# Karlsruhe Institute of Technology (KIT) at TRECVID 2009

*Hazım Kemal Ekenel, Arne Schumann, Hua Gao, Rainer Stiefelhagen*

Computer Science Department, Karlsruhe Institute of Technology (KIT)
Adenauerring 2, Karlsruhe 76131, Germany
{ekenel,arne.schumann,hua.gao,stiefel}@ira.uka.de
Web page: http://cvhci.ira.uka.de/

***Abstract*—In this paper, we present the high-level feature detection system developed by the Computer Vision for Human-Computer Interaction Lab (CVHCI) at Karlsruhe Insitute of Technology (KIT) for the TRECVID 2009 evaluation. In our previous two participations, the feature detection system relied exclusively on global features. This year, a completely new system with the focus on local low-level features has been developed. The new system supports temporal sampling as well as spatial partitioning. Local SURF descriptors are computed for grayscale images and different color spaces. The local descriptors are transformed into a more compact histogram representation using a Bag of Words approach. Color Moments and Texture Wavelets are the only two global features remaining in this new system. For each low-level feature and concept in the evaluation, a support vector machine was trained using a grid search scheme based on video-constrained cross-validation. Finally, multiple scores are fused using a simple weighted fusion approach.**

## I. Introduction

In our third TRECVID participation the goal was to include a number of new developments into our concept detection system. Most notably, the new system focuses on local descriptors and no longer on global features. Due to the larger number of concepts describing events or activities in the TRECVID 2009 evaluation, temporal sampling has been included in the system, as well as a number of other options, such as dense sampling, spatial partitioning or video-constrained cross-validation. The importance of those options was investigated in a number of experiments.

This paper is organized as follows: Section II explains the five main segments of the system in more detail. Experimental results for different configurations are presented in Section III and Section IV draws some conclusions and gives ideas for further improvements and future work.

## II. The 2009 System

Our high-level feature detection system, as depicted in Figure 1 consists of five major parts. The *data extraction* segment offers options to extract key images from each shot of the video data. The *feature extraction* segment is responsible for computing a selected number of low-level features and storing them as data vectors. Those vectors are then passed on to the *data reduction* section. Here, the vectors are processed in order to reduce their number and dimension while keeping as much information as possible. The reduced representation is passed to the *machine learning* segment where a support vector machine (SVM) is trained for each low-level feature and target concept. Finally, the *score fusion* part combines multiple classification scores for the same concept into one.

### *A. Data Extraction*

The input to the data extraction component is either the official TRECVID 2009 dataset or an additional dataset from the Quaero project, that is annotated for the same features. For each shot in the shot reference, frames are extracted from the video files using either a standard *keyframe extraction* or *temporal sampling*.

### *B. Low-level Features*

The low-level feature extraction component focuses mostly on local features. For keypoint detection, the blob-based Hessian Determinant keypoint detector of the SURF [3] implementation can be used, as well as the Hessian Affine [12] keypoint detector, the MSER [13] keypoint detector or a dense sampling strategy that uniformly samples the image at a fixed pixel distance and assigns a fixed scale to the sampled points.

Three different kinds of descriptors can be computed for each keypoint. The 128-dimensional, gray-level *SURF descriptors* are used to represent intensity information around the keypoints. In order to include color information into the feature vectors, a separate descriptor can be computed on each channel of the RGB color space. Those three descriptors are then concatenated to a final 384-dimensional feature vector. In order to investigate the importance of color space selection, another local color descriptor based on the opponent color space has been implemented. The three different channels of this color space are computed as follows:

$$I = \frac{R+G+B}{3} \tag{1}$$

$$O_1 = \frac{R+G-2B}{4} + 0.5 \tag{2}$$

$$O_2 = \frac{R-2G+B}{4} + 0.5 \tag{3}$$

Since images with larger homogeneous areas tend to have very few keypoints and are thus not very well described by local

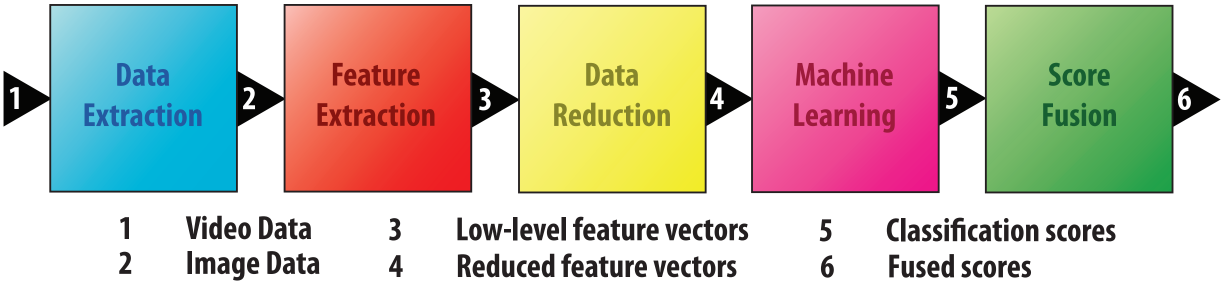

Fig. 1. The five main segments of the 2009 HLF detection system.

features, two global features from the previous system have been included.

*Color moments* of the first order (mean), second (variance) and third order (skewness) are computed on a $3 \times 3$ grid to represent the color distribution in the image. This results in a 81-dimensional feature vector. A $4 \times 4$ grid of *Haar wavelets* as described in [1], [2] is used to capture texture information. A level-4 analysis with 12 subbands results in a feature vector with 192 dimensions.

### C. Data Reduction

The data reduction segment is only used for the local low-level features. The variable number of local descriptors in each image as well as the high dimension of the combined descriptors from one image are both problematic for machine learning. A *Bag of Words* (BoW) approach helps transforming the local features into a single, fixed-length histogram for each image.

In a first step, a large number of local descriptors is used to find cluster centers and create a so-called visual vocabulary. We use a clustering approach based on hierarchical k-means for vocabulary creation. The clustering algorithm used for creating the visual vocabularies was kindly provided by the Learning and Recognition in Vision (LEAR) team at Institut National de Recherche en Informatique et en Automatique (INRIA).

For a single image, each local descriptor can then be assigned to a cluster. Two strategies are available to do this. In *hard assignment* a local descriptor is assigned to exactly one cluster. In *soft assignment* a local descriptor can be assigned to more than one cluster. The latter option is usually more successful in preserving the information contained in the local descriptors. Those assignments are then recorded in a histogram.

Another option available in this segment is *spatial partitioning*. In order to include information about the spatial distribution of the local descriptors, the image is split in three different ways: $1{\times}1$, $1{\times}3$ and $2{\times}2$. For each of these partitions a separate BoW-histogram is computed. The histograms are then concatenated to form the final data vector.

### D. Machine Learning

The model learning section consists mostly of a modified version of LIBSVM. L1-SVMs with a radial-basis kernel are used for classification. Instead of class labels, the modified version returns scores, which correspond to the distance of the feature vector to the hyperplane.

For training, this section offers a cross-validation implementation, that partitions the video data into $n$ separate parts. Following findings in [11], all shots of the same video will be assigned to the same partition to increase validation accuracy. Each vector of the training data is predicted once. Those prediction scores are ranked and evaluated using the average precision. The best performing parameter combination is chosen for the system.

### E. Fusion

The multitude of different combinations of algorithms that can be applied to each image results in a large number of SVM classifiers for each image. The scores produced by those classifiers have to be fused into a single final score for each concept. The fusion is done using a simple weighted sum of all available scores. The weights of those scores are determined using the performance of the classifiers during cross-validation.

### F. Baseline System

The baseline configuration of the system only uses the official TRECVID data and one keyframe per shot. For each keyframe local gray-level descriptors are computed using the Hessian Determinant keypoint detector. Those descriptors are then transformed into BoW histograms. With a 500-dimensional visual vocabulary and all three spatial partitioning options this results in three data vectors of 500, 1500 and 2000 dimensions, respectively. For each of those vectors, a SVM predicts a concept score and the three scores are fused into a final prediction using weighted fusion.

## III. Experiments

The performance of the new system was evaluated in three kinds of experiments. At first, the results from five different runs of a preliminary version of the system were submitted to the TRECVID 2009 evaluation. With the final version of the system, similar runs were performed on the TRECVID 2009 data and evaluated using the ground-truth provided by NIST. And finally, the system's performance was evaluated on the TRECVID 2008 data and compared to the predecessor system [2].

| Run | Mean infAP |
|---|---|
| UKA_BLS | 0.026 |
| UKA_GL | 0.037 |
| UKA_DE | 0.008 |
| UKA_ALL | 0.017 |
| UKA_ALL+OPP | 0.007 |

TABLE I
MEAN PRECISIONS OF THE FIVE OFFICIAL RUNS ON THE 2009 DATA.

| Experiment | Mean infAP |
|---|---|
| EXP_BLS | 0.034 |
| EXP_GL | 0.042 |
| EXP_OPP | 0.036 |
| EXP_TS | 0.038 |
| EXP_Q | 0.045 |
| EXP_ALL | 0.048 |

TABLE II
MEAN PRECISIONS OF FURTHER EXPERIMENTS ON THE 2009 DATA.

If not otherwise indicated, the keypoint detector used for feature extraction is Hessian Determinant.

### *A. Results of the TRECVID 2009 Submissions*

Five different runs were submitted to the TRECVID 2009 evaluations. The *mean inferred average precisions* (mean infAP) [10] of those runs are listed in Table I.

1) For the UKA_BLS run only the baseline system was used.
2) The UKA_GL contains baseline system scores fused with the two sets of scores for the global features. This combination of local descriptors and global features resulted in our best mean infAP.
3) The UKA_DE was performed using gray-level SURF descriptors for densely sampled keypoints at a fixed scale. Again, the scores were fused with those of the baseline system.
4) The UKA_ALL run is a combination of all above runs. The baseline scores are fused with global features and densely sampled local descriptors.
5) The UKA_ALL+OPP run adds local color descriptors using the opponent color space to run 4.

The baseline performance of our system was 0.026. Fusing the baseline scores with the scores of the two global features increased the performance to a value of 0.037. The performance of the remaining three runs suffered because of an error in the computation of the fusion weights.

### *B. Further Experiments on the 2009 Data*

Due to the limited amount of time only a small part of the possible configurations of the new system could be tested and submitted to the TRECVID 2009 evaluation. After the ground truth for the test data became available, we conducted a number of further experiments on the 2009 data to evaluate the final implementation of the system:

1) EXP_BLS - A run of the final implementation of the baseline system on the TRECVID 2009 data.
2) EXP_GL - Fusion of the global feature scores from UKA_GL with the baseline scores from EXP_BLS using the corrected fusion implementation.
3) EXP_OPP - Fusion of detection scores for local color descriptors using the opponent color space with those from EXP_BLS.
4) EXP_TS - A run using the baseline configuration with a temporal sampling of three frames for each shot instead of a single keyframe.
5) EXP_Q - A run using the baseline system with the addition of more positive samples for each concept from the Quaero dataset.
6) EXP_ALL - A fusion of the scores from EXP_TS, EXP_GL and EXP_Q.

The mean infAP values for all experiments are listed in Table II.

With the final implementation of the system, the baseline precision increased to a value of 0.034. The correctly computed weights in the fusion part of the system lead to a significant increase in the fused scores for global features and opponent color features.

Using additional frames for each shot to increase detection performance for those concepts that describe events or activities boosted the mean infAP to 0.038. The greatest increase however was achieved using additional positive samples in the training data.

A fusion of the baseline scores with all additional scores from experiments 2-5 lead to our best precision on the TRECVID 2009 data with a value of 0.048. A direct comparison of the the runs UKA_BLS, UKA_GL, EXP_BLS and EXP_ALL can be seen in Table III.

### *C. Comparison to the 2008 System*

In order to compare this new system to our system from 2008, another set of experiments was conducted on the TRECVID 2008 data:

1) EXP8_BLS - Run of the baseline system on the 2008 data.
2) EXP8_GL - A combination of the baseline system and global features.
3) EXP8_OPP - A combination of the baseline system and local opponent-color-descriptors.
4) EXP8_Q - A run of the baseline system using additional positive samples from the Quaero dataset.
5) EXP8_ALL - A combination of experiments EXP8_Q, EXP8_GL and EXP8_OPP.
6) EXP8_RGB - A combination of the baseline system with local rgb-color-descriptors.
7) EXP8_DE - Baseline configuration using densely sampled keypoints at a fixed scale.
8) EXP8_3KP - Gray-level SURF features computed for three different keypoint descriptors (Hessian Determinant, Hessian Affine and MSER). For each keypoint detector a separate 500-dimensional BoW histogram was computed. The histograms were concatenated to one

| Concept | UKA_BLS | UKA_GL | EXP_BLS | EXP_ALL |
|---|---|---|---|---|
| Classroom | 0.007 | 0.020 | 0.008 | 0.021 |
| Chair | 0.013 | 0.038 | 0.015 | 0.023 |
| Infant | 0.001 | 0.001 | 0.001 | 0.001 |
| Traffic_intersection | 0.038 | 0.067 | 0.041 | 0.056 |
| Doorway | 0.040 | 0.050 | 0.042 | 0.048 |
| Airplane_flying | 0.031 | 0.007 | 0.034 | 0.042 |
| Person-playing-a-musical-instrument | 0.018 | 0.038 | 0.023 | 0.045 |
| Bus | 0.004 | 0.009 | 0.007 | 0.012 |
| Person-playing-soccer | 0.049 | 0.106 | 0.058 | 0.108 |
| Cityscape | 0.072 | 0.071 | 0.079 | 0.101 |
| Person-riding-a-bicycle | 0.009 | 0.009 | 0.009 | 0.011 |
| Telephone | 0.002 | 0.003 | 0.008 | 0.012 |
| Person-eating | 0.000 | 0.000 | 0.001 | 0.001 |
| Demonstration_Or_Protest | 0.005 | 0.005 | 0.005 | 0.009 |
| Hand | 0.008 | 0.008 | 0.014 | 0.014 |
| People-dancing | 0.015 | 0.164 | 0.107 | 0.173 |
| Nighttime | 0.082 | 0.046 | 0.087 | 0.097 |
| Boat_Ship | 0.081 | 0.049 | 0.085 | 0.105 |
| Female-human-face-closeup | 0.051 | 0.062 | 0.060 | 0.084 |
| Singing | 0.001 | 0.005 | 0.005 | 0.012 |
| Mean | 0.0264 | 0.038 | 0.034 | 0.048 |

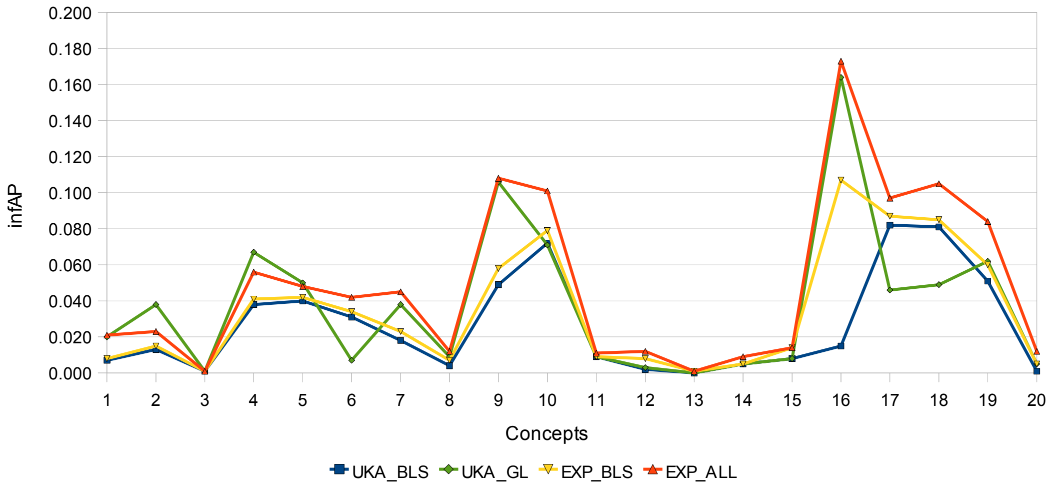

TABLE III
INFERRED AVERAGE PRECISIONS OF DIFFERENT RUNS FOR EACH CONCEPT.

| Experiment | Mean infAP |
|---|---|
| EXP8_BLS | 0.0694 |
| EXP8_GL | 0.0701 |
| EXP8_OPP | 0.0698 |
| EXP8_Q | 0.0729 |
| EXP8_ALL | 0.0731 |
| EXP8_RGB | 0.0695 |
| EXP8_DE | 0.0623 |
| EXP8_3KP | 0.0775 |
| UKA08 | 0.0389 |

TABLE IV
MEAN PRECISIONS OF EXPERIMENTS ON THE 2008 DATA.

vector for each spatial partition. For the partitions 1x3 and 2x2 only Hessian Determinant and Hessian Affine are used to reduce the dimension of the resulting vectors.

Table IV shows the mean infAP values for those experiments compared to the best score of our 2008 system (UKA08).

Again, the additional positive samples result in a big performance increase. The greatest increase was however achieved using additional keypoint detectors[1]. All experiments with the new, local feature based system have a significantly higher performance than the predecessor system.

## IV. CONCLUSION

Overall, the new system was an improvement over the previous version. However, the top results of this year's evaluation show that there is still a lot of room for improvements.

The system still offers many more configurations that will be evaluated. Further experiments will include dense sampling in combination with local color features, finding the optimal number of frames for temporal sampling, evaluating the RGB-descriptors against the Opponent-descriptors and a SVM parameter search on a more dense grid.

[1] The binaries for the Hessian Affine and MSER keypoint detectors were downloaded from http://www.robots.ox.ac.uk/ vgg/research/affine/detectors.html#binaries

Future additions to the system could be an additional keypoint detector for the local features in order to increase robustness to certain types of images, a cross-domain learning approach using images from the web or additional data reduction techniques that are also suitable for the global features. Finally, due to the large number of different system configurations there are still a number of experiments to conduct that might further improve the detection accuracy.

## V. Acknowledgements

We would like to thank the Quaero project for providing the additional training data as well as the Learning and Recognition in Vision (LEAR) team at Institut National de Recherche en Informatique et en Automatique (INRIA) for providing their clustering algorithm for local descriptors.

This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

## References

[1] H. K. Ekenel, M. Fischer, H. Gao, K. Kilgour, J. S. Marcos, and R. Stiefelhagen, ”Universität Karlsruhe (TH) at TRECVID 2007”, *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2007.

[2] H. K. Ekenel, H. Gao, and R. Stiefelhagen, ”Universität Karlsruhe (TH) at TRECVID 2008”, *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2008.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, ”SURF: Speeded Up Robust-Features”, 2008.

[4] Chih-Chung Chang and Chih-Jen Lin, ”LIBSVM: a library for support vector machines”, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[5] S. Chang, J. He, Y. Jiang, A. Yanagawa, E. Zavesky, E. El Khoury, and C. Ngo, ”Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search”, *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2008.

[6] A. Natsev, J. R. Smith, J. Teic’, L. Xie, R. Yan, W. Jiang, and M. Merler, ”IBM Research TRECVID-2008 Video Retrieval System”, *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2008.

[7] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, et al., ”The MediaMill TRECVID 2008 Semantic Video Search Engine”, *NIST TRECVID Workshop*, Gaithersburg, USA, Nov. 2008.

[8] F. Jurie, and B. Triggs, ”Creating Efficient Codebooks for Visual Recognition”, 2005.

[9] Y.-C. Jiang, C.-W. Ngo, J. Yang, ”Towards optimal bag-of-features for object categorization and semantic video retrieval”, ACM CIVR, 2007.

[10] E. Yilmaz, and J.A. Aslam, ”Inferred AP: Estimating Average Precision with Incomplete Judgements”, 2006.

[11] J.C. can Gemert, C.G.M. Snoek, C.J. Veenman, and A.W.M. Smeulders, ”The Influence of Cross-Validation on Video Classification Performance”, 2006.

[12] K. Mikolajezyk, C. Schmid, ”Scale and affine invariant interest point detectors”, 2004.

[13] J. Matas, O. Chum, M. Urban, T. Pajdla, ”Robust wide baseline stereo from maximally stable extremal regions”, 2002