Fast Discriminative Linear Models for Scalable Video Tagging

Roberto Paredes
Instituto Tecnológico de Informática
Universidad Politecnica de Valencia
rparedes@dsic.upv.es

Adrian Ulges
Image Understanding and Pattern Recognition Research Group
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern
adrian.ulges@dfki.de

Thomas Breuel
Image Understanding and Pattern Recognition Research Group
Technical University Kaiserslautern and DFKI, Kaiserslautern
tmb@iupr.dfki.de

Abstract

While video tagging (or "concept detection") is a key building block of research prototypes for video retrieval, its practical use is hindered by the computational effort associated with learning and detecting thousands of concepts. Support vector machines (SVMs), which can be considered the standard approach, scale poorly since the number of support vectors is usually high. In this paper, we propose a novel alternative that offers the benefits of rapid training and detection. This linear-discriminative method is based on the maximization of the area under the ROC. In quantitative experiments on a publicly available dataset of web videos, we demonstrate that this approach offers a significant speedup at a moderate performance loss compared to SVMs, and also outperforms another well-known linear-discriminative method based on a Passive-Aggressive Online Learning (PAMIR).

Work supported by the Spanish projects TIN2008-04571 and Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and German Research Foundation (DFG), project MOONVID (BR 2517/1-1)

1. Introduction

This paper addresses *video tagging*, the automatic annotation of video data with textual descriptions of generic concepts appearing in it, like objects, locations, and activities. The task (as illustrated in Figure 1) has also been referred to as "High-level feature extraction" [18] or "concept detection" [25]. It poses a difficult challenge due to strong intra-class variation and the enormous size of tag vocabularies. Yet, though current systems do not give a performance sufficient for a fully automatic labeling, the approach is considered a key component of modern video search prototypes [19].

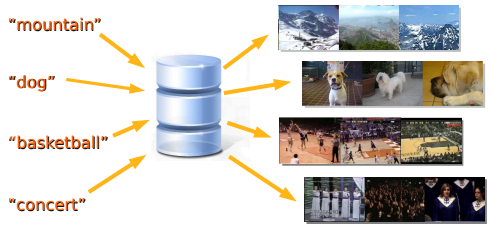

Figure 1. An illustration of concept detection: detectors corresponding to semantic concepts mine video collections for material of interest.

Conventionally, concept detection is cast as a binary classification problem, i.e. the system decides for each target concept whether it is present or not. To do so, multimodal features describing the content of a video are extracted, and machine learning techniques – typically Support Vector Machines (SVMs) [17] – are used to infer *scores* indicating concept presence.

For concept detection to gain attraction, tag vocabularies must cover a wide range of potential user interest – for

example, Chang et al. [4] reported significantly improved search results by increasing the number of concepts learned, and Hauptmann et al. [10] estimated about $5,000$ concepts to be required for high-quality video search. Correspondingly, while current detectors utilize static vocabularies of a few hundred of concepts [24], we would like to scale systems to significantly more concepts and larger datasets.

This poses a challenge to the underlying machine learning techniques: a fast and scalable training procedure is required, and to mine large-scale video databases for thousands of concepts, rapid classification is mandatory. While SVMs – which can be considered state-of-the-art [23, 24, 25] – offer a high accuracy, they do not satisfy these requirements. Due to strong intra-class variation, the number of support vectors is high and the solutions learned tend to degenerate to nearest neighbor models [25], which makes classification and learning slow.

Therefore, in this paper we propose a model that adopts the idea of maximum margin optimization from SVMs, but offers a substantially simpler and faster learning and classification. The approach determines a linear decision boundary by a maximization of the area under the ROC.

In quantitative experiments on a publicly available dataset of real-world web videos, the proposed approach achieves an accuracy comparable to SVMs while giving a more than 500-fold speedup of training and classification.

2 Related Work

Concept detection is targeted at automatically inferring the presence of semantic concepts (like objects, locations, or activities) from the audiovisual content of a video stream. Given a vocabulary of concepts $c_1, .., c_n$ and an input video $\mathcal{X}$, the task is to estimate concept scores $sc(c_1, \mathcal{X}), .., sc(c_n, \mathcal{X})$. Note that – in contrast to some image annotation systems [2] – no multiclass decision is made, i.e. concept presence and not concept prominence is judged.

Concept detection is strongly related to tasks like scene recognition and object category recognition, since it includes concepts related to scene types (e.g., "desert") and object categories (e.g., "airplane"). However, it can be considered a more general challenge – a priori, the range of potential target concepts is unlimited (though usually some basic constraints are imposed like utility for the user and feasibility of detection [7]).

The majority of concept detection systems are variations of a core architecture introduced in [6]. In this processing pipeline, input videos are segmented into shots using shot boundary detection [15]. For each shot, a variety of features is extracted, including image-based descriptions of representative key-frames, like color histograms or color moments and texture descriptors [23] or patch-based representations [22]. Other modalities like motion [9] and speech [12] have also been investigated, but will be omitted in this paper.

Features extracted from the video shots form the input to machine learning techniques (typically, SVMs [17]) which estimate scores indicating concept presence. A final fusion step combines scores for different keyframes and modalities, and can also make use of correlations between the presence of different concepts (like "road" and "car") [14].

A key effort in video tagging research is the annual TRECVID evaluation campaign [18]. Since 2002, concept detection has been addressed in TRECVID's "High-level Feature Extraction" task, providing researchers with datasets, annotations, and standard evaluation procedures. It is fair to say that research effort on concept detection focuses in TRECVID, with a community of over 30 research groups participating. Aside from TRECVID itself, other efforts towards standardization and comparability of results are being made by sharing intermediate results like features, annotations, and trained concept models (e.g., [24]).

3 Video Classification

In the following, we assume that a video $\mathcal{X}$ is represented by a set of key-frames $x_1, .., x_n$. Thus, concept detection is performed at key-frame level and a later fusion is applied in order to get an unique evidence for the whole video. Usually a *score* is assigned to each pair key-frame x_i and category c and the total score at video level can be computed by fusing the scores of the different key frames of that video:

$$sc(c, \mathcal{X}) = \mathit{Fusion}(sc(c, x_1), \cdots, sc(c, x_n)) \qquad (1)$$

where $\mathit{Fusion}(\cdot)$ is a function that takes all the key-frames scores as arguments.

In the present work we propose the following avg+max fusion:

$$sc(c, \mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} sc(c, x_i) \; + \max_{1<=i<=n} sc(c, x_i) \qquad (2)$$

This fusion scheme has consistently shown better empirical results than the simpler average fusion over all our experiments. Better fusion schemes and approaches could be applied for obtaining the video score but are beyond the scope of this paper.

4 Linear approaches

In order to be able to learn *thousands* of concepts a concept detection approach should have a very fast learning capabilities.

To achieve this, we propose to learn a very simple model, namely just a linear one. We adopt two approaches, the first

follows the same ideas introduced recently by Grangier and Bengio [8] where they have used an efficient training procedure by adapting the Passive-Aggressive Online Learning algorithm [5], we will refer to this approach as PAMIR. The second approach follows the ideas proposed by Paredes and Vidal [27] and Villegas and Paredes [26] for minimum error classification and maximum area under the ROC respectively.

Both approaches are adapted to the present problem, video classification. Given a key-frame x_i we propose to obtain a key-frame score as follows:

$$sc(c, x_i) = \mathbf{w}_c \mathbf{x}_i \tag{3}$$

where, $\mathbf{w}_c$ is a weight vector associated to concept c and $\mathbf{x}_i$ is the vectorial representation of key-frame x_i.

This simple score computation allows a very fast concept detection phase. Moreover, the model ($\mathbf{w}_c$) is very simple, low-dimensional, and compact, such it can be learned from limited amounts of training samples.

4.1 PAMIR model

In [8] the authors proposed to find a discriminative projection from the visual representation space to the textual representation space. In our case, such projection is performed to a one dimensional space (concept space) and is accomplished by means of a *weight vector* $\mathbf{w}_c$.

4.1.1 Index optimisation

In order to find a suitable vector $\mathbf{w}_c$ that leads to high values of the score for videos where the concept c appears and low values of the score for videos where the concept c does not appears, we propose to maximize the following discriminative index [8]:

$$J(\mathbf{w}_c) = \sum_{\forall \mathbf{x}_p \in X_p} \sum_{\forall \mathbf{x}_n \in X_n} (\mathbf{w}_c \mathbf{x}_p - \mathbf{w}_c \mathbf{x}_n) \tag{4}$$

where, $\mathbf{x}_p \in X_p$ is a key-frame that belongs to a video where the target concept appears (*positive* key-frame) while $\mathbf{x}_n$ is a key-frame that belongs to a video where the target concept does not appear (*negative* key-frame). So for each concept, X_p is the set of all the positive key-frames, X_n is the set of all the negative key-frames and $(\mathbf{x}_p, \mathbf{x}_n)$ is any possible pair of positive and negative key-frames.

The maximization of the index 4 involves a costly optimisation procedure since all possible pairs $(\mathbf{x}_p, \mathbf{x}_n)$ have to be considered, and the total number of pairs is $\mid X_p \mid \mid X_n \mid$. In order to find a suitable vector $\mathbf{w}_c$ within a reasonable time interval, we follow the same ideas as presented in [8]. We propose to find the weight vector $\mathbf{w}_c$ using an online iterative procedure as a result of the following expression:

$$\mathbf{w}_c^i = \underset{\mathbf{w}_c}{\operatorname{argmin}} \frac{1}{2} \mid\mid \mathbf{w}_c - \mathbf{w}_c^{i-1} \mid\mid^2 \; + \; \mathcal{C}\, l(\mathbf{w}_c; \mathbf{x}_p, \mathbf{x}_n) \tag{5}$$

where i is the iteration and the cost function $l(\cdot)$ is the hinge loss function:

$$l(\mathbf{w}_c; \mathbf{x}_p, \mathbf{x}_n) = \begin{cases} 0 & \mathbf{w}_c(\mathbf{x}_p - \mathbf{x}_n) > 1 \\ 1 - \mathbf{w}_c(\mathbf{x}_p - \mathbf{x}_n) & otherwise \end{cases}$$

This index has two different terms. The first term, $\frac{1}{2} \mid\mid \mathbf{w}_c - \mathbf{w}_c^{i-1} \mid\mid^2$ expresses the differences between the previous value of the weight vector $\mathbf{w}_c^{i-1}$ and the next solution $\mathbf{w}_c^i$, such that $\mathbf{w}_c^i$ is enforced to be close to $\mathbf{w}_c^{i-1}$. This is a desirable smoothness property of online learning algorithms. In order to avoid an *aggressive* behaviour of the approach the second term has a parameter $\mathcal{C}$ that controls the trade-off between increasing the discriminative power of the model and reducing the distance between the new and previous weight vector. The iterative procedure starts with $\mathbf{w}_c^0 = 0$, and at each iteration a random pair $(\mathbf{x}_p, \mathbf{x}_n)$ is evaluated. For this pair the new weight vector $\mathbf{w}_c^i$ is obtained by solving Equation 5. Based on [5] the solution to this equation is:

$$\mathbf{w}_c^i = \mathbf{w}_c^{i-1} + \Gamma^i (\mathbf{x}_p - \mathbf{x}_n) \tag{6}$$

where the Lagrange multiplier Γ^i is:

$$\Gamma^i = \min \left\{ \mathcal{C}, \frac{l(\mathbf{w}_c; \mathbf{x}_p, \mathbf{x}_n)}{\mid\mid \mathbf{x}_p - \mathbf{x}_n \mid\mid^2} \right\} \tag{7}$$

It is important to note that when the loss $l(\mathbf{w}_c; \mathbf{x}_p, \mathbf{x}_n)$ is zero no model update is performed and the PAMIR model gets important computational savings.

The iterative procedure is stopped after a predefined number of iterations (which is usually much lower than the total number of overall pairs).

4.2 MaxROC model

In [26] the authors proposed to find a linear projection for score fusion. In order to find a suitable projection the authors proposed to maximize the area under the ROC (AROC) of a binary problem. This linear projection could be learned for each video category and can be considered again a projection from the visual representation space to the textual representation space.

4.2.1 Index optimisation

In order to find a suitable projection $\mathbf{w}_c$ for a particular category c we proposed as presented in [26] to maximize the AROC for the binary problem defined by the positive and

negative key-frames for this particular category. An analytical expression of this AROC is the following:

$$J(\mathbf{w}_c) = \frac{1}{\mid X_p \mid\mid X_n \mid} \sum_{\forall \mathbf{x}_p \in X_p} \sum_{\forall \mathbf{x}_n \in X_n} step(\mathbf{w}_c \mathbf{x}_p - \mathbf{w}_c \mathbf{x}_n) \tag{8}$$

where $step(\cdot)$ is the step function centered at 0.

This index is optimized following a gradient descent approach. To this end the index must be derivable and the step function is substituted by the $sigmoid(\cdot)$ function:

$$S_\beta(z) = \frac{1}{1 + \exp(-\beta z)} \, . \tag{9}$$

and the derivative of the sigmoid function is:

$$sigm'(z) = \frac{\beta e^{\beta(1-z)}}{(1 + e^{\beta(1-z)})^2} \, . \tag{10}$$

$sigm'(z)$ is a "windowing" function which is maximum for $z = 1$ and vanishes for $|z - 1| \gg 0$. If β is large, then $sigm'(z)$ approaches the Dirac delta function, conversely, if β is small, then $sigm'(z)$ is approximately constant for a wide range of values of z.

Finally the index gradient is:

$$\frac{\partial J(\mathbf{w}_c)}{\partial \mathbf{w}_c} = \frac{1}{T} \sum_{\forall \mathbf{x}_p \in X_p} \sum_{\forall \mathbf{x}_n \in X_n} sigm'(\mathbf{w}_c \mathbf{x}_p - \mathbf{w}_c \mathbf{x}_n)(\mathbf{x}_p - \mathbf{x}_n) \tag{11}$$

and the gradient update is:

$$\mathbf{w}' = \mathbf{w} + \mu \frac{\partial J(\mathbf{w}_c)}{\partial \mathbf{w}_c} \tag{12}$$

where $T = \mid X_p \mid\mid X_n \mid$ and μ is the learning rate.

Again to optimize such index entails to consider every possible pairs $(\mathbf{x}_p, \mathbf{x}_n)$. In order to find a suitable vector $\mathbf{w}_c$ within a reasonable time we are going to adopt a stochastic gradient descent approach. Pairs $(\mathbf{x}_p, \mathbf{x}_n)$ are randomly selected from the pool of all the possible pairs and the linear projection $\mathbf{w}_c$ is updated an online manner. Empirical results show that the randomization does not affect significantly to the final result. In order to improve even further the computational behaviour of the approach we propose to discard the gradient updates of such pairs for which the derivative of the sigmoid $sigm'(\mathbf{w}_c \mathbf{x}_p - \mathbf{w}_c \mathbf{x}_n)$ is not greater that a certain threshold thr_s, canceling and saving small changes.

5 Experiments and Results

This allows to scale concept detection up to thousands of concepts and large-scale datasets.

In order to asses the performance of the proposed linear approaches, we compare them with Support Vector Machines (SVMs) [17], which can be considered the standard approach in concept detection and are used in most current systems [19, 24]. We first describe the experimental setup (including datasets and performance measures) and then discuss quantitative results.

5.1 Dataset: Youtube-22Concepts

While the focus of concept detection has been mostly on TV material [18], web video – which is publicly available from portals like YouTube – offers a free information source and application area for large-scale visual learning, and that has just recently entered the focus of computer vision research [13, 21]. In contrast to TV-based benchmarks – which require a manual annotation and the synthesis of concepts of interest – web video is publicly available at a large scale, and textual annotations are provided by users during upload, which allows to study video tagging on real-world data for a well-defined target function (namely the tagging behavior of YouTube users) [7]. For this reason, we evaluate the proposed approach on a dataset of web videos, namely the *youtube-22concepts* dataset[1] (which has also been made publicly available for research purposes).

The dataset consists of 2,200 real-world online video clips for 22 concepts, including objects ("helicopter"), events ("interview"), sports ("basketball"), and locations ("beach"). The clips were downloaded using the YouTube API[2] together with tags given by YouTube users. For evaluation, the stratified default split of the clips into 75% training and 25% testing was used. The overall length of the dataset is about 194 hrs. Each clip was represented by a set of key-frames using an adaptive clustering approach [1], which gives about 97,000 key-frames for the whole dataset.

5.2 Setup

Bag-of-visual-words features [22] were extracted using a dense regular sampling over several scales, which gave about 3,600 patches per key-frame. These were described using standard SIFT descriptors [16] and clustered to a vocabulary of 2,000 visual words using K-means (these features are also publicly available[3]). This feature representation can be considered state-of-the-art in concept detection [22]. Its extraction can be done in real-time [20]. It should also be kept in mind that feature extraction is only run once per image, while classification may be applied for

[1] http://sites.google.com/a/iupr.com/iupr-image-and-video-computing/youtube-22-concepts-dataset

[2] www.youtube.com/dev

[3] http://users.dsic.upv.es/ iaprtc5/data/YT-22Concepts.tgz

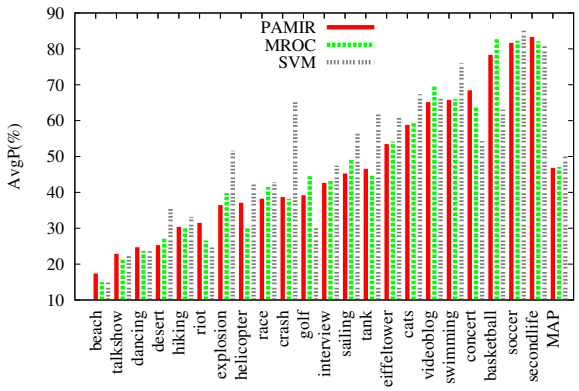

Figure 2. Results of Video Tagging for SVM and linear approaches

thousands of concepts in practice. Correspondingly, classifier speed can be considered a critical bottleneck and will be the focus of this experiment.

To measure annotation performance, we rank videos according to their score and compute the *average precision* of this ranked retrieval list, which corresponds to the area under the recall-precision curve. These average precisions for all concepts were again averaged, obtaining the *mean* average precision (MAP). This can be considered a standard measure [18].

In our experiments, we test SVMs using the libsvm standard C implementation [3]. We test a user-defined χ^2 kernel and the standard linear kernel. The cost parameter $\mathcal{C}$ and the scale parameter γ of the χ^2 kernel were estimated using a grid search cross-validation [11] maximizing average precision (training could be faster when omitting this step – however, in this case a decrease in classification accuracy is to be expected).

For the PAMIR and MROC models, in-house C implementations were used. The number of iterations of the PAMIR algorithm was set to 10^6. The parameters of MROC were fixed to $\mu = 0.001$ and $thr_s = 0.1$, the number of iterations was also 10^6.

5.3 Results

Figure 2 shows the results of the proposed approaches and SVMs for each one of the 22 concepts, and the MAP over all concepts. For some concepts the SVM results are significantly better than the linear models. In such cases, probably the representation method does not lead to a *linear* separation between the concept presence and concept absence objects. On the other hand, for some concepts the linear methods get significantly better results than the SVM. In such cases, probably the linear separation of the concept space is achieved by the representation scheme and the linear methods, being simpler, get a higher generalization than the SVMs considering that in the present scenario the SVMs tend to degenerate to nearest neighbor models [25].

Method / #samples		MAP	Time (secs)
SVM(χ^2)	200	25.5%	94
SVM(χ^2)	1500	42.5%	525
SVM(χ^2)	9000	52.3%	16540
SVM(lin.)	1500	36.2%	117
PAMIR	9000	46.8%	37
MROC	9000	47.0%	24

Table 1. MAP results and time (secs) required for training one concept for SVM and linear models

Table 1 shows the MAP and the training time for all the methods. We tested SVMs for different sizes of training sets, and also for different kernel choices. Our results show that the simple linear models give a competitive concept detection accuracy. Slight improvements can be achieved by training SVMs on very large datasets, but this slows down training significantly. On the other hand, if reducing the training set size, the accuracy of SVM-based detectors drops significantly below the one of the proposed models. It is remarkable that the iterative training of the proposed models is significantly simpler compared to the batch optimization of SVMs, which suffers from a high number of support vectors (in fact, on average far over 60% of our training samples are support vectors, which corresponds to earlier observations in concept detection by Yang and Hauptman [25]).

The computational benefits of the linear methods are not only limited to the training process but apply to the detection phase as well. Table 2 shows the time needed by SVM and linear approaches to detect a particular concept in *all* the test key-frames. Again, classifier speed depends on the number of training samples (which corresponds to the number of support vectors), with the proposed linear approaches being up to 10,000 times faster than SVM (note that the linear SVM implementation could hypothetically achieve a speed comparable to the proposed linear methods, but this does not hold for the LIBSVM implementation).

Overall, our results indicate that the proposed linear approaches are an attractive alternative compared to SVMs, which – depending on the training set size – are either significantly slower or reach a lower detection accuracy.

Further, the MROC model presents computational benefits, as the experiments show the model update rejection of MROC leads to more model update savings than the PAMIR model. The experiments show that PAMIR saves a total of 69% of model updates while MROC saves a total of 76% but still leading to better MAP results.

The significant speedup obtained allows us to extend our video tagging approach to deal with a significantly higher

Method / training samples		Time (secs)
SVM(χ^2)	200	89
SVM(χ^2)	1500	456
SVM(χ^2)	9000	2040
SVM(lin.)	1500	228
linear	9000	0.2

Table 2. Time (secs) for computing the score of *all* test key-frames.

number of concepts in the future.

6 Conclusions

In this paper, we have demonstrated that very simple linear models lead to a concept detection technique that gives (at a slight performance loss) a 500 times faster training and classification than standard techniques. This result allows video tagging with comparable to state-of-the-art performance and with a significant speedup, which opens the possibility to learn and detect thousands of concepts.

The proposed linear model, MROC, has shown some benefits compared to the PAMIR model. The model update rejection rate is higher and the MAP result is even better.

One important issue to study is the possibility of other features faster than SIFT in order to speed up the feature extraction process as well.

References

[1] D. Borth, A. Ulges, C. Schulze, and T. M. Breuel. Keyframe Extraction for Video Tagging and Summarization. In *Proc. Informatiktage 2008*, pages 45–48, 2008.

[2] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[4] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *Proc. TRECVID Workshop (unreviewed workshop paper)*, Nov. 2006.

[5] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwart, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research (JMLR)*, 7, 2006.

[6] A. A. et al. IBM Research TRECVID-2003 Video Retrieval System. In *Proc. TRECVID Workshop (unreviewed workshop paper)*, Nov. 2003.

[7] M. N. et al. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[8] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(8):1371–1384, 2008.

[9] A. Haubold and M. Naphade. Classification of video events using 4-dimensional time-compressed motion features. In *Proc. CIVR*, pages 178–185, July 2007.

[10] A. Hauptmann, R. Yan, and W. Lin. How many High-Level Concepts will Fill the Semantic Gap in News Video Retrieval? In *Proc. CIVR*, pages 627–634, Jul 2007.

[11] C. Hsu, C. Chang, and C. Lin. A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[12] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition. In *Proc. Int. Conf. Semantics and Digital Media Technology*, pages 78–90, Dec. 2007.

[13] IEEE Computer Society, Computer Vision and Pattern Recognition Workshops. *Proc. First Internet Vision Workshop*, June 2008.

[14] W. Jiang, S.-F. Chang, and A. Loui. Context-Based Concept Fusion with Boosted Conditional Random Fields. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2007.

[15] R. Lienhart. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. *Int. J. of Img. and Graph.*, 1(3):469–286, 2001.

[16] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[17] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[18] A. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In *Proc. CIVR*, pages 11–17, July 2005.

[19] C. Snoek and M. Worring. Concept-based Video Retrieval. *Found. Trends in Inf. Retrieval*, 4(2):215–322, 2009.

[20] J. Uijlings, A. Smeulders, and R. Scha. Real-Time Bag of Words, Approximately. In *Proc. CIVR*, July 2009.

[21] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning Automatic Concept Detectors from Online Video. *Comp. Vis. Img. Underst. (in press)*, 2009.

[22] K. van de Sande, T. Gevers, and C. Snoek. A Comparison of Color Features for Visual Concept Classification. In *Proc. CIVR*, pages 141–150, July 2008.

[23] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video Diver: Generic Video Indexing with Diverse Features. In *Proc. MIR*, pages 61–70, Sept. 2007.

[24] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Technical report, Columbia University, March 2007.

[25] J. Yang and A. Hauptmann. (Un)Reliability of video concept detection. In *Proc. CIVR*, pages 85–94, July 2008.

[26] M. Villegas and R. Paredes Score Fusion by Maximizing the Area Under the ROC Curve In *4th Iberian Conference on Pattern Recognition and Image Analysis*, volume 5524 of LNCS, pages 473–480, June 2009

[27] R. Paredes and E. Vidal Learning weighted metrics to minimize nearest-neighbor classification error *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2006 (Vol. 28, No. 7) pp. 1100-111