The Fraunhofer IDMT at ImageCLEF 2011 Photo Annotation Task

Karolin Nagel, Stefanie Nowak, Uwe Kühhirt and Kay Wolter

Fraunhofer Institute for Digital Media Technology (IDMT)
Ehrenbergstr. 31, 98693 Ilmenau, Germany
uwe.kuehhirt@idmt.fraunhofer.de, research@stefanie-nowak.de

Abstract. This paper presents the participation of the Fraunhofer IDMT in the ImageCLEF 2011 Photo Annotation Task. Our approach is focused on text-based features and strategies to combine visual and textual information. First, we apply a pre-processing step on the provided Flickr tags to reduce noise. For each concept, tf-idf values per tag are computed and used to construct a text-based descriptor. Second, we extract RGB-SIFT descriptors using the codebook approach. Visual and text-based features are combined, once with early fusion and once with late fusion. The concepts are learned with SVM classifiers. Further, a post-processing step compares tags and concept names to each other. Our submission consists of one text-only and four multi-modal runs. The results show, that a combination of text-based and visual-features improves the result. Best results are achieved with the late fusion approach. The post-processing step only improves the results for some concepts, while others worsen. Overall, we scored a Mean Average Precision (MAP) of 37.1% and an example-based F-Measure (F-ex) of 55.2%.

Keywords: image annotation, multi-modal fusion, tag features

1 Introduction

The ImageCLEF 2011 Photo Annotation Task challenges participants to evaluate their multi-label image annotation approaches on a set of Flickr images with the goal to achieve the most accurate annotation of these images. The images belong to 99 different concepts. These range from scene descriptions such as place and time over abstract categories, e.g., *partylife* to very specific concepts such as *dog* or *car*. This year's newly added concepts focus on emotions that the images convey, e.g., *happy* or *melancholic*. In addition to the images and concept associations, the participants are provided with the Flickr user tags and EXIF data of the images. A detailed overview of the data set and the task can be found in [1].

Our main objective to solve this task is to explore how tags can be combined with visual features in order to optimize the annotation result.

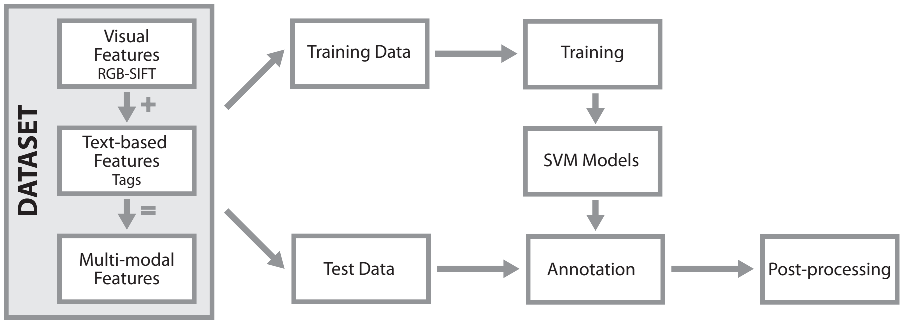

Fig. 1: Overview of the annotation system.

2 System Overview

In Figure 1, an overview of our annotation system is shown. We use visual and textual information of the training data to learn models. These are then employed to annotate the test data. Afterwards, a post-processing step is applied. The following sections describe each step in detail.

2.1 Feature Extraction

Visual Features: As our focus lies on the text-based features and the combination of different modalities, we only use one visual descriptor. The baseline makes use of dense-sampled RGB-SIFT descriptors [2]. These scale-invariant features describe the form and shape of a region around a certain pixel using edge orientation histograms [3]. They are extracted on a 6 pixel wide grid and post-processed with a k-means algorithm to generate a dictionary which contains 2,000 visual words.

Text-based Features: We use the Flickr user tags to construct text-based feature descriptors. As tagging on Flickr is relatively free, tags exist in different languages and word variations. In order to reduce this redundancy, we pre-process the tags prior to the generation of textual features. First, all Flickr user tags are translated into English by using the Google Translate API [4]. Afterwards, tags are stemmed with the help of the Porter Stemming Algorithm [5] in order to merge word variations like *explorer* – *explored* into one tag.

We employ a supervised approach which learns tag frequencies on the concepts of the training set. Similar to the group of Meiji University [6], concept-based tf-idf weights [7] are assigned to each tag. A tag's term frequency (tf) is detected by counting the number of times the tag occurs in a certain concept. The document frequency (df) term is equivalent to the fraction of concepts the tag t appears in, as shown in Equation 1. Therefore, tags that appear very often in only a few concepts get higher weights assigned than tags that appear fairly often in many concepts:

$$df_t = \frac{\text{number of concepts with tag } t}{\text{total number of concepts}}. \tag{1}$$

Finally, the inverse document frequency (*idf*) is calculated as $log(df_t)$.

For each concept, the tf-idf values of the tags of an image are summed up. This leads to a feature vector containing 99 elements with scores normalized in the range of $[0; 1]$. These features are then employed in the learning stage.

2.2 Concept Learning and Annotation

For each concept, a SVM with RBF kernel is learned using the one-against-all strategy and optimized with the concept-based F-Measure on the training set. To combine visual and textual features, we employ two different approaches: early fusion and late fusion.

For the early fusion approach, both, visual and text-based features, are considered simultaneously to learn the SVM models. The late fusion approach learns SVM models for each modality separately and then combines the classification results using the geometric mean.

2.3 Post-processing

To further optimize the annotation result, we apply a simple post-processing step. Each image's tags are again translated and stemmed and afterwards compared to the concept names, which are stemmed as well. In case a concept consist of more than one word, the tags are compared to each of these words. If a tag and at least one word of the concept match, the image is assigned to that concept.

3 Submission

We submitted five different runs in total. One run uses only textual information, the other four runs make use of multi-modal information sources.

- Tags only
- Early fusion of RGB-SIFT and tags
- Early fusion of RGB-SIFT and tags with post-processing step
- Late fusion of RGB-SIFT and tags
- Late fusion of RGB-SIFT and tags with post-processing step

4 Results and Discussion

The results are evaluated with concept-based and example-based performance measures. Detailed information about the evaluation process can be found in [1].

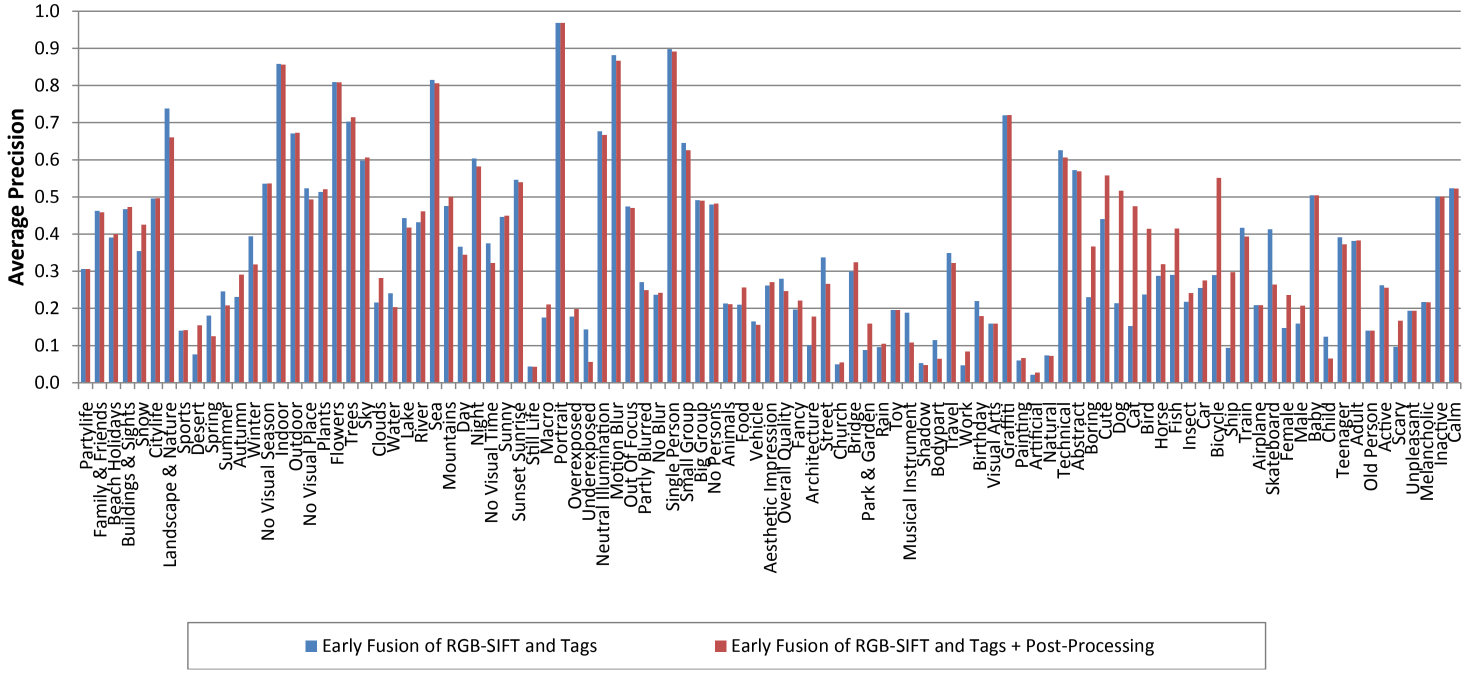

Fig. 2: Comparison of the early fusion results with and without post-processing.

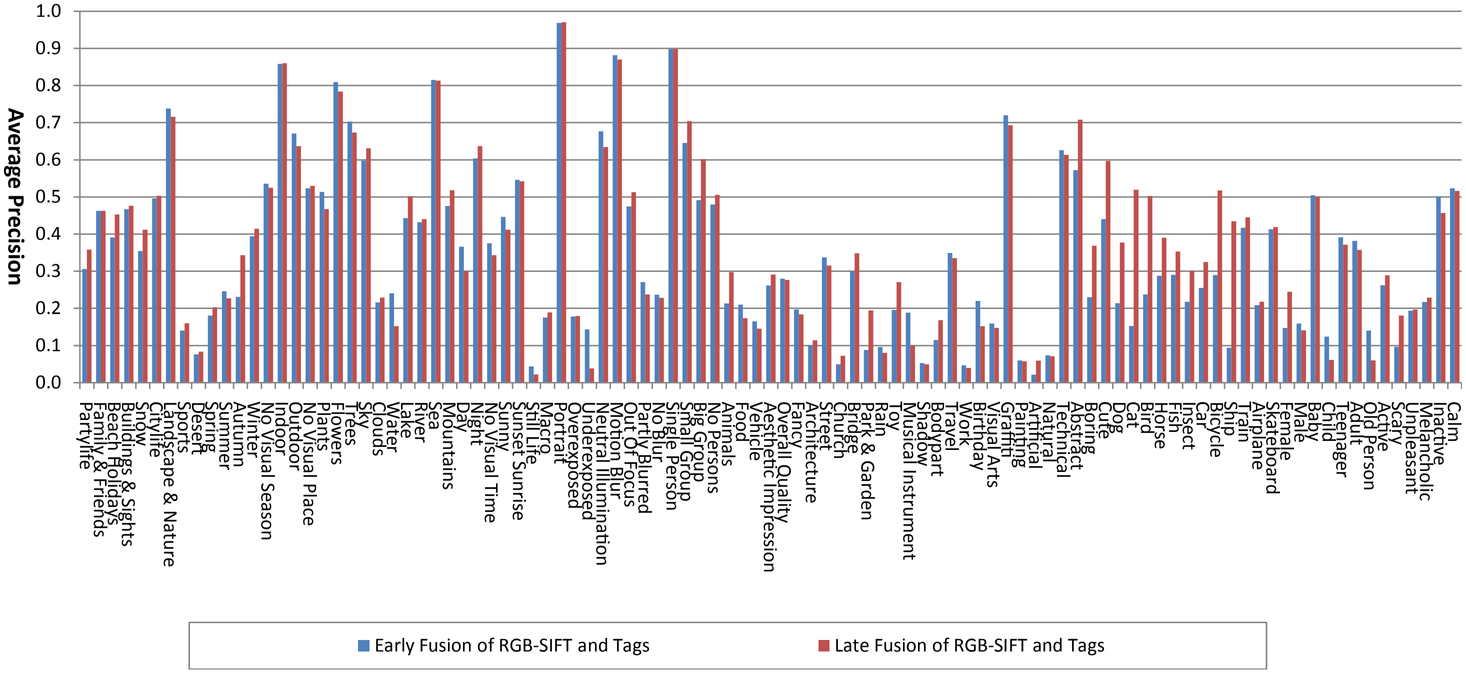

Fig. 3: Comparison of the results of early fusion and late fusion.

Table 1: Results of the runs for the evaluation per conceptin terms of MAP. The best run is marked in bold letters.

Run	**MAP**
Tags	0.3257
Early fusion RGB-SIFT & tags	0.3465
Early fusion RGB-SIFT & tags + post-processing	0.3613
Late fusion RGB-SIFT & tags	**0.3710**
Late fusion RGB-SIFT & tags + post-processing	0.3652

4.1 Evaluation per Concept

In Table 1, the final scores for the concept-based evaluation with the MAP are presented. Overall, our system scored a best run of 37.1% MAP for the multimodal approach. The text-only approach results in a MAP of 32.6%.

The late fusion approach outperformed the early fusion one by about 3% (37.1% versus 34.7%). The post-processing step does not improve the result of the late fusion approach, though it increases the results for the early fusion run. Figure 2 shows that the post-processing actually works well for some concepts, while the detection performance for others worsens. Concepts that suffer the most from the post-processing step are those whose names consist of more than one word, e.g., *park or garden*, *small group* or *old person*. Meanwhile, concepts like *cat*, *horse*, *airplane*, or *skateboard* improve significantly. The main reason for this is the rather simple approach of the post-processing step. The consideration of composite concepts should help to improve the performance.

For most of the concepts, early and late fusion perform quite similarly. The main difference can be found for the concepts *abstract*, *boring* and *cute* as well as the different kinds of animals and vehicles. Here, late fusion outperforms early fusion, as can be seen in Figure 3.

4.2 Evaluation per Example

Table 2 shows the overall results of the example-based evaluation. Best results are achieved with a late fusion of RGB-SIFT and tag features and the post-processing step, scoring an F-Measure of 55.2%. Early fusion of RGB-SIFT and tags resulted in the best Semantic R-Precision (SR-Precision) with 71.3%.

Using the example-based F-Measure, late fusion performs slightly better than early fusion, whereas the results for early fusion are better using the SR-Precision. Furthermore, the post-processing step seems to improve the results marginally.

Table 2: Results of the runs for the evaluation per example. Evaluation measures are the F-ex and the SR-Precision. The best run is marked in bold letters.

Run	**F-ex**	**SR-Precision**
Tags	0.5254	0.6767
Early fusion RGB-SIFT & tags	0.5413	**0.7128**
Early fusion RGB-SIFT & tags + post-processing	0.5416	0.7121
Late fusion RGB-SIFT & tags	0.5512	0.7014
Late fusion RGB-SIFT & tags + post-processing	**0.5519**	0.7014

5 Conclusions

The first participation of Fraunhofer IDMT in the ImageCLEF Photo Annotation Task reveals promising results. Using our textual descriptor in combination with one visual descriptor, we achieve annotation results that can compete well with other systems. The textual features work especially well for rather specific concepts that describe objects in an image. A combination of different textual and visual features is likely to result in a very stable annotation.

Future work will consider relations between tags as well as concepts more intently. Additionally, the inclusion of more visual features and text-based descriptors will be a main objective.

Acknowledgements

This work was partly supported by grant 01MQ07017 of the German research program THESEUS funded by the Ministry of Economics.

References

1. Nowak, S., Nagel, K., Liebetrau, J.: The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In: CLEF 2011 working notes, Amsterdam, The Netherlands. (2011)
2. Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1582–1596
3. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2) (November 2004) 91–110
4. Google Code: Google Translate `http://translate.google.de/`, last check: 05 Aug 2011.

5. Porter, M.: An algorithm for suffix stripping. Program: electronic library and information systems **14**(3) (1993) 130–137
6. Motohashi, N., Izawa, R., Takagi, T.: Meiji University at the ImageCLEF2010 Visual Concept Detection and Annotation Task: Working notes. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)