DBRIS at ImageCLEF 2012 Photo Annotation Task

Magdalena Rischka and Stefan Conrad

Institute of Computer Science
Heinrich-Heine-University of Duesseldorf
D-40204 Duesseldorf, Germany
{rischka,conrad}@cs.uni-duesseldorf.de

Abstract. For our participation in the ImageCLEF 2012 Photo Annotation Task we develope an image annotation system and test several combinations of SIFT-based descriptors with bow-based image representations. Our focus is on the comparison of two image representation types which include spatial layout: the spatial pyramids and the visual phrases. The experiments on the training and test set show that image representations based on visual phrases significantly outperform spatial pyramids.

Keywords: SIFT, bow, spatial pyramids, visual phrases

1 Introduction

This paper presents our participation in the ImageCLEF 2012 Photo Annotation Task. The ImageCLEF 2012 Photo Annotation Task is a multi-label image classification challenge: given a training set of images with underlying concepts the aim is to detect the presence of these concepts for each image of a test set using an annotation system based on visual or textual features or a combination of both. Detailed information on the task, the training and test set of images, the concepts and the evaluation measures can be found in the overview paper [1]. Our automatic image annotation system bases only on visual features. We focus on the comparison of two image representations which regard spatial layout: the spatial pyramid[4] and the visual phrases[3]. The spatial pyramid is very popular and often used, especially in the context of scene categorization, whereas visual phrases seem to pass out of mind in the literature.

The remainder of this paper is organized as follows: in section 2 we describe the architecture and the technical details of our image annotation system, in section 3 we present the evaluation on the training and the test set and discuss the results to end with a conclusion in section 4.

2 Architecture of the DBRIS image annotation system

The architecture of our automatic image annotation system together with the methods used in each step is illustrated in figure 1. To obtain the image representation of the training and test images, local features are extracted by applying the Harris-Laplace detector and the SIFT[5] descriptor in different color variants. The extracted local features are then summarized to the bag-of-words (bow) image representation as well as the image representations spatial pyramid[4] and visual phrases[3]. For the classifier training and classification steps we use an KNN-like classifier with one representative per concept. In the following subsections we describe each step in detail.

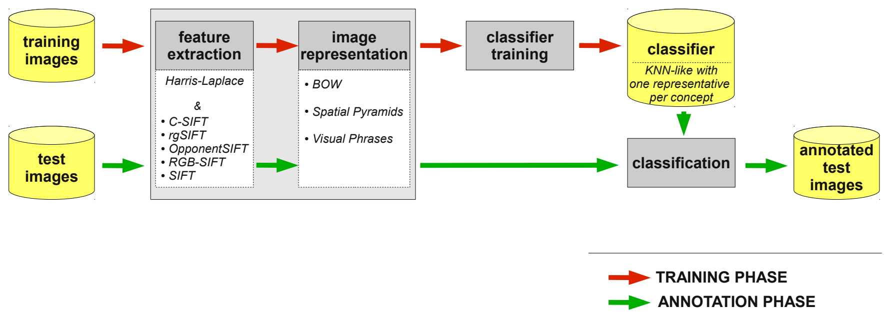

Fig. 1. Architecture of the DBRIS image annotation system

2.1 Features

For the choice of local features we refer to the evalution of color descriptors presented in [2]. We adopt the features C-SIFT, rgSIFT, OpponentSIFT, RGB-SIFT and SIFT as they are shown to perform best on the evaluation's underlying image benchmark, PASCAL VOC Challenge 2007. To extract these features with the Harris-Laplace point sampling strategy as the base, we use the color descriptor software [2].

2.2 Image representations

For each of the features, we quantize its descriptor space (225.000 descriptors) into 500 and 5000 visual words using K-Means. The visual words serve as a basis for the BoW, spatial pyramid and visual phrases representations. The representations are created in the common way using hard assignment of image features to visual words. We use the spatial pyramid constructions `1x3`, `1x1+1x3` and `1x1+2x2+4x4` in a weighted and unweighted version. To construct visual phrases we follow [3] and define a *visual phrase* as a pair of adjacent visual words. Assume an image contains the keypoints $kp_a = \{(x_a, y_a), scale_a, orient_a, descr_a\}$ and $kp_b = \{(x_b, y_b), scale_b, orient_b, descr_b\}$ with their assigned visual words vw_i and

vw_j, respectively. Then the image contains the visual phrase $vp_{ij} = \{vw_i, vw_j\}$ if the Euclidean distance of the keypoints' location in the image satisfy the term

$$EuclideanDistance((x_a, y_a), (x_b, y_b)) < max(scale_a, scale_b) \cdot \lambda \quad (1)$$

We set $\lambda = 3$. Analogously to the bow representation an image is represented by a histogram of visual phrases. Furthermore we create a representation combining bow with visual phrases, weighting bow with a value of 0.25 and the visual phrases histogram with 0.75. Table 1 summarizes all image representations with their number of dimensions we used in combination with each feature.

image representation	number of dimensions
`bow`	5.000
`sp 1x3`	15.000
`sp 1x1+1x3`	20.000
`sp 1x1+2x2+4x4`	105.000
`sp 1x1+2x2+4x4 w`	105.000
`vp`	125.250
`bow & vp`	130.250

Table 1. Image representations

2.3 Classifier

We use a KNN-like classifier, where concepts are not represented by the set of the corresponding images, but only by one representative. The representative of a concept is obtained by averaging the image representations of all images belonging to the concept. To classify a test image the similarities between the test image and the representatives of all concepts are determined. As similarity function we use the histogram intersection. To receive binary decisions on the membership to the concepts, we set an image-dependent threshold: a concept is present in the test image if the similarity between the test image and the concept is equal or greater than 0.75 times the maximum of the similarities of the test image to all concepts.

3 Evaluation

In the following we describe two evaluations: firstly we present the results of our experiments made on the training set. Secondly we discuss the evaluation of the five runs submitted to ImageCLEF.

3.1 Training and classification on the training set

To train and evaluate the DBRIS image annotation system, we split the training set of images into two disjoint parts (of size 7500), whereby both parts contain almost equal size of images for each concept. For each training and test pair we train the classifier on the one part and then use this classifier to classify the other part of images. The evaluation results are then averaged over the two training and test pairs.

In the first experiment we train one image annotation system for each of the 35 combinations of descriptors and image representations. Figure 2 shows the results in terms of MiAP values (averaged over all concepts). Comparing the systems with regard to the descriptors we observe an almost identical performance behaviour as shown in [2]. Except for the rgSIFT combined with the visual phrases based image representations, C-SIFT outperforms all the other descriptors in every image representation. The worst results are obtained by the SIFT descriptor. When we consider the image representations, we can see that the image representations based on visual phrases perform significantly better than the other ones for all descriptors. In the case of the descriptors C-SIFT, rgSIFT and OpponentSIFT the representations `vp` and `bow & vp` achieve similar values. When using RGB-SIFT and SIFT, the `bow & vp` representation is the better choice of the two. Bow and the representations based on spatial pyramid differ slightly from each other. Which one to choose depends on the descriptor used.

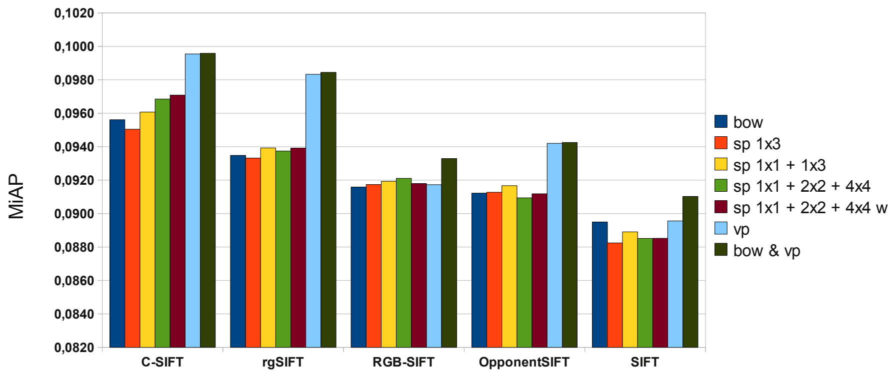

Fig. 2. MiAP values for each combination of descriptor and image representation

In the next experiment we join all descriptors into one annotation system, i.e. for each of the seven image representations we train an image annotation system whose classifier consists of five classifiers corresponding to the five descriptors. At the classification step, the similarities between the test image and the concept representatives obtained in each of the five classifiers are averaged over these five classifiers. The binary decisions on the membership to the concepts are calculated in the same way as described in section 2.3. Furthermore we create a configuration which combines the five descriptors with the image representations `sp 1x1+1x3`, `sp 1x1+2x2+4x4 w`, `bow & vp` and `vp`. The annotation system with this configuration consists of 20 classifiers (5 descriptors x 4 representations) and is called `combined`. The MiAP values for all configurations are shown in figure 3. The performances of the image representations behave similar to the progress at the C-SIFT descriptor in figure 2, but the MiAP values are lower and comparable with the rgSIFT results. The combination of more representations improve the performance of the bow and the spatial pyramids, but the image representations based on visual phrases still achieve better results.

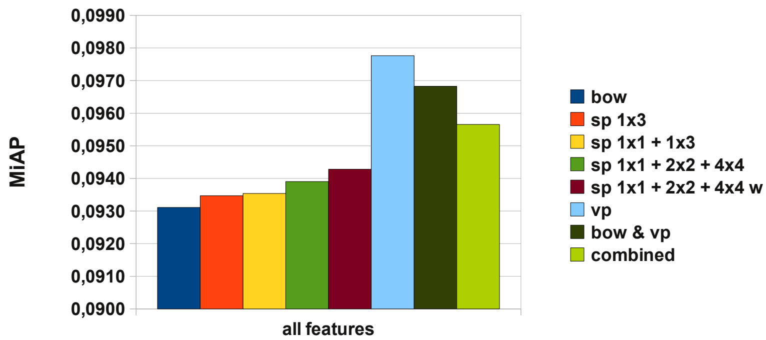

Fig. 3. MiAP values for each image representation

3.2 Classification of the test set

For the classification of the test set, we train the classifier on the whole training set. For the five submission runs we choose five of the image representations from the second experiment presented in section 3.1: `sp 1x1+2x2+4x4 w` as run DBRIS 1, `combined` as DBRIS 2, `sp 1x1+1x3` as DBRIS 3, `vp` as DBRIS 4 and `bow & vp` as DBRIS 5. Figure 4 and figure 5 present the results of the configurations for each concept (MiAP values) and as averages (MiAP, MnAP, GMiAP, GMnAP) over all concepts. Best values or values which are significantly better than others within a certain concept are highlighted in green. To evaluate the image representations as a whole, firstly we consider the averages MiAP, MnAP, GMiAP, GMnAP in figure 5. The image representations `vp` and `bow & vp` yield the best values again, followed by `combined`, `sp 1x1+2x2+4x4 w` and `sp 1x1+1x3`. These results reflect the evaluation in figure 3. When we consider the concepts with their concept categories, we can see that there are some concept categories where the image representations based on visual phrases dominate. These concept categories are *quantity*, *age*, (*gender*) and *view*. These observations have also been made in the experiments on the training set. Other concept categories which yield best results with the visual phrases on the training set are *relation* and *setting*. A possible reason for the success of the visual phrases in these concept categories can be that these concepts contain a lot of pictures of persons. Visual phrases can catch human features like eyes, mouth, etc. better than the spatial pyramids because they work on a finer level. The success of the visual phrases in the concept category *water* can not be confirmed by the experiments on the training set. As visual phrases are popular for object detection tasks, it is surprising that these image representations fail in the concept category *fauna*. The best results in the concept category *fauna* are achieved with the image representations based on spatial pyramids. Spatial pyramids are also successful in *sentiment* and *transport*.

4 Conclusion

At the end we want to summarize the experiences we gathered in the experiments. The best performing descriptor, which is C-SIFT in the experiments, yields better performance than joining all descriptors together. For the choice on image representations, the image representations based on visual phrases significantly outperform the spatial pyramids and the bow representation. The evaluation shows that visual phrases are especially appropriate for concepts dealing with persons. Although visual phrases are often used in object detection tasks, they are also successful in scene categorization.

References

1. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. CLEF 2012 working notes, Rome, Italy (2012)
2. van de Sande, K. E. A., Gevers, T., Snoek, C. G. M.: Evaluating Color Descriptors for Object and Scene Recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32 (9), pp. 1582-1596, (2010) `http://www.colordescriptors.com`
3. Zheng, Qing-Fang and Gao, Wen: Constructing visual phrases for effective and efficient object-based image retrieval. In: ACM Trans. Multimedia Comput. Commun. Appl., vol 5, 1, art. 7 (2008)
4. S. Lazebnik, C. Schmid and J. Ponce: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York (2006), vol. 2, pp. 2169 - 2178
5. Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004

	concept	sp 1x1+2x2+4x4 w	combined	sp 1x1+1x3	vp	bow & vp
		DBRIS 1	DBRIS 2	DBRIS 3	DBRIS 4	DBRIS 5
0	timeofday_day	0,4714	0,4412	0,4399	0,4687	0,4699
1	timeofday_night	0,0498	0,0506	0,0463	0,0453	0,0458
2	timeofday_sunrisesunset	0,0384	0,0484	0,0386	0,056	0,0563
3	celestial_sun	0,0238	0,0274	0,0247	0,0271	0,0276
4	celestial_moon	0,0068	0,0068	0,0068	0,0068	0,0068
5	celestial_stars	0,0025	0,0025	0,0025	0,0025	0,0025
6	weather_clearsky	0,0715	0,0723	0,0718	0,0715	0,0713
7	weather_overcastsky	0,0517	0,0482	0,0497	0,0472	0,0479
8	weather_cloudysky	0,1355	0,2272	0,1647	0,2312	0,2312
9	weather_rainbow	0,0026	0,0092	0,0035	0,0029	0,0028
10	weather_lightning	0,0132	0,0133	0,0132	0,0131	0,013
11	weather_fogmist	0,0175	0,0128	0,0163	0,1016	0,0172
12	weather_snowice	0,0203	0,0185	0,0182	0,0178	0,0153
13	combustion_flames	0,0039	0,0044	0,004	0,0046	0,0046
14	combustion_smoke	0,0049	0,0048	0,0048	0,0048	0,0048
15	combustion_fireworks	0,0052	0,0039	0,0048	0,0035	0,0043
16	lighting_shadow	0,0824	0,0762	0,0795	0,0756	0,0767
17	lighting_reflection	0,0332	0,0332	0,0333	0,035	0,0347
18	lighting_silhouette	0,0341	0,0341	0,0344	0,0422	0,0467
19	lighting_lenseffect	0,0429	0,0517	0,0445	0,0542	0,0497
20	scape_mountainhill	0,0602	0,1355	0,0692	0,0918	0,0931
21	scape_desert	0,0189	0,053	0,0346	0,097	0,0958
22	scape_forestpark	0,2221	0,257	0,2415	0,2138	0,2113
23	scape_coast	0,1782	0,1725	0,1785	0,1744	0,1773
24	scape_rural	0,0505	0,0615	0,0541	0,0675	0,07
25	scape_city	0,1104	0,118	0,1123	0,1114	0,1112
26	scape_graffiti	0,0444	0,0493	0,0456	0,0446	0,0444
27	water_underwater	0,0161	0,0092	0,0127	0,0125	0,019
28	water_seaocean	0,0513	0,0486	0,0516	0,0552	0,0559
29	water_lake	0,0127	0,013	0,0136	0,0156	0,0156
30	water_riverstream	0,0507	0,0492	0,0603	0,1249	0,1263
31	water_other	0,0328	0,0363	0,0331	0,0353	0,0346
32	flora_tree	0,3332	0,321	0,3395	0,3376	0,3379
33	flora_plant	0,0585	0,0715	0,0649	0,0717	0,0707
34	flora_flower	0,0778	0,0832	0,0792	0,0984	0,0985
35	flora_grass	0,2075	0,2603	0,2248	0,1724	0,1707
36	fauna_cat	0,0226	0,0179	0,0154	0,0163	0,0145
37	fauna_dog	0,0506	0,0451	0,0476	0,0498	0,0485
38	fauna_horse	0,0166	0,0183	0,0164	0,0146	0,0146
39	fauna_fish	0,0089	0,0051	0,0055	0,0059	0,006
40	fauna_bird	0,0345	0,0339	0,038	0,0326	0,0331
41	fauna_insect	0,0156	0,0213	0,0178	0,018	0,0167
42	fauna_spider	0,0047	0,0045	0,0051	0,0046	0,005
43	fauna_amphibianreptile	0,0067	0,0072	0,0075	0,0069	0,0071
44	fauna_rodent	0,015	0,0136	0,0156	0,0147	0,0147
45	quantity_none	0,7234	0,7399	0,738	0,7234	0,7233
46	quantity_one	0,2774	0,2324	0,247	0,2798	0,2798
47	quantity_two	0,05	0,0498	0,0499	0,052	0,0518
48	quantity_three	0,019	0,0188	0,0189	0,0205	0,0207
49	quantity_smallgroup	0,0535	0,0551	0,0531	0,0595	0,0592
50	quantity_biggroup	0,0579	0,0622	0,0589	0,0664	0,0661

Fig. 4. Results of the submitted runs 1 (in MiAP)

concept		sp 1x1+2x2+4x4 w	combined	sp 1x1+1x3	vp	bow & vp
		DBRIS 1	DBRIS 2	DBRIS 3	DBRIS 4	DBRIS 5
51	age_baby	0,0084	0,0087	0,0085	0,009	0,0091
52	age_child	0,0362	0,0346	0,0342	0,0371	0,0367
53	age_teenager	0,0484	0,0399	0,0481	0,1173	0,1177
54	age_adult	0,3153	0,2604	0,2854	0,2996	0,2986
55	age_elderly	0,024	0,0435	0,0264	0,0267	0,0273
56	gender_male	0,1927	0,1923	0,1922	0,2043	0,203
57	gender_female	0,269	0,2158	0,2395	0,2353	0,2347
58	relation_familyfriends	0,0775	0,0763	0,0774	0,0832	0,0828
59	relation_coworkers	0,0257	0,032	0,0275	0,0338	0,0321
60	relation_strangers	0,1295	0,0516	0,0835	0,0663	0,072
61	quality_noblur	0,735	0,7375	0,7359	0,723	0,7232
62	quality_partialblur	0,3039	0,2484	0,3037	0,3073	0,3076
63	quality_completeblur	0,0083	0,0083	0,0083	0,0094	0,0085
64	quality_motionblur	0,0208	0,0232	0,0228	0,0204	0,0205
65	quality_artifacts	0,0235	0,0199	0,0206	0,0208	0,0205
66	style_pictureinpicture	0,0207	0,0231	0,0228	0,0212	0,0195
67	style_circularwarp	0,0157	0,0154	0,0155	0,0161	0,0159
68	style_graycolor	0,0371	0,114	0,0859	0,0904	0,0903
69	style_overlay	0,0398	0,0399	0,0399	0,0392	0,0392
70	view_portrait	0,1448	0,163	0,1447	0,2032	0,2031
71	view_closeupmacro	0,1684	0,1722	0,1703	0,1722	0,1736
72	view_indoor	0,1434	0,1469	0,1436	0,1624	0,1624
73	view_outdoor	0,4595	0,4303	0,4208	0,4527	0,4532
74	setting_citylife	0,1762	0,1831	0,1759	0,1814	0,1809
75	setting_partylife	0,0375	0,0421	0,0398	0,0422	0,0429
76	setting_homelife	0,07	0,0698	0,0704	0,0763	0,0763
77	setting_sportsrecreation	0,0362	0,037	0,0361	0,0368	0,0368
78	setting_fooddrink	0,0821	0,0974	0,0766	0,1064	0,1006
79	sentiment_happy	0,1198	0,1859	0,1819	0,1247	0,123
80	sentiment_calm	0,1601	0,171	0,1624	0,1703	0,1719
81	sentiment_inactive	0,0949	0,0951	0,0954	0,0944	0,0943
82	sentiment_melancholic	0,062	0,0614	0,0612	0,0666	0,0642
83	sentiment_unpleasant	0,0535	0,0504	0,0515	0,0486	0,0489
84	sentiment_scary	0,0389	0,0309	0,0358	0,0333	0,0328
85	sentiment_active	0,1657	0,0899	0,1202	0,1207	0,166
86	sentiment_euphoric	0,017	0,0182	0,0176	0,0186	0,0182
87	sentiment_funny	0,1521	0,1543	0,1527	0,1121	0,1121
88	transport_cycle	0,0336	0,0309	0,0301	0,031	0,0298
89	transport_car	0,0719	0,0677	0,0718	0,0651	0,0674
90	transport_truckbus	0,013	0,0103	0,012	0,0107	0,0112
91	transport_rail	0,0147	0,0147	0,0154	0,0143	0,0148
92	transport_water	0,0693	0,0508	0,0687	0,0669	0,068
93	transport_air	0,0057	0,0059	0,0058	0,0056	0,0057
map_n		0,0774	0,081	0,0788	0,0818	0,0818
map_i		0,0927	0,0938	0,0925	0,0976	0,0972
gmap_n		0,0355	0,0374	0,0363	0,0385	0,0385
gmap_i		0,0441	0,0454	0,0445	0,0476	0,047

Fig. 5. Results of the submitted runs 2 (in MiAP)