MediaMill at TRECVID 2013: Searching Concepts, Objects, Instances and Events in Video

C.G.M. Snoek[†‡], K.E.A. van de Sande[†‡], D. Fontijne[‡], A. Habibian[†], M. Jain[*], S. Kordumova[†], Z. Li[†], M. Mazloom[†], S.L. Pintea[†], R. Tao[†], D.C. Koelma[†‡], A.W.M. Smeulders[†‡]

[†]ISLA, University of Amsterdam
Amsterdam, The Netherlands

[‡]Euvision Technologies B.V.
Amsterdam, The Netherlands

http://www.mediamill.nl

Abstract

In this paper we summarize our TRECVID 2013 [15] video retrieval experiments. The MediaMill team participated in four tasks: concept detection, object localization, instance search, and event recognition. For all tasks the starting point is our top-performing bag-of-words system of TRECVID 2008-2012, which uses color SIFT descriptors, average and difference coded into codebooks with spatial pyramids and kernel-based machine learning. New this year are concept detection with deep learning, concept detection without annotations, object localization using selective search, instance search by reranking, and event recognition based on concept vocabularies. Our experiments focus on establishing the video retrieval value of the innovations. The 2013 edition of the TRECVID benchmark has again been a fruitful participation for the MediaMill team, resulting in the best result for concept detection, concept detection without annotation, object localization, concept pair detection, and visual event recognition with few examples.

1 Task I: Concept Detection

Our concept detection approach builds on previous editions of the MediaMill semantic video search engine [19,20]. New this year is our convolutional neural network and concept detection without annotation. Since deep learning is critically dependent on labeled examples and suffers from noisy and incomplete annotations, as common in TRECVID [1,18], we manually extended the collaborative annotations.

Color Difference Coding Our baseline concept detection system uses a bag-of-words with color point descriptors only. For point sampling we rely on dense sampling, with an interval distance of six pixels and sampled at multiple scales. We used a spatial pyramid of 1x1 and 1x3 regions in our experiments. We used a mixture of SIFT, TSIFT, and C-SIFT descriptors [23]. We compute the descriptors around points obtained from dense sampling, and reduce the dimensionality with principal component analysis. We encode the color descriptors with the aid of difference coding using Fisher vectors with a Gaussian Mixture Model codebook [17]. For efficient storage we perform product quantization [6] on the features. The classifier is a linear SVM, which we apply on either the keyframe or on a maximum of six frames per shot.

Convolutional Neural Network Our deep learning concept detection system is a convolutional neural network with 8 layers with weights [8]. The input is raw pixel data, the output are concept scores. The network is trained using error back propagation. However, in contrast to ImageNet, there are too few labeled examples in the TRECVID SIN 2013 set for deep learning to be effective. We studied how additional examples from ImageNet [2] can be exploited to better train our networks. To improve the results, we took a network that had already been trained on ImageNet and retrained it for the 60 TRECVID 2013 SIN concepts. Similar to our color difference coding baseline we apply the network on either the keyframe or on a maximum of six frames per shot.

Learning from Social Media Learning video concept detectors from social-tagged media sources, such as Flickr images and YouTube videos, has the potential to address a wide variety of concept queries for video search. The focus of our TRECVID 2013 investigations for the no-annotation task are experiments with a video search engine which is capable of learning concept detectors from social media [7]. For each of the 60 concepts defined in the Semantic Indexing task we harvest positive examples from Flickr using two strategies (see runs below.) Our total pool is 1 million images for tag relevance training. The training set for concept detectors is 200k consisting of images tagged with the 60 concepts. We compute tag relevance on the 200k using the 1 million for neighbor voting [9]. We subselect relevant positive and negatives from the 200k to a maximum of 4,000 top-ranked examples [7]. The negatives were sampled from the other 59 categories of the 200k set using negative

[*]Visiting from INRIA, Rennes, France.

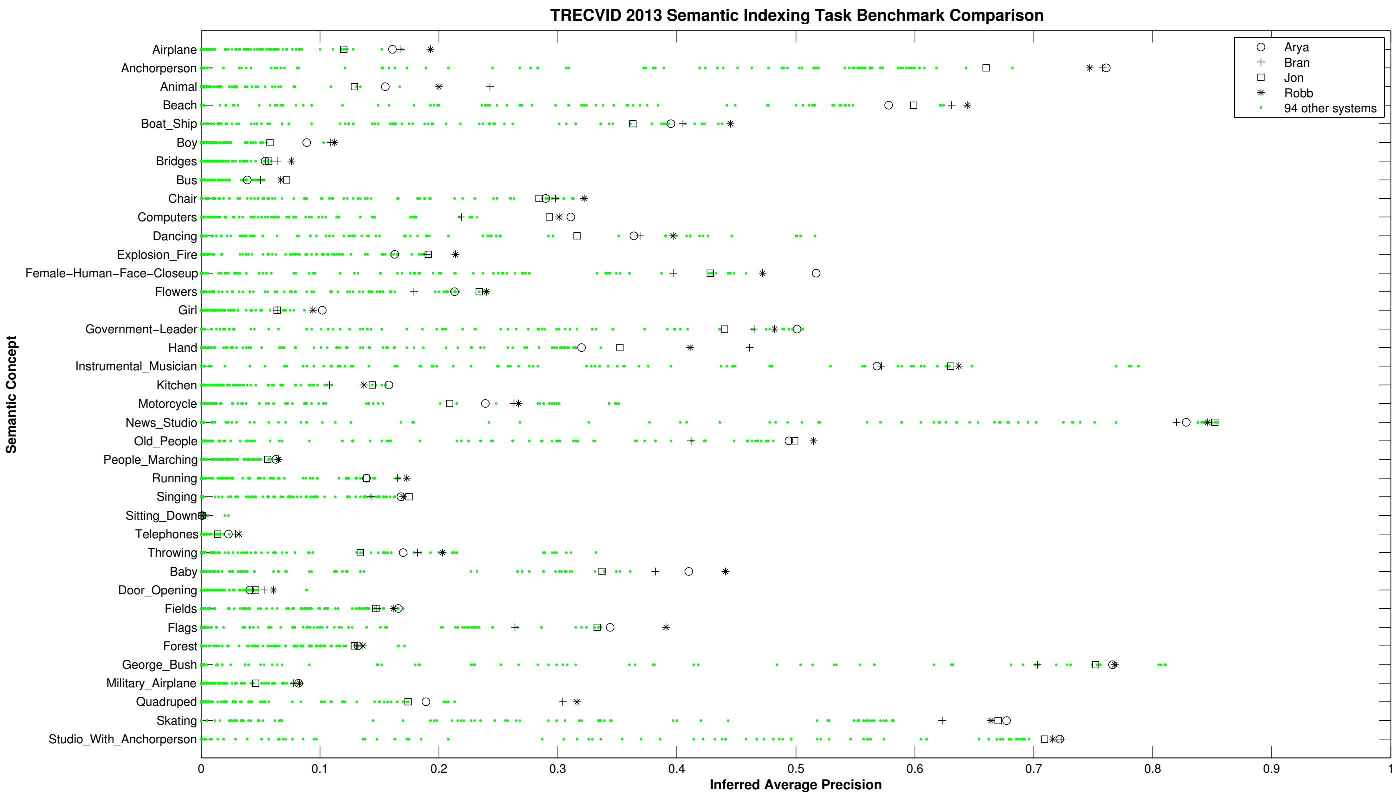


Figure 1: Comparison of MediaMill video concept detection experiments with other concept detection approaches in the TRECVID 2013 Semantic Indexing task.

bootstrap [10]. As the implementation for the final concept detectors we rely on color difference coding.

1.1 Submitted Runs with Annotation

We submitted four runs in the regular SIN task and two runs in the concept pair task. We summarize our regular SIN task submission in Figure 1.

UvA-Jon is our baseline run. It is based on color difference coding with multiple frames. It achieves an mAP of 0.286 and is the best performer for 2 out of 38 concepts. This run came out forth in terms of overall system performance.

UvA-Bran is our deep learning baseline based on multiple keyframe per shot. It achieves an mAP of 0.296 and is the best performer for 4 out of 38 concepts. While deep learning outperforms the more traditional color difference coding the overall difference is small. We also observed this during our development experiments which motivated us for a hybrid approach.

UvA-Arya is our hybrid system that fuses deep learning and color difference coding by a simple weighted average obtained by cross-validation. This run is based on classifying single keyframes per shot only. It achieves an mAP of 0.300 and is the best performer for 6 out of 38 concepts. Although this hybrid system relies on a single keyframe only it is still able to outperform both the multi-frame deep learning and color difference coding system.

UvA-Robb is similar to UvA-Arya, but it is based on multiple-frames per shot. It achieves an mAP of 0.321 and is the best performer for 15 out of 38 concepts. As expected the multi-frame variant adds an additional jump in accuracy and ends as first system in terms of overall mAP. This run also formed the basis for our runs in the concept pair task, where we simply used a combination of concept scores by sum (*UvA-Rickon*) or multiplication (*UvA-Shaggydog*). These runs came out second and first, overall, in the concept pair detection task.

1.2 Submitted Runs without Annotation

The no annotation task has two versions, $type-E$ to collect training data, and $type-F$ to collect training data using a query text built manually from the concept name and definition. We submit one run for each type.

UvA-Sansa In this $type-E$ run we harvest the positive examples by querying the Flickr API using query text automatically derived from the concept name and definition. For concepts not commonly appearing as tags in Flickr, like *Anchorperson* and *Quadruped*, we rely on query expansion with

Wikipedia as a source. Within the no-annotation $type - E$ condition, this run obtains the best AP for 32 out of 38 concepts and ranks first overall with an mAP of 0.048.

UvA-Lady In this $type - F$ run we simply rely on the concept name and definition and manually define a set of keywords to query the Flickr API. Within the no-annotation $type - F$ condition, this run obtains the best AP for 34 out of 38 concepts and ranks first overall with an mAP of 0.046.

2 Task II: Object Localization

We perceive object localization in video as a supervised learning problem. So we require bounding box annotations for objects of interest. We have refined a subset of the global image annotations for 10 (global) concepts to object-level by adding their bounding boxes.

Selective Search Rather than relying on exhaustive scanning of the image with boxes at multiple scales and aspect ratio's, we prefer (fast) selective search by Uijlings *et al*. [22]. Selective search generates a restricted set of about 2,000 object box hypothesis per image, using several hierarchical segmentations, which are independent of the object category. Since the number of boxes is restricted we can exploit computationally expensive bag-of-words features. For point sampling we rely on dense sampling at multiple scales. We used a fine spatial pyramid and a mixture of SIFT, TSIFT, and C-SIFT descriptors [23]. We compute the descriptors around points obtained from dense sampling, and reduce them all with principal component analysis. We encode the color descriptors with the aid of hard assignment with a codebook of 4,096 elements. For efficient storage we perform product quantization [6] on the features. The classifier is an SVM with fast intersection kernel approximation proposed by Maji *et al*. [11]. Following the convention from the object detection literature we perform negative mining of hard examples [22].

2.1 Submitted Runs

UvA-Snow This run is based on the ranking provided by *UvA-Jon*, using color difference coding only. In the top-1,000 shots we localize objects. It ranks 4th in terms of the iframe f-score metric.

UvA-Summer This run is based on the ranking provided by *UvA-Bran* using deep learning only. In the top-1,000 shots we localize objects. This run obtains the best overall iframe precision and fscore, as well as the highest mean-pixel fscore, recall and precision.

UvA-Nymeria This run is based on the ranking provided by *UvA-Arya*, which combined color difference coding with deep learning on keyframe level. In the top-1,000 shots we localize objects. This one has the lowest iframe recall among all our four runs, indicating once more the importance of multi-frame sampling for video.

UvA-Greywind This run is based on the ranking provided by *UvA-Robb*, which is similar to *UvA-Nymeria* but then with a maximum of 6 frames per shot analyzed. In the top-1,000 shots we localize objects. This run has the highest iframe recall among all our four submissions.

3 Task III: Instance Search

We address instance search with a two-step procedure. The sampled video frames, with a sampling rate of 2 frames per shot, are first ranked by a global search which considers the entire image in the analysis. The initial ranking is then refined by a localized search which evaluates per frame many bounding boxes holding candidates for the target. The boxes are generated using selective search [22]. We apply the reranking step on the top 100,000 frames of the initial result. We consider the maximum frame score as the shot score and rank the video shots for the final evaluation.

For the global search, we use Fisher vectors with a Gaussian Mixture Model vocabulary containing 256 components [17]. The Fisher vectors are reduced to 4,096 dimensions with principal component analysis. For the localized analysis, we adopt a much larger vocabulary with 20,000 clusters, and do not apply any dimension reduction. The combination of localized search and difference coding with large vocabularies poses heavy demands on the memory and computation. To handle this, we use a point-indexed representations and decomposed similarity measures, achieving efficient storage and evaluation of many boxes. We refer to the work of Tao *et al*. [21] for the details on the efficient instance search.

3.1 Submitted Runs

UvA-Shifu This run is purely based on the global search. We consider two settings of extracting local descriptors, one with SIFT around interest points detected by Hessian-Affine detector [16] and one with SIFT, TSIFT and C-SIFT around densely sampled points. The two sets are fed into the Fisher vector separately. Finally, the frame score is a weighted sum of two scores. This run has an mAP of 0.052.

UvA-Bajie This run only considers interest points with SIFT. Reranking is applied. This run scores an mAP of 0.108.

UvA-Shaseng This run is similar to the *UvA-Bajie*, but learns a classifier instead of querying by similarity, in both the global search and localized search. A linear SVM classifier is learned using the provided query frames as positive examples and randomly sampled 100,000 frames as negative. We blindly consider the sampled frames as negative without manual checking. This run achieves an mAP of 0.113.

UvA-Wukong This run is similar to *UvA-Shaseng*, but here we use both interest points and densely sampled points as in *UvA-Shifu* for the global search to get the initial result. The reranking is the same as *UvA-Shaseng*. This run scores an mAP of 0.127.

4 Task IV: Event Recognition

Our event recognition system is founded on two representations, one based on low-level multimedia features and one based on a representation of concepts [4,12,13]. In addition, within the SESAME team [3, 14], we also investigate together with SRI International and the University of Southern California several additional multimedia approaches to video event detection.

Multimedia Encoding The system computes a product quantized Fisher vector [6,17] of SIFT, TSIFT, and C-SIFT descriptors [23] on two frames per second. The Fisher vectors are averaged per video to obtain a single feature vector per video. As our audio features, we extract Mel-frequency cepstral coefficients (MFCCs) over a 10ms window. The derivatives of the MFCCs and the second derivative are also computed. The MFCC features are difference coded with Fisher vectors using a Gaussian Mixture Model. As motion features we compute MBH [24] and HOG descriptors along the motion trajectories. Fisher encoding is used to aggregate them followed by power normalization as in [5]. A linear SVM is trained for each feature and event and applied on the videos to obtain confidence scores.

Semantic Encoding The systems uses 346 concept detectors from the TRECVID 2012 SIN task and 1,000 concept detectors from the 2012 ImageNet Large Scale Visual Recognition Challenge to classify two frames per second using the Fisher vector of SIFT, TSIFT, and C-SIFT descriptors. We use three variants to encode the concepts per video, approach one is based on simple averaging [4], approach two is based on difference coding, approach three is based on concept selection via cross entropy [12,13]. We also experimented with a new semantic fusion approach. To learn events we rely on an SVM with χ^2 kernel.

4.1 Submitted Runs

Pre-Specified 100Ex This run combines low-level audio and motion features and semantic encoding with manifold difference coding and semantic fusion. It scores an mAP of 0.281. Our audio-only and visual-only variants score 0.059 and 0.260 respectively.

Pre-Specified 10Ex This run is similar to the 100ex condition, but here we add concept selection into the fusion. It scores an mAP of 0.015. Our audio-only and visual-only variants score 0.026 and 0.140 respectively. Overall, this run came out second in the *Pre-Specified 10Ex* condition, and is the best visual-only system.

AdHoc 100Ex This run maps the low-level audio, visual, motion features and the difference coded concepts into a mid-level semantic representation to perform fusion at the semantic level. It scores an mAP of 0.253. Our audio-only and visual-only variants score 0.056 and 0.238 respectively.

AdHoc 10Ex This run is similar to the *AdHoc 100Ex* condition, but uses less examples. It scores an mAP of 0.143. Our audio-only and visual-only variants score 0.027 and 0.137 respectively. Overall, this run came out third in the *adHoc 10Ex* condition, and is the best visual-only system without relevance feedback.

5 Highlights

We summarize the highlights of our 2013 TRECVID participation in Figure 2.

Acknowledgments

The authors are grateful to NIST and the TRECVID coordinators for the benchmark organization effort. We thank Edgar Meij for sharing code. This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

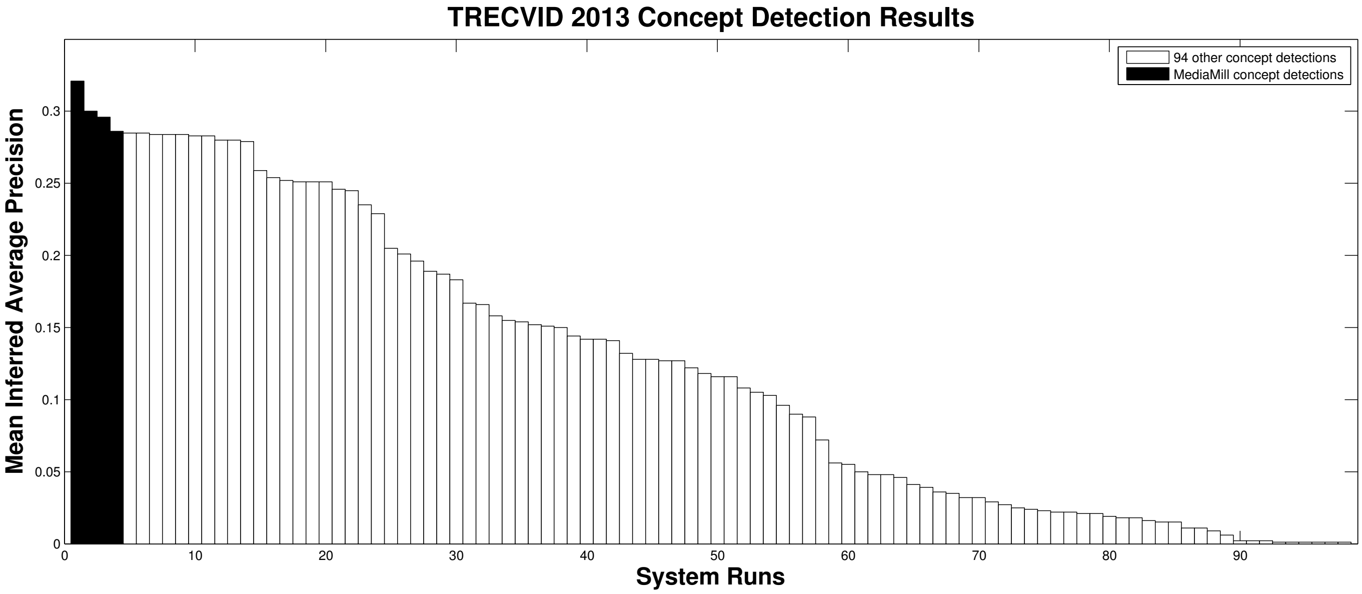
References

[1] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *ECIR*, 2008.

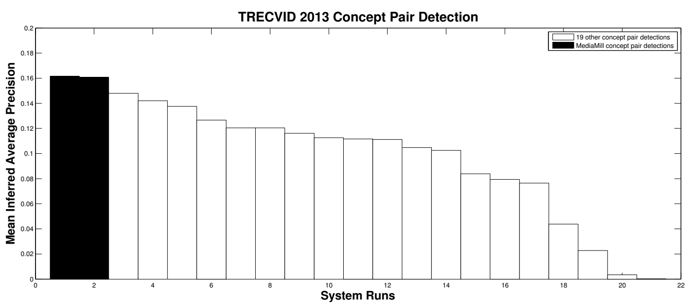
[2] A. Berg, J. Deng, S. Satheesh, H. Su, and F.-F. Li. ImageNet large scale visual recognition challenge 2011, 2011. http://www.image-net.org/challenges/LSVRC/2011.

[3] R. B. Bolles et al. The 2013 SESAME multimedia event detection and recounting system. In *TRECVID Workshop*, 2013.

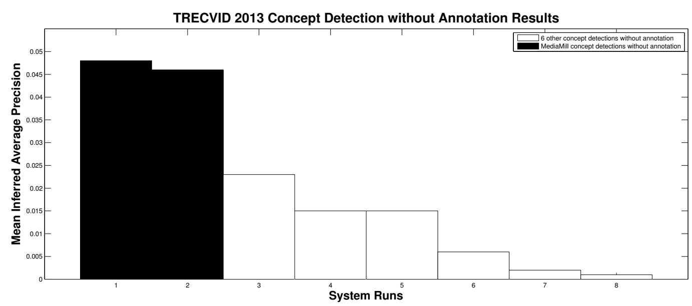
[4] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.



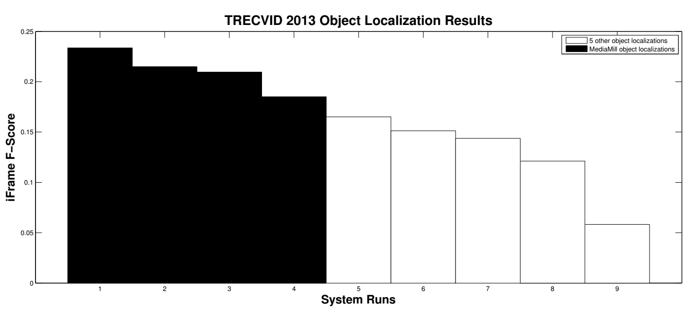
(a) Concept Detection



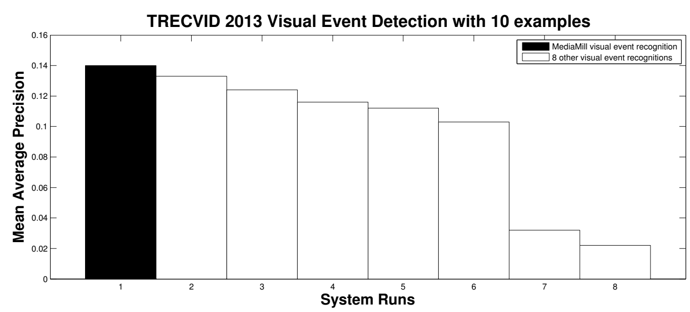
(b) Concept Pair Detection



(c) Concept Detection without Annotation



(d) Object Localization



(e) Visual Event Recognition with ten examples

Figure 2: Comparison of MediaMill video retrieval experiments with other approaches in the TRECVID 2013 benchmark for (a) concept detection, (b) concept pair detection, (c) concept detection without annotation, (d) object localization, and (e) visual event recognition with only ten examples. MediaMill is best overall performer for all five tasks.

[5] M. Jain, H. Jégou, and P. Bouthemy. Better motion for better action recognition. In *CVPR*, 2013.

[6] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 33(1):117–128, 2011.

[7] S. Kordumova, X. Li, and C. G. M. Snoek. Evaluating sources and strategies for learning video concepts from social media. In *CBMI*, 2013.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[9] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE TMM*, 11(7):1310–1322, 2009.

[10] X. Li, C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. Bootstrapping visual categorization with relevant negatives. *IEEE TMM*, 15(4):933–945, 2013.

[11] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[12] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. G. M. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[13] M. Mazloom, A. Habibian, and C. G. M. Snoek. Querying for video events by semantic signatures from few examples. In *ACM Multimedia*, 2013.

[14] G. K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. E. A. van de Sande, A. W. M. Smeulders, and C. G. M. Snoek. Evaluating multimedia features and fusion for example-based event detection. *MVA*, 2013.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *TRECVID Workshop*, 2013.

[16] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009.

[17] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[18] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR*, 2006.

[19] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 43(6):76–78, 2010.

[20] C. G. M. Snoek, K. E. A. van de Sande, A. Habibian, S. Kordumova, Z. Li, M. Mazloom, S.-L. Pintea, R. Tao, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2012 semantic video search engine. In *TRECVID Workshop*, 2012.

[21] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders. Locality in generic instance search from one example. In *CVPR*, 2014.

[22] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[23] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 32(9):1582–1596, 2010.

[24] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.