University of Marburg at TRECVID 2009: High-Level Feature Extraction

Markus Mühling[1,2], Ralph Ewerth[1,2], Thilo Stadelmann[1,2], Bing Shi[2],
and Bernd Freisleben[1,2]

[1] *SFB/FK615, University of Siegen, D-57068 Siegen, Germany*
[2] *Dept. of Math. and Computer Science, University of Marburg, D-35032 Marburg, Germany*
{muehling, ewerth, stadelmann, shib, freisleb}@informatik.uni-marburg.de

Abstract

In this paper, we summarize our results for the high-level feature extraction task at TRECVID 2009. Our last year's high-level feature extraction system relied on low-level features as well as on state-of-the-art approaches for camera motion estimation, text detection, face detection and audio segmentation. Based on the observation that the use of face detection results improved the performance of several face related concepts, we have incorporated further specialized object detectors. Using specialized object detectors trained on separate public data sets, object-based features are generated by assembling detection results to object sequences. A shot-based confidence score and additional features, such as position, frame coverage and movement, are computed for each object class. The object detectors are used for two purposes: (a) to provide retrieval results for concepts directly related to the object class (such as using the boat detector for the concept boat), (b) to provide object-based features as additional input for the SVM-based concept classifiers. Thus, other related concepts can also profit from object-based features. Furthermore, we investigated the use of SURF (Speeded Up Robust Features). The use of object-based features improved the high-level feature extraction results significantly. Our best run achieved a mean inferred average precision of 9.53%.

1. Structured Abstract

The results of our participation in the high-level feature extraction task are presented in this section in the form of the requested structured abstract. In the following sections, we describe our system for high-level feature extraction along with the experimental results. In section 2 we describe the extracted features focusing on object-based features and interest point features. The entire high-level feature extraction system is discussed in detail in section 3, while the experimental results are presented in section 4. Section 5 concludes the paper.

"What approach or combination of approaches did you test in each of your submitted runs?"

The following six runs of categories "A" and "C" were submitted:

- A_Marburg1: Baseline;
- A_Marburg2: Baseline plus shot-based confidence scores for the detected object classes;
- A_Marburg3: Marburg2 plus further object-based features such as position, frame coverage and movement;
- A_Marburg4: Bag-of-features approach using scale and rotation invariant interest point features (SURF);
- C_Marburg5: Marburg4, returning object detection results for directly related concepts;
- C_Marburg6: Marburg3, returning object detection results for directly related concepts.

"What, if any significant differences (in terms of what measures) did you find among the runs?"
"Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?"

We extended our baseline system using specialized object detectors trained on separate public data sets for the following object classes: "airplane", "bicycle", "boat", "bus", "car", "chair" and "person". Adding shot-based confidence scores for each object class led to a relative performance improvement of 12.1%. In a second experiment, we supplemented our feature set with additional object-based features like position, frame coverage and movement derived from object sequences. This approach considering additional object-based features improved the previous system

from 8.88% to 9.53% mean inferred average precision and achieved our best result for the high level feature extraction task. Based on this system, we performed a further experiment returning object retrieval results for directly related concepts, like using the chair detector for the concept "chair". This combination of using object detectors for directly related concepts and concept classifiers exploiting object-based features for the remaining concepts achieved no performance gain compared to the previous system neither on the set of related concepts nor on the entire concept set. While the concepts "airplane-flying", "bus" and "person-riding-a-bicycle" were significantly boosted using direct object retrieval results, the performance of the concepts "boat_ship" and "chair" clearly dropped.
Furthermore, we performed two runs based on interest point features using SURF (Speeded Up Robust Features). These runs could not achieve the performance of the baseline system.

"Overall, what did you learn about runs/approaches and the research question(s) that motivated them?"

The experiments revealed that the approaches exploiting object-based features improved the overall high-level feature extraction results significantly. Not only concepts that directly correspond to an object class profited from the additional object-based features, but almost all concepts profited from these features. Using direct object retrieval results, we achieved the second best result among all submitted runs for the concept "person-riding-a-bicycle" and also "airplane-flying" and "bus" were pushed under the top six teams. The approaches using scale- and rotation-invariant interest point features could not unfold their potential.

2. Feature Extraction

Our video analysis system automatically extracts several low-level features as well as mid-level features, which are the result of state-of-the-art algorithms in the field of camera motion estimation [6], text detection [9], face detection [16], object detection [7] and audio segmentation. Compared to our last year's system, we additionally extracted object-based features [13] and investigated scale- and rotation-invariant interest point features using SURF [2]. In this section, we describe our novel features while the remaining features were already described in detail in our last year's TRECVID paper [14]. In section 2.1, we present our object-based features, followed by the interest point features in section 2.2.

2.1 Object-based Features

State-of-the-art object detection approaches [7][9][16] are utilized to find object appearances for the following object classes: "airplane", "bicycle", "boat", "bus", "car", "chair", "face", "overlaid text" and "person".
Using these object detectors trained on separate public data sets, detection results are assembled to object sequences, and a shot-based confidence score as well as further features, like position, frame coverage and movement, are computed. The main components are described below.

2.1.1 Object Detection

In addition to face and text detection already used in our last year's system, a state-of-the-art object detection approach provided by Felzenswalb et al. [8] and released in conjunction with the PASCAL Visual Object Classes (VOC) Challenge 2008 [5] is used. The object models, built on the development data of the VOC challenge, which consists of images obtained from the "flickr" website, are available online at [http://people.cs.uchicago.edu/~pff/latent/]. The approach uses discriminatively trained mixtures of deformable part models and is an extension of their previous work [7]. The models consist of a global template that covers the whole object, several smaller part templates, and a model describing the spatial arrangement of the smaller parts. The templates are based on histograms of gradient features.

2.1.2 Object Sequence Generation

Due to lack of time and the huge amount of video data we used different strategies to assemble object detections to object sequences. For faces, a tracking procedure based on Intel's OpenCV library [12] is used to assemble object appearances in subsequent frames of a shot. First, in the detected object region of a preceding frame, a feature detector is applied to find points of interest that are suitable for tracking. For this purpose, pixels with the highest eigenvalues are selected and tracked in the next frame using the optical flow computation method of Bouguet [3]. An object is tracked successfully if the ratio of tracked feature points within the detected object region of the next frame is above a predefined percentage value.
For the remaining objects, we used agglomerative single linkage clustering to assemble object detections within a shot to object sequences, which is computationally much faster than the previous approach. For this purpose, the distance between two detected object regions considers position, size, frame

number and detection score. The clustering process stops if no more clusters can be merged due to overlapping object regions or if a predefined threshold is exceeded.
Text detection results are treated differently. Under the assumption that overlaid text is constant in position and size, text detection results are assembled to sequences, if the overlaid text is detected at approximately the same position and size for several subsequent I-Frames, otherwise it is discarded.

2.1.3 Derived Features

The previously generated object sequences are used to extract the following features. First, a shot-based confidence score for each object class is calculated. The computation of confidence scores is slightly different for face, text and the remaining object sequences. For face sequences, the average number of detection hits per sequence is computed. If several face sequences exist, then the maximum average value is chosen as confidence score. For text sequences, the accumulated frame coverage of all text elements is used. The remaining object sequences are treated similarly to face sequences using detection scores instead of detection hits.
Second, further features are derived from the object sequences. The first feature is the number of object sequences per class. Furthermore, for each object, the sequence with the highest average confidence score per shot is selected and the following features are extracted: average object position, average frame coverage and movement. Movement describes the maximum distance between two object positions. For face sequences, the percentage of detected profile faces and the ratio of sequence length versus shot length are calculated additionally. Due to the absence of an appropriate confidence score for text sequences, the detected text areas are used to derive the following features: number of appearing text elements, average text position and average, maximum and accumulated text frame coverage.

2.2 Interest Point Features

In a bag-of-features approach we investigated the use of scale- and rotation-invariant interest point features. Our approach is based on the OpenCV implementation, called "Speeded Up Robust Features" (SURF), which is an enhancement of the SIFT features [11] focusing on computational performance. This scale- and rotation-invariant interest point detector and descriptor relies on integral images and thus can be computed and compared much faster. The interest points are detected using a Hessian matrix-based measure and described by 128-dimensional feature vectors, orientation and scale. The distance between two interest point descriptors is computed using the Euclidean distance between the feature vectors ignoring orientation and scale.
Interest point descriptors extracted from keyframes are clustered to build a visual vocabulary using a K-Means algorithm. Due to the huge number of interest points we only used keyframes from positive labeled training shots to construct the visual vocabulary. Using this vocabulary, the shots are described as a feature vector indicating the presence of each visual word. The histograms are generated by mapping the bag of interest point descriptors from a keyframe to the visual words. Instead of just increasing the nearest neighbor, we used a soft-weighting scheme [10]. We built a visual vocabulary consisting of 1000 visual words to obtain 1000-dimensional feature vectors that serve as the input for our SVM-based concept classifiers.

3. High-Level Feature Extraction System

The goal of the proposed system is to learn models for the high-level semantic features based on the extracted audiovisual low-level and mid-level features. In our baseline system, we concatenated the multi-modal low-level and mid-level features in an early fusion scheme and fed them directly into a support vector machine with radial basis function kernel using the implementation provided by the libSVM library [4]. To reduce the unbalance of positive and negative training samples, which concerns nearly all concepts, we reduced the number of negative instances by sub-sampling. For each concept, we used only every fourth negative training sample. The sub-sampling of negative instances not only accelerates the process of building the concept model but most notably led to clearly better results in terms of mean inferred average precision last year.
This year, we supplemented the feature set of the baseline system using object-based features described in section 2.1 (see Figure 1). The shot-based confidence scores for the detected object classes as well as further features derived from object sequences serve as additional input for the SVM-based concept classifiers. Besides using object detectors to derive additional input for the concept classifiers, they can also be used to directly return results for the concepts "airplane-flying", "boat_ship", "bus", "chair" and "person-riding-a-bicycle". These results are ranked according to the previously described shot-based confidence scores.

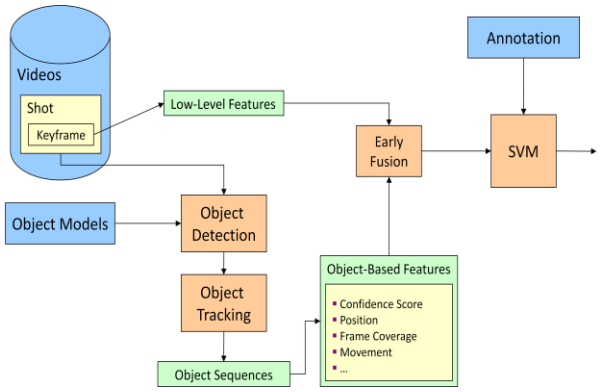

Figure 1: High-level feature extraction system using object-based features.

The ranking of the object retrieval results for the concept "airplane-flying" additionally considers the object-based feature "movement", and the results for the concept "person-riding-a-bicycle" are sorted according to a combination of confidence scores for person and bicycle.

Furthermore, we focused on interest point features using SURF. The 1000-dimensional feature vectors described in section 2.2 were directly fed into a SVM to build the final concept classifiers. This system solely relies on interest point features.

4. Experimental Results

In this section, we present our results for the high-level feature extraction task. We submitted four runs of category "A" and two runs of category "C". The high-level feature extraction experiments were evaluated by the TRECVID team [15] using the inferred average precision measure suggested by Aslam et al. [1].

Last year's high-level feature extraction system (A_Marburg1) served as a basis for our experiments this year. Interestingly, the performance comparison based on the subset of concepts also tested in 2008 reveals a performance decrease from 8.08% in 2008 to 5.35% in 2009 in terms of mean inferred average precision.

In a first experiment (A_Marburg2), we added shot-based confidence scores for each object class to the feature set of our baseline system, which led to a relative performance improvement of 12.1%. In a second experiment (A_Marburg3), we supplemented our feature set with additional object-based features like position, frame coverage and movement derived from object sequences. This approach considering additional object-based features improved the previous system from 8.88% to 9.53% mean inferred average precision and achieved our best overall result for the high-level feature extraction task.

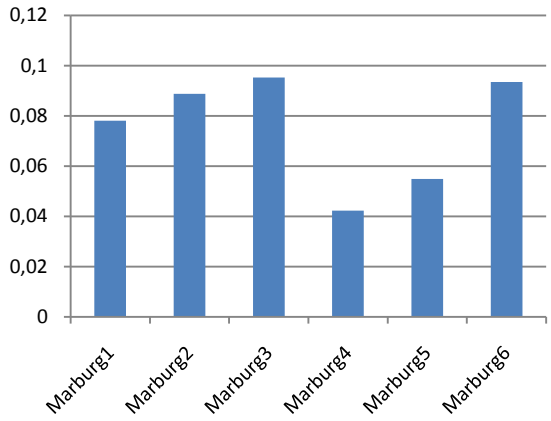

Figure 2: Overview of the results of our six runs in terms of mean inferred average precision.

Based on this system, we performed a further experiment (C_Marburg6) returning object retrieval results for directly related concepts, such as using the chair detector for the concept "chair". This combination of concept classifiers exploiting object detection results as additional features on the one hand and returning object detection results for directly related concepts on the other hand achieved no performance gain compared to the previous system neither on the entire concept set nor on the set of directly related concepts. While the concepts "airplane-flying", "bus" and "person-riding-a-bicycle" were strongly boosted by directly returning object detection results, the performance of the concepts "boat_ship" and "chair" clearly dropped (see Figure 3). Nevertheless, using direct object retrieval results, we achieved the second best result among all

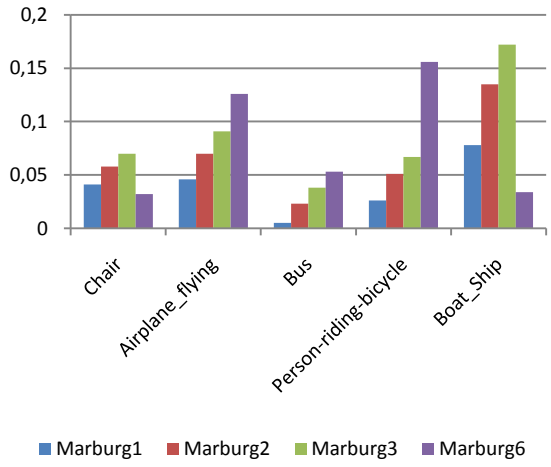

Figure 3: Comparison of our four runs exploiting object detection results on the set of directly related concepts in terms of inferred average precision.

submitted runs for the concept "person-riding-a-bicycle" and also "airplane-flying" and "bus" were pushed under the top six teams. Furthermore, we performed two runs based on scale- and rotation-invariant interest point features using SURF. These runs (A_Marburg4 and C_Marburg5) could not achieve the performance of the baseline system. Figure 2 shows the results of all our submitted runs in terms of mean inferred average precision.

5. Conclusions

In this paper, we have presented our experiments for the high-level feature extraction task. Based on the observation that the use of face detection results in our last year's system improved the performance of several related concepts, specialized object detectors for further object classes have been incorporated. Using object detectors trained on separate public data sets, detection results were assembled to object sequences, and a shot-based confidence score as well as several further features, like position, frame coverage and movement, were computed. The experiments revealed that the approaches exploiting object-based features improved the overall high-level feature extraction results significantly. Almost all concepts and not only concepts that directly correspond to the object classes profited from the additional object-based features.

The system based on interest point features using SURF could not unfold its potential. Further experiments and a comparison to SIFT features are necessary for a better understanding of the system behavior.

Finally, our best run based on object-based features obtained a mean inferred average precision of 9.53%.

6. Acknowledgements

This work is financially supported by the Deutsche Forschungsgemeinschaft (SFB/FK 615, Teilprojekt MT).

7. References

1. Aslam, J. A., Pavlu V., and Yilmaz, E. Statistical Method for System Evaluation Using Incomplete Judgments. In *Proceedings of the 29th ACM SIGIR Conference*, Seattle, 2006, pp. 541-548.
2. Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. SURF: Speeded-Up Robust Features. *Computer Vision and Image Understanding*, 110, 3, 2008, pp. 346-359.
3. Bouguet, J.-Y. Pyramidal Implementation of the Lucas Kanade Feature Tracker. In *OpenCV Documentation, Intel Corporation*, Microprocessor Labs, 1999.
4. Chang, C.-C. and Lin, C.-J. LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., and Zisserman A. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. www.pascal-network.org/challenges/VOC/voc2008/.
6. Ewerth, R., Schwalb, M., Tessmann, P., and Freisleben, B. Estimation of Arbitrary Camera Motion in MPEG Videos. In *Proc. of the 17th International Conference on Pattern Recognition*, Vol. 1, Cambridge, UK, 2004, pp. 512-515.
7. Felzenszwalb, P., McAllester, D., and Ramanan, D. A Discriminatively Trained, Multiscale, Deformable Part Model. In *Proceedings of the IEEE Computer Vision and Pattern Recognition 2008*, pp. 1-8.
8. Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. Discriminatively Trained Mixtures of Deformable Part Models, PASCAL VOC Challenge 2008, http://www.pascal-network.org/challenges/VOC /voc2008/.
9. Gllavata J., Ewerth R., and Freisleben B. Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In *Proceedings of 17th Int. Conference on Pattern Recognition*, Vol. 1, Cambridge, UK, 2004, pp. 425-428.
10. Jiang, Y., Ngo, C., and Yang, J. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proceedings of the 6th ACM Int'l Conference on Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 494-501.
11. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal of Computer Vision*, 60, 2, 2004, pp. 91-110.
12. OpenCV, Open Computer Vision library, http://sourceforge.net/projects/opencvlibrary.
13. Mühling, M., Ewerth, R., and Freisleben, B. Improving Semantic Video Retrieval via Object-Based Features, In *Proc. of the 3rd IEEE Int'l Conference on Semantic Computing*, Berkeley, USA, 2009, pp. 109-115.
14. Mühling, M., Ewerth, R., Stadelmann, T., Shi B., and Freisleben, B. University of Marburg at TRECVID 2008: High Level Feature Extraction. In *Online Proceedings of TRECVID Conference Series 2008*: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs. org.html
15. Smeaton, A. F., Over, P., and Kraaij, W. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, California, USA, 2006, pp. 321-330.
16. Viola, P. and Jones, M. J. Robust Real-Time Face Detection. In *International Journal of Computer Vision*, 57(2), Kluwer Academic Publishers, Netherlands, 2004, pp. 137-154.