Multimodal Video Concept Detection via Bag of Auditory Words and Multiple Kernel Learning

Markus Mühling, Ralph Ewerth, Jun Zhou, and Bernd Freisleben

Department of Mathematics & Computer Science, University of Marburg
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{muehling,ewerth,zhouj,freisleb}@informatik.uni-marburg.de

Abstract. State-of-the-art systems for video concept detection mainly rely on visual features. Some previous approaches have also included audio features, either using low-level features such as mel-frequency cepstral coefficients (MFCC) or exploiting the detection of specific audio concepts. In this paper, we investigate a bag of auditory words (BoAW) approach that models MFCC features in an auditory vocabulary. The resulting BoAW features are combined with state-of-the-art visual features via multiple kernel learning (MKL). Experiments on a large set of 101 video concepts from the MediaMill Challenge show the effectiveness of using BoAW features: The system using BoAW features and a support vector machine with a χ^2-kernel is superior to a state-of-the-art audio approach relying on probabilistic latent semantic indexing. Furthermore, it is shown that an early fusion approach degrades detection performance, whereas the combination of auditory and visual bag of words features via MKL yields a relative performance improvement of 9%.

Keywords: Visual concept detection, video retrieval, bag of words, bag of auditory words, audio codebook, multiple kernel learning.

1 Introduction

The detection of audiovisual concepts in video shots is an essential prerequisite for semantic video retrieval, navigation and browsing. State-of-the-art systems concentrate on high-level features serving as intermediate descriptions to bridge the "semantic gap" between data representation and human interpretation. Hauptmann et al. [6] stated that less than 5000 concepts, detected with a minimum accuracy of 10% mean average precision, are sufficient to provide search results comparable to text retrieval in the World Wide Web.

Current approaches mainly focus on visual features based on local keypoints and scale-invariant feature transform (SIFT) descriptors [15] that currently achieve top performance in visual recognition tasks. Such descriptors are clustered to create a visual vocabulary (codebook), where the cluster centers are regarded as "visual words". Similar to the representation of documents in the field of text retrieval, an image or a video shot can then be represented as a bag

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 40–50, 2012.

of visual words (BoVW) by mapping local descriptors to the visual vocabulary. In some previous approaches, audio features are used for visual concept detection, either by using low-level features such as mel-frequency cepstral coefficients (MFCCs) or by using detection results of specific audio events such as silence, speech, music and noise as mid-level features for subsequent training of video concept classifiers.

In this paper, we leverage the bag of words approach for audio features to enhance video concept detection and propose multiple kernel learning (MKL) as the appropriate fusion scheme for these bag of auditory words (BoAW) and state-of-the-art BoVW features. First, MFCC audio features are extracted from each video shot. Then, an auditory vocabulary is created via k-means clustering. This vocabulary or codebook, respectively, is then exploited to describe and represent a shot via a histogram (bag) of auditory words. These histograms are used to train audio models for video concepts using support vector machines (SVM) and to finally classify video shots based on these models. Experimental results show that a χ^2-kernel is more appropriate for BoAW features than a radial basis function (RBF) kernel, and the proposed system relying on the auditory vocabulary significantly outperforms a state-of-the-art approach that uses probabilistic latent semantic indexing (pLSA). In addition, BoAW features are combined with state-of-the-art visual features (visual vocabulary based on dense sampled SIFT descriptors) via MKL. In contrast to an early fusion approach, the system relying on MKL for fusing auditory and visual features clearly improves a state-of-the-art concept detection system.

The paper is organized as follows. Section 2 discusses related work. Section 3 describes the construction of the auditory vocabulary and the multimodal concept detection system. Experimental results are presented in section 4. Section 5 concludes the paper and outlines areas for future research.

2 Related Work

In recent years, researchers have shifted their attention to generic video concept detection systems, since the development of specialized detectors for hundreds or thousands of concepts seems to be infeasible. Continuous progress has been reported in the field of visual concept detection using bag of (visual) words approaches (BoW). The top 5 official runs at the TRECVid 2010 semantic indexing task rely on BoW representations [20].

In addition to the visual modality, the audio signal of videos carries important information that can help to improve the performance of generic video concept detection systems. Most of the approaches that incorporate audio information directly use additional low-level features such as MFCCs, ΔMFCCs, pitch, zero-crossing rate, energy, or log-power to classify semantic concepts [1][5][14]. For example, Bredin et al. [1] have extracted low-level features including MFCCs and their derivatives to build Gaussian mixture models (GMM) for each of the semantic concepts.

In other approaches, the results of audio event detectors are used as additional mid-level features. Besides acoustic events such as speech, non-speech, background and gender, Snoek et al. [21] detected the occurrence of 16 additional audio events such as "child-laughter", "baby-crying", "airplane-propeller", "sirens", "traffic-noise", "car-engine", "dog-barking", or "applause" and used the results as additional inputs for concept classifiers like SVMs. Inspired by classical text document analysis, Lu and Hanjalic [16] try to automatically determine these audio elements by regarding them as natural clusters of the audio data. Between 2 and 20 elements are discovered per audio document using an iterative spectral clustering method.

The audio concept classification framework used by Feki et al. [4] first removes segments of silence and then separates the audio signal into speech, music and environmental sound. The environmental sound segments are further classified using a time-frequency analysis based on MFCC features. For video concept detection, visual features and the previously described audio classification results are fed into a fuzzy reasoning system to fuse the different modalities [3].

Jiang et al. [9] have introduced a novel representation called short-term audio-visual atoms. Audio features based on a matching pursuit representation [17] of the audio signal and region-based color, texture, edge, and motion features are combined and a joint audio-visual codebook is built using multiple instance learning.

Inoue et al. [7] have used a statistical framework to combine visual and audio features for video concept detection. The distribution of SIFT descriptors for each shot are described by GMMs, and a SVM with a GMM-kernel that compares GMMs was used for training and classification. In addition, hidden Markov models (HMM) were built for each concept based on audio features, including MFCCs, log-power and the correponding derivatives. The final classification result is a weighted combination of log likelihood ratios from the audio models and from the SIFT GMMs. Using additional audio features, the results for 20 semantic concepts on documentary films could be improved from 15% mean average precision to 16.4%. At the TRECVid [20] evaluation in 2010, the GMM kernel was also applied for MFCC features [8] resulting in a noticeable relative performance improvement for several concepts like "singing", "dancing", "cheering" or "animal".

Peng et al. [18] have proposed a method that performs an audio-only analysis of the video data and investigates the use of an audio pLSA model for video concept detection. An audio vocabulary based on MFCC features from acoustically homogenous segments is built and the latent audio topics are discovered using pLSA. Each shot is then described by the probabilities of the discovered latent topics and classified by a SVM. Results are reported on 85 hours of news videos for 10 concepts from the MediaMill Challenge. Diou et al. [2] have combined BoW audio features based on MFCCs with visual features in an early fusion scheme. However, for the 30 evaluated concepts of the TRECVid 2010 semantic indexing task, the additional use of BoW audio features clearly decreased the performance from 4.5% to 3.5% mean inferred average precision.

The bag of auditory words approach has recently been successfully applied in the fields of music information retrieval and multimedia event detection. Riley et al. [19] have represented songs as a bag of auditory words showing robust results for a variety of signal distortions and Jiang et al. [12] combined bag of words representations for audio and visual features using a late fusion scheme to detect events like "making a cake" or "assembling a shelter".

3 Concept Detection System

In this section, our approach for multimodal video concept detection is presented. We describe the BoAW approach and the MKL framework that is proposed as an appropriate fusion scheme for the combination of BoAW and BoVW features. The application of the BoW representation to auditory features is presented in Section 3.1. SVMs have proven to be powerful for visual concept detection [20] and they are used to build audio models and to classify video shots based on these models. Besides the linear and the RBF-kernel, the χ^2-kernel is applied due to the representation of features as histograms. The used kernel functions for the SVMs are described in Section 3.2. The state-of-the-art visual features and the proposed MKL framework to combine the feature representations of both modalities are presented in Section 3.3.

3.1 Bag of Auditory Words

Since the BoW representation based on local SIFT descriptors achieves superior performance in the field of visual concept detection [20], we leverage the BoW paradigm for audio features. Using a time-frequency analysis of the audio signal, 12-order MFCCs (Mel-Frequency Cepstral Coefficients) are extracted from audio frames of 20 ms length with an overlap of 50%. Thus a video shot is represented as a set of 12-dimensional MFCC vectors. Based on these MFCC vectors, an auditory vocabulary is generated using the k-means clustering algorithm, and the final cluster centers can be interpreted as "auditory words". Similar to the representation of documents in the field of text retrieval, a video shot can then be represented as a bag of auditory words that are the results of a vector quantization process using the generated vocabulary or codebook, respectively. Finally a shot is described as a histogram, counting the occurences of auditory words. To diminish the quantization loss during histogram generation, a soft-weighting scheme similar to the one proposed by Jiang et al. [10] is used. Instead of mapping a MFCC vector only to its nearest neighbor, the top K nearest auditory words are selected. Using an auditory vocabulary of N auditory words, the importance of an auditory word t in a shot is represented by the weights of the resulting histogram bins $w = [w_1, \ldots, w_t, \ldots, w_N]$ with

$$w_t = \sum_{i=1}^{K} \sum_{j=1}^{M_i} sim(j, t), \tag{1}$$

where M_i is the number of MFCC vectors whose i-th nearest neighbor is the auditory word t.

3.2 Kernel Choice

Since SVMs are used to train audio models and to finally classify video shots, a kernel function needs to be specified. A kernel function intuitively measures the similarity between two data instances. Commonly used kernels are the linear and the radial basis function (RBF) kernel:

$$k_{linear}(x, y) = x^T y, \tag{2}$$

$$k_{RBF}(x, y) = e^{-\gamma \sum_i (x_i - y_i)^2}. \tag{3}$$

Since histogram representations are used in the proposed approach, we also apply the χ^2-kernel. It is based on the corresponding histogram distance:

$$k_{\chi^2}(x, y) = e^{-\gamma \chi^2(x,y)} \quad with \quad \chi^2(x, y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}. \tag{4}$$

Jiang et al. [11] have used the χ^2-kernel successfully for BoVW features in the context of visual concept detection. In their study, the χ^2-kernel has outperformed both the linear and the RBF-kernel.

3.3 Multimodal Fusion

In a multimodal fusion setting, BoAW features are combined with state-of-the-art visual features. For visual features we use the BoVW representation and extract densely sampled local SIFT descriptors from the keyframes using the implementation of the Vision Lab Features Library (VLFEAT) [24]. Color information is integrated using RGB-SIFT, where the SIFT descriptors are computed independently for the three channels of the RGB color model (red, green, blue). Thus, the final feature vector is the concatenation of the individual descriptors. Based on these local descriptors, a global visual vocabulary is generated using the k-means algorithm. Each keyframe or shot, respectively, is described as a histogram indicating the presence of each "visual word". Again, the previously described soft-weighting scheme is applied to consider the similarities of the local descriptors to the codebook entries.

The easiest way to combine BoAW and BoVW features is the early fusion scheme. Using this method, visual and audio features are simply concatenated and directly fed into a SVM. A more sophisticated approach to combine the capabilities of different modalities is MKL. It is applied to find an optimal kernel weighting

$$k_{multimodal} = \alpha \cdot k_{audio} + \beta \cdot k_{visual} \quad with\ \alpha \geq 0,\ \beta \geq 0 \tag{5}$$

where the kernel functions k_{audio} and k_{visual} take both feature modalities into account. We use the l_2-norm to control the sparsity of the weights α and β for audio and visual features, respectively. Throughout our experiments, we use the MKL framework provided by the Shogun library [23] in combination with the SVM implementation of Joachims [13], called SVM^{light}.

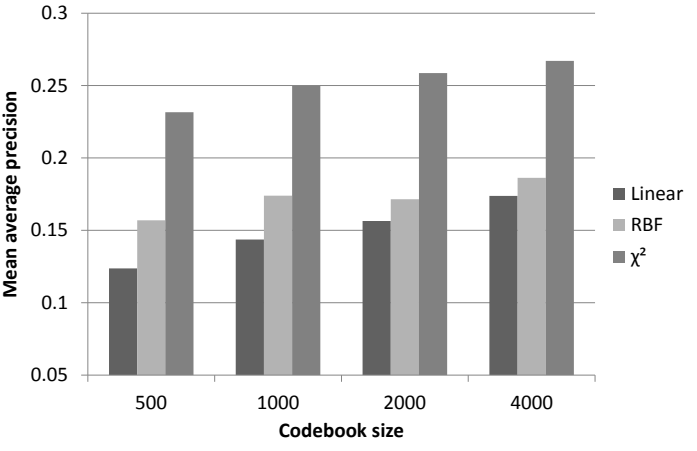

Fig. 1. Performance evaluation of different kernel functions and codebook sizes using BoAW features

4 Experimental Results

In this section, the performance impact of BoAW features in the field of video concept detection is investigated. For this purpose, the MediaMill Challenge [22] is used. It offers a dataset based on the TRECVid 2005 [20] training set with an extensive set of 101 annotated concepts, including objects, scenes, events and personalities. It consists of 86 hours of news videos containing 43,907 completely annotated video shots. These shots are divided into a training set of 30,993 shots and a test set of 12,914 shots.

4.1 Evaluation Criteria

To evaluate the concept retrieval results, the measure of average precision (AP) is used. For each concept, the implemented system returns a list of ranked shots, which is used to calculate the average precision as follows:

$$AP(\rho) = \frac{1}{|R|} \sum_{k=1}^{N} \frac{\left|R \cap \rho^k\right|}{k} \psi(i_k) \tag{6}$$

where $\rho^k = i_1, i_2, \ldots, i_k$ is the ranked shot list up to rank k, N is the length of the ranked shot list, R ist the set of relevant shots and $\left|R \cap \rho^k\right|$ is the number of relevant shots in the top k of ρ. The function $\psi(i_k) = 1$ if $i_k \in R$ and 0 otherwise. To evaluate the overall performance, the mean AP score is calculated by taking the mean value of the average precisions for the individual concepts. Furthermore, the official partial randomization test in the TRECVid evaluation [20] is used to determine whether our system is significantly better than a reference system, or whether the difference is only due to chance.

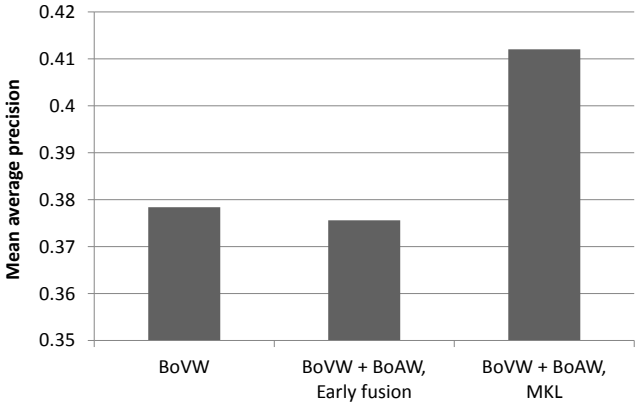

Fig. 2. Performance evaluation of BoAW features in a multimodel setting using early fusion and MKL

4.2 Results

We have performed several experiments to investigate the performance impact of BoAW features both alone and in combination with visual features.

In a first experiment based on an audio-only analysis of the data, different auditory vocabulary sizes and kernel methods have been taken into account. We have compared the linear, RBF and χ^2-kernel in combination with codebook sizes between 500 and 4000 auditory words. The experimental results are presented in Figure 1. The χ^2-kernel significantly outperformes the linear as well as the RBF-kernel. Using 4000 auditory words, the χ^2-kernel yields 43.3% improvement compared to the RBF-kernel. A larger vocabulary also has a positive impact on the overall performance. In combination with the χ^2-kernel, a vocabulary size of 4000 auditory words achieves a mean AP of 26.7% compared to 23.2% for 500 words. Based on these results, the χ^2-kernel and a vocabulary size of 4000 auditory words are used exclusively in the experimental evaluations below.

In a second experiment, we have investigated the impact of BoAW features in a multimodal concept detection system. The state-of-the-art baseline system performs a visual-only analysis of the data using dense sampled RGB-SIFT descriptors with a vocabulary of 4000 "visual words". Both modalities, visual and audio features, are combined using MKL and by using a simple early fusion scheme. In order to save computation time, we have trained the models using a reduced number of negative training samples per concept. The results of the two different fusion strategies are presented in Figure 2. While the early fusion strategy causes a slight performance decrease, the fusion of visual and audio features via MKL achieves a relative performance improvement of 8.9% compared to the baseline system. In total, 31 concepts yield a relative performance improvement of more than 10%. In particular, the concepts representing personalities profit

Table 1. Performance comparison between the visual only baseline system and the multimodal system using MKL, showing average precision values of concepts with relative performance improvements of at least 18%

AP [%]	BoVW	BoVW+BoAW
Motorbike	0.3	4.1
Cycling	13.7	91.7
Racing	11.4	52.3
Bicycle	17.6	80.0
Baseball	0.7	1.6
Natural disaster	8.7	18.0
Boat	18.3	34.1
Golf	36.4	51.3
Waterbody	36.9	49.4
Aircraft	16.6	21.8
Football	54.6	70.6
River	69.9	89.8
Entertainment	55.0	70.2
Sports	49.3	62.2
Table	10.9	13.7
Food	52.6	64.2
Basketball	54.6	65.7
Soccer	72.4	85.6

from the additional audio features, increasing the mean AP for this group of concepts from 9.2% to 11.1%. Further concepts with relative improvements of at least 18% are shown in Table 1.

4.3 Discussion

The experiments indicate that the kernel choice is a critical decision for the performance of the BoAW approach. While the RBF-kernel concentrates on the largest histogram differences due to the quadratic exponential decay, the χ^2-kernel considers the bins more equally. This seems to be beneficial regarding the large intra-class variations of audio signals. Keeping in mind that the ground truth annotation of the 101 semantic concepts is based upon a visual inspection of the video shots, the BoAW approach achieves an impressive performance of 26.7% mean AP on the MediaMill Challenge. The performance is even significantly better than the baseline system provided by the MediaMill Challenge with 21.6% mean AP, which uses local as well as global texture information. The state-of-the-art approach of Peng et al. [18] relying on audio pLSA attained a mean AP of approximately 20.7% on a subset of 10 concepts from the MediaMill Challenge. On the same subset, we achieve a superior performance of 26.8% mean AP using BoAW features, yielding a relative improvement of approximately 30%. Besides the mean AP, Peng et al. displayed AP scores for half of the ten concepts. For these concepts, performance comparisons between the BoAW method and the audio pLSA approach are shown in Figure 3.

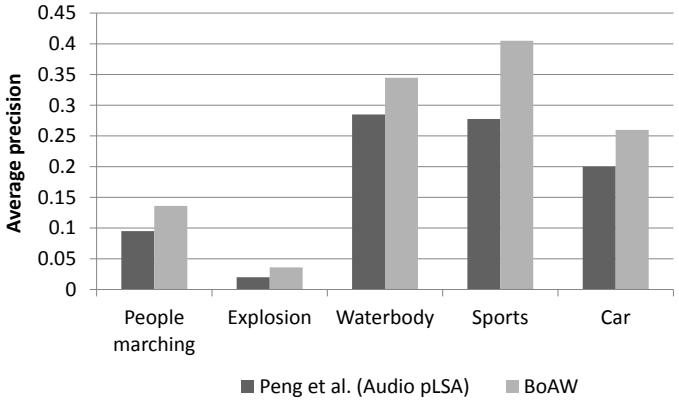

Fig. 3. Performance comparison between the BoAW method and the audio pLSA approach of Peng et al. [18]

Via MKL, additional BoAW features clearly improve the performance of a state-of-the-art video concept detection system that relies on visual features only. The weak performance of the early fusion strategy confirms the results of Diou et al. [2] at the TRECVid 2010 challenge, where the additional use of BoW audio features in an early fusion scheme clearly decreased the performance. This is not surprising, since audio information is more or less important depending on the semantic concept. While "racing" or "motorbike", for example, are characterized by engine noise, there is no discriminative audio information for concepts such as "house" or "gras". In this case, audio features can be even misleading for the classification process. MKL instead of early fusion learns optimized kernel weights that provide information about the relevance of both modalities for the discrimination of semantic concept classes. Hence, audio features are more or less considered depending on the corresponding concept.

5 Conclusions

In this paper, we have presented a bag of auditory words approach for video concept detection that models MFCC features in an auditory vocabulary. This vocabulary is used to describe video shots via histograms of auditory words. SVMs are employed to build the audio models and to finally classify the video shots. Experimental results on a large set of 101 semantic video concepts have shown the effectiveness of the proposed approach. Using BoAW features in combination with the χ^2-kernel yields almost 45% improvement compared to the RBF-kernel.

The proposed system relying on BoAW features outperforms a state-of-the-art audio approach that uses pLSA [18] and is even significantly better than the baseline system provided by the MediaMill Challenge, which used local as well as global texture features.

Furthermore, the resulting BoAW features are combined with visual features via MKL. Using MKL instead of an early fusion scheme significantly improves the results of a state-of-the-art video concept detection system that relies on visual features only.

Areas for future work are the integration of temporal information beyond the scope of audio frames and the investigation of features based on the matching pursuit method instead of MFCCs.

Acknowledgements. This work is supported by the German Ministry of Education and Research (BMBF, D-Grid) and by the German Research Foundation (DFG, PAK 509).

References

1. Bredin, H., Koenig, L., Farinas, J.: IRIT @ TRECVid 2010: Hidden Markov Models for Context-aware Late Fusion of Multiple Audio Classifiers. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
2. Diou, C., Stephanopoulos, G., Delopoulos, A.: The Multimedia Understanding Group at TRECVID-2010. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
3. Elleuch, N., Zarka, M., Feki, I., Ammar, A.B.E.N., Alimi, A.M.: REGIMVID at TRECVID 2010: Semantic Indexing. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
4. Feki, I., Ammar, A.B., Alimi, A.M.: Audio Stream Analysis for Environmental Sound Classification. In: International Conference on Multimedia Computing and Systems (2011)
5. Gorisse, D., Precioso, F., Gosselin, P., Granjon, L., Pellerin, D., Rombaut, M., Bredin, H., Koenig, L., Lachambre, H., Khoury, E.E., Vieux, R., Mansencal, B., Zhou, Y., Benois-Pineau, J., Jégou, H., Ayache, S., Safadi, B., Quénot, G., Benoît, A., Lambert, P.: IRIM at TRECVID 2010: Semantic Indexing and Instance Search. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
6. Hauptmann, A., Yan, R., Lin, W.H.: How Many High-Level Concepts Will Fill the Semantic Gap in News Video Retrieval? In: International Conference on Image and Video Retrieval, pp. 627–634. ACM, New York (2007)
7. Inoue, N., Saito, T., Shinoda, K., Furui, S.: High-Level Feature Extraction Using SIFT GMMs and Audio Models. In: 20th International Conference on Pattern Recognition, pp. 3220–3223. IEEE (2010)
8. Inoue, N., Wada, T., Kamishima, Y., Shinoda, K., Kim, I., Byun, B., Lee, C.H.: TT+GT at TRECVID 2010 Workshop. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
9. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.: Short-Term Audio-Visual Atoms for Generic Video Concept Classification. In: 17th ACM International Conference on Multimedia, pp. 5–14. ACM Press, New York (2009)
10. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In: International Conference on Image and Video Retrieval, pp. 494–501. ACM, New York (2007)
11. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.G.: Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study. IEEE Transactions on Multimedia 12, 42–53 (2010)

12. Jiang, Y.G., Zeng, X., Ye, G., Bhattacharya, S., Ellis, D., Shah, M., Chang, S.F.: Columbia-UCF TRECVID 2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
13. Joachims, T.: Text Categorization With Support Vector Machines: Learning With Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
14. Li, H., Bao, L., Gao, Z., Overwijk, A., Liu, W., Zhang, L.F., Shoou-I, Y., Chen, M.Y., Florian, M., Hauptmann, A.: Informedia @ TRECVID 2010. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2010 (2010)
15. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
16. Lu, L., Hanjalic, A.: Audio Keywords Discovery for Text-Like Audio Content Analysis and Retrieval. IEEE Transactions on Multimedia 10(1), 74–85 (2008)
17. Mallat, S., Zhang, Z.: Matching Pursuits With Time-Frequency Dictionaries. IEEE Transactions on Signal Processing 41(12), 3397–3415 (1993)
18. Peng, Y., Lu, Z., Xiao, J.: Semantic Concept Annotation Based on Audio PLSA Model. In: 17th ACM International Conference on Multimedia (MM 2009), pp. 841–844. ACM Press, New York (2009)
19. Riley, M., Heinen, E., Ghosh, J.: A Text Retrieval Approach to Content-based Audio Retrieval. In: 9th International Conference of Music Information Retrieval, pp. 295–300 (2008)
20. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation Campaigns and TRECVid. In: 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York (2006)
21. Snoek, C.G.M., van de Sande, K.E.A., Rooij, O.D., Huurnink, B., Uijlings, J.R.R., Liempt, M.V., Bugalho, M., Trancoso, I., Yan, F., Tahir, M.A., Mikolajczyk, K., Kittler, J., de Rijke, M., Geusebroek, J.M., Gevers, T., Worring, M., Smeulders, A.W.M., Koelma, D.C.: The MediaMill TRECVID 2009 Semantic Video Search Engine. In: TREC Video Retrieval Evaluation Workshop, TRECVid 2009 (2009)
22. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In: ACM International Conference on Multimedia, pp. 421–430. ACM, New York (2006)
23. Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F., Binder, A., Gehl, C., Franc, V.: The SHOGUN Machine Learning Toolbox. Journal of Machine Learning Research 99, 1799–1802 (2010)
24. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), `http://www.vlfeat.org/`