The University of Amsterdam's Concept Detection System at ImageCLEF 2011

Koen E. A. van de Sande and Cees G. M. Snoek
Intelligent Systems Lab Amsterdam, University of Amsterdam
Software available from: http://www.colordescriptors.com

Abstract

The University of Amsterdam participated in the photo annotation task and the concept-based retrieval task of ImageCLEF 2011. In the per-image evaluation of the photo annotation task, we achieve the highest score overall. For the concept-based retrieval task, we submitted the best visual-only run. For the concept-based retrieval task, we considered three ways to perform visual retrieval: fully automatic, human topic mapping and human topic inspection. For a fully automatic system, including more random negatives to train a topic model improves results. For a human selecting relevant concepts to the topic, multiplication fusion works better than summation. For human topic inspection, a relevance feedback scheme on the train data gives an 8-fold increase in the number of positive examples per topic. Depending on the topic, the human topic mapping (best for 21 topics) and inspection (best for 17 topics) give the best results. An oracle fusion of the different methods would increase MAP from 0.100 for our best run to 0.128 overall.

1 Introduction

The University of Amsterdam participated in the photo annotation task and the concept-based retrieval task of ImageCLEF 2011. The Large-Scale Visual Concept Detection Task [5] evaluates visual concept detectors. The concepts used are from the personal photo album domain: *beach holidays*, *snow*, *plants*, *indoor*, *mountains*, *still-life*, *small group of people*, *portrait.* For more information on the dataset and concepts used, see the overview paper [5]. Our participation in the last two years, in ImageCLEF 2009/2010, focussed on increasing the robustness of the individual concept detectors based on the bag-of-words approach, and less on the per-image evaluation.

Last years experiments [6–9, 11] emphasize in particular the role of visual sampling, the value of color invariant features, the influence of codebook construction, and the effectiveness of kernel-based learning parameters. This was successful, resulting in the best visual only run for the photo annotation task in terms of MAP. Speedups using parallel computing were investigated in [10, 12].

In 2009, the per-image evaluation suggested that the assignment of concept tags to images leaves room for improvement. The primary evaluation metric used in 2010 and beyond for the per-image evaluation was the average example-based F-measure. We have looked into optimizing this measure with our system.

A new task for this year is the concept-based retrieval task. By extending the test set to 200,000 images, this ensures that systems need to have reasonable computation times. Another difference in this task is that there are no predefined concepts, but a collection of 40 *topics*. These topics are typically combinations of several existing ImageCLEF concepts, but can have complex boolean expressions within them. They come in the form of a textual description and up to 5 example images.

2 Photo Annotation

Our concept detection system is an improved version of the system from the ImageCLEF book [4], where we have performed additional experiments [8] which give insight into the effect of different sampling methods, color descriptors and spatial pyramid levels within the bag-of-words model. Our runs this year roughly correspond to *Harris-Laplace and dense sampling every 6 pixels (multi-scale) with 4-SIFT* and *Harris-Laplace and dense sampling every pixel (single-scale) with 4-SIFT* from this book chapter [8]. However, instead of 4-SIFT, we only consider three ColorSIFT variants this year. One of these three is an optimized color descriptor which allows these three to perform as good as 4-SIFT. Please refer to the cited papers[1] for implementation details of the system.

To achieve better results in the per-image evaluation, where we need to perform a binary assignment of a tag to an image, we use the probabilistic output of the SVM. In a cross-validation experiment, we have found a threshold of 0.3 to be good for most concepts: the default threshold of 0.5 would be too conservative when evaluating with an example-based F-measure where precision and recall are weighted equally. Optimizing the threshold on a per-concept basis instead of a single threshold was found to be less stable. Instead of a single parameter, 99 parameters need to be chosen (one per concept), and this estimation is done on the data of a single concept (instead of over 99 concepts).

New this year is our inclusion of textual information based on the image tags. As a textual representation of the image, we use a binary vector signaling whether a tag is present or absent among the provided Flickr tags. We select all words which occur at least 25 times. Tags consisting of multiple words, split by spaces are turned into multiple words. Also, words consisting of only digits are discarded. This gives us a lexicon of 1008 words. The binary feature vectors are L2-normalized.

[1] Papers available from http://www.colordescriptors.com

Table 1: Overall results of the our runs evaluated over all concepts in the Photo Annotation task with Average Precision.

Run name	Type	AP
Core	Visual	0.368
CoreA	Visual	0.375
CoreFast	Visual	0.364
Multimodal-CoreA	Visual+Tags	**0.433**
Multimodal-CoreA-MKL	Visual+Tags	0.415

2.1 Photo Annotation Runs

We have submitted five different runs. All runs use both Harris-Laplace and dense sampling with the SVM classifier.

- `Core`. Harris-Laplace and dense sampling every 6 pixels (multi-scale) with 3-SIFT.
- `CoreA`. Harris-Laplace and dense sampling every pixel (single-scale) with 3-SIFT.
- `CoreFast`. Harris-Laplace and dense sampling every 6 pixels (multi-scale) with 3-SIFT and fast intersection kernel [2]: instead of a χ^2 kernel, this run allows classification of test images whose computation time is independent of the number of support vectors.
- `Multimodal-CoreA`. Combination of the `CoreA` visual features with our text features; equally weighed at the SVM kernel level.
- `Multimodal-CoreA-MKL`. Combination of the `CoreA` visual features with our text features; weighed at the kernel level by multiple kernel learning.

2.2 Evaluation Per Concept

In table 1, the overall scores for the evaluation of concept detectors are shown. The features with sampling at every pixel instead of every 6 pixels perform better (0.375 versus 0.368), which is similar to the result obtained in [8]. The use of a fast intersection kernel SVM [2] slightly reduces accuracy (0.368 to 0.364), but brings significant speed gains (useful for the concept-based retrieval task). The two final runs perform better than the others by including the textual modality, as was seen in ImageCLEF last year, for example in [3]. We confirm that including textual information based on the image tags improves results by 0.05 MAP. Indeed, numerous images are tagged directly with the name of a concept, or a synonym thereof (e.g. *Graffiti* or *Sky*). It should come as no surprise that this information is highly relevant for those concepts.

Table 2: Results using the per-image evaluation measures for our runs in the Photo Annotation Task. Measures are the average example-based F-measure and SR-precision.

Run name	Type	F-measure	SR-precision
Core	Visual	0.608	0.732
CoreA	Visual	0.612	0.734
CoreFast	Visual	0.605	0.730
Multimodal-CoreA	Visual+Tags	**0.622**	**0.742**

2.3 Evaluation Per Image

For the per-image evaluation, overall results are shown in table 2. Our emphasis on optimizing the threshold for tag assignment has resulted in the best overall run in terms of example-based F-measure and SR-precision over all submissions.

3 Concept-Based Retrieval

The use of topics in the concept-based retrieval task, instead of concepts, poses a new problem to concept detection: what do we use as a starting point? Each topic has up to 5 example images, which could also be used to start visual retrieval. Since the topics are primarily combinations of several existing ImageCLEF concepts, we could use existing concept detectors. However, to do the latter fully automatic, we would need language parsing tools with support for boolean logic. An alternative is to add a 'manual' component to the system where a human maps topics to existing topics. But, a human can go a step further in their inspection of the topic. The concept-based retrieval task states that the training set of the annotation task (8,000 images annotated with 99 visual concepts) can be used to train the concept detectors. Therefore, we have extended the formulation of the topic by using relevance feedback on this training set.

Overall, we have explored 3 approaches:

- **Fully automatic retrieval**. We use only the provided example images as positive examples to train a new concept detector. We combine these positive examples with either 10, 33 or 100 random negatives from the photo annotation train set. These are runs `auto10`, `auto33` and `auto100`.
- **Human topic mapping**. A human reads the topic and then selects relevant concept(s). For run `1concept`, the human can only select a single concept. For `2conceptsum` and `2conceptmul`, the human can select two concepts. The probability scores of these concepts are then combined using either summation or multiplication.

- **Human topic inspection**. A human can give quick feedback on whether images are relevant for a certain topic. Therefore, we have taken the concept models trained for the fully automatic retrieval, and applied them to the training set. A human was then given up to 7.5 minutes per topic to check the top ranked images for additional positive examples, and allowed to mark negative examples as well. Besides the output from the fully automatic system, the human was also allowed to look at the positive examples for one of the 99 existing concepts, and get additional positives from there. We also include a run with 100 negatives randomly added besides the negatives selected by a human.

The concept detectors used for concept-based retrieval are trained using the `Core` system from the photo annotation task, unless the word *fast* is in the name. In the latter case, the `CoreFast` system was used. It is of interest to note that we have only used visual information for the concept-based retrieval, where other participants have also included information from the tags.

3.1 Results

In Figure 1, we show results for our 3 concept-based retrieval approaches. For the fully automatic system, including more random negatives improves results. The fully automatic system achieves 0.043 MAP with 100 negative examples. Additional negative examples might improve results further, but this also increases the chances that there are true positives among the random negatives. For the human concept mapping, selecting two concepts (where possible) results in a large improvement over selecting a single concept. This is expected, as the topics are designed to be boolean combinations of existing concepts Topics which directly map to a single concept have been left out on purpose. When combining two concepts, the multiplication fusion (0.089 MAP) works better than the summation fusion (0.080 MAP). For the human topic inspection, results are much better than the automatic system: the number of positives has increased to 42 on average, and 228 negatives have been selected. We find that including 100 random negatives still improves results; apparently the negatives selected by a human are not sufficient. To check whether selecting negatives is necessary at all, an interesting experiment would be to leave out the negatives selected by the human completely, and to only use random negatives. See also [1].

The human concept mapping achieves the best results for 21 out of 40 topics. The human concept inspection achieves the best results for 17 out of 40 topics. Had we used the best approach per topic (oracle fusion), we would have increased MAP from 0.100 for our best run to 0.128 overall. Further analysis is needed to determine the relationship between how closely the topic maps to existing concepts, accuracy and the specificity of the topic.

	Fully automatic			Human topic mapping					Human topic inspection			
Topic	auto10	auto33	auto100	Concept #1	Concept #2	1concept	2conceptmul	2conceptsum	normal	+100neg	fast	fast+100neg
1 Graffiti on buildings/walls	0,017	0,074	0,062	Graffiti	Building_Sights	0,082	0,022	0,000	0,173	0,235	0,184	**0,253**
2 Toy vehicle	0,001	0,000	0,002	car	Toy	0,000	**0,023**	0,003	0,000	0,000	0,000	0,000
3 1 person doing sports at sea	**0,123**	0,057	0,044	Sea	Single_Person	0,002	0,002	0,002	0,008	0,009	0,007	0,008
4 Airplane in the sky	0,045	0,000	0,024	airplane	Sky	0,125	**0,185**	0,166	0,036	0,067	0,051	0,091
5 Rider on horse	0,000	0,000	0,000	horse		**0,029**	**0,029**	**0,029**	0,023	0,027	0,021	0,024
6 Cyclist	0,000	0,000	0,000	bicycle		0,053	0,053	0,053	0,050	0,054	0,054	**0,059**
7 Mountains with sky during night	0,000	0,006	0,006	Night	Mountains	0,000	**0,133**	0,086	0,081	0,085	0,088	0,092
8 Fish in water	0,000	0,000	0,000	fish	Water	**0,016**	0,000	0,000	0,007	0,007	0,007	0,008
9 Desert scenery	0,056	0,095	0,097	Desert		0,154	0,154	0,154	0,192	0,213	0,204	**0,216**
10 1 person playing music instrument	0,001	0,005	0,008	MusicalInstrument		**0,089**	**0,089**	**0,089**	0,012	0,029	0,012	0,034
11 Animal in snow	0,000	0,036	0,021	Animals	Snow	0,001	**0,158**	0,116	0,077	0,096	0,093	0,111
12 Snowy winter landscape	0,044	0,056	0,072	Snow	Trees	**0,138**	0,117	0,116	0,093	0,096	0,096	0,097
13 Female person(s) doing sports	0,000	0,000	0,000	Sports		**0,011**	**0,011**	**0,011**	0,000	0,002	0,000	0,002
14 Cities at night with cars	0,000	**0,100**	0,080	Night	Citylife	0,023	0,042	0,042	0,036	0,036	0,033	0,034
15 Sea sunset or sunrise	0,155	0,132	0,137	Sunset_Sunrise		0,040	0,040	0,040	0,231	**0,232**	0,223	0,225
16 Outside view of a church	0,000	0,000	0,098	Church	Outdoor	0,300	0,398	**0,417**	0,344	0,344	0,369	0,367
17 Waters in autumn	0,007	0,016	0,012	Autumn	Water	0,005	0,155	**0,159**	0,110	0,153	0,109	0,155
18 Female old person	0,000	0,001	0,001	female	old_person	0,004	**0,012**	0,002	0,000	0,001	0,000	0,001
19 Close-up of trees	0,045	0,071	0,076	Trees		0,117	0,117	0,117	0,133	0,128	**0,136**	0,128
20 Trains indoor	0,013	0,009	0,006	train	Indoor	0,009	**0,058**	0,018	0,001	0,001	0,001	0,001
21 Scary dog(s)	0,000	0,000	**0,007**	dog		**0,007**	**0,007**	**0,007**	0,006	0,006	**0,007**	0,006
22 Portrait that is out of focus	0,052	0,022	0,022	Portrait	Out_of_focus	0,002	**0,074**	0,046	0,072	0,051	0,071	0,053
23 Bridges not over water	0,000	0,000	0,000	Bridge		0,000	0,000	0,000	**0,004**	0,002	0,003	0,003
24 Funny baby	0,000	0,001	0,003	Baby		**0,036**	**0,036**	**0,036**	0,006	0,012	0,006	0,016
25 Melancholic photos in rain	0,018	0,053	0,145	Rain		0,149	0,149	0,149	0,236	0,238	0,250	**0,256**
26 Houses in mountains	0,000	0,000	0,000	Mountains	Building_Sights	0,005	0,126	**0,132**	0,003	0,001	0,003	0,001
27 Family holidays at the beach	0,000	0,093	0,076	Beach_Holidays		0,090	0,090	0,090	0,083	0,104	0,092	**0,111**
28 Fireworks	0,000	0,384	0,415	Night	Outdoor	0,006	0,005	0,005	0,375	0,389	0,404	**0,423**
29 Close-up of flowers with raindrops	0,000	0,000	0,000	Flowers	Rain	0,002	**0,011**	0,002	0,005	0,006	0,005	0,006
30 Cute toys arranged as still-life	0,000	0,002	0,002	Toy	Still_Life	**0,155**	0,108	0,101	0,044	0,056	0,044	0,060
31 Ship/boat on a river	0,004	0,003	0,003	ship	River	0,002	**0,025**	0,017	0,012	0,020	0,012	0,024
32 Underexposed photos of animals	0,001	0,000	0,000	Animals	Underexposed	0,025	0,035	0,036	0,062	0,063	**0,066**	0,065
33 Cars and motion blur	0,000	0,128	0,108	car	Motion_Blur	0,009	**0,522**	0,453	0,307	0,317	0,340	0,345
34 Unpleasant insects	0,000	0,000	0,001	insect		0,046	0,046	0,046	0,046	**0,048**	0,046	0,045
35 Close-up of bird	0,000	0,000	0,026	bird		0,081	0,081	0,081	0,103	0,103	**0,108**	0,106
36 Scary shadows of people	0,042	0,000	0,054	Shadow		0,069	0,069	0,069	0,089	0,088	**0,097**	0,094
37 Painting of person(s)	0,003	0,030	0,018	Painting	Single_Person	0,044	0,026	0,000	0,087	0,092	0,099	**0,105**
38 Birthday or wedding cake	0,000	0,000	0,001	Food	Partylife	0,005	0,021	0,020	0,030	0,030	0,030	**0,033**
39 House surrounded by garden	0,084	0,073	0,071	Building_Sights	Park_Garden	0,000	**0,116**	0,094	0,060	0,062	0,061	0,072
40 Close-up of bodypart	0,011	0,017	0,022	bodypart		0,207	0,207	0,207	0,233	0,258	0,230	**0,261**
MAP	**0,018**	**0,037**	**0,043**			**0,053**	**0,089**	**0,080**	**0,087**	**0,094**	**0,092**	**0,100**
#pos	5	5	5						42	42	42	42
#neg	10	33	100						228	328	228	328

Figure 1: Results for the concept-based retrieval task. Every row corresponds to a topic; the maximum MAP score per row has a yellow background. At the bottom, the average number of positive/negative examples per topic model is listed (where relevant).

4 Conclusion

The submissions from our visual concept detection system in the ImageCLEF 2011 photo annotation task have resulted in the best run in the per-image evaluation. In the concept-based retrieval task, it was the best visual-only system. For the concept-based retrieval task, we considered three ways to perform visual retrieval: fully automatic, human topic mapping and human topic inspection. For a fully automatic system, including more random negatives to train a topic model improves results. For a human selecting relevant concepts to the topic, multiplication fusion works better than summation. For human topic inspection, a relevance feedback scheme on the train data gives an 8-fold increase in the number of positive examples per topic. Depending on the topic, the human topic mapping (best for 21 topics) and inspection (best for 17 topics) give the best results. An oracle fusion of the different methods would increase MAP from 0.100 for our best run to 0.128 overall.

5 Acknowledgements

This work is supported by STW SEARCHER, FES COMMIT, and the IARPA via Department of Interior National Business Center contract number D11PC20067. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Social negative bootstrapping for visual categorization. In *ACM International Conference on Multimedia Retrieval*, 2011.

[2] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] T. Mensink, G. Csurka, F. Perronnin, J. Sánchez, and J. Verbeek. Lear and xrces participation to visual concept detection task - imageclef 2010. In *Working Notes for the CLEF 2010 Workshop*, 2010.

[4] H. Mueller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF*, volume 32 of *Lecture Notes in Computer Science: The Information Retrieval Series*. Springer, 2010.

[5] S. Nowak, K. Nagel, and J. Liebetrau. The clef 2011 photo annotation and concept-based retrieval tasks. In *Working Notes of CLEF 2011*, 2011.

[6] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, E. Gavves, D. Odijk, M. de Rijke, T. Gevers, M. Worring, D. C. Koelma, and A. W. M. Smeulders. The mediamill trecvid 2010 semantic video search engine. In *Proceedings of the TRECVID Workshop*, 2010.

[7] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665–681, 2010.

[8] K. E. A. van de Sande and T. Gevers. *University of Amsterdam at the Visual Concept Detection and Annotation Tasks*, chapter 18, pages 343–358. Volume 32 of *The Information Retrieval Series: ImageCLEF* [4], 2010.

[9] K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. The university of amsterdam's concept detection system at imageclef 2009. In *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2010.

[10] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Accelerating visual categorization with the gpu. In *ECCV Workshop on Computer Vision on GPU*, 2010.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering visual categorization with the GPU. *IEEE Transactions on Multimedia*, 13(1):60–70, 2011.