Chapter 18
University of Amsterdam at the Visual Concept Detection and Annotation Tasks

Koen E. A. van de Sande and Theo Gevers

Abstract Visual concept detection is important to access visual information on the level of objects and scene types. The current state-of-the-art in visual concept detection and annotation tasks is based on the bag-of-words model. Within the bag-of-words model, first points are sampled according to some strategy, then the area around these points are described using color descriptors. These descriptors are then vector-quantized against a codebook of prototypical descriptors, which results in a fixed-length representation of the image. Based on these representations, visual concept models are trained. In this chapter, we discuss the design choices within the bag-of-words model and their implications for concept detection accuracy.

18.1 Introduction

Robust image retrieval is highly relevant in a world that is adapting to visual communication. Online services like Flickr show that the sheer number of photos available online is too much for any human to grasp. Many people place their entire photo album on the internet. Most commercial image search engines provide access to photos based on text or other metadata, as this is still the easiest way for a user to describe an information need. The indices of these search engines are based on the filename, associated text or tagging. This results in disappointing retrieval performance when the visual content is not mentioned, or properly reflected in the associated text. In addition, when the photos originate from non-English speaking countries, such as China, or Germany, querying the content becomes much harder.

To cater for robust image retrieval, the promising solutions from the literature are in majority concept-based, see *e.g.* the overview in Snoek and Worring (2009), where detectors are related to objects, like *trees*, scenes, like a *desert*, and people, like *big group*. Any one of those brings an understanding of the current content. The

University of Amsterdam, Science Park 107, 1098 XG Amsterdam, e-mail: ksande@uva.nl

concepts in such a lexicon allows users to query on presence or absence of visual content elements, *e.g*. a semantic entry into the data.

The Large-Scale Visual Concept Detection Task of ImageCLEF 2009, discussed in Nowak and Dunker (2009), evaluates 53 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 Chapter .

The current state-of-the-art in visual concept detection and annotation tasks is based on the bag-of-words model (van de Sande et al (2010); Marszałek et al (2007); Snoek et al (2009); Wang et al (2007)). Within the bag-of-words, first points are sampled according to some strategy, then the area around these points are described using color descriptors. These descriptors are then vector-quantized against a codebook of prototypical descriptors, which results in a fixed-length representation of the image. Based on these representations, visual concept models are trained.

Based on our previous work on concept detection (van de Sande et al (2010); Snoek et al (2008); Uijlings et al (2009)), participation of the University of Amsterdam within ImageCLEF has focused on improving the robustness of the visual features used in concept detectors.

Systems with the best performance in image retrieval (van de Sande et al (2010); Marszałek et al (2007)) and video retrieval (Snoek et al (2008); Wang et al (2007)) use combinations of multiple features for concept detection. The basis for these combinations is formed by good color features and multiple point sampling strategies. In this chapter, we discuss the design choices within the bag-of-words model and their implications for concept detection accuracy. We focus especially on the effect of these choices on the large-scale visual concept detection and annotation task from ImageCLEF 2009 and ImageCLEF@ICPR 2010.

The remainder of this chapter is organized as follows. Section 18.2 defines components in our concept detection pipeline. Section 18.3 details our experiments and results. Finally, in Section 18.6, conclusions are drawn.

18.2 Concept Detection Pipeline

We perceive concept detection as a combined computer vision and machine learning problem. The first step is to represent an image using a fixed-length feature vector. Given a visual feature vector x_i, the aim is then to obtain a measure, which indicates whether semantic concept C is present in photo i. We may choose from various visual feature extraction methods to obtain x_i, and use a supervised machine learning approach to learn the appearance relation between C and x_i. The supervised machine learning process is composed of two phases: training and testing. In the first phase, the optimal configuration of features is learned from the training data. In the second phase, the classifier assigns a probability $p(C|x_i)$ to each input feature vector for each semantic concept C.

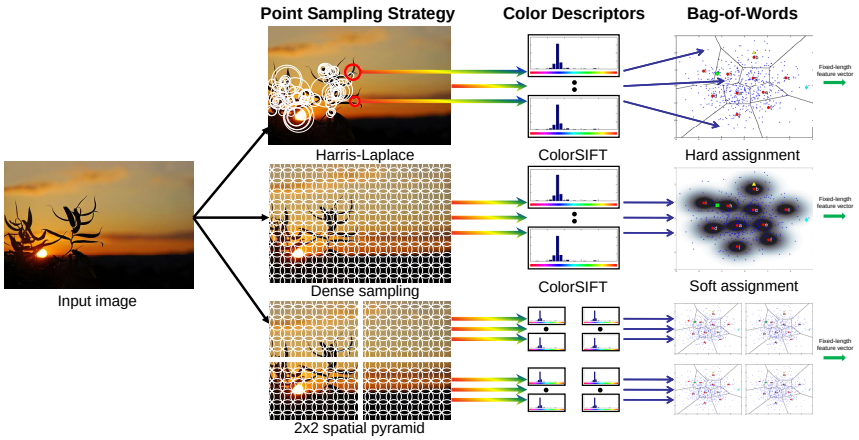

Fig. 18.1: University of Amsterdam's ImageCLEF 2009 concept detection scheme. The scheme serves as the blueprint for the organization of Section 18.2.

18.2.1 Point Sampling Strategy

The visual appearance of a concept has a strong dependency on the viewpoint under which it is recorded. Salient point methods Tuytelaars and Mikolajczyk (2008) introduce robustness against viewpoint changes by selecting points, which can be recovered under different perspectives. Another simpler solution is to use many points, which is achieved by dense sampling.

In the context of concept classification, two classes of concepts are identified: objects and scene types. Dense sampling has been shown to be advantageous for scene type classification, since salient points do not capture the entire appearance of an image. For object classification, salient points can be advantageous because they ignore homogenous areas in the image. If the object background is not highly textured, then most salient points will be located on the object or the object boundary.

We summarize our sampling approach in Figure 18.1: Harris-Laplace and dense point selection, and a spatial pyramid.[1]

Harris-Laplace point detector

In order to determine salient points, Harris-Laplace relies on a Harris corner detector. By applying it on multiple scales, it is possible to select the characteristic scale of a local corner using the Laplacian operator, discussed in Tuytelaars and Mikolajczyk (2008). Hence, for each corner the Harris-Laplace detector selects a

[1] Software to perform point sampling, color descriptor computation and the hard and soft assignment is available from `http://www.colordescriptors.com`.

scale-invariant point if the local image structure under a Laplacian operator has a stable maximum.

Dense point detector

For concepts with many homogenous areas, like scenes, corners are often rare. Hence, for these concepts relying on a Harris-Laplace detector can be suboptimal. To counter the shortcoming of Harris-Laplace, random and dense sampling strategies have been proposed by Fei-Fei and Perona (2005) and Jurie and Triggs (2005). We employ dense sampling, which samples an image grid in a uniform fashion using a fixed pixel interval between regions. To study the effect of different parameter choices for dense sampling, we investigate three different settings:

- An interval of 6 pixels and sample at a singe scale ($\sigma = 1.2$).
- An interval of 6 pixels and sample at multiple scales ($\sigma = 1.2$ and $\sigma = 2.0$).
- An interval of 1 pixel, *e.g*. sample every pixel with a singe scale ($\sigma = 1.2$).

Spatial pyramid

Both Harris-Laplace and dense sampling give an equal weight to all keypoints, irrespective of their spatial location in the image. To overcome this limitation, Lazebnik et al (2006) suggest to repeatedly sample fixed subregions of an image, *e.g*. 1x1, 2x2, 4x4, *etc*., and to aggregate the different resolutions into a so called spatial pyramid. Since every region is an image in itself, the spatial pyramid can be used in combination with both the Harris-Laplace point detector and dense point sampling, as was done in van de Sande et al (2008), for example. For the ideal spatial pyramid configuration, Lazebnik et al (2006) claims 2x2 is sufficient, Marszałek et al (2007) suggests to include 1x3 also. We investigate multiple divisions of the image in our experiments.

18.2.2 Color Descriptor Extraction

In the previous section, we addressed the dependency of the visual appearance of semantic concepts on the viewpoint under which they are recorded. However, the lighting conditions during photography also play an important role. van de Sande et al (2010) analyzed the properties of color descriptors under classes of illumination changes within the diagonal model of illumination change, and specifically for data sets consisting of Flickr images. In ImageCLEF, the images used also originate from Flickr. Here we use the four color descriptors from the recommendation table in van de Sande et al (2010). The descriptors are computed around salient points obtained from the Harris-Laplace detector and dense sampling. For the color descriptors in Figure 18.1, each of those four descriptors can be inserted.

SIFT

The SIFT feature proposed by Lowe (2004) describes the local shape of a region using edge orientation histograms. The gradient of an image is shift-invariant: taking the derivative cancels out offsets (see van de Sande et al (2010) for details). Under light intensity changes, *i.e*. a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Because the SIFT feature is normalized, the gradient magnitude changes have no effect on the final feature. To compute SIFT features, we use the version described by Lowe (2004).

OpponentSIFT

OpponentSIFT describes all the channels in the opponent color space using SIFT features. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color information in the image. The feature normalization, as effective in SIFT, cancels out any local changes in light intensity.

C-SIFT

The C-SIFT feature uses the C invariant by Geusebroek et al (2001), which can be intuitively seen as the gradient (or derivative) for the normalized opponent color space $O1/I$ and $O2/I$. The I intensity channel remains unchanged. C-SIFT is known to be scale-invariant with respect to light intensity. See Burghouts and Geusebroek (2009) and van de Sande et al (2010) for detailed evaluation.

RGB-SIFT

For the RGB-SIFT, the SIFT feature is computed for each *RGB* channel independently. Due to the normalizations performed within SIFT, it is scale-invariant, shift-invariant, and invariant to light color changes and shift (see van de Sande et al (2010) for details).

18.2.3 Bag-of-Words model

We use the well-known bag-of-words model, also known as codebook approach, see *e.g*. Leung and Malik (2001); Jurie and Triggs (2005); Zhang et al (2007); van Gemert et al (2010); van de Sande et al (2010). First, we assign visual descriptors to discrete codewords predefined in a codebook. Then, we use the frequency distribution of the codewords as a feature vector representing an image. We construct a codebook with a maximum size of 4096 using k-means clustering. An important is-

sue is *codeword assignment*. An comparison of codeword assignment is presented in van Gemert et al (2010). Here we only discuss two codeword assignment methods:

- **Hard assignment**. Given a codebook of codewords, the traditional codebook approach assigns each descriptor to a single best representative codeword in the codebook. Basically, an image is represented by a histogram of codeword frequencies describing the probability density over codewords.
- **Soft-assignment**. The traditional codebook approach may be improved by using soft-assignment through kernel codebooks. A kernel codebook uses a kernel function to smooth the hard-assignment of image features to codewords. Out of the various forms of kernel-codebooks, we selected *codeword uncertainty* based on its empirical performance, shown in van Gemert et al (2010).

Each of the possible sampling methods from Section 18.2.1 coupled with each visual descriptor from Section 18.2.2, and an assignment approach results in a separate visual codebook. An example is a codebook based on dense sampling of RGB-SIFT features in combination with hard-assignment. Naturally, various configurations can be used to combine multiple of these choices. By default, we use hard assignment in our experiments. Soft assignment is only used when explicitly stated. For simplicity, we employ equal weights in our experiments when combining different features.

18.2.4 Machine Learning

The supervised machine learning process is composed of two phases: training and testing. In the first phase, the optimal configuration of features is learned from the training data. From all machine learning approaches on offer to learn the appearance relation between C and x_i, the support vector machine by Vapnik (2000) is commonly regarded as a solid choice. We use the LIBSVM implementation by Chang and Lin (2001) with probabilistic output as described in Lin et al (2007). The parameter of the support vector machine we optimize is C. In order to handle imbalance in the number of positive versus negative training examples, we fix the weights of the positive and negative class by estimation from the class priors on training data. It was shown by Zhang et al (2007) that in a codebook-approach to concept detection the earth movers distance and χ^2 kernel are to be preferred. We employ the χ^2 kernel, as it is less expensive in terms of computation.

In the second machine learning phase, the classifier assigns a probability $p(C|x_i)$ to each input feature vector for each semantic concept C, *i.e.* the trained model is applied to the test data.

18.3 Experiments

Experiments in this section are performed using the MIR-Flickr dataset (see Chapter) using the 53 concepts annotated for ImageCLEF 2009. Concept models are trained using the ImageCLEF 2009 Photo Annotation task training set. Results are computed using the labels provided after the contest for 3000 images from the test set. In subsequent years, these images became part of the training set. All results in this section will use the Average Precision metric which will be the standard for the large-scale visual concept detection and annotation task from ImageCLEF 2010 onward. Higher average precision scores imply better accurate concept detection.

18.3.1 Spatial Pyramid Levels

In Table 18.1, we show results for different subdivisions of the image based on the spatial pyramid framework. These results are obtained for an intensity SIFT descriptor sampled with an interval of 6 pixels at two different scales and hard codebook assignment.

Table 18.1: Overall results of using different pyramid levels on the 3000 images from ImageCLEF 2009 whose annotations have been made available. Evaluated over 53 concepts in the Photo Annotation task using Mean Average Precision (MAP). The third column lists the number of concepts for which that row obtains the highest AP relative to all other rows (excluding the bottom row with the 1x1, 2x2, 1x3 combination). In the fourth row, the bottom row is included when determining the row with the highest AP.

Pyramid subdivisions	Average Precision	#concepts with highest AP	#concepts with highest AP
1x1	0,415	0	0
2x2	0,411	3	2
3x3	0,398	3	3
1x2	0,409	1	1
1x3	0,421	5	3
2x1	0,398	1	1
3x1	0,396	1	1
1x1, 2x2	0,427	10	6
1x1, 1x2	0,427	6	6
1x1, 1x3	0,432	14	7
1x1, 2x1	0,420	6	5
1x1, 3x1	0,421	4	4
1x1, 2x2, 1x3	0,433	not included	14

Inspecting the results, we see in the first seven rows that only a pyramid with horizontal bars, 1x3, obtains higher overall AP than just using the full image (1x1

subdivision). Therefore, we introduce combinations of two pyramid subdivisions, one of which is always 1x1. Here, we see that 1x1+2x2, 1x1+1x2 and 1x1+1x3 obtain the highest overall scores. Looking at per-concept results (not shown here), 1x1+2x2 obtains the highest AP compared to other subdivisions for 10 concepts (see third column), and 1x1+1x3 for 14 concepts. The subdivision of 1x1+1x2 obtains the highest score for only 6 concepts. Based on these results and its popularity in the PASCAL VOC Everingham et al (2010), we introduce the combination of 1x1+2x2+1x3, which combines the two best subdivisions with two parts. If we then recount the number of concepts for which a row obtains the highest AP (see the fourth column), this new combination obtains the highest score for 14 concepts and the highest AP of all rows.

In terms of priority, the 1x3 subdivision (horizontal bars) is the most important, then the 1x1 subdivision (the whole image), and finally a 2x2 division. This raises the question as to why the 1x3 subdivision works so well on the MIR-Flickr dataset. A possible explanation is the way photographers work: they attempt to center the object of interest, have a straight horizon which is approximately in the middle of the image. Therefore, the top horizontal bar of a 1x3 division will probably be sky, the middle horizontal bar will contain the object of interest plus the horizon, and the bottom bar will contain the ground.

Based on these results, we draw the conclusion that using a combination of spatial subdivisions of 1x1, 2x2 and 1x3 is a good choice for the MIR-Flickr dataset. The experiments in the remainder of this section will use exactly these subdivisions.

18.3.2 Point Sampling Strategies and Color Descriptors

In Figure 18.2, results are shown for 4 different point sampling methods and 4 different color descriptors. Inspecting these results, dense sampling clearly outperforms the Harris-Laplace salient points. Sampling at two scales instead of a single scale at an interval of 6 pixels is better. However, when the sampling interval is set to 1 pixel, *e.g.* every pixel is described, performance at a single scale exceeds the 6-pixel multi-scale results. These observations are consistent across all descriptors. The main drawback of sampling every pixel is that 36 times more descriptors are extracted per image, which results in a significant increase in feature extraction times. A possible solution to the computational load is to use software specifically optimized for dense sampling, as is done in Uijlings et al (2009).

When comparing the performance of different descriptors in Figure 18.2, we observe that the RGB-SIFT descriptor yields the highest performance on the MIR-Flickr dataset. The presence of rather 'artistic' photographs with large variations in lighting conditions in the dataset might explain why the illumination color-invariant descriptor gets the best results. Ordering the descriptors by their performance, the OpponentSIFT descriptor is in second place, followed by SIFT and finally C-SIFT.

In conclusion, a dense sampling strategy and the RGB-SIFT descriptor together give the best results for a single feature on the MIR-Flickr dataset.

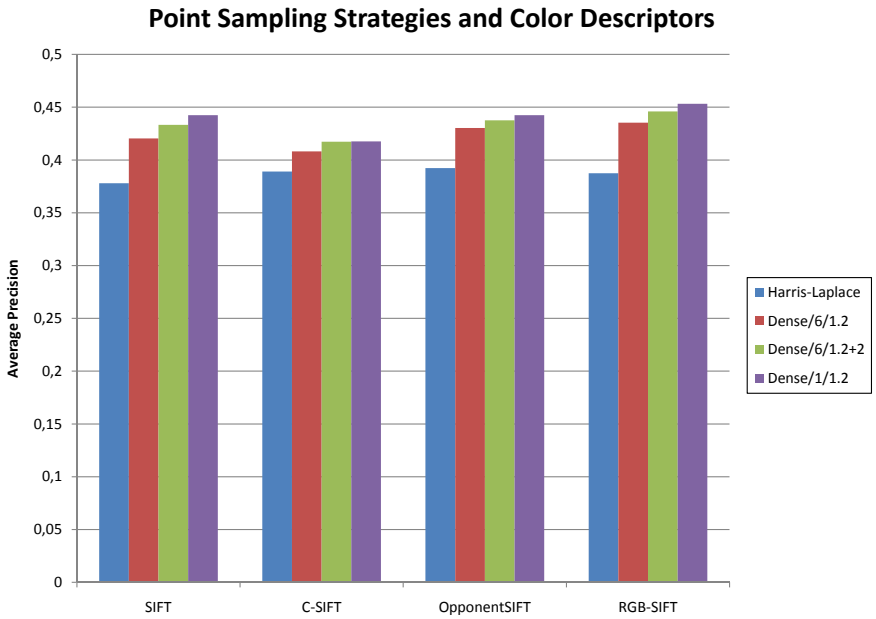

Fig. 18.2: Performance of color descriptors using either Harris-Laplace salient points or dense sampling. The dense sampling has an interval of either 1 or 6 pixels and is carried out at one (1.2) or two scales (1.2 and 2).

18.3.3 Combinations of Sampling Strategies and Descriptors

Complete concept detection systems do not use a single feature, as was done in the previous section, but use combinations of different features. The more features are added, the higher the performance becomes. In Figure 18.3 and Table 18.2, results of different system configurations are shown. The baseline is a single feature, the densely sampled SIFT descriptor with a 1x1, 2x2 and 1x3 spatial pyramid, *i.e.* the best result from the spatial pyramid experiment. From the comparative experiment of point sampling strategies and descriptor, we know the RGB-SIFT descriptor is the best individual descriptor for the MIR-Flickr dataset. Therefore, results for this descriptor are also listed.

The best combinations typically used in the University of Amsterdam concept detection system are as follows:

- Combination of 8 features: Harris-Laplace salient points paired with each of the 4 ColorSIFT descriptors, and densely sampled points at multiple scales with an interval of 6 pixels, also paired with each of the 4 ColorSIFT descriptors.
- Combination of 8 features: Harris-Laplace salient points paired with each of the 4 ColorSIFT descriptors, and densely sampled points at a single scale at every

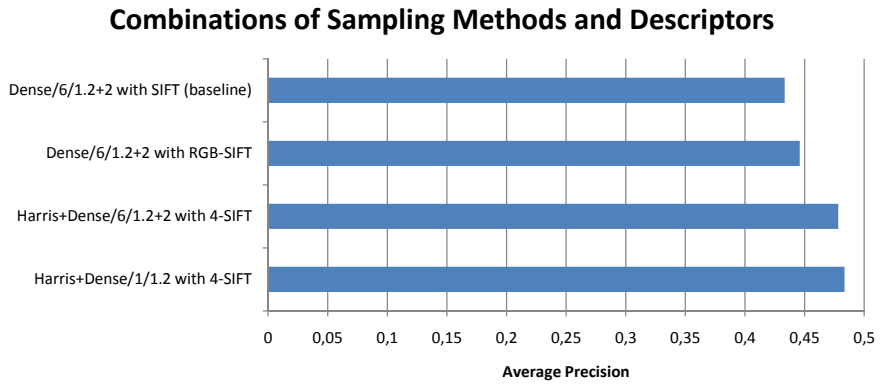

Fig. 18.3: Performance of different combinations of multiple point sampling strategies and multiple descriptors. Numeric results are given in Table 18.2.

pixel, *e.g.* an interval of 1 pixel, also paired with each of the 4 ColorSIFT descriptors.

The results in Figure 18.3 and Table 18.2 show that the first combination is a relative improvement of 7% over the RGB-SIFT feature (absolute difference 0.032). The second combination is an improvement of 8% (absolute difference 0.038). These differences are significant in benchmark evaluations, and they show that using different color descriptors together is not redundant, because performance improved by combining them.

It is important to realize that about 90% of the state-of-the-art performance can be obtained by using the densely sampled RGB-SIFT feature with an interval of 6 pixels. The computational effort to extract this feature instead of applying the full feature combination is 8 times lower when compared to the first combination, and 25 times lower compared to the second combination. For datasets orders of magnitude larger than the MIR-Flickr dataset, choosing the single best feature might be more practical.

Table 18.2: Performance of different combinations of multiple point sampling strategies and multiple descriptors. A visualization of the results is given in Figure 18.3.

Combinations of sampling strategies and descriptors	Average Precision
Dense sampling every 6 pixels (multi-scale) with SIFT (baseline)	0,433
Dense sampling every 6 pixels (multi-scale) with RGB-SIFT	0,446
Harris-Laplace and dense sampling every 6 pixels (multi-scale) with 4-SIFT	0,478
Harris-Laplace and dense sampling every pixel (single-scale) with 4-SIFT	**0,484**

In conclusion, by combining different sampling strategies and descriptors, a performance improvement of up to 8% is possible. At the same time, when feature extraction becomes a computational bottleneck, picking a single good feature can already give up to 90% of the performance of the state-of-the-art.

18.3.4 Discussion

Based on the experiments in this section, we have found that the combination of spatial pyramid subdivisions of 1x1, 2x2 and 1x3 is a good choice for the MIR-Flickr dataset. This confirms similar observations made on the PASCAL VOC in Everingham et al (2010).

In terms of point sampling strategy, dense sampling strategy gives the best results for on the MIR-Flickr dataset. This is related to the large number of scene concepts to be annotated: dense sampling has been shown to be advantageous for scene type classification, since salient points do not capture the entire appearance of an image. For dense sampling, it holds that denser sampling almost always gives higher accuracy than coarser sampling. The decision on how dense to sample should be made based on the available compute resources.

Which descriptor gives the highest performance depends on the dataset used: in van de Sande et al (2010) the C-SIFT descriptor gives the highest accuracy on PASCAL VOC. On the MIR-Flickr dataset, the RGB-SIFT descriptor, which is invariant to illuminant color changes, gives the highest average precision.

By combining different sampling strategies and descriptors, as is done in all concept detection systems aiming for high accuracy, a performance improvement of up to 8% is possible given the features in this chapter. When limited compute resources are available, picking a single good feature, *e.g.* refraining from the use of combinations, can already give up to 90% of the performance of the current state-of-the-art.

18.4 ImageCLEF 2009

This section reports on the official ImageCLEF 2009 results of our concept detection system. Our focus on invariant visual features for concept detection in ImageCLEF 2009 was successful. It has resulted in the top ranking for the large-scale visual concept detection task in terms of both EER and AUC.

All runs submitted to ImageCLEF 2009 use both Harris-Laplace, dense sampling with an interval of 6 pixels at two scales, the SVM classifier and a spatial pyramid with 1x1, 2x2 and 1x3 subdivisions. We do not use the EXIF metadata provided for the photos.

- **OpponentSIFT**: single color descriptor with hard assignment.
- **2-SIFT**: uses OpponentSIFT and SIFT descriptors.

- **4-SIFT**: uses OpponentSIFT, C-SIFT, RGB-SIFT and SIFT descriptors. This run is equal to the first combination of the combination experiment in Section 18.3.3.
- **Soft 4-SIFT**: uses OpponentSIFT, C-SIFT, RGB-SIFT and SIFT descriptors with soft assignment. The soft assignment parameters have been taken from our PASCAL VOC 2008 system van de Sande et al (2010).

In table 18.3, the overall scores for the evaluation of concept detectors are shown. We note that the 4-SIFT run with hard assignment achieves not only the highest performance amongst our runs, but also over all other runs submitted to the Large-Scale Visual Concept Detection task.

Table 18.3: Overall results of the our runs evaluated over all concepts in the Photo Annotation task using the equal error rate (EER) and the area under the curve (AUC).

Run name	Codebook	Average EER	Average AUC
4-SIFT	Hard-assignment	**0.2345**	**0.8387**
Soft 4-SIFT	Soft-assignment	0.2355	0.8375
2-SIFT	Hard-assignment	0.2435	0.8300
OpponentSIFT	Hard-assignment	0.2530	0.8217

In table 18.4, the Area Under the Curve scores have been split out per concept. We observe that the three aesthetic concepts have the lowest scores. This comes as no surprise, because these concepts are highly subjective: even human annotators only agree around 80% of the time with each other. For virtually all concepts besides the aesthetic ones, either the Soft 4-SIFT or the Hard 4-SIFT is the best run. This confirms our beliefs that these (color) descriptors are not redundant when used in combinations. Therefore, we recommend the use of these 4 descriptors instead of 1 or 2. The difference in overall performance between the Soft 4-SIFT or the Hard 4-SIFT run is quite small. Because the soft codebook assignment smoothing parameter was directly taken from a different dataset, we expect that the soft assignment run could be improved if the soft assignment parameter was selected with cross-validation on the training set. Together, our runs obtain the highest Area Under the Curve scores for 40 out of 53 concepts in the Photo Annotation task (20 for Soft 4-SIFT, 17 for 4-SIFT and 3 for the other runs). This analysis has shown us that our system is falling behind for concepts that correspond to conditions we have included invariance against. Our method is designed to be robust to unsharp images, so for *Out-of-focus*, *Partly-Blurred* and *No-Blur* there are better approaches possible. For the concepts *Overexposed*, *Underexposed*, *Neutral-Illumination*, *Night* and *Sunny*, recognizing how the scene is illuminated is very important. Because we are using invariant color descriptors, a lot of the discriminative lighting information is no longer present in the descriptors. Again, there should be better approaches possible for these concepts, such as estimating the color temperature and overall light intensity.

Table 18.4: Results per concept for our runs in the Large-Scale Visual Concept Detection Task using the Area Under the Curve. The highest score per concept is highlighted using a grey background. The concepts are ordered by their highest score.

Concept	4-SIFT	Soft 4-SIFT	2-SIFT	Opp.SIFT
Clouds	**0,958**	0,958	0,951	0,945
Sunset-Sunrise	0,953	**0,954**	0,947	0,946
Sky	0,945	**0,948**	0,935	0,930
Landscape-Nature	**0,944**	0,942	0,940	0,936
Sea	**0,935**	0,930	0,932	0,926
Mountains	**0,934**	0,931	0,930	0,922
Lake	0,911	0,903	**0,912**	0,900
Beach-Holidays	0,906	**0,907**	0,898	0,884
Trees	**0,903**	0,902	0,892	0,881
Water	0,901	**0,903**	0,892	0,886
Night	**0,898**	0,895	0,895	0,892
River	**0,897**	0,889	0,891	0,883
Outdoor	0,890	**0,896**	0,879	0,871
Food	**0,895**	0,895	0,881	0,877
Desert	**0,891**	0,865	0,891	0,884
Building-Sights	0,880	**0,882**	0,873	0,861
Big-Group	**0,881**	0,877	0,870	0,858
Plants	0,877	**0,881**	0,853	0,839
Flowers	0,868	**0,875**	0,846	0,836
Autumn	**0,870**	0,866	0,863	0,849
Portrait	**0,865**	0,864	0,857	0,846
Underexposed	0,858	**0,859**	0,857	0,854
No-Persons	0,850	**0,858**	0,837	0,826
Partly-Blurred	**0,852**	0,852	0,845	0,830
Winter	0,843	**0,846**	0,832	0,828
Snow	**0,846**	0,845	0,829	0,825
Day	0,841	**0,845**	0,831	0,824
No-Blur	0,843	**0,845**	0,836	0,823

Concept	4-SIFT	Soft 4-SIFT	2-SIFT	Opp.SIFT
No-Visual-Time	0,833	**0,835**	0,822	0,815
Indoor	0,830	**0,835**	0,823	0,810
Familiy-Friends	**0,834**	0,834	0,822	0,813
Partylife	**0,834**	0,834	0,831	0,819
Vehicle	**0,832**	0,832	0,832	0,822
Animals	0,818	**0,828**	0,811	0,797
Citylife	**0,826**	0,826	0,819	0,813
Still-Life	0,824	**0,825**	0,808	0,795
Spring	**0,822**	0,801	0,812	0,791
Canvas	**0,817**	0,810	0,803	0,790
Summer	**0,813**	0,813	0,791	0,782
Macro	**0,812**	0,791	0,805	0,795
No-Visual-Season	0,805	**0,806**	0,794	0,782
Small-Group	0,792	**0,795**	0,784	0,776
Single-Person	0,792	**0,795**	0,780	0,769
Out-of-focus	**0,792**	0,781	0,784	0,774
No-Visual-Place	**0,789**	0,786	0,781	0,779
Overexposed	**0,788**	0,782	0,777	0,771
Neutral-Illumination	0,778	**0,783**	0,775	0,774
Sunny	0,763	**0,765**	0,744	0,741
Motion-Blur	0,744	**0,747**	0,725	0,710
Sports	**0,695**	0,695	0,679	0,673
Aesthetic-Impression	0,658	**0,662**	0,657	0,657
Overall-Quality	0,656	0,656	0,653	**0,658**
Fancy	0,565	0,559	0,580	**0,583**
Average	**0,8387**	**0,8375**	**0,8300**	**0,8217**

18.4.1 Evaluation Per Image

For the hierarchical evaluation, overall results are shown in table 18.5. When compared to the evaluation per concept, the Soft 4-SIFT run is now slightly better than the normal 4-SIFT run. While our method provides the best run for the per-concept evaluation, for the hierachical evaluation measure, several other participants perform better. Discussion at the workshop has shown that exploiting the hierachical nature of the concepts used is an interesting future direction.

18.4.2 Conclusion

The focus on invariant visual features for concept detection in ImageCLEF 2009 was successful. It resulted in the top ranking for the large-scale visual concept detection task in terms of both EER and AUC. For 40 individual concepts, the highest performance of all submissions to the task was obtained. For the hierarchical evaluation, how to exploit the hierachical nature of the concepts is still an open question.

Table 18.5: Results using the hierarchical evaluation measures for our runs in the ImageCLEF 2009 Large-Scale Visual Concept Detection Task.

		Average Annotation Score	
Run name	Codebook	with agreement	without agreement
Soft 4-SIFT	Soft-assignment	**0.7647**	**0.7400**
4-SIFT	Hard-assignment	0.7623	0.7374
2-SIFT	Hard-assignment	0.7581	0.7329
OpponentSIFT	Hard-assignment	0.7491	0.7232

18.5 ImageCLEF@ICPR 2010

The visual concept detection and annotation task has been part of a contest for the 2010 ICPR conference. The MIR-Flickr dataset is used with the 53 concepts annotated for ImageCLEF 2009. However, this time there are 8000 labelled images available for training, and the test set consists of 13000 images.

University of Amsterdam submitted two runs to the ICPR contest: the two good combinations which were identified in Section 18.3.3. In table 18.6, the overall scores for the evaluation of concept detectors are shown. The first run is equal in terms of features to the best run submitted to ImageCLEF 2009. The second run, with more densely sampled SIFT, achieves higher accuracy. Compared to all other runs submitted to the task, the University of Amsterdam runs are ranked in second place. University of Surrey achieved the highest overall accuracy. Their system uses similar visual features within a bag-of-words model, but uses improved machine learning algorithms.

Table 18.6: Overall results of the our runs evaluated over all concepts in the ImageCLEF@ICPR 2010 Photo Annotation task using the equal error rate (EER) and the area under the curve (AUC).

Run contents	Average EER	Average AUC
Harris-Laplace and dense sampling every 6 pixels (multi-scale) with 4-SIFT	0.2214	0.8538
Harris-Laplace and dense sampling every pixel (single-scale) with 4-SIFT	0.2182	0.8568
University of Surrey (enhanced machine learning)	**0.2136**	**0.8600**

18.6 Conclusion

The current state-of-the-art in visual concept detection and annotation tasks is based on the bag-of-words model. Within this model, we have identified several design

choices which lead to higher classification accuracy. Participation in the ImageCLEF Photo Annotation benchmarks was successful, and this participation was based on the following conclusions: (1) In terms of point sampling strategy, dense sampling gives the best results due to the large number of scene concepts to be annotated. (2) Increasing sampling density improves accuracy. (3) Spatial pyramid subdivisions of 1x1, 2x2 and 1x3 are a good choice for datasets in general. (4) The descriptor which gives the highest performance depends on the dataset used; for the MIR-Flickr dataset, the RGB-SIFT descriptor is recommended. (5) By combining different sampling strategies and descriptors, a performance improvement of up to 8% is possible given the features in this chapter.

Finally, when limited compute resources are available, picking a single good feature, *e.g.* refraining from the use of combinations, can already give up to 90% of the performance of the current state-of-the-art.

References

Burghouts GJ, Geusebroek JM (2009) Performance evaluation of local color invariants. Computer Vision and Image Understanding 113:48–62

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Everingham M, Van Gool L, Williams C, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88(2):303–338, DOI http://dx.doi.org/10.1007/s11263-009-0275-4

Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, vol 2, pp 524–531

van Gemert JC, Veenman CJ, Smeulders AWM, Geusebroek JM (2010) Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(7):1271–1283, DOI http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.132

Geusebroek JM, van den Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(12):1338–1350

Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: IEEE International Conference on Computer Vision, pp 604–610

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, vol 2, pp 2169–2178

Leung TK, Malik J (2001) Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 43(1):29–44

Lin HT, Lin CJ, Weng RC (2007) A note on Platt's probabilistic outputs for support vector machines. ML 68(3):267–276

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Marszałek M, Schmid C, Harzallah H, van de Weijer J (2007) Learning object representations for visual object class recognition. Visual Recognition Challenge workshop, in conjunction with IEEE ICCV

Nowak S, Dunker P (2009) Overview of the clef 2009 large scale visual concept detection and annotation task. In: CLEF working notes 2009, Corfu, Greece

van de Sande KEA, Gevers T, Snoek CGM (2008) A comparison of color features for visual concept classification. In: ACM International Conference on Image and Video Retrieval, pp 141–150

van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9)

Snoek CGM, Worring M (2009) Concept-based video retrieval. Foundations and Trends in Information Retrieval 4(2):215–322

Snoek CGM, van de Sande KEA, de Rooij O, Huurnink B, van Gemert JC, Uijlings JRR, *et al*.(2008) The MediaMill TRECVID 2008 semantic video search engine. In: Proceedings of the TRECVID Workshop

Snoek CGM, van de Sande KEA, de Rooij O, Huurnink B, Uijlings JRR, van Liempt M, Bugalho M, Trancoso I, Yan F, Tahir MA, Mikolajczyk K, Kittler J, de Rijke M, Geusebroek JM, Gevers T, Worring M, Koelma DC, Smeulders AWM (2009) The MediaMill TRECVID 2009 semantic video search engine. In: Proceedings of the TRECVID Workshop

Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3):177–280

Uijlings JRR, Smeulders AWM, Scha RJH (2009) Real-time bag-of-words, approximately. In: ACM International Conference on Image and Video Retrieval

Vapnik VN (2000) The Nature of Statistical Learning Theory, 2nd edn

Wang D, Liu X, Luo L, Li J, Zhang B (2007) Video diver: generic video indexing with diverse features. In: ACM International Workshop on Multimedia Information Retrieval, Augsburg, Germany, pp 61–70

Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2):213–238