Sample Selection, Category Specific Features and Reasoning

Eugene Mbanya, Sebastian Gerke, Christian Hentschel, and Patrick Ndjiki-Nya

Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute
{eugene.mbanya|christian.hentschel|sebastian.gerke|
patrick.ndjiki-nya}@hhi.fraunhofer.de

Abstract. In this paper we present our approach to the 2011 ImageClef PhotoAnnotation task, which is based on the well known bag-of-words model. We investigated an approach for selecting the most informative training samples per concept for classification and the impact of fusing the OpponentSIFT feature with the GIST feature which calculates global image statistics, on scene-based concepts. We also incorporated a post-classification processing step, which refined classification results based on rules of inference and exclusion between concepts. The different approaches provided classification gains when compared to the standard bag-of-words model using only the OpponentSIFT feature.

1 Introduction

The ImageClef Photo Annotation Task is an annual competition, which draws interest from the Computer Vision research community, with the aim of addressing the problem of efficient annotation of large scale image collections. The task achieves this by inviting competition from different research institutions to provide solutions for the automatic classification of photos taken from the Flickr[1] community into different categories.

The 2010 ImageClef PhotoAnnotation Task [6] focused on providing annotations to a testset of 10000 images using 93 concepts. These concepts were mostly object-based e.g. *dog*, *cat*, scene-based e.g. *landscape*, *beach*, event-based e.g. *work*, *travel*, quality-based e.g. *blurry*, *overexposed* or representation-based e.g. *portrait*, *art*. Our experiments [5] in last year's task were based on the standard bag-of-words model for image classification. We performed a feature fusion of the OpponentSIFT [9] local feature and the no-reference objective image sharpness measure [3], which showed classification gains with the most gains occuring among the quality-based concepts. We also performed a post-classification step, which was based on the observation that many of the categories could be seen as being related to one other, thereby enabling the inference or exclusion of other categories.

In this year's task, the list of concepts was extended to 99 including sentiment-based concepts e.g. *sad* and *happy*. Our experiments in this year's task built on

[1] http://blog.flickr.net/en/2009/10/12/4000000000/

our approaches of last year. We performed a feature fusion of the OpponentSIFT descriptor and the GIST descriptor, with the target being to improve the classification performance of scene-based concepts, since the GIST descriptor has successfully been applied to scene classification and recognition [8]. Computation of the descriptor involves accumulating image statistics over the entire scene rather than over local regions. We also evaluated an approach to select the most informative training samples from the training set per category to train the classifiers, which resulted in qualitative as well as runtime performance gains. We will refer to this approach as *Smart Sampling (SS)* in the rest of this paper. Finally we used optimized versions of our post-classification processing algorithms of last year's task in this year's detection system.

In the following, section 2 describes our concept detection system in more detail and outlines the differences to our system of last year while section 3 summarizes this paper by giving some results and providing an outlook into future work.

2 Concept Detection System

Our concept detection system for this year's task was an optimized version of last year's system, which integrated a fusion of the OpponentSIFT descriptor and the GIST descriptor, the SS approach for effcient training sample selection per category and an optimized post-classification processing step. For a detailed description of last year's system architecture, including details of the OpponentSIFT feature refer to [5].

The spatial envelope model was initially introduced in [8] as a low-dimensional representation of a scene and successfully applied for k-nearest-neighbor scene classification. The idea was to represent the dominant spatial structure of a scene rather than applying segmentation or any kind of processing of individual objects or regions. The authors propose a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. The spatial envelop model was later termed the GIST descriptor following Friedman's [4] definition of a scene gist; an abstract representation of the scene that spontaneously activates memory representations of scene categories (a city, a mountain, etc.), which is essentially what is captured by the spatial envelope model [7]. Extraction of GIST is performed by filtering the image by a bank of Gabor filters. The image is split into a 4×4 regular grid and the Gabor filter responses are averaged over each block.

We use the implementation presented in [2] which takes a squared gray-level image of fixed size as input. All images are rescaled initially to 256×256 irrespective of their aspect ratios. We obtain a GIST feature vector of 512 dimensions per image, which is significantly less than the bag of keypoints vectors at even the smallest grid resolution. Moreover, the GIST features can be computed much more efficiently as there are no codebook and histogram computation steps involved. Fusion of the GIST descriptor and the OpponentSIFT descriptor was done analogously to our fusion of the OpponentSIFT descriptor and the no-reference objective image sharpness measure in last year's system.

We further applied a Smart Sampling optimization step to all classifiers. Smart Sampling selects the most informative training images from the training set in order to train a classifier faster and more efficiently [1]. This occurs in an iterative process, whereby the training set is divided into a number of subsets and the classifier is trained using one of the subsets and used to classify another subset. For every further iteration, the training set is composed of the previous iteration's training images and classified images having a classifier confidence of $|c| < 1$, and with classification performed on a new subset. Through this process, we select training samples from the whole training set which lie in close proximity to the separating hyperplanes of the classifier for each category. This led to an increase in runtime and qualitative performance.

Finally, we performed further tests to optimize the *Exclusion* and *Inference* rules, which we used in the post-classification processing step of last year's system, while adapting them to the newly added categories of this year's task. Changes in the category list of this year, led to the need for modifications of the post-classification processing algorithms. Last year it was possible to identify groups of categories, with each image being able to belong to only one category in each group. With the exclusion of categories such as *No_Visual_Season* this year, the same groups as in last year could no longer be identified. Consequently, the following equation which was assumed to hold for all groups of excluding categories was invalidated.

$$\bigcup_{p \in P} C_p = I \tag{1}$$

The equation was modified to

$$\bigcup_{p \in P} C_p \subseteq I \tag{2}$$

which further leads to different update rules for confidences. Rather than

$$c'(i,p) = \begin{cases} c(i,p) & \text{if } c(i,p) > c(i,q) \quad \forall q \in P \setminus p \\ 0 & \text{else} \end{cases} \tag{3}$$

where the maximum confidence for a category is maintained and all other confidences in a group of excluding categories are set to 0, the following update rule was used.

$$c'(i,p) = \begin{cases} c(i,p) & \text{if } c(i,p) > c(i,q) \text{ and } c(i,q) > 0.73105 \quad \forall q \in P \setminus p \\ 0 & \text{else} \end{cases} \tag{4}$$

The treshold 0.73105 was chosen because it is the value of the sigmoid function applied to 1.0, i.e. the exclusion rule is only used if the maximum category confidence for a given image is outside the tube around the separating hyperplane.

For category inference, the system from last years submission was maintained. However, the rule confidence threshold (i.e. the threshold that toggles if a rule is used) was tuned. Last year a threshold of 0.99 was used, meaning that a rule was only used if its confidence in the training set was at least 0.99. We optimized this threshold to maximize the Example-based F-Measure. The maximum gain was reached using a value of 0.63 for this threshold. The update rule was modified analogously to the exclusion update rule given before, by applying rules only if the confidence of the rules' left-hand side category confidence was above 0.73105 (i.e. outside of the tube around the separating hyperplane).

3 Results and Summary

We submitted 5 different runs. All runs use the OpponentSIFT histograms as baseline. The first run (*OpSIFT*) uses the OpponentSIFT feature alone. Another run (*OpSIFT+Excl+Inf+SS*) uses the Smart Sampling optimization and applies category inference and exclusion as a post-classification processing step. A third run (*OpSIFT+Gist*) does no post-classification processing but uses the GIST feature as an additional feature. The fourth run (*OpSIFT+SS*) uses the Smart Sampling optimization step in addition to the OpponentSIFT feature. A final fifth run (*OpSIFT+SS+Inf*) uses the OpponentSIFT features with the Smart Sampling optimization step and applies the inference rule on the classification results.

Three different evaluation measures were computed. For evaluating the classification performance per concept the Mean Average Precision (MAP) was used. The evaluation per example was performed using the example-based F-Measure (F-Ex) and the Semantic R-Precision (SR-Precision).

Table 1 shows the average scores achieved for each measure. For all evaluation measures, all runs with extensions to the baseline OpponentSIFT method resulted in performance gains.

Run-Configuration	MAP	Avg. F-Ex	SR-Precision
OpSIFT	0.325111	0.579904	0.71262264
OpSIFT+GIST	0.335234	0.588042	0.71764547
OpSIFT+SS	0.326483	0.580804	0.71251965
OpSIFT+Excl+Inf+SS	0.325981	0.580729	0.71252900
OpSIFT+SS+Inf	0.325981	0.580729	0.71252900

Table 1. Average evaluation scores for all submitted runs. Highlighted values show the run, which obtained the best score for a specific evaluation measure.

In order to observe the influence of the GIST feature on scene-based concepts, we compared the MAP for those concepts out of the overall 99 concepts we considered as scene-based concepts ('Building Sights', 'Citylife', 'Landscape/ Nature','Indoor','Outdoor','Mountains', 'Sunset/ Sunrise', 'Park/ Garden', 'Beach/ Holidays') with their baseline (using only the OpponenSIFT descriptor) values.

This is depicted in table 2. We observe that using the GIST feature together with the OpponenSIFT feature yields a gain of 0.00807 compared to the baseline alone.

Category	OpSIFT	OpSIFT+GIST
Park/Garden	0.381190	0.401956
Sunset/Sunrise	0.722446	0.716504
Mountains	0.496681	0.489736
Outdoor	0.875141	0.881579
Indoor	0.573047	0.581117
Landscape/Nature	0.769700	0.778197
Citylife	0.502739	0.515869
Building Sights	0.518642	0.545948
Beach/Holidays	0.369467	0.370691
Mean Average Precision	**0.578772**	**0.586844**

Table 2. Average Precision for categories where the GIST feature has been used. Best scores are highlighted.

Table 3 shows the detailed average precision for each category individually. For those categories where the results differ, the best performing run is highlighted in the table. In terms of MAP per category, the exclusion of categories often performed worse than the other runs. For categories especially, where the average precision was already low, the exclusion rule worsened the results. We attributed this to a lack of reliability of the SVM confidence outputs. The SVM classifier of bad performing categories had a very small output range, e.g. values ranging from 0.94 to 0.96. In such cases, the number of support vectors used for the categories was usually near the total number of training samples. Using these categories for inference or exclusion of other categories yielded a significant decrease in performance, propagating the error introduced by one categorie's classifier to other categories. In the future, a check for the reliability of classifier outputs should be performed to avoid such error propagation. This also holds for the category inference post-processing rule, which also suffers from this problem.

Category	OpSIFT	OpSIFT+SS	OpSIFT+SS+I+E	OpSIFT+GIST	OpSIFT+SS+I
Partylife	0.234868	0.244337	0.244337	0.232401	0.244337
Family_Friends	0.476125	0.477097	0.477097	0.494923	0.477097
Beach_Holidays	0.369467	0.357638	0.357638	0.370691	0.357638
Building_Sights	0.518642	0.525117	0.525117	0.545948	0.525117
Snow	0.127158	0.135231	0.135231	0.147183	0.135231
Citylife	0.502739	0.504007	0.504007	0.515869	0.504007
Landscape_Nature	0.7697	0.767596	0.767596	0.778197	0.767596
Sports	0.141025	0.141975	0.141975	0.147944	0.141975
Desert	0.03989	0.039747	0.039747	0.04092	0.039747
Spring	0.119698	0.157181	0.157181	0.155588	0.157181
Summer	0.226242	0.22794	0.22794	0.283984	0.22794
Autumn	0.293937	0.300913	0.300913	0.31722	0.300913
Winter	0.181803	0.203416	0.203416	0.188093	0.203416
Indoor	0.573047	0.575148	0.575148	0.581117	0.575148
Outdoor	0.875141	0.873909	0.873742	0.881579	0.873742

Plants	0.695391	0.694401	0.694414	0.707835	0.694414
Flowers	0.383228	0.387561	0.387561	0.378102	0.387561
Trees	0.601623	0.598829	0.598829	0.614786	0.598829
Sky	0.863894	0.863283	0.863979	0.867138	0.863979
Clouds	0.802869	0.80461	0.80461	0.816009	0.80461
Water	0.60001	0.601554	0.601554	0.614244	0.601554
Lake	0.265445	0.254102	0.254102	0.277511	0.254102
River	0.210948	0.210205	0.210205	0.231157	0.210205
Sea	0.486808	0.4743	0.4743	0.481429	0.4743
Mountains	0.496681	0.487793	0.487793	0.489736	0.487793
Day	0.837934	0.839187	0.823052	0.844487	0.823052
Night	0.519675	0.514734	0.514734	0.522686	0.514734
Sunny	0.426355	0.429615	0.429615	0.431574	0.429615
Sunset_Sunrise	0.722446	0.721848	0.721848	0.716504	0.721848
Still_Life	0.313663	0.31001	0.31001	0.339996	0.31001
Macro	0.463493	0.467147	0.467147	0.4712	0.467147
Portrait	0.6062	0.604458	0.604458	0.630568	0.604458
Overexposed	0.152449	0.163292	0.163292	0.172349	0.163292
Underexposed	0.22352	0.243862	0.243862	0.23928	0.243862
Neutral_Illumination	0.976283	0.976023	0.95715	0.976305	0.95715
Motion_Blur	0.218029	0.187167	0.187167	0.20733	0.187167
Out_of_focus	0.171646	0.163928	0.163928	0.159413	0.163928
Partly_Blurred	0.700893	0.702884	0.702884	0.725587	0.702884
No_Blur	0.891267	0.890623	0.889673	0.897333	0.889673
Single_Person	0.504431	0.51182	0.51182	0.544427	0.51182
Small_Group	0.274982	0.265547	0.265547	0.311041	0.265547
Big_Group	0.362615	0.366965	0.366965	0.37098	0.366965
No_Persons	0.888193	0.887454	0.873173	0.893866	0.873173
Animals	0.394545	0.401927	0.401927	0.433562	0.401927
Food	0.462321	0.45861	0.45861	0.456744	0.45861
Vehicle	0.438679	0.434949	0.434949	0.454827	0.434949
Aesthetic_Impression	0.258722	0.259471	0.259471	0.284088	0.259471
Overall_Quality	0.216586	0.230788	0.230788	0.222498	0.230788
Fancy	0.164452	0.160123	0.160123	0.171753	0.160123
Architecture	0.265967	0.266071	0.266071	0.268401	0.266071
Street	0.332411	0.333852	0.333852	0.348156	0.333852
Church	0.16725	0.144814	0.144814	0.09339	0.144814
Bridge	0.059348	0.068853	0.068853	0.052069	0.068853
Park_Garden	0.38119	0.380053	0.380053	0.401956	0.380053
Rain	0.005603	0.005029	0.005029	0.007794	0.005029
Toy	0.21895	0.210614	0.210614	0.221367	0.210614
MusicalInstrument	0.039009	0.054212	0.054212	0.057238	0.054212
Shadow	0.092445	0.103842	0.103842	0.101155	0.103842
bodypart	0.224537	0.226855	0.226855	0.232947	0.226855
Travel	0.118831	0.115688	0.115688	0.116576	0.115688
Work	0.037324	0.04159	0.04159	0.040592	0.04159
Birthday	0.00923	0.011126	0.011126	0.010305	0.011126
Visual_Arts	0.333149	0.373254	0.373254	0.34425	0.373254
Graffiti	0.062694	0.107677	0.107677	0.050603	0.107677
Painting	0.189725	0.192701	0.192701	0.194181	0.192701
artificial	0.143548	0.146887	0.146887	0.130711	0.146887
natural	0.708399	0.708505	0.708534	0.712678	0.708534
technical	0.057725	0.060137	0.060137	0.058513	0.060137
abstract	0.019408	0.0212	0.0212	0.019378	0.0212
boring	0.074566	0.072623	0.072623	0.073969	0.072623
cute	0.591775	0.5908	0.5908	0.594248	0.5908
dog	0.263146	0.251378	0.251378	0.288769	0.251378
cat	0.069199	0.059024	0.059024	0.069874	0.059024
bird	0.184953	0.171641	0.171641	0.193626	0.171641
horse	0.117973	0.13016	0.13016	0.105762	0.13016
fish	0.020769	0.032466	0.032466	0.021992	0.032466
insect	0.104334	0.092262	0.092262	0.178416	0.092262
car	0.321748	0.312561	0.312561	0.354207	0.312561
bicycle	0.166077	0.178552	0.178552	0.234617	0.178552
ship	0.162083	0.1085	0.1085	0.090505	0.1085
train	0.170937	0.158154	0.158154	0.184375	0.158154
airplane	0.135721	0.129396	0.129396	0.144941	0.129396
skateboard	0.005289	0.003328	0.003328	0.002331	0.003328

female	0.428183	0.432995	0.432995	0.452765	0.432995
male	0.206412	0.202545	0.202545	0.202445	0.202545
Baby	0.156655	0.160898	0.160898	0.184934	0.160898
Child	0.12056	0.14846	0.14846	0.149927	0.14846
Teenager	0.230589	0.233345	0.233345	0.243365	0.233345
Adult	0.494042	0.483964	0.483964	0.511625	0.483964
old_person	0.057175	0.060688	0.060688	0.053173	0.060688
happy	0.354886	0.352248	0.352248	0.383667	0.352248
funny	0.292249	0.307143	0.307143	0.345606	0.307143
euphoric	0.054367	0.053485	0.053485	0.059926	0.053485
active	0.23966	0.259608	0.259608	0.268043	0.259608
scary	0.177867	0.181647	0.181647	0.185067	0.181647
unpleasant	0.202828	0.201428	0.201428	0.214772	0.201428
melancholic	0.257974	0.260325	0.260325	0.255496	0.260325
inactive	0.500738	0.50328	0.50328	0.512244	0.50328
calm	0.510715	0.515642	0.515642	0.527191	0.515642
Mean Average Precision	**0.325111**	**0.326483**	**0.325981**	**0.335234**	**0.325981**

Table 3: Average Precision per category. Scores are only highlighted, when any of the extensions provided a performance increase or decrease.

Acknowledgements This work was supported in part by the German Federal Ministry of Economics and Technology under the project THESEUS (01MQ07018).

References

1. Antoine Bordes, Seyda Ertekin, Jason Weston, and Lon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.
2. Matthijs Douze, Herv Jgou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of gist descriptors for web-scale image search. *Proceeding of the ACM International Conference on Image and Video Retrieval CIVR 09*, page 1, 2009.
3. R. Ferzli and L.J. Karam. A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *Image Processing, IEEE Transactions on*, 18(4):717 –728, April 2009.
4. A Friedman. Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology General*, 108(3):316–355, 1979.
5. Eugene Mbanya, Christian Hentschel, Sebastian Gerke, Mohan Liu, Andreas Nrnberger, and Patrick Ndjiki-Nya. Augmenting bag-of-words - category specific features and concept reasoning. 2010.
6. S. Nowak and M. Huiskes. New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In *Working Notes of CLEF 2010*, Padova, Italy, 2010.
7. A Oliva. Gist of the scene. *Elsevier*, chapter 1:251–257, 2005.
8. Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, May 2001.
9. K.E.a. van de Sande, T. Gevers, and C. G.M. Snoek. A comparison of color features for visual concept classification. *Proceedings of the 2008 international conference on Content-based image and video retrieval - CIVR '08*, page 141, 2008.