DFKI-IUPR participation in TRECVID'09 High-level Feature Extraction Task

Damian Borth
Department of Computer Science, University of Kaiserslautern
d_borth@cs.uni-kl.de

Markus Koch
Department of Computer Science, University of Kaiserslautern
m_koch@cs.uni-kl.de

Adrian Ulges
German Research Center for Artificial Intelligence (DFKI)
adrian.ulges@dfki.de

Thomas M. Breuel
University of Kaiserslautern and DFKI
tmb@cs.uni-kl.de

Abstract

Run No.	Run ID	Run Description	infMAP (%)
training on TV09 data (type: A)			
1	IUPR-VW-TV	SIFT visual words with SVMs	8.5
2	IUPR-ADAPT-TV	SIFT visual words with PA1SD	5.1
combined training on YouTube and TV09 data (type: C)			
3	IUPR-VW+TT-TV	SIFT visual words with SVMs, fused with TubeTagger concept detection scores	8.3
4	IUPR-ADAPT-YT	SIFT visual words with PA1SD, trained on YouTube, adapted to TV09	5.1
training on YouTube data (type: c)			
5	IUPR-VW-YT	SIFT visual words with SVMs	3.2
6	IUPR-VW+TT-YT	SIFT visual words with SVMs, fused with TubeTagger concept detection scores	3.2

Similar to our TRECVID participation in 2008 [23], our main motivation in TRECVID'09 is to use web video as an alternative data source for training visual concept detectors. Web video material is publicly available at large quantities from portals like YouTube, and can form a noisy but large-scale and diverse basis for concept learning. Unfortunately, web-based concept detectors tend to be inaccurate when applied to different target domains (e.g., TRECVID data [24]). This "domain change" problem is the focus of this year's TRECVID participation.

We tackle it by introducing a highly-efficient linear discriminative approach, where a model is initially learned on a large dataset of YouTube video and then adapted to TRECVID data in a highly efficient on-line fashion. Results show that this cross-domain learning approach (infMAP 5.1%) (1) outperforms SVM detectors purely trained on YouTube (infMAP 3.2%), (2) performs as good as the linear discriminative approach trained directly on standard TRECVID'09 development data (infMAP 5.1%), but (3) is outperformed by an SVM trained on standard TRECVID'09 development data (infMAP of 8.5%).

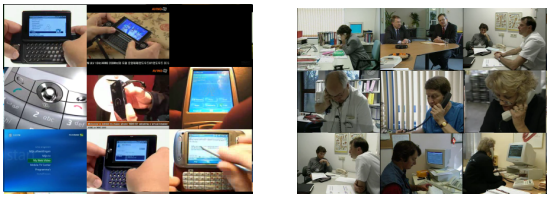
1. Introduction

Recently, the usage of socially tagged images and video as training sources for semantic concept detection (or "high-level feature extraction") has become more and more prominent [17, 23, 24]. Such data is publicly available at large scale from online portals like Flickr or YouTube and is associated with a noisy but rich corpus of tags, comments and ratings that are provided by their online communities.

Utilizing this information (e.g. web video) can help to reduce or even neglect the demand for expert labeled datasets, which currently serve as a foundation for supervised machine learning training, the underlying technology of current concept detection systems [4, 18, 26]. Being able to automatically learn new concepts from web video can increase concept vocabularies immensely and make content based video retrieval systems scalable and capable to cover user's information need [8]. Furthermore, large scale acquisition of expert labeled datasets [2, 14, 19] is a time-consuming and therefore cost intensive effort leading to static datasets which cannot adapt. This results in missing new emerging concepts of interest like "Michael Jackson", "Inauguration Day" or "iPod" [1] in their corpus. Additionally, the focus on only a few annotated video collections like e.g the "Sound & Vision" dataset currently used in TRECVID might limit the variability of retrieval results since the detectors work well on this dataset (or similar one) but generalize poorly on others as demonstrated in [27]. Web video, on the other hand, is providing a more diverse dataset being able to train more general detectors performing better when applied on previously unknown datasets [20].

On the downside, the usage of web videos as training material for concept detection systems faces new challenges as its quality strongly dependents on user generated tags. First, such weakly labeled video clips are often subjectively annotated, unreliable and coarse containing a great amount of non-relevant content (only a fraction between 20%-50% is relevant as estimated in [21]).

Second, in a setup where concept detectors are trained on web video and afterwards applied to a particular domain like the "Sound & Vision" dataset used in TRECVID, we are facing the so-called domain change problem: a significant discrepancy of the visual characteristics between given domains. This is illustrated in Figure 1: imagine training a detector for the concept "Telephone" on YouTube data (which shows mostly close-ups) and applying it on TV broadcast data (here, the TRECVID'08 dataset [11]), which shows mostly office telephones. Obviously, the web-based detectors will perform poorly on this particular dataset as already has been reported during TRECVID'08 [24]. So far, this challenge has been addressed using *cross-domain learning* techniques in the context of switches between different TV channels [29, 6] or TRECVID datasets (from TV05 to TV07) [9, 28] but not in the context of web-based learning.



(a) (b)

Figure 1. Frames from YouTube (a) and TRECVID (b) videos tagged with Telephone. The domain change leads leads to a different visual appearance of the concepts.

This raises the question whether concept detectors trained on the web video domain (here, YouTube) can be successfully *adapted* to another (here, Sound & Vision). Our participation in TRECVID's High-level Feature Task aims to answer this question in introducing a light-weight linear discriminative adaptation approach using bag-of-visual-words features and being able to (1) perform domain adaptation of its model learned from YouTube to the special domain of TRECVID's Sound & Vision data and (2) showing that this adaptation approach is highly efficient in terms of dealing with potentially huge datasets. Additionally, we provide several control runs with state-of-the-art SVMs for comparison. The paper first describes the acquired YouTube dataset and its characteristics. After this, the different approaches are outlined and results of the runs are provided.

[1] top ranked searches 2009 by "Google Insights for Search" for web search, news search and product search respectively

Table 1. Queries for Training Set Acquisition from YouTube.

concept	YouTube query	YouTube category
Classroom	classroom & school -secret	-
Chair*	chair & office & bürostuhl -wheel -trailer	HowTo&Style
Infant*	baby & kleine babys -driver -ride	People&Blog
Traffic Intersection*	traffic & intersection & strassen & kreuzung	Autos&Vehicles
Doorway*	türen & öffnen & doors & gates	People&Blog & Entertainment
Airplane flying	airplane & flying -jefferson -indoor -school -kids	Autos&Vehicles
Person playing a musical instrument*	instrument & learn to play	Music
Bus	bus -van -suv -vw -ride	Autos&Vehicles
Person playing soccer*	people playing soccer & fussball spielen	Sport
Cityscape	cityscape -slideshow -emakina	Travel&Places
Person riding a bicycle*	riding bicycle & fahrrad	Sports
Telephone	phone & device	-
Person eating*	food eating contest & essuen und kochen	Entertainment
Dem._Or_Prot.	protesting	-
Hand	hand & daft	-
People dancing*	people dancing & learn to dance	Sports
Nighttime	by & night	Travel&Places
Boat_Ship	ship & (queen | freedom | royal)	Autos&Vehicles
Female human face closeup*	female videoblog & girl makeup	People&Blog & HowTo&Style
Singing	singing & (gospel | choir)	-

*new concept for 2009 evaluation

2. Datasets

To evaluate the potential of domain adaptation for concept detectors when trained on freely available web video and adapted to TRECVID's Sound & Vision data, we differentiate between to datasets: first, a collection of video clips downloaded from YouTube (referred to as YOUTUBE) where user generated tags serve as annotations for training. The other video collection is the standard TRECVID'09 Sound & Vision development data (referred to as TRECVID) with high-quality expert annotations. In terms of cross-domain learning terminology, the YOUTUBE data is our *source domain*, whereas TRECVID defines our *target domain* to which we want to adapt.

To download videos from YouTube, first we have to make use of the YouTube API[2] for meta-data retrieval of potential video clips. This is done using a textual query like "food eating contest" embedded in the API call. This step is critical because the quality of the trained concept detectors strongly depends on manually selecting proper keywords used in the YouTube query. Given the fact, that every minute 20 hours of new video material is uploaded to YouTube [10], it can be understood that a quality control of retrieved video material at this scale is impossible. However, to make sure that simple misinterpretations does not occur like for the concept "cityscape" not to download videos of various 3d software tutorials, where a "cityscape" is modeled, we perform two manual refinements during keyword selection (a complete list of final queries is given in Table 1):

1. YouTube is organizing videos in categories like "Sport" or "Autos&Vehicles". For some concepts, we enhanced the query with a canonical category, which restricted the list of retrieved videos to this category. For example, by choosing the category "People&Blog" we could improve the quality of video material for the concept "female human face closeup" in getting more video clips of closeup faces.

2. Queries were additionally refined by inspecting of YouTube search results and accordingly adding or excluding additional keywords. For example, for the concept "person eating" we added the keyword "food" or for the concept "chair" we excluded the term "wheel".

After defining proper queries and retrieving meta-data of potential video clips, we downloaded 150 videos for each new concept from YouTube. To reduce data load we only downloaded the first 3 minutes of each clip resulting in a training set of about 120 hours of total length.

Figure 2 is displaying random sample keyframe from both training sets *YOUTUBE* and *TRECVID* for the representative concepts "person_playing_soccer", "traffic_intersection", "person_eating" and "female_human_face_closeup". It can be seen that while YOUTUBE grasps the

[2]http://www.youtube.com/dev

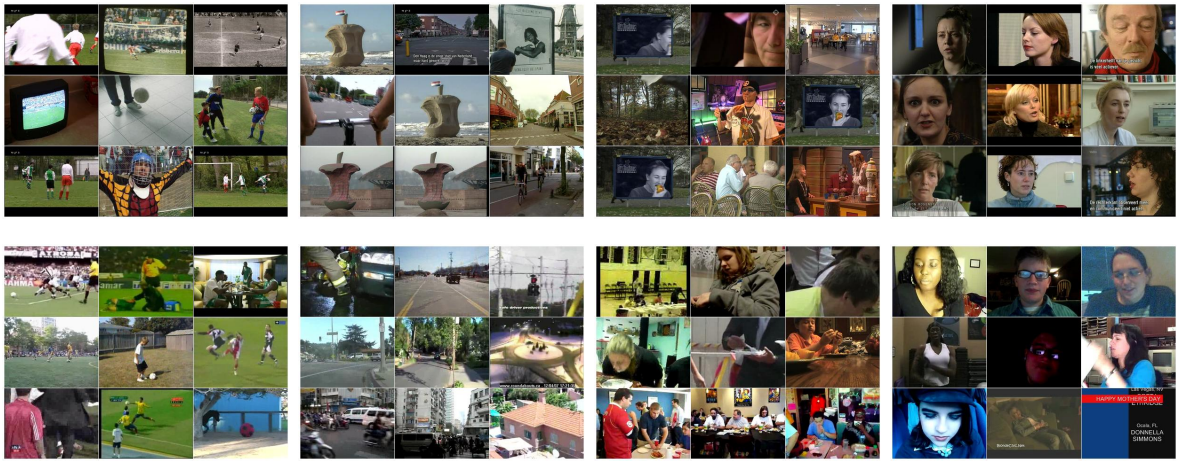


Figure 2. An illustration of randomly selected key frames from the *TRECVID* (top) and *YOUTUBE* (bottom) training sets. The concepts are "person_playing_soccer", "traffic_intersection", "person_eating", and "female_human_face_closeup". While the *TRECVID* dataset shows high annotation quality, the material downloaded from YouTube contains some amount of junk.

concept definition it contains some amount of non-relevant content as already reported in [23]. Note that this non-relevant material will also be used as positive samples in our concept detector training. Especially, for the concept "telephone", where the initial example in Figure 1 is taken from we observe a significant shift in the domain from video clips of smart phone in YOUTUBE to office scenarios in TRECVID data. Remember, that we are reusing the old web material from the 2008 evaluation for the 10 concepts that are kept in the 2009 evaluation.

3. Approach

The goal of our TRECVID participation is the evaluation of a domain adaptation technique based on a light-weight linear discriminative approach where a model is trained on the rich collection of YouTube videos and afterwards adapted to the particular domain of TRECVID's Sound & Vision data in a highly efficient on-line fashion. Additionally, we perform control experiments comparing adaptation results with state-of-the-art SVMs classifiers. As input for both classifier we use SIFT visual words. However, we also evaluate a new feature set consisting of concept detection scores provided by the TubeTagger detector [22].

3.1 Keyframe Extraction

Regarding shot representation we are extracting keyframes for each video/shot. Here, we deal differently with the given datasets:

1. For the *YOUTUBE* data, keyframe extraction is performed according to a change detection scheme [22] providing $53k$ keyframes for the entire dataset ($19k$ keyframes were already available from used concepts in the TRECVID'08 evaluation), which corresponds to an average of ca. 18 keyframes per YouTube video clip.

2. For the *TRECVID* data, the standard shot boundary reference was used for temporal segmentation and a intra-shot diversity based approach for keyframe extraction [3]. For each shot, a K-Means clustering is performed over MPEG7 Color Layout Descriptors [13] extracted from all frames the number of clusters is fitted using the Bayesian Information Criterion [16]. For each cluster the frame closest to the cluster center is chosen as a keyframe.

3.2 Features

From all keyframes the following visual features are extracted:

- **Visual Words (SIFT)**: Visual words are extracted by performing a dense regular sampling of SIFT features [12] at several scales, obtaining ca. $3,600$ features per keyframe. Features are clustered to $2,000$ visual words using K-Means forming "bag-of-visual-words" descriptors.

- **TubeTagger Semantic Feature Space**: TubeTagger is a system which performs a visual learning on YouTube clips, allowing it to distinguish between 233 semantic concepts (see [22] for more details).

  We exploit the semantic knowledge of the system by applying the already existing classifiers to the extracted keyframes, and combining the resulting scores into one 233-dimensional feature vector for each keyframe. Each entry hereby corresponds to the affinity towards one semantic concept known to TubeTagger i.e representing it in context of the TubeTagger concept vocabulary. For example, the TubeTagger detectors for "street" and "vehicle" might be a valuable for classification of an unknown concept like "car".

3.3 Statistical Models

Two different statistical models are used:

- **Support Vector Machines**: Support vector machines (SVMs) are a standard approach for concept detection and form the core of numerous concept detection systems [25, 26]. We used the LIBSVM [5] implementation with a χ^2 kernel, which has empirically been demonstrated to be a good choice for histogram features [30]:

$$K(x,y) = e^{-\frac{d_{\chi^2}(x,y)^2}{\gamma^2}} \quad (1)$$

  where $d_{\chi^2}(.,.)$ is the χ^2 distance. γ and the SVM cost of misclassifications C were estimated separately for each concept using a grid search over the 3-fold cross-validated average precision. A problem is that training sets are *imbalanced*, i.e. the number of negative samples outnumbers the number of positive ones. Those setups cause problems for many classifiers, including SVMs [1]. To overcome this problem, the dominant class is subsampled to obtain roughly balanced training sets. For the TRECVID based runs, 5 SVMs were trained on small-scale training sets with 400 negative samples randomly sampled from the *TRECVID* set, and the results were fused using a simple averaging. For the YouTube-based runs (where significantly more positive training samples were available), we used 3000 positive and 6000 negative training examples from the *YOUTUBE* data set.

  In all cases, SVM scores were mapped to probability estimates using the LIBSVM standard implementation.

- **Passive Aggressive Online Learning**: Passive Aggressive Online Learning was first introduced in context of image retrieval [7], but also proved to be an valid alternative to SVMs when dealing with large scale concept detection of video clips [15]. This approach describes a highly efficient linear discriminative classifier which optimizes the area under the Receiver Operating Characteristic (ROC) curve in projecting keyframes to a one dimensional space (concept space) by means of a *weight vector* w_c.

  The one dimensional concept space is optimized by maximizing the following criterion using an online iterative procedure (more details in [7, 15]):

$$J(w_c) = \sum_{\forall x_p \in X_p} \sum_{\forall x_n \in X_n} (w_c x_p - w_c x_n) \quad (2)$$

  where $x_p \in X_p$ is a keyframe that does show the concept (positive keyframe) while $x_n \in X_n$ is a keyframe that does not show the concept (negative keyframe). For each concept X_p is the set of all positive keyframe, X_n is the set of all negative keyframes and (x_p, x_n) is any possible pair of positive and negative keyfames. Learning during training is done in optimizing this index in a way that positive keyframes are projected to high values in this concept space whereas negative keyframes are projected to low ones.

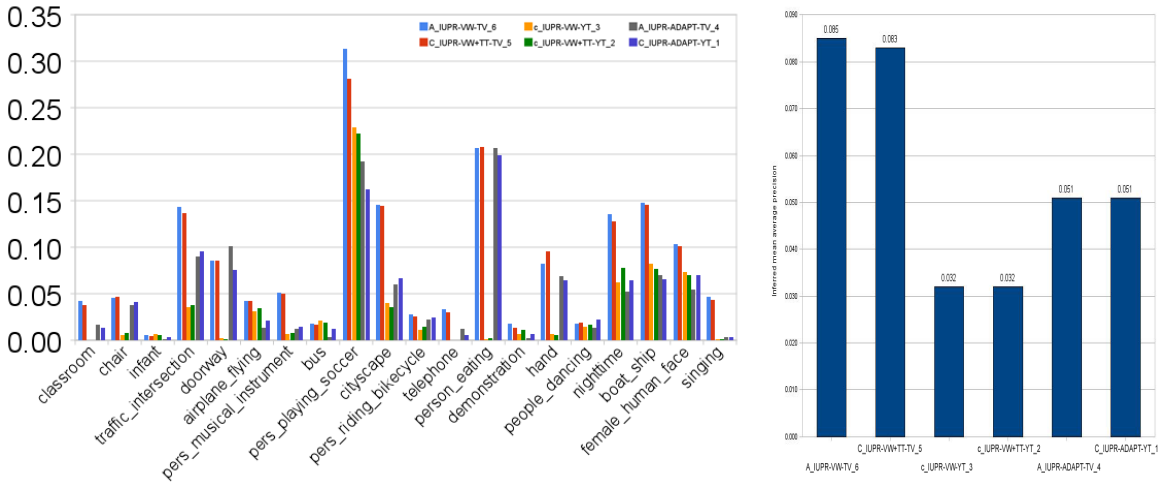


Figure 3. Quantitative results for all IUPR runs (the first two runs are using a SVM trained on the TRECVID'08 standard data, the middle two are using a SVM trained on YouTube data and the last two runs are using the PAMIR approach). left**: per-concept results.** right**: the mean inferred average precision per run.**

3.4 Domain Adaptation

Domain adaptation is based on an extension of the Passive Aggressive Online Learning approach. The idea is to utilize the online learning properties of the algorithm for adaptation. Initially, for each concept a model is trained on YOUTUBE data forming a weight vector $w_c^{YOUTUBE}$. Each of this weight vectors is modified in a second training sequence performed on TRECVID data. Finally, for each concept an adapted model represented by the weight vector $w_c^{ADAPTED}$ is created, which is now used for final classification of the TRECVID'09 test data. Note that this is possible because the models of the proposed online learning algorithm is entirely defined by its weight vectors w_c. Such an adaptation would be not suitable for SVMs, where cross-domain learning is performed differently and therefore is also computational more expensive.

3.5 Late Fusion

Finally, scores obtained from several keyframes for each shot and feature scores for each keyframe (in Runs 3 and 6) are fused:

- Having several keyframes for each shot, the corresponding scores are simply averaged, providing a single score for each shot and feature.
- For fusing different features, we perform a weighted sum fusion whereas concept-specific weights are learned using a grid search maximizing average precision on the TRECVID 2007 test set, using the TRECVID 2007 development set as training set. After re-training on the TRECVID 2009 development data, these weights are used to fuse the different features into a final concept score.

4 Results

We submitted a total of 6 runs: 2 runs trained on *TRECVID* data, 2 trained on a combination of *YOUTUBE* and *TRECVID* and 2 runs trained entirely on *YOUTUBE* data:

1. **A_IUPR-VW-TV** In this run, we used the SVM approach in combination with SIFT visual word features trained on TRECVID data.
2. **A_IUPR-ADAPT-TV** This run serves as a control for the adaptation approach (Run 4). Here, we trained with the PAMIR approach [15] over SIFT visual words features on TRECVID data.
3. **C_IUPR-VW+TT-TV** In contrast to Run 1, we fused detection scores with SVM results trained on TRECVID data using the TubeTagger semantic features.
4. **C_IUPR-ADAPT-YT** Here, we perform domain adaptation with the Passive Aggressive

Online Adaptation approach over the SIFT visual word features i.e. training on YOUTUBE and adaptation on TRECVID.

5. **c_IUPR-VW-YT** Same as in Run 1 but using YOUTUBE data as a training source.

6. **c_IUPR-VW+TT-YT** Same as in Run 3 but using YOUTUBE data as training source.

Quantitative results are displayed in Figure 3 showing that cross-domain learning with the proposed approach is in general possible. The Passive Aggressive Online Adaptation provides a infMAP of 5.1% which is outperforming the SVM detector purely trained on YOUTUBE (infMAP of 3.2%, 3.2% for run 5 and 6) and showing an as good performance as trained with the PAMIR approach on standard TRECVID'09 data (infMAP of 5.1%). However, the linear adaptation approach is performing worse than a SVM trained on standard TRECVID'09 data (infMAP of 8.5%, 8, 3% for run 1 and 3). Similar results are already reported by [15] when comparing PAMIR to SVMs. However, the PAMIR approach is especially suitable in large-scale setups like web video because of its 500-fold speedup against SVMs.

A potential performance gain could be reached in fine-tuning the only hyper-parameter of the PAMIR approach: the number of iterations optimizing the index in Equation 2. Experiments on the TRECVID development data indicate that the conservative setup used for the TRECVID benchmark experiments (in average $250k$ iterations for initial YouTube based training and $125k$ iterations for adaptation) is not optimal for all concepts. For example, for the concept "person playing soccer" the adaptation to TRECVID data damages concept detector performance, whereas a way higher number if adaptation iterations is necessary for concepts with a high amount of redundancy and duplicates like e.g. "person eating" and "traffic intersection".

5 Acknowledgements

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), project MOONVID (BR 2517/1-1). Additionally, we would like to thank Roberto Paredes from the Universidad Politécnica de Valencia for kindly providing an implementation of the PAMIR model.

References

[1] R. Akbani, S. Kwek, and N. Japkowicz. Applying Support Vector Machines to Imbalanced Datasets. In *Proc. Europ. Conf. Machine Learning*, pages 39–50, 2004.

[2] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *Proc. Europ. Conf. on Information Retrieval*, pages 187–198, 2008.

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

[4] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang. IBM Research TRECVID-2007 Video Retrieval System. In *Proc. TRECVID Workshop*, 2007.

[5] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001.

[6] L. Duan, I.W. Tsang, D. Xu, and T.S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proc. Int. Conf. Machine Learning*. ACM New York, NY, USA, 2009.

[7] 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.

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

[9] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui. Cross-domain Learning Methods for High-level Visual Concept Classification. In *ICIP*, pages 161–164, 2008.

[10] R. Junee. Zoinks! 20 Hours of Video Uploaded Every Minute! The YouTube Blog; available from `http://www.youtube.com/blog?entry=on4EmafA5MA` (retrieved: May'09), May 2009.

[11] W. Kraaij and P. Over. TRECVID-2008 High-Level Feature Task: Overview. In *Proc. TRECVID Workshop*, November 2008.

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

[13] B. Manjunath, J.-R. Ohm, V. Vasuvedan, and A. Yamada. Color and Texture Descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[14] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[15] R. Paredes, A. Ulges, and T. Breuel. Fast Discriminative Linear Models for Scalable Video Tagging. In *Proc. Int. Conf. on Machine Learning and Applications*, 2009.

[16] G. Schwarz. Estimating the Dimension of a Model. *Ann. of Stat.*, 2(6):461–464, 1978.

[17] A. Setz and C. Snoek. Can Social Tagged Images Aid Concept-Based Video Search? In *Proc. Int. Conf. on Multimedia and Expo*, pages 1460–1463, 2009.

[18] C. Snoek, M. Worring, O. de Rooij, K. van de Sande, R. Yan, and A. Hauptmann. VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems. *IEEE MultiMedia*, 15(1):86–91, 2008.

[19] C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, and A. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proc. Int. Conf. on Multimedia*, pages 225–226, October 2006.

[20] A. Ulges. *Visual Concept Learning from User-tagged Web Video.* PhD thesis, University of Kaiserslautern, Germany, 2009.

[21] A. Ulges, D. Borth, and T. Breuel. Visual Concept Learning from Weakly Labeled Web Videos. In *Video Search and Mining*. Springer-Verlag, 2009.

[22] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger YouTube-based Concept Detection. In *Proc. Int. Workshop on Internet Multimedia Mining*, December 2009.

[23] A. Ulges, M. Koch, C. Schulze, and T. Breuel. Learning TRECVID'08 High-level Features from YouTube™. In *Proc. TRECVID Workshop*, November 2008.

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

[25] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video Diver: Generic Video Indexing with Diverse Features. In *Proc. Int. Workshop Multimedia Information Retrieval*, pages 61–70, September 2007.

[26] 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, 2007.

[27] J. Yang and A. Hauptmann. (Un)Reliability of Video Concept Detection. In *CIVR*, pages 85–94, July 2008.

[28] J. Yang and A.G. Hauptmann. A framework for classifier adaptation and its applications in concept detection. 2008.

[29] J. Yang, R. Yan, and A. Hauptmann. Cross-Domain Video Concept Detection using Adaptive SVMs. In *Proc. Int. Conf. on Multimedia*, pages 188–197, September 2007.

[30] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. J. Comput. Vis.*, 73(2):213–238, 2007.