Semantic Contexts and Fisher Vectors for the ImageCLEF 2011 Photo Annotation Task

Yu Su and Frédéric Jurie

GREYC, University of Caen, France
firstname.lastname@unicaen.fr

Abstract. This paper describes the participation of UNICAEN/GREYC to the ImageCLEF 2011 photo annotation task. The proposed approach uses visual image features and binary annotations of concepts only. In this approach, the annotations are predicted by SVM classifiers trained separately for each concept. The classifiers take Bag-of-Words histograms and fisher vectors representations as inputs, both being combined at the decision level. Furthermore, contextual information is also embedded into the Bag-of-Words histograms to enhance their performance. The experimental results show that the combination of Bag-of-Words histograms and Fisher vectors brings significant performance increase (e.g. 4% for Mean Average Precision). Furthermore, the results of our best-run rank in top 3 for both concept and image level evaluations.

Keywords: Image classification, Photo annotation, Bag-of-Words model, Semantic context, Fisher Vectors

1 Introduction

The aim of the ImageCLEF 2011 photo annotation task is to automatically assign to each image a set of concepts taken from a list of 99 possible pre-defined visual concepts. In this task, the participants are given 8000 training images associated with the corresponding 99 binary labels, each of which corresponds to a visual concept, as well as the photo tagging ontology, EXIF data and Flickr user tags. In the test phase, the participants are requested to give to each test image the labels of all the visual concepts describing the image. The evaluation of performance is done at concept and image levels. For the former, Mean Average Precision (MAP) is computed for each concept. For the latter, F-Measure (F-ex) and the Semantic R-Precision (SR-Precision) are computed for each image. For more details on this task, please refer to [7] .

In our participation, we did not use photo tagging ontology, EXIF data and Flickr user tags. Our results are only based on visual image features. Specifically, we extracted different types of local features (e.g. SIFT) from images and then adopted Bag-of-Words (BoW) model to aggregate local features into a global image descriptor. Our participation is mainly inspired by the work of Su and Jurie [11], which proposed to embed some contextual information into

the BoW model. In addition, some improvements over [11] are also proposed. Indeed, Fisher Vectors (FV) have been reported to give good performance on both object recognition and image retrieval tasks [9]. Thus, we also computed FV from images and combined them with the context-embedded BoW histograms at decision level, i.e, training classifiers for Fisher Vectors and context-embedded BoW histograms separately and combining classifiers by averaging their outputs. As to photo annotation, the above process is performed for each concept independently and the averaged classifier outputs are used as the confidences of concept occurrence.

The organization of this paper is as follows: In section 2, we describe local features used in our method. Then, we explain how to combine BoW model with both semantic contexts (section 3) and FV (section 4). Experimental evaluation is given in section 5, followed by a conclusion in the last section.

2 Visual Features

In our method, 6 kinds of visual features are extracted from each image, which are introduced in the following paragraph. Before feature computation, the images are scaled to be at most 300×300 pixels, with their original aspect ratios maintained. Except for LAB features which encode color information, color images are first transformed to grayscale.

SIFT Vector quantized SIFT descriptors [6] are computed for 5000 image patches with randomly selected positions and scales (with scales from 16 to 64 pixels), and are quantized to 1024 *k-means* centers.

HOG HOG descriptors [3] are densely extracted on a regular grid at step of 8 pixels. On each node of the grid a 31 dimensional descriptor is computed and then 2×2 neighboring descriptors are concatenated to form a descriptor of 124 dimensions. HOG features are finally vector quantized to 256 *k-means* centers.

Textons Texton descriptors [12] are generated by computing the output of 36 Gabor filters with different scales and orientations for each pixel, and then quantized to 256 *k-means* centers.

SSIM Self-similarity descriptors [10] are computed on a regular grid at step of five pixels. Each feature is obtained by computing the correlation map of a patch of 5×5 in a window with radius of 40 pixels, then quantizing it in 3 radial bins and 10 angular bins, obtaining 30 dimensional descriptor vectors. Self-similarity features are finally quantized to 256 *k-means* centers.

LAB LAB descriptors [4] are computed for each pixel and then quantized to 128 *k-means* centers.

Canny Canny edge descriptors [1] are computed for each pixel and then quantized to 8 orientation bins.

Finally, concatenating all BoW histograms gives a 1928-dimensional feature vector which can describe an image or a image region.

3 Image Representation by Embedding Semantic Contexts into BoW Model

In this section, we first review how do we define the semantic contexts and embed them into BoW model as introduced in [11]. Then we introduce our improvements over this method.

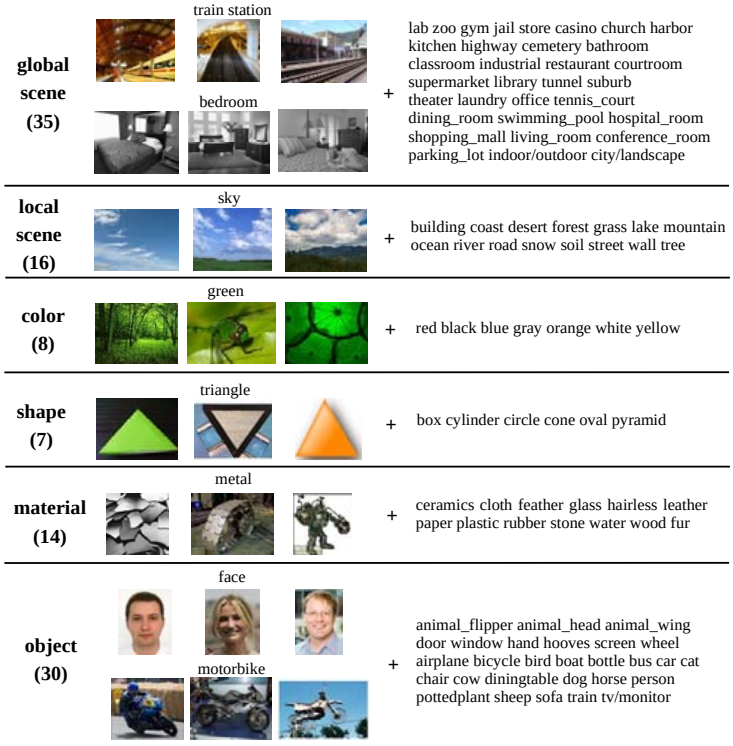

Fig. 1. Grouped semantic contexts and some illustrative training images [11]. The values between brackets are the number of semantic contexts within corresponding groups.

3.1 Semantic Context

In [11], 110 semantic contexts are defined by hand with the intention of providing abundant semantic information for image description. (see Fig. 1). Two types of contexts are distinguished: *global* contexts including global scenes and *local* contexts including local scenes, colors, shapes, materials and objects.

For each semantic context, we learn a SVM classifier with linear kernel (hereafter called as context classifiers). For the *global* contexts, the classifiers are

learned on whole images described by BoW histograms. For the *local* contexts, the classifiers are learned on some randomly sampled image regions described again by BoW histograms. The training images are automatically downloaded from Google image search by using the name of context as query. After the manual annotation, about 400 relevant images are reserved for each context. They are used as positive images for the corresponding context while images from the other contexts are considered as negatives.

In test phase, images (for *global* contexts) or regions (for *local* contexts) are input to context classifiers and a sigmoid function is used to transform the original decision values to probabilities (refer to [2]).

3.2 Embedding Semantic Contexts into BoW model

Assume that, for an image I, a set of local features $f_i, i = 1, \ldots, N$ are extracted from it, where N is the number of local features. The BoW model consists of V visual words $v_j, j = 1, \ldots, V$. The traditional BoW feature for v_j measures the occurrence probability of v_j on image I, say $p(v_j|I)$. In practice, $p(v_j|I)$ is usually computed by:

$$p(v_j|I) = \frac{1}{N} \sum_{i=1}^{N} \delta(f_i, v_j), \tag{1}$$

where

$$\delta(f_i, v_j) = \begin{cases} 1 \text{ if } j = \arg\min\limits_{j=1,\ldots,V} d(f_i, v_j) \\ 0 \text{ else} \end{cases} \tag{2}$$

and d is a distance function (e.g., the $L2$ norm).

Marginalizing $p(v_j|I)$ over different local contexts gives:

$$p(v_j|I) = \sum_{k=1}^{C} p(v_j|c_k, I) p(c_k|I), \tag{3}$$

where c_k is the k-th context, C is the number of *local* contexts (75 in our case), $p(v_j|c_k, I)$ is the context-specific occurrence probability of v_j on image I, $p(c_k|I)$ is the occurrence probability of context c_k on image I.

On the other hand, the second term of Eq. 3, which gives the distribution of different contexts on image I, can also provide rich information to describe the image, as shown by [13]. For example, knowing an image is composed of one third of *sky*, one third of *sea* and one third of *beach*, brings a lot of information regarding the content of this image. At the end, images are eventually represented by multiple context-specific BoW histograms, i.e., $p(v_j|c_k, I)$ and a vector of context-occurring probabilities, i.e., $p(c_k|I)$.

In [11], $p(v_j|c_k, I)$ is constructed by modeling the probabilistic distribution of context c_k on image I which is estimated by dividing image I into a set of regions I_p and predicting the occurrence probabilities of c_k for each region (by

using context classifiers). By denoting $I_p(f_i)$ the set of image regions which cover the local feature f_i, we define:

$$p(v_j|c_k, I) = \frac{1}{N}\sum_{i=1}^{N}\delta(f_i, v_j)p(c_k|I_p(f_i)), \tag{4}$$

where $p(c_k|I_p(f_i)$ can be considered as the weight of local feature f_i. In practice, $p(c_k|I_p(f_i))$ is computed by averaging the outputs of the context classifier (for c_k) on $I_p(f_i)$.

As to $p(c_k|I)$, it can be easily computed by averaging the outputs of the context classifiers (for c_k) on all image regions in I_p. This process is similar to the computation of $p(c_k|I_p(f_i))$ in previous subsection. In addition, we also represent image I by the occurrence probabilities of *global* contexts. These probabilities are computed by running the corresponding context classifiers on the whole image. Finally, an image is represented by concatenating the occurrence probabilities of both *global* and *local* contexts, i.e.,

$$(p(c_1|I), \ldots, p(c_C|I), p(c_{C+1}|I), \ldots, p(c_{C'}|I)),$$

where C' is the number of all contexts (110 in our case) and C is the number of *local* contexts (75 in our case). We call this image descriptor as semantic features.

3.3 Our improvements over [11]

The above subsection reviewed the process of constructing context-specific BoW histograms introduced in [11]. For our participation to the ImageCLEF 2011 photo annotation task, some improvements over this method are proposed. First, we learn a specific vocabulary for each semantic context rather than use a uniform vocabulary for all contexts as in [11]. Second, instead of selecting a single context for each visual word as in [11], we train a classifier for each context-specific BoW histogram and then combine all the classifiers. Detailed implementation of these two improvements are given in the following paragraph.

In the traditional vocabulary learning process, local features extracted from a set of images are randomly (or uniformly) sampled and then vector quantized to get visual words. Differently, when learning our context-specific vocabulary, the sampling of local features is based on the distribution of this context on images. Specifically, more local features are sampled at the image regions with higher context-occurring probabilities (brighter image regions in Fig. 2). In practice, this process is implemented by assigning each local feature f_i a probability $p(c_k|I_p(f_i))$ (defined in section 3.2) and sampling local features based on their probabilities, which is formulated as follows.

$$s(f_i) = \begin{cases} 1 \text{ if } p(c_k|I_p(f_i) \geq r_i \\ 0 \text{ else} \end{cases} \tag{5}$$

where $s(f_i)$ indicates whether the local feature f_i is selected or not and r_i are random numbers which are uniformly sampled between 0 and 1.

After sampling local features for each context, k-means is used to build multiple context-specific vocabularies. An image is then represented by multiple context-specific BoW histograms. The construction of context-specific BoW histogram is the same as that in 3.2 (see Eq.4)

Concatenating all the context-specific BoW histograms leads to a very high dimensional feature vector (in our case 1928×75=144,600D). Thus, we train a classifier for each context-specific BoW histogram and combine the classifiers by averaging their outputs.

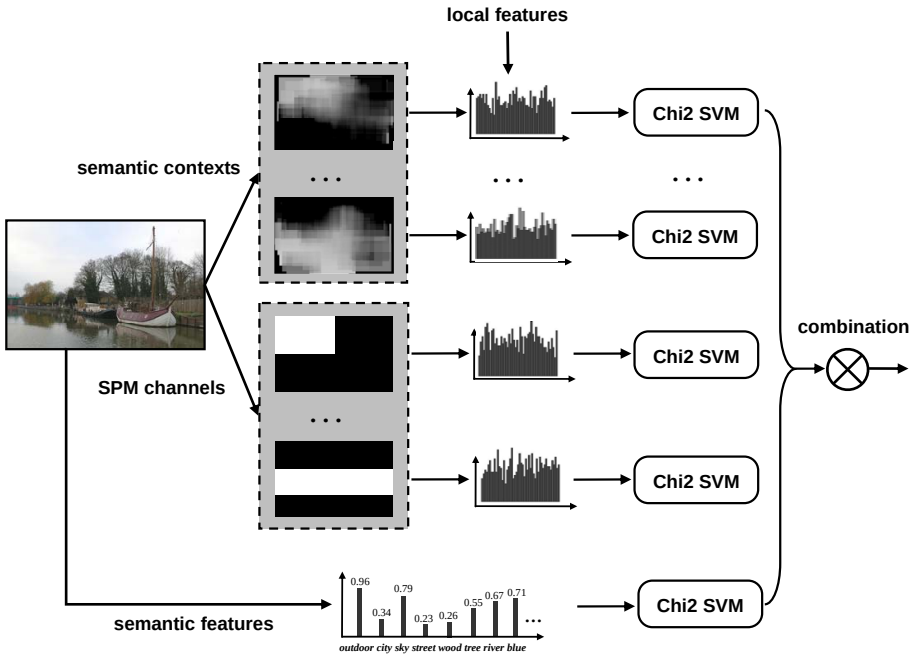

Fig. 2. Combination of BoW model and semantic contexts. For an image, multiple saliency maps are generated by both context classifiers and SPM channels, with which multiple BoW histograms are constructed by weighting local features according to saliency maps. After that, multiple classifiers are learned, each of which corresponds to a BoW histogram. In addition, the occurrence probabilities of semantic contexts (also referred as semantic features) are predicted for the image, for which a classifier is learned. Finally, all the classifiers are combined by averaging their outputs.

Recall that the way we embed contextual information into BoW model is based on weighting local features (see Eq.4). It is similar to the well-known spatial pyramid matching (SPM) [5] which divides an image into grids and build a histogram for each grid. This process can be also considered as weighting local features: for certain grid, the weights of the local features within it is set to 1

and the weights of other local features are set to 0. Although less flexible than context-based weights, the binary weights in SPM are more stable which is also favorable. Thus, we also train classifiers for BoW histograms of SPM channels. In our method, a three level pyramid, 1×1, 2×2, 3×1 (totally 8 channels) is used. It is worthwhile to point out that, different from traditional SPM, we learn a specific vocabulary for each SPM grid based on local features within this grid.

Finally, we train a classifier for the semantic features and combine it with the classifiers for context-specific BoW histograms and SPM channels by averaging their outputs. For both BoW histograms and semantic features, classifiers are learned by SVM with chi-square kernel. The whole process is illustrated in Fig.2.

4 Image Representation by Fisher Vectors

Similar to the BoW model, Fisher Vectors [8] can also be used to aggregate local features into a global descriptor which is called Fisher Vectors (FV). FV can be considered as an extension of BoW histograms. They encodes how the parameters of the model should be changed to represent the image, rather than only consider the number of occurrences of each visual word as in BoW model. In our participation, we adopted the improved FV as introduced in [9] which is shown to outperform BoW histogram on some large-scale image retrieval tasks.

For an image, we computed the FV for each kind of local features except Canny for which no actual visual word exist. As in [9], a three level pyramid, 1×1, 2×2, 3×1 (8 channels in total) is used to enhance the performance of fisher vector.

For SIFT and HOG descriptors, PCA is used to reduce the dimension of descriptors to 64. For SIFT descriptors, a 64-centroid Gaussian mixture model (GMM) is computed to construct fisher vector whose dimensionality is therefore $64 \times 64 \times 8 \times 2 = 65,536$. For HOG, Texton, SSIM and LAB descriptors, 64-centroid GMMs are learned therefore the dimensionalities of fisher vector for these descriptors are 16,384, 9,216, 7,680 and 768 respectively. Then we train a classifier (SVM with linear kernel) for each fisher vector and combine all classifiers by averaging their outputs. The whole process is illustrated in Fig.3. Please note that the semantic contexts are not used for this representation.

5 ImageCLEF Evaluation

In our participation, we submitted four runs to the photo annotation task. In this section, we describe these runs and compare their performances with other visual-only runs.

5.1 Description of Our Runs

Run 1: MultiFeat_Chi2SVM In this run, context-specific BoW histograms, each of which corresponds to a semantic context, as well as the semantic features

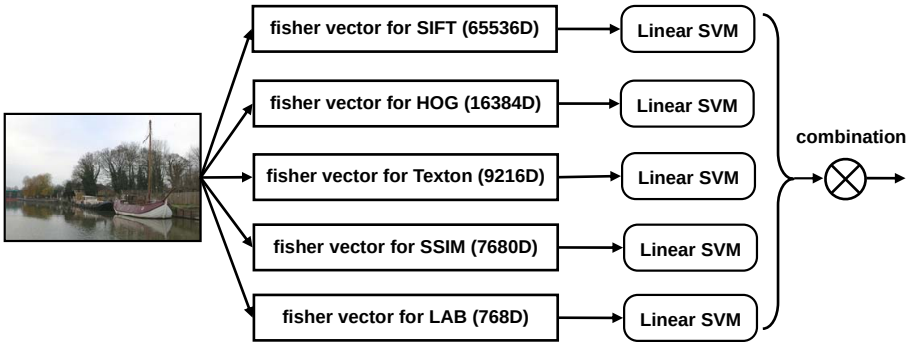

Fig. 3. Image representation and classification based on fisher vectors of multiple types of local features. The values in brackets are the dimensionalities of corresponding Fisher Vectors. After training a classifier for each Fisher Vectors, multiple classifiers are combined by averaging their outputs.

are used to describe images. As illustrated in Fig.2, we trained separated classifiers (SVMs with chi-square kernel) for both context-specific BoW histograms and semantic features and then combine them by averaging their outputs.

Run 2: BoW+FisherKernel In this run, we combined all the classifiers in run 1 and classifiers for fisher vectors of different features (refer to Fig.3). The combination is performed by averaging the outputs of all classifiers.

Run 3: SVMOutput In this run, the confidences of all 99 concepts obtained from run 2 are used as a new image descriptor. A classifier (SVM with chi-square kernel) is learned on this descriptor and used to give the confidences of concepts. By doing so, we hope to benefit from the correlation of different concepts.

Run 4: BoW+FisherKernel+SVMOuput In this run, we averaged the confidences obtained from run 1, 2 and 3.

In our participation, we used the implementation of LIBSVM [2] to learn SVM classifier. The value of the SVM parameter C and the normalization factor γ of chi-square kernel are determined by fivefold cross-validation. As to the image regions used for learning local context classifiers and generating saliency maps, on each image we sampled 100 regions with random positions and scales (with scales from 20% to 40% of the image size).

For concept level evaluation, the classifier outputs are used as confidences directly. For image level evaluation, the real valued confidences are binarized by a threshold which is determined by fivefold cross-validation for each run.

5.2 Results of Our Runs

The performances (MAP, F-ex and SR-Precision) of our runs are listed in Table 1. It can be concluded that the performance of context-specific BoW histograms is significantly enhanced by combining them with fisher vectors. It is worthwhile to point out that, according to our experiments on training data, the performance of fisher vectors alone is comparable to that of context-specific BoW histograms. Another conclusion drawn from Table 1 is that using classifier outputs as new features does not bring any improvement. Thus we need to design more powerful methods to utilize the correlation of concepts.

Runs	MAP	F-ex	SR-Precision
MultiFeat_Chi2SVM	34.2	56.0	69.4
BoW+FisherKernel	**38.2**	**60.0**	**72.7**
SVMOutput	34.5	49.1	65.1
BoW+FisherKernel+SVMOuput	**38.2**	59.2	72.5

Table 1. MAP, F-ex and SR-Precision of our runs. For each measure, values in bold indicate the best performance of 4 runs.

For more detailed result, Fig.4 gives the MAPs of 99 concepts obtained from Run 2. For some concepts, the MAPs are very low, e.g. less than 10%. The reason is either that the concept is hard to predict (e.g. *abstract*) or that the number of training samples is quite small (e.g. only 12 images are annotated with *skateboard*).

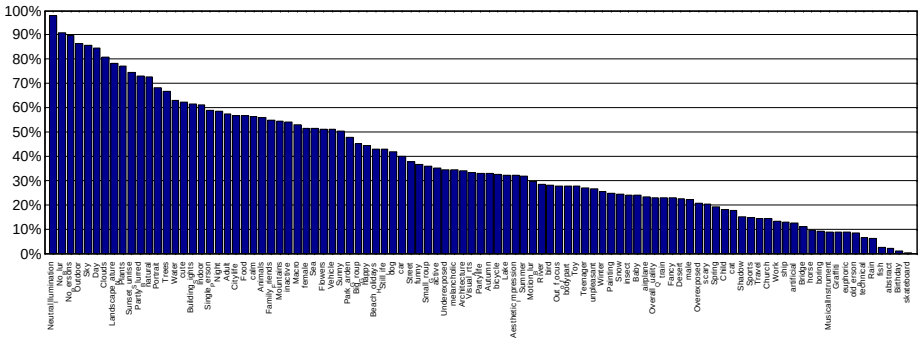

Fig. 4. The MAPs of all 99 concepts obtained from Run 2.

Finally, we compare our best run (Run 2: BoW+FisherKernel) with the best runs (visual-only) of several competitors. It can be seen from Table 2 that no run gave the best result for both concept and image level evaluation. For concept level

evaluation (MAP as performance measure), TUBFI_scores performed best, while for image level evaluation (F-ex and SR-Precision as performance measures), ISIS_runpa-UvA-coreA performed best. Our best run ranks in the second place for both MAP and F-ex and the third place for SR-Precision.

Runs	MAP	F-ex	SR-Precision
BPACAD_bpacad_avg_cns	36.7	56.8	72.9
ISIS_runpa-UvA-coreA	37.5	**61.2**	**73.4**
LIRIS_4visual_model_4	35.5	53.9	72.5
TUBFI_scores	**38.8**	55.2	62.1
Our best run	38.2	59.2	72.5

Table 2. MAP, F-ex and SR-Precision of our runs. For each measure, values in bold indicate the best performance of 4 runs. Our best run ranks in the second place for both MAP and F-ex and the third place for SR-Precision.

6 Conclusion

In our participation to the ImageCLEF photo annotation task, multiple visual features were used for representing the images. We embedded contextual information into the traditional Bag-of-Words model and further combined it with fisher vector which has been shown to have good performance on image classification and retrieval tasks. The evaluation results showed that the performance of the Bag-of-Words model can be significantly enhanced by combining it with semantic contexts and fisher vector. Our best run gave 38.2%, 59.2% and 72.5% for MAP and F-ex and SR-Precision respectively, while the best results of visual-only runs are 38.8%, 61.2% and 73.4% respectively.

Acknowledgement

This work was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

References

1. Canny, J.: A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6), 679–698 (1986)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)

4. Hunter, R.: Photoelectric color difference meter. JOSA 48(12), 985–993 (1958)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
7. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. CLEF 2011 working notes (2011)
8. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2006)
9. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification (2010)
10. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR (2007)
11. Su, Y., Jurie, F.: Visual word disambiguation by semantic contexts. In: ICCV (2011)
12. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision 62(1), 61–81 (2005)
13. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal on Computer Vision 72(2), 133–157 (2007)