University of Marburg at TRECVID 2011: Semantic Indexing Task

Markus Mühling, Khalid Ballafkir, Ralph Ewerth, and Bernd Freisleben
Department of Mathematics & Computer Science
University of Marburg
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{muehling, ewerth, freisleb}@informatik.uni-marburg.de,
ballafki@students.uni-marburg.de

ABSTRACT

In this paper, we summarize our results for the semantic indexing task at TRECVID 2011. Last year, we showed that the use of object detection results as additional mid-level features improved the overall performance of a bag-of-visual-words (BoVW) approach. This year, we repeated the experiment on a large concept vocabulary of 346 classes. In addition, we investigated whether feature descriptions of object regions can also improve the concept detection performance. Due to the large number of face-related concepts, like "adult", "female", "male", "dark skinned person", "first lady", "glasses", or "arafat", BoVW features are extracted from face regions and are used as an additional feature representation. Furthermore, a new post-processing scheme is introduced, that leads to a rescoring of shots based on concept relations. The experiments showed that the use of additional object-based features significantly improved the concept detection performance. Further improvements are attained using region-based BoVW features and relation-based rescoring. Altogether, our best run achieved a mean inferred average precision of 12.3% and we submitted the best results for the concepts "overlaid text" and "two persons".

1. INTRODUCTION

The results of our participation in the semantic indexing task (also known as high-level feature extraction task) are presented in this section in form of the requested structured abstract. In the following sections, we present our system for semantic indexing along with the experimental results. In Section 2, the the different feature types of the semantic indexing system are described. The multiple kernel learning (MKL) framework and the relation-based rescoring scheme are explained in Section 3 and 4, respectively. The experimental results are discussed in Section 5, while Section 6 concludes the paper.

"What approach or combination of approaches did you test in each of your submitted runs?"
Considering the complete list of 346 concepts the following four runs of category "A" were submitted:

- F_A_Marburg1_4: Baseline (BoVW with densely sampled RGB-SIFT features and spatial pyramids)
- F_A_Marburg2_3: Baseline plus object-based features using MKL
- F_A_Marburg3_2: Marburg2 plus BoVW features of face regions
- F_A_Marburg4_1: Marburg3 plus relation-based rescoring

"What, if any significant differences (in terms of what measures) did you find among the runs?" and "Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?"
The supplementation of state-of-the-art BoVW features with object-based features is investigated. The different feature representations are combined using MKL. The runs using additional object-based features significantly improved the overall performance of the baseline system. The relative performance improvement of Marburg2 compared to the baseline system amounts to 9.1%. Besides confidence scores of object detectors, we used an additional BoVW-based feature representation for face regions. Most face-related concepts, like "old people" or "speaking to camera", could be slightly improved by adding these features. Furthermore, a post-processing framework that rescores shots based on concept relations is introduced. The relative performance improvement of the relation-based rescoring framework obtained 3.7% compared to the reference system. Overall, the last run (Marburg4) achieved our best result with 12.3% mean inferred average precision compared to 10.7% of the baseline system, which is a relativ performance improvement of 14.8%.

"Overall, what did you learn about runs/approaches and the research question(s) that motivated them?"
The experiments revealed the usefulness of object-based features. The additional use of object-detection results as additional mid-level features yielded significant performance improvements. Many concepts clearly profited from the additional object-based features, such as "anchor person", "car", "quadruped", "streets", "speaking to camera", "table", "traffic" or "two people". Further improvements of the overall performance were obtained by adding BoVW-based features for face regions and by applying relation-based rescoring. In particular, the concepts "news" and "studio with anchorperson" took advantage of the relation-based rescoring framework. In comparison to other teams we achieved the best results for the concepts "overlaid text" with 14.2% inferred average precision and "two people" with 6.7%.

2. CONCEPT DETECTION SYSTEM

Since state-of-the-art semantic concept detection systems mainly rely on the BoVW approach, our current baseline system employs this feature representation. Based on the success of object-based features in our last years' systems [8][9][10], we incorporated again the results of specialized object detectors trained on separate public data sets. In addition, BoVW features are extracted from face regions, since a large set of concept classes is related to the object class "face" or "person", respectively. As an appropriate fusion scheme, MKL is used to combine different feature representations. Furthermore, we introduced a post-processing framework to integrate concept relations. While the extracted BoVW features are introduced in Section 2.1, the object-based features, including additional BoVW features from detected face regions, are presented in Section 2.2. The MKL framework is described in Section 3 followed by the relation-based rescoring scheme in Section 4.

2.1 Bag-of-Visual-Words

A dense sampling strategy is performed to extract SIFT [7] descriptors, because the sparse representation using keypoint detectors like Harris-Laplace or DoG is often insufficient to describe natural images. To extract dense sampled SIFT features, the Vision Lab Features Library (VLFEAT) [14] is used. It provides a fast algorithm for the calculation of a large number of SIFT descriptors of densely sampled features of the same scale and orientation. The SIFT descriptor geometry is specified by the number and size of the spatial bins and the number of orientation bins. A sampling step size of 5 pixels, 8 orientation bins, and 4x4 spatial bins with a spatial bin size of 5 pixels were used. Thus, the resulting keypoint descriptors form a 128-dimensional feature vector. Similar to the representation of documents in the field of text retrieval, an image is represented by a bag of visual words. Therefore a visual vocabulary is needed, which is generated from a set of training images. The extracted keypoint descriptors are clustered in their feature space using K-means and the cluster centers are interpreted as visual words. Due to the huge amount of keypoints per image we only used 10 positively labeled training shots or keyframes, respectively, per concept to construct a 2000-dimensional vocabulary. Based on this vocabulary, histograms are generated per shot by mapping the extracted SIFT descriptors to the visual words. Instead of just considering the nearest neighbor, a soft-weighting scheme similar to the one of Jiang et al. [4] is used to reduce the quantization loss.

Color Information

Color information is integrated using RGB-SIFT descriptors. The SIFT descriptors are computed independently for the three channels of the RGB color model. Thus, the final keypoint descriptor is the concatenation of the individual descriptors, resulting in a 3x128-dimensional feature vector. Due to the normalization during the SIFT feature extraction, the RGB-SIFT descriptor is equal to the transformed color SIFT descriptor, and is therefore invariant against light intensity and color changes or shifts, respectively [13].

Spatial Information

Since all geometric information gets lost during histogram generation, a spatial pyramid representation is additionally used. Histograms of visual words are calculated for the whole image as well as for a spatial image partitioning of 2x2 regions, resulting in a concatenated feature vector of 10000 dimensions.

2.2 Object-based Features

State-of-the-art object detectors [3][15] are utilized to find object appearances for the following 21 object classes: "face", "aeroplane", "bicycle", "bird", "boat", "bottle", "bus", "car", "cat", "chair", "cow", "dining table", "dog", "horse", "motorbike", "person", "potted plant", "sheep", "sofa", "train", "tvmonitor". The object detectors, trained on separate public data sets, are applied to the keyframe images and shot-based confidence scores as well as further derived features are computed. Frontal faces are detected using the Adaboost-based approach provided by the OpenCV library [2], which is an implementation of the approach suggested by Viola and Jones [15] with Lienhart's extensions [6]. Since this approach usually reports many detections for a face of slightly different sizes and positions, an average rectangle is computed based on the reported detections, and the number of detections is used as a confidence score. For each shot, the number of faces, the average and maximum confidence score, as well as the average and the maximum size of the detected bounding boxes are used as features. The remaining object classes are detected using an approach based on deformable part models [3]. Each object model consists of six components, which intuitively corresponds to the different views of an object. Per shot the maximum component-based detection scores for each object class are used as mid-level features, resulting in 120 feature values. Together with the face detection results, we obtain a 125 dimensional feature vector.

Furthermore, histograms of visual words are extracted from detected face regions. The face class is chosen exemplarily due to the very good face detection results and the large number of face-related concepts. Therefore, an additional face-related codebook of 1000 visual words is constructed. For the codebook as well as for the histogram generation, the face regions are scaled to 50x50 pixels and dense sampled RGB-SIFT descriptors are extracted with a sampling step size of 2 pixels, 8 orientation bins, and a spatial bin size of 4. The integration of color and the soft weighting scheme are employed as described in Section 2.1. If several face regions are detected in an image, they are summarized in a single histogram.

3. MULTIPLE KERNEL LEARNING

Object-based features and BoVW features are combined using MKL. This fusion strategy tries to find an optimal kernel weighting

$$k_{combined} = \alpha \cdot k_{bovw} + \beta \cdot k_{obj} \quad with\ \alpha \geq 0,\ \beta \geq 0 \qquad (1)$$

where the kernel functions k_{bovw} and k_{obj} take into account both feature modalities. For all feature representations the χ^2-kernel is used to measure the similarity between two data instances. It is based on the corresponding histogram distance:

$$k_{\chi^2}(x, y) = e^{-\gamma\chi^2(x,y)} \qquad (2)$$

with

$$\chi^2(x, y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}. \qquad (3)$$

Since the l_2-norm gained the best results in our last year's system this norm is exclusively used to control the sparsity of the kernel weights in the MKL framework. Throughout our experiments, we use the MKL framework provided by the Shogun library [12] in combination with the SVM implementation of Joachims [5], called SVM^{light}.

4. RELATION-BASED RESCORING

In a post-processing step, concept relations are exploited in order to improve the concept detection results. Two types of relations between the semantic concepts are provided by the organizers of the semantic indexing task: implications and exclusions. The relation "A $\Rightarrow$ B" for example is valid, if concept A is a subclass or specification, respectively, of concept B. This relation is also true if concept B is a part of concept A. Besides implications, two concepts can exclude each other such as "Indoor" and "Outdoor". The given set of concept relations contains 427 implications and 559 exclusions. The relations are implemented in a simple post-processing framework by adding and subtracting scores from the individual concept detection results. The two types of relations are processed as follows:

A implies B

The relation "A $\Rightarrow$ B" is realized by taking the positive shots of concept A into account and increasing the corresponding scores of concept B by this value (see lines 1-7 of Listing 1). Additionally the logically implicated relation "$\neg$B $\Rightarrow$ $\neg$A" is considered and the confidence scores of concept A are reduced for shots with a negative score for concept B (see lines 1-7 of Listing 1). But this relation is only applied if the detector of concept B predicted at least one positive shot on the test set (see line 9 of Listing 1). Otherwise, we do not trust the corresponding results of concept B.

Listing 1: A implies B.

```
1  foreach shot in shots
   do
3    if (score_A(shot) > 0)
     then
5      score_B(shot) += score_A(shot)
     fi
7  done

9  if ( max_{shot ∈ shots} (score_B(shot)) > 0)
   then
11   foreach shot in shots
     do
13     if (score_B(shot) < 0)
       then
15       score_A(shot) += score_B(shot)
       fi
17   done
   fi
```

A excludes B

Exclusions are bidirectional, which means that "A excludes B" as well as "B excludes A" is valid. For both directions a simple kind of confidence prediction as already described in the previous section is performed (see lines 1 and 9 of Listing 2) in order to decide whether the relation is useful. The relation "A excludes B" is realized by subtracting the confidence scores of concept A from the corresponding scores of concept B and vice versa (see Listing 2).

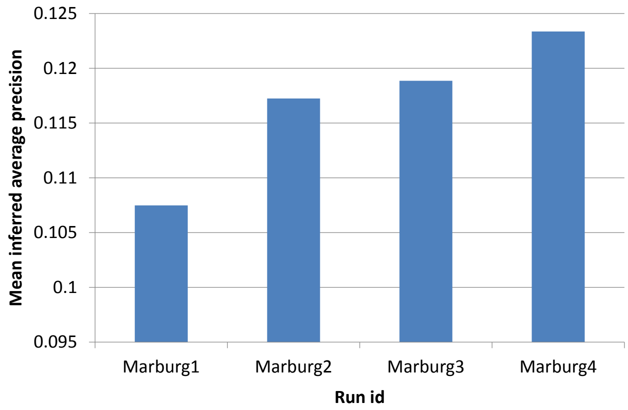

Figure 1: Performance evaluation of the four submitted runs in terms of mean inferred average precision.

Listing 2: A excludes B.

```
   if ( max_{shot ∈ shots} (score_A(shot)) > 0)
2  then
     foreach shot in shots
4    do
       score_B(shot) -= score_A(shot)
6    done
   fi
8
   if ( max_{shot ∈ shots} (score_B(shot)) > 0)
10 then
     foreach shot in shots
12   do
       score_A(shot) -= score_B(shot)
14   done
   fi
```

5. EXPERIMENTAL RESULTS

In this section the experimental results of the four submitted runs are presented. These runs are "full" submissions and belong to the category "A". 50 out of the 346 semantic concepts were evaluated by the TRECVID team [11] based on the inferred average precision measure suggested by Aslam et al. [1]. Additionally, the official partial randomization test [11] is used to determine whether a system is significantly better than the reference system. Figure 1 shows the results of the four submitted runs in terms of mean inferred average precision. The BoVW approach in combination with spatial pyramids served as a baseline system for our experiments (Marburg1). In the second experiment (Marburg2), the BoVW features are combined with object-based features using MKL. This approach considering additional object-based features was significantly better than the baseline system at a significance level of 0.01. Many concepts clearly profited from the additional object-based features, for example, "anchor person", "car", "quadruped", "streets", "speaking to camera", "table", "traffic" or "two people". In a further experiment the feature representations are supplemented by histograms of visual words from detected face regions. These region-based BoVW features are again combined with the previous representations using MKL. This run (Marburg3) achieved only slight performance improvements compared to Marburg2. Most face-related concepts could be slightly improved like "old people" or "speaking to camera". In the last run (Marburg4), concept relations are exploited in a post-processing step to improve the detection results. Compared to the reference system (Marburg3), a relative performance improvement of 3.8% is achieved. In particular, the concepts "news" and "studio with anchorperson" took advantage of the relation-based rescoring. Using all extensions, the last run achieved our best overall performance for the semantic indexing task improving our baseline system from 10.7% to 12.3% mean inferred average precision.

6. CONCLUSIONS

In this paper, we presented our experiments for the semantic indexing task. Based on the success of object-based features in our last year's system, object detection results are again incorporated as additional mid-level features. Due to the large number of face-related concepts, region-based BoVW features are additionally extracted from face regions. The different feature representations are combined using MKL. The experiments revealed that the approaches employing additional object-based features significantly improved the overall performance of the baseline system. The region-based BoVW features could also achieve slight performance improvements. Furthermore, concept relations are exploited in a post-processing step to improve the detection results. The relation-based rescoring framework further improved the results and yielded our best overall performance with a mean inferred average precision of 12.3%. In comparison to other teams, we achieved the best results for the concepts "overlaid text" with 14.2% inferred average precision and "two people" with 6.7%. Overall, we were among the five best teams for the concepts "car", "female human face", "overlaid text", "two people" and "text".

7. ACKNOWLEDGMENTS

This work is financially supported by the German Research Foundation (DFG, PAK 509, Project MT) and the German Ministry of Education and Research (BMBF, D-Grid Initiative, Project MediaGrid).

8. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A Statistical Method for System Evaluation using Incomplete Judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 541–548, New York, NY, USA, 2006. ACM.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade Object Detection with Deformable Part Models. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2241–2248. IEEE Computer Society, 2010.

[4] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 494–501, NY, USA, 2007. ACM.

[5] T. Joachims. Text Categorization With Support Vector Machines: Learning With Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142. Springer, 1998.

[6] R. Lienhart, L. Liang, and A. Kuranov. A Detector Tree of Boosted Classifiers for Real-time Object Detection and Tracking. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, volume 1, pages 277–280, Baltimore, Maryland, USA, 2003.

[7] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[8] M. Mühling, R. Ewerth, and B. Freisleben. Improving Semantic Video Retrieval via Object-Based Features. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 109–115, 2009.

[9] M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2009: High-Level Feature Extraction. In *Proceedings of TREC Video Retrieval Evaluation Workshop (TRECVid'09)*, 2009.

[10] M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2010: Semantic Indexing. In *Proceedings of TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, 2010.

[11] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, New York, USA, 2006. ACM.

[12] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, 99:1799–1802, 2010.

[13] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A Comparison of Color Features for Visual Concept Classification. In *Proceedings of the International Conference on Content-Based Image and Video Retrieval*, pages 141–150, NY, USA, 2008. ACM.

[14] A. Vedaldi and B. Fulkerson. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/, 2008.

[15] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.