TubeTagger – YouTube-based Concept Detection

Adrian Ulges, Markus Koch
IUPR Research Group
German Research Center for Artificial Intelligence (DFKI) GmbH
D-67663 Kaiserslautern, Germany
{*adrian.ulges,markus.koch*}*@dfki.de*

Damian Borth, Thomas M. Breuel
Department of Computer Science
University of Kaiserslautern
D-67663 Kaiserslautern, Germany
{*d_borth,tmb*}*@cs.uni-kl.de*

***Abstract*—We present TubeTagger, a concept-based video retrieval system that exploits web video as an information source. The system performs a *visual* learning on YouTube clips (i.e., it trains detectors for semantic concepts like "soccer" or "windmill"), and a *semantic* learning on the associated tags (i.e., relations between concepts like "swimming" and "water" are discovered). This way, a text-based video search free of manual indexing is realized.**

We present a quantitative study on web-based concept detection comparing several features and statistical models on a large-scale dataset of YouTube content. Beyond this, we report several key findings related to concept learning from YouTube and its generalization to different domains, and illustrate certain characteristics of YouTube-learned concepts, like *focus of interest* and *redundancy*. To get a hands-on impression of web-based concept detection, we invite researchers and practitioners to test our web demo[1].

***Keywords*-information retrieval; image databases; pattern recognition;**

I. Introduction

Over the last years, *concept-based video retrieval* [1] has evolved as an exciting research area. It realizes a text-based search of video databases by substituting a manual indexing with automatic visual detectors that mine video collections for semantic concepts, like objects ("car"), locations ("desert"), or activities ("interview"). This approach has proven highly effective and is now implemented in several research systems [2], [3], [4].

One problem with concept detection, however, is that the machine learning techniques underlying in require training data for large-scale concept vocabularies and semantic relations. Particularly, training samples for a visual learning of concepts have been acquired manually so far [5], which is a time-consuming and cost-intensive process. This poses severe limitations: the number of concepts remains limited, the insufficient scale of training sets gives rise to overfitting, and adapting to changes of user's information needs (like new concepts of interest) remains difficult.

Another important trend over the last years has been the enormous growth of web-based video platforms, like YouTube, Vimeo, or blinkx. These have not only become sources of information and entertainment to millions of users, but also offer a novel kind of knowledge base for the machine interpretation of multimedia data, coupling huge amounts of content with user-generated annotations, ratings, and categorizations.

Figure 1. TubeTagger Concept Learning from YouTube.

In this paper, we present a system called *TubeTagger* that links concept-based video retrieval and web video. Its key idea is to employ web video platforms like YouTube as a source of training data for two kinds of learning:

Visual Learning: TubeTagger employs web video content to train concept detectors – for example, to learn the visual appearance of the concept "soccer", result clips of a corresponding YouTube search form positive examples in the training set.

Semantic Learning: TubeTagger learns to link concepts from tag co-occurrence statistics. For example, the system discovers that the terms "swimming" and "water" are related as they appear together frequently. This allows to map users' text queries to the vocabulary of learned concepts.

TubeTagger realizes a concept-based retrieval without a tedious manual annotation of training samples and semantic relations (learning a concept only needs to be triggered with a textual YouTube query). This way, TubeTagger can be used in three applications (which are all realized in our web demo[1], and are also illustrated in Figure 2):

Automatic Deep Tagging: While YouTube tags are only made on clip level, TubeTagger can automatically detect concepts *within* a video, and so directly guide users to certain events of interest.

[1] http://www.dfki.uni-kl.de/~ulges/tubetagger

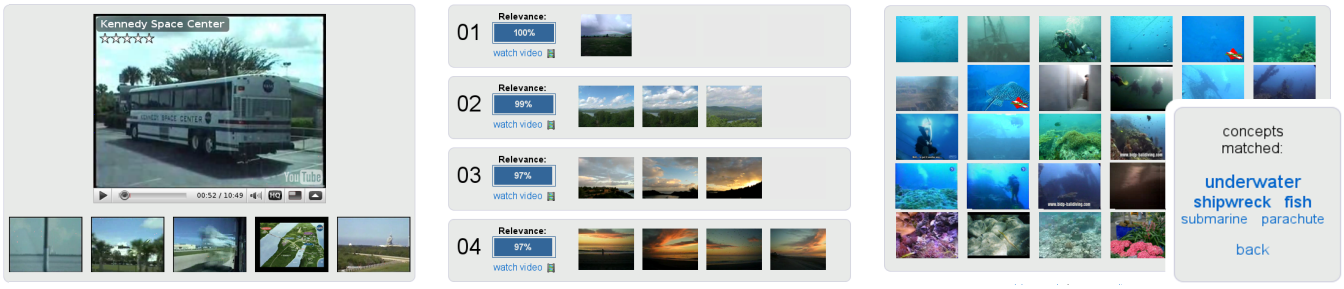

Figure 2. left: **automatic deep tagging** – TubeTagger finds the concept "bus" in a 10-minute clip. center: **tag recommendation** – TubeTagger suggests the tag "clouds" for some YouTube clips. right: **text-based video search** – TubeTagger matches the query "diving" to learned concepts like "underwater", "shipwreck", and "fish".

Tag Recommendation: TubeTagger can suggest additional labels for YouTube clips to overcome tag sparseness.

Text-based Video Search: Finally, TubeTagger can also use the learned semantic relations to match detected concepts to textual user queries, and so realize a text-based search of video databases free of manual indexing.

II. Related Work

Concept-based video retrieval has evolved as a novel research field over the last decade [6] (for an overview, see [1]). Research in the area focuses in TRECVID [7], an evaluation campaign for video retrieval in which a variety of concept detection systems have been developed [2], [3], [4]. So far, the standard approach is to employ expert-labeled training examples for concept learning. The resulting ground truth annotations are shared in TRECVID and other related efforts [8].

Replacing such a tedious manual annotation with an widely automatic learning from web content has only recently gained traction as a new research direction, driven by the rise of large-scale image and video sharing services like Flickr and YouTube. Web-based *image* content has already been studied quite intensively: for example, Fergus et al. [9] learn visual models of object categories from Google's image search and filter junk material using a topic model (for this purpose, Wnuk et al. [10] and Li et al. [11] propose a nearest neighbor analysis in feature space). Kennedy et al. [12] identify concepts suitable for additional web-based training material, and try to automate this decision.

When it comes to web *video* content, less contributions can be found. Zelnik-Manor et al. [13] and Schindler et al. [14] have presented studies on shot boundary detection and categorization of web video content, and emphasized the difficulty of the domain due to enormous content variance and weakness of labels. Other contributions are targeted at an automatic categorization of web videos based on their visual content *and* associated tags [15].

Finally, the generalization capabilities of web video-based detectors to different domains (e.g., to TV content) have been addressed. Setz and Snoek [16] and Ulges et al. [17] have applied web-based detectors to TRECVID datasets and reported that annotations on the target domain lead to a better accuracy, but that web-based detectors are effective for a bootstrapping on novel domains.

Finally, some approaches have been suggested for a *semantic* learning from web-based sources (i.e., textual information is employed and not image and video content as above). Haubold and Natsev [18] use web-based text corpuses for semantic reasoning, and point out that web-based text information is more large-scale and up-to-date. Yang et al.'s *Web 2.0 Dictionary* [19] follows a similar approach, constantly updating its tag correlations from the web. In TubeTagger, we have adopted such a semantic learning from web video, but also combined it with a visual learning as discussed above.

III. System Setup

Concept learning in TubeTagger is illustrated in Figure 1: YouTube material is downloaded (which is described in Section III-A) and employed to train visual concept detectors, which can later be used to compute *scores* indicating concept presence in previously unseen video content (Section III-B). In parallel, concept co-occurrences are learned from YouTube tags (Section III-C).

A. Data Acquisition

To learn a concept, the user provides a textual query to the YouTube API, which returns a list of videos matching this query. For training, we download a certain number N of clips per concept (this number will be investigated later in the experiments). These clips serve as positive examples for training the target concept. Negative samples are drawn from other videos *not* tagged with the concept.

To improve the quality of downloaded material, we *refine* the text query to the YouTube API. This is done by inspecting the first YouTube result page and iteratively adding additional terms and category information to the query. For example, to download training content for the concept "rainbow", the query "rainbow beautiful" was used, and only videos from the YouTube category "travel&places" were downloaded.

B. Visual Learning

TubeTagger performs concept detection on the basis of keyframes (moderate improvements of accuracy can be achieved by integrating other feature modalities such as motion information [17], which are omitted here).

Keyframes are extracted using a simple change detection and fed to a concept detection system. For each concept, a binary classification problem is formulated: all keyframes sampled from videos tagged with the target concept are used as positive training samples, keyframes from other clips as negative ones. As a feature representation, well-known bag-of-visual-words descriptions are used [20]: Patches are regularly sampled at several scales in the frame, described using SIFT [21] or SURF [22], matched to a 2000-dimensional codebook of patch categories, and finally aggregated in histograms. The resulting visual descriptors are fed to a machine learning method – we tested several statistical classifiers (output scores of all models are mapped to probabilities [23] afterwards):

1) Support Vector Machines (SVMs) [24], which are a state-of-the-art approach in concept detection [1], [3]. A χ^2 kernel was used, with parameters fitted by three-fold cross-validation.
2) Passive-Aggressive Online Learning (PAMIR) [25], [26], a linear model based on margin maximization. Training is done using an efficient online algorithm[1].
3) Maximum Entropy (MAXENT): we also test an approach based on the maximum entropy principle [27]. The posterior is modeled in a log-linear fashion, and the decision boundary is estimated using an iterative scaling algorithm.

C. Semantic Learning

Concept-based video retrieval only gains traction when learned concepts can be matched to a wide range of textual queries made by the user. For this purpose, TubeTagger learns relations between concepts from tag co-occurrence statistics. For each concept $t \in T$, a bag-of-words representation is extracted using counts of tags from the associated video clips, obtaining a histograms h_t. A user query q is represented by a similar histogram h_q. q is then mapped to learned concepts by computing weights as inner products: $w(q,t) := < h_t, h_q >$. The five concepts with the highest weights T_5 are chosen as potential matches, and the score of a keyframe x for the query q is computed by a weighted sum fusion:

$$P(q|x) = \sum_{t \in T_5} \frac{w(q,t)}{\sum_{t \in T_5} w(q,t)} P(t|x)$$

[1] an implementation of the model was kindly provided by Roberto Paredes from the Universidad Politécnica de Valencia.

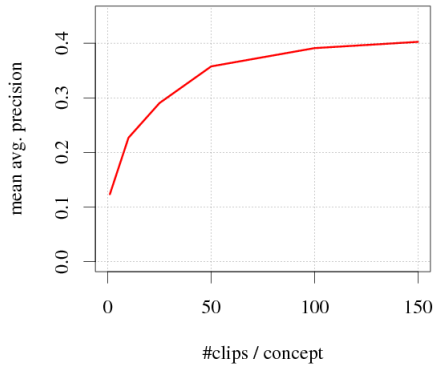

Figure 3. Benchmarking how many YouTube clips are required for concept learning. It can be seen that detector accuracy (using the PAMIR model) stabilizes when using between 100 and 150 clips for training.

IV. Concept Detection Experiments

We have evaluated TubeTagger on a dataset of $1,200$ hrs. (ca. $750,000$ keyframes) of YouTube content. 233 concepts (ranging from "airplane" to "wrestling") were manually selected with respect to feasibility of detection and availability of appropriate YouTube training material. For these concepts, queries to the YouTube API were formulated as described in Section III-A, and 150 clips per concept were downloaded for system training and 50 for testing.

We benchmark TubeTagger in a "video" mode, in which the system suggests tags for YouTube clips in the test set. Keyframe-level concept scores (see Section III-B) averaged, obtaining clip-level concept scores. Over these, we measure the mean average precision (MAP), using the original YouTube tags as ground truth on the test set.

Number of Training Videos: In a first experiment, we investigate how much YouTube training content is required for an accurate concept detection. Ten test concepts were chosen. For each concept, a random training sets of a certain number N of clips was randomly sampled and keyframes were extracted as positive training samples (negative samples were obtained by randomly sampling three times as many keyframes from other YouTube clips). TubeTagger was trained on this material (using the PAMIR approach) and then applied to a held-out test set of $200,000$ keyframes. Results – averaged over 10 runs of random training set resampling – are illustrated in Figure 3. It can be seen that training on a single clip is only slightly better than random guessing (MAP 10%). When using more training clips, performance increases until finally saturating at $100 - 150$ clips per concept. These observations were consistent over all 10 test concepts and were correspondingly expected to generalize well to other concepts. For efficiency reasons, we used 100 training clips (ca. $2,000$ positive keyframes) in the following experiments.

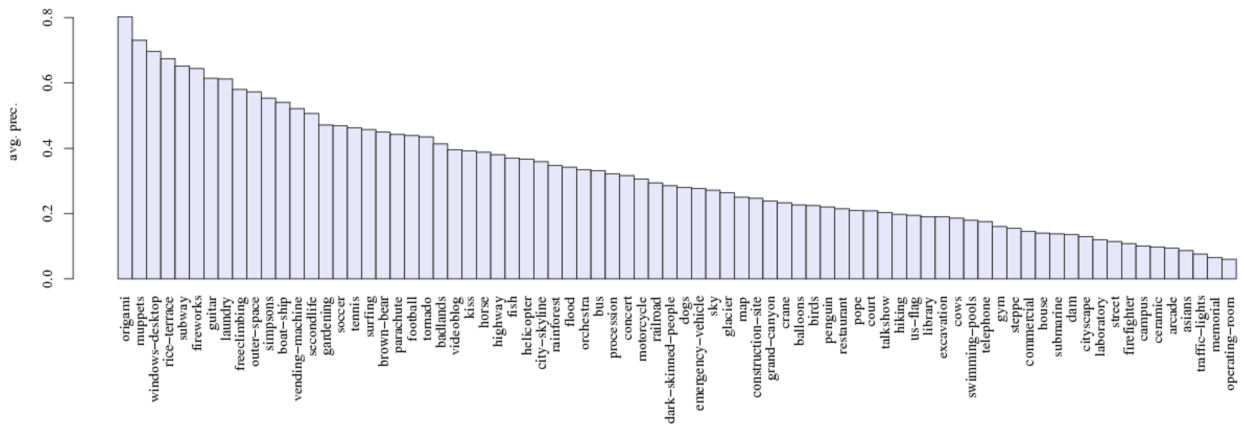

Figure 4. Detection results for a subsample of 78 representative concepts from our 233 test concepts.

Table I
EVALUATING DIFFERENT FEATURE TYPES AND CLASSIFIERS ON YOUTUBE CONTENT. A COMBINATION OF SIFT VISUAL WORDS WITH SVMS PERFORMS BEST.

	model		
feature type	SVMs	PAMIR	MAXENT
SURF	20.4	15.4	14.1
SIFT	**22.4**	18.4	15.5

Evaluating Features and Classifiers: We have also tested TubeTagger on a subset of 81 concepts in a similar setup as above (testing on a dataset of 50 clips per concept. Different statistical classifiers were tested, as well as SIFT and SURF visual word features. Results are illustrated in Table I: the best system (SVMs+SIFT) achieves a 18-fold improvement over random guessing (1.2%). It can be seen that SIFT gives a consistent improvement over SURF, and that SVMs outperform the simpler linear models. Yet, these alternatives may be interesting in a practical setting: particularly, SURF and PAMIR lead to significant speedups [26], which may be of vital importance in a real-world setting.

Full Test - 233 Concepts: Finally, we test TubeTagger on a large-scale test set of 233 concepts. Concept-wise test sets were compiled of 50 positive and 2,000 negative clips. Sample results on concept level are illustrated in Figure 4. We see that the overall results are – though far from human accuracy – promising: a mean average precision of 32.2% is achieved, and all concepts are significantly better detected than by a random guessing (MAP 2.4%).

V. INSIGHTS IN WEB-BASED CONCEPT DETECTION

In the following, we discuss key observations made during an in-depth inspection of the concept-wise results on the full set of 233 test concepts (the user is invited to validate and extend these findings by testing the web demo[1]). As already shown in Figure 4, detector performance varies strongly between concepts, ranging from 80.2% ("origami") to 6.0% ("operating-room"). An in-depth inspection of results revealed that in fact there seem to be different kinds of concepts:

"Good" Concepts: For some concepts, web-based concept detection works well in a sense that rich training content can be obtained from YouTube (for example, "boat/ship", avg. prec. 54.1%). Such "good" concepts can be characterized by the fact that a broad community of YouTube users records, edits, and uploads material. Often, they are inherently "interesting" or "spectacular", like scenic views ("mountain"), sights ("pyramids"), or sports ("basketball").

"Redundancy" Concepts: For other concepts, we find YouTube material to be adequate but not of a sufficient diversity. For example, see the concept "drummer" (avg. prec. 77.0%) in Figure 5 (center): here, TubeTagger has only learned from three specific series of drum lessons, and correspondingly the system overfits to this content. Note that making use of such redundancy is useful within the domain (for example, when dealing with new upcoming videos by the same user) but leads to a poor generalization to other domains (for more information on this issue and quantitative results, please refer to a previous publication of ours [17])

"Bad" Concepts: Finally, for other concepts we obtain neither a sufficient quantity of data nor a sufficient diversity. These concepts tend to be associated with everyday locations and objects, which might appear in YouTube content but are not regularly used as a tag, like "fence" (avg. prec. 8%), "gas station" (avg. prec. 10%), or "shopping mall" (avg. prec. 9.9%). For these concepts, training sets (and with them detection performance) are poor.

Focus of Interest: Another important aspect is that – to be used as a tag – a concept must be in the focus of attention: YouTube users will not use "house" as a tag if a house only appears somewhere in the background, but because the house is of particular interest to them. Correspondingly, we observed two effects: first, concept instances in YouTube

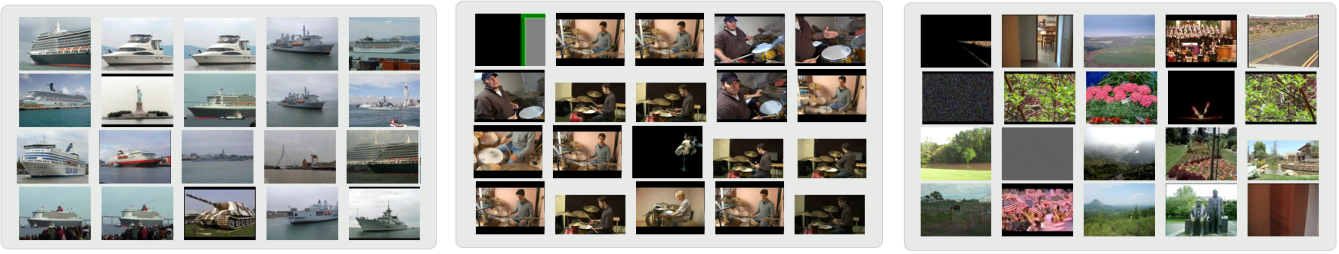

Figure 5. TubeTagger detection results for different kinds of concept: **left**: a "good" concepts ("boat-ship"), showing a high diversity of good-quality training material. **center**: a "redundancy" concept ("drummer"): training material is good quality, but of a low diversity. **right**: a "bad concept" ("fence"), for which we obtain no appropriate training material and detection quality is poor.

Figure 6. Concepts in YouTube-based training sets are usually in the focus of interest (illustrated here for "telephone"): clips at YouTube (left) shows close-ups, while expert-labeled TRECVID samples (right) show phones in the background of office scenes.

training sets have a certain tendency to be *outstanding* or special: for example, material for the concept "fountain" shows mostly spectacular watershows in Las Vegas. Second, concepts of interest are usually shown in *close-ups*. This is illustrated in Figure 6, where YouTube "telephone" training material shows only close-ups of phones, while TRECVID content (which is expert-annotated) shows office scenes with phones in the background. Obviously, the YouTube-based detector will work poorly on TRECVID data (see a previous publication of ours [17] for a quantitative evaluation of YouTube detectors on different domains).

VI. Web Demo

Results on the full 233 concepts can be browsed in our web demo[1] (note that no textual annotations on the test set – particularly, no YouTube tags – were used). The demo shows TubeTagger in three different applications (see Figure 2):

Automatic Deep Tagging: When searching for a concept, TubeTagger returns a list of test keyframes ranked by their concept score. By clicking on one of these frames, the user is directed to the associated scene within a YouTube clip. This way, we can perform a frame-accurate concept detection beyond coarse tags, as illustrated in Figure 2 (left). We found this feature particularly useful for concepts that are object- or event-related.

Tag Recommendation: When working in "video" mode, TubeTagger fuses keyframe scores to video level by a simple averaging, and employs them for tag recommendation. This is illustrated in Figure 2 (center), where TubeTagger suggests the tag "clouds" for some concepts.

Text Search: Finally, TubeTagger can also answer text queries by mapping them to known concepts, using the semantic learning of tag relationships (Section III-C). For an example, see Figure 2 (right), where the result for the query "diving" has been aggregated from matching concepts like "underwater", "shipwreck", and "fish". Although this simple correlation-based model does not truly learn *semantic* relations (like "is-part" and "is-subcategory"), we found it very useful in finding "good" concepts for a query.

VII. Discussion

We have presented TubeTagger, a concept-based video retrieval system that performs a widely unsupervised visual and semantic learning from YouTube. Our results and observations indicate that web material does have the potential to overcome the scalability problem in concept learning. Yet, several issues remain to be addressed.

First, we have observed that not all concepts are suitable for learning from web video content. This raises the question whether automatic strategies for selecting "good" concepts can be successful (as has already been proposed by Kennedy et al. [12] for the image domain).

A second issue is the investigation of strategies to improve the quality of downloaded YouTube training material. Here, the textual queries made by the user have been formulated manually so far (and by inspecting the retrieved YouTube material). It is an interesting question whether a better support of query formulation can be achieved.

Finally, we plan to improve the system with respect to scalability issues: using 16 cores, our current prototype learns 250 concepts in a week, with feature extraction being the most important bottleneck to be addressed (here, solutions for speed-up exist [28]).

Acknowledgment

This work was supported by the German Research Foundation (DFG), project MOONVID (BR 2517/1-1).

REFERENCES

[1] C. Snoek and M. Worring, "Concept-based Video Retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[2] 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*, Oct. 2006, pp. 225–226.

[3] 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*, Sep. 2007, pp. 61–70.

[4] A. Yanagawa, W. Hsu, and S.-F. Chang, "Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors," Columbia University, Tech. Rep., 2007.

[5] L. Kennedy, A. Hauptmann, M. Naphade, J. Smith, and S.-F. Chang, "LSCOM Lexicon Definitions and Annotations Version 1.0," ADVENT Technical Report, Columbia University, Tech. Rep., 2006.

[6] M. Naphade, S. Basu, J. Smith, C.-Y. Lin, and B. Tseng, "Modeling Semantic Concepts to Support Query by Keywords in Video," in *Proc. Int. Conf. Image Processing*, 2002, pp. 145–148.

[7] A. Smeaton, "Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience," in *Proc. Int. Conf. Image and Video Retrieval*, July 2005, pp. 11–17.

[8] 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*, vol. 13, no. 3, pp. 86–91, 2006.

[9] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," *Computer Vision*, vol. 2, pp. 1816–1823, 2005.

[10] K. Wnuk and S. Soatto, "Filtering Internet Image Search Results Towards Keyword Based Category Recognition," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[11] X. Li, C. Snoek, and M. Worring, "Learning Tag Relevance by Neighbor Voting for Social Image Retrieval," in *Proc. Int. Conf. on Multimedia Information Retrieval*, October 2008, pp. 180–187.

[12] L. Kennedy, S.-F. Chang, and I. Kozintsev, "To Search or to Label?: Predicting the Performance of Search-based Automatic Image Classifiers," in *Int. Workshop Multimedia Information Retrieval*, October 2006, pp. 249–258.

[13] L. Zelnik-Manor, S. Zanetti, and P. Perona, "A Walk Through the Web's Video Clips," in *Proc. First Internet Vision Workshop*, Jun. 2008, pp. 1–7.

[14] G. Schindler, L. Zitnick, and M. Brown, "Internet Video Category Recognition," in *Proc. First Internet Vision Workshop*, Jun. 2008, pp. 1–7.

[15] L. Yang, J. Liu, X. Yang, and X.-S. Hua, "Multi-modality Web Video Categorization," in *Proc. Int. Workshop Multimedia Information Retrieval*, September 2007, pp. 265–274.

[16] A. Setz and C. Snoek, "Can Social Tagged Images Aid Concept-Based Video Search?" in *Proc. Int. Conf. on Multimedia & Expo*, June 2009, pp. 1460–1463.

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

[18] A. Haubold and A. Natsev, "Web-based Information Content and its Application to Concept-based Video Retrieval," in *Proc. Int. Conf. Image and Video Retrieval*, July 2008, pp. 437–446.

[19] Q. Yang, X. Chen, and G. Wang, "Web 2.0 Dictionary," in *Proc. Int. Conf. on Image and Video Retrieval*, 2008, pp. 591–600.

[20] J. Sivic and A. Zisserman, "Video Google: Efficient Visual Search of Videos," in *Toward Category-Level Object Recognition*. Springer-Verlag New York, Inc., 2006, pp. 127–144.

[21] D. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Int. Conf. Computer Vision*, September 1999, pp. 1150–1157.

[22] H. Bay, T. Tuytelaars, and L. van Gool, "SURF: Speeded Up Robust Features," in *Proc. Europ. Conf. Computer Vision*, May 2006, pp. 404–417.

[23] T.-F. Wu, C.-J. Lin, and R. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.

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

[25] D. Grangier and S. Bengio, "A Discriminative Kernel-based Model to Rank Images from Text Queries," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1371–1384, 2008.

[26] R. Paredes, A. Ulges, and T. Breuel, "Fast Discriminative Linear Models for Scalable Video Tagging," in *Proc. Int. Conf. on Machine Learning and Applications (accepted for publication)*, December 2009.

[27] T. Deselaers, D. Keysers, and H. Ney, "Discriminative Training for Object Recognition using Image Patches," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, Jun. 2005, pp. 157–162.

[28] J. Uijlings, A. Smeulders, and R. Scha, "Real-Time Bag of Words, Approximately," in *Proc. Int. Conf. Image and Video Retrieval*, 2009.