, more than one keyframe is extracted for performing the high-level feature extraction. NIST also provides the English ASR output and machine-translated transcripts for those non English video materials, which are used for text retrieval. We use the 24 multimedia search topics developed by NIST for our experiments. Details of the topics can be seen in [5].

The performances are evaluated by non-interpolated average precision (AP) and mean average precision (MAP) criteria. Non-interpolated average precision is calculated by computing the precision after every retrieved relevant shot and then averaging these precisions over the total number of retrieved relevant/correct shots in the collection for that topic/feature or the maximum allowed result set (whichever is smaller). Average precision favors highly ranked relevant documents. It allows comparison of result sets of different sizes. The topic averages are averaged across all topics to create the non-interpolated mean average precision (MAP). See the TREC-10 Proceedings appendix on common evaluation measures for more information [9].

### 4.2 Experiment Results

We present four runs of automatic retrieval. The descriptions of the four runs are as follows and the evaluation results are shown in Figure 2, compared with the median effect of automatic retrievals in TRECVID2005.

Run1: This run uses only text retrieval on ASR/MT.

Run2: This run uses high-level feature score weighted sum method, using the feature scores of queries as weights directly.

Run3: This run uses high-level feature score weighted sum method, using PMIWS to calculate weights.

Run4: This run integrates high-level features and text retrieval result with PMIWS.

Observed from Figure 2, using point mutual information as weights of high-level features (Run3) is obviously better than using the feature scores of queries as weights (Run2). We can also find that high-level features are usable for topics which have correlated specific concepts. If the correlation is tight and the correlated specific concepts are well-performing, the results of using high-level features (Run3) can be even better than results of text retrieval (Run1) like 0155(map), 0165(basketball), 0168(road, car), 0170(building), 0171(goal).

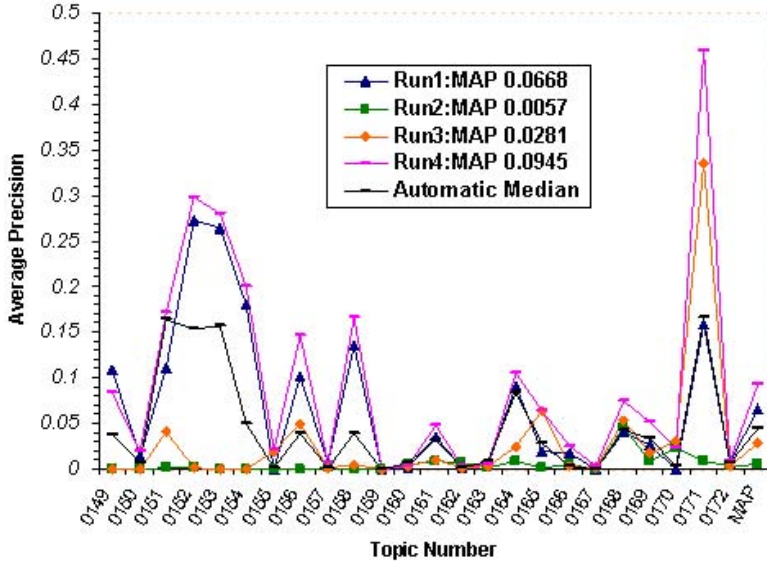


Fig. 2. Performance evaluation results

This validates the effectiveness of PMIWS for judging the relevance of high-level features to queries.

The comparison of individual retrieval runs (Run1 and Run3) and fusion runs (Run4) are also clear in Figure 2. For almost all of the topics, the result of fusion (Run4) is better than the results of using text only (Run1) and using high-level features only (Run3). Note for some topics like the named topics, high-level features are useless by themselves but useful for fusion. This can be explained by regarding high-level features as revisers, which correct the temporal mismatch of the text and the visual content of shots. The MAP of fusion (Run4) is 41.5% higher than using text only (Run1) and 236.3% higher than using high-level features only (Run3). And it is much better than the median effect of automatic retrievals in TRECVID2005 for all topics. This exhibits the effectiveness of PMIWS for multi-modal fusion.

## 5 Conclusions and Future Works

In this paper, we investigate the issue of how to make good use of high-level semantic features in video retrieval. We focus on determining the relevance of all the semantic features to the query. To achieve this, a method called point-wise mutual information weighted scheme (PMIWS) is presented, which has the following advantages:

- 1) The method can reflect the exact relevance of semantic features to queries and assigned reasonable weights, by considering the prior distributions of high-level features and referring to information theory.

2) The method gives an integrated view of fusing semantic features and text. Experiments demonstrate that the fusion retrieval has great improvements on individual retrievals. The idea can be extended for other modalities.

3) The weights of semantic features are calculated automatically based on the semantic feature extraction. It is scalable with the advances of high-level feature extraction.

In the future, we will focus our work on the following aspects to make this method a more effective and practical one:

1) The performance of our high-level feature extraction is not adequately well, which greatly restricts the search performance. Even for the positive samples and a correlated high-level feature of the same query, the feature extraction results are not always consistent. So we have to further improve the performance of high-level feature extraction. Meanwhile we should consider the radiabilities of semantic feature detectors more carefully.

2) Taking text of the query as another feature  $Q_{text}$  is somewhat crude. We will go further to analyze text of the query and then get more meaningful text features. We will also use PMIWS to fuse the other modalities.

3) Lack of positive examples remains a big problem. Interactive retrieval can be a great help to this with human interaction, and we will investigate the application of PMIWS in an interactive setting to achieve better results.

## Acknowledgement

This work is supported by National Natural Science Foundation of China (60135010), National Natural Science Foundation of China(60321002) and the Chinese National Key Foundation Research & Development Plan (2004CB318108).

## References

1. AWM Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.12, pp.1349-1380, December 2000
2. Apostol Natsev, Milind R. Naphade, John R. Smith: Semantic representation: search and mining of multimedia content. *KDD 2004*: 641-646
3. Michael G. Christel and Alexander G. Hauptmann: The Use and Utility of High-Level Semantic Features in Video Retrieval. *CIVR 2005*, 134-144
4. C. D. Manning and H. Schütze : *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999
5. Paul Over, Tsveta Ianeva, Wessel Kraaij, Alan Smeaton: *TRECVID 2005 - An Introduction*. Proceedings of TRECVID2005
6. Arnon Amir, Janne Argillandery, Murray Campbellz, Alexander Hauboldz, Giridharan Iyengar, Shahram Ebadollahiz, Feng Kangz, Milind R. Naphadez, Apostol (Paul) Natsevz, John R. Smithz, Jelena Tesicz, Timo Volkmer: IBM Research TRECVID-2005 Video Retrieval System. Proceedings of TRECVID2005

7. Jinhui Yuan, Huiyi Wang, Lan Xiao, Dong Wang, Dayong Ding, Yuanyuan Zuo, Zijian Tong, Xiaobing Liu, Shuping Xu, Wujie Zheng, Xirong Li, Zhangzhang Si, Jianmin Li, Fuzong Lin, Bo Zhang: Tsinghua University at TRECVID 2005. Proceedings of TRECVID2005
8. G.Iyengar, P. Duygulu, S. Feng, P. Ircing, SP Khudanpur, D. Klakow, MR Krause, R.Manmatha, HJ Nock, D. Petkova, B. Pytlik, P. Virga Pages: Joint Visual-Text Modeling for Automatic Retrieval of Multimedia Documents. Proceedings of the 13th ACM international conference on Multimedia, 2005, 21-30
9. TREC-10 Proceedings Appendix on Common Evaluation Measures. <http://trec.nist.gov/pubs/trec10/appendices/measures.pdf>
10. R. Yan, J. Yang, and A. G. Hauptmann: Learning query-class dependent weights in automatic video retrieval. In Proceedings of ACM Multimedia 2004: 548-555, Oct. 2004.
11. Horst Eidenberger, C. Breiteneder: Semantic Feature Layers in Content-Based Image Retrieval. Proceedings IEEE International Conference on Control, Automation, Robotic and Vision, Singapore, 2002
12. Milind R. Naphade, Lyndon Kennedy, John R. Kender, Shih-Fu Chang, John R. Smith, Paul Over, Alex Hauptmann: A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. IBM Research Report RC23612 (W0505-104), May, 2005
13. J. Platt: Probabilities for SV machines. In Advances in Large Margin Classifiers, pages 61-74. MIT Press, 2000.
14. Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lexing Xie, Akira Yanagawa, Eric Zavesky, Dong-Qing Zhang: Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. Proceedings of TRECVID2005
15. Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao and Gang Wang: TRECVID 2005 by NUS PRIS. Proceedings of TRECVID2005
16. C.G.M. Snoek, J.C. van Gemert, J.M. Geusebroek, B. Huurnink, D.C. Koelma, G.P. Nguyen, O. de Rooij, F.J. Seinstra, A.W.M. Smeulders, C.J. Veenman, M. Worring: The MediaMill TRECVID 2005 Semantic Video Search Engine. Proceedings of TRECVID2005
17. Markus Koskela, Jorma Laaksonen, Mats Sjoberg, Hannes Muurinen: PicSOM Experiments in TRECVID 2005. Proceedings of TRECVID2005
18. A.G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, Y. Zhang: CMU Informedia's TRECVID 2005 Skirmishes. Proceedings of TRECVID2005