Visual Word Disambiguation by Semantic Contexts

Yu Su
GREYC, University of Caen, France
yu.su@unicaen.fr

Frédéric Jurie
GREYC, University of Caen, France
frederic.jurie@unicaen.fr

Abstract

This paper presents a novel schema to address the polysemy of visual words in the widely used bag-of-words model. As a visual word may have multiple meanings, we show it is possible to use semantic contexts to disambiguate these meanings and therefore improve the performance of bag-of-words model. On one hand, for an image, multiple context-specific bag-of-words histograms are constructed, each of which corresponds to a semantic context. Then these histograms are merged by selecting only the most discriminative context for each visual word, resulting in a compact image representation. On the other hand, an image is represented by the occurrence probabilities of semantic contexts. Finally, when classifying an image, two image representations are combined at decision level to utilize the complementary information embedded in them. Experiments on three challenging image databases (PASCAL VOC 2007, Scene-15 and MSRCv2) show that our method significantly outperforms state-of-the-art classification methods.

1. Introduction

Image classification, including object and scene classification, is a central area in computer vision research. Among the recent advances made on this topic, perhaps the most significant one is representing images by the statistics of local features, in particular the introduction of the bag-of-words (BoW) model [22] in which local features extracted from an image are first mapped to a set of visual words. An image is then represented as a histogram of visual word occurrences. Combined with powerful classifiers such as the Support Vector Machine, the BoW model has demonstrated impressive performances on several challenging image classification tasks [4, 8, 30].

Words in natural languages are frequently polysemous. One usual example is *crane*, meaning either a bird or a construction equipment according to the context of use. So, in the literature of natural language processing, lots of efforts were made to disambiguate words based on their contexts (e.g., [16, 33]). Polysemy is also critical for visual words:

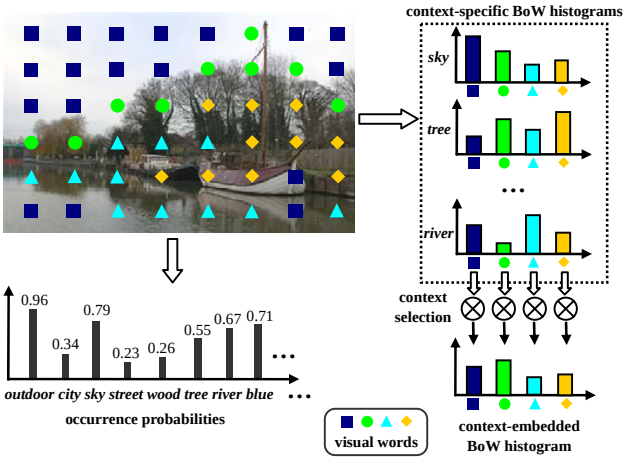


Figure 1. Brief overview of our method. See text for more details.

as only local information is encoded, the same visual word could be used to construct different types of objects. As a simple example, we could easily imagine that the same image structure, e.g., a 'window' like visual word, could be interpreted as a 'car window' or a 'plane window' depending on the average color of the local background. Surprisingly, the disambiguation of visual word has been studied only marginally.

The role of context in natural language motivates us to put a special emphasis on disambiguating visual words by the contextual information extracted from images. Although recent literature on utilizing context is abundant [7, 10, 12, 20, 32], when a high-performance image classification system is required in practice, people almost always use the basic BoW model or its variants [5]. In other words, the use of context remains an open problem. In this paper, we show that the contextual information can be used to significantly boost the performance of BoW model.

The main idea of our method is illustrated in Fig. 1. For an image, we first construct multiple BoW histograms, each of which corresponds to a context. That means the same visual word would have different occurrence frequencies when different contexts are considered. For example, in Fig. 1, the occurrence frequency of the visual word denoted by 'square'

is higher in context *sky* than in *tree*, because this visual word often appears in sky areas. By embedding contextual information, the visual words in each single histogram are less ambiguous. Considering the huge dimensionality if these context-specific histograms were all used, we propose a dimensionality reduction method by selecting only the most discriminative context for each visual word. The resultant histogram is called as *context-embedded BoW histogram* which has the same dimensionality as the standard BoW histogram. This is the key contribution of our paper.

Furthermore, we show that the occurrence probabilities of contexts (see Fig. 1), also provide useful information to describe images. Finally, when classifying images, both image representations (context-embedded BoW histogram and occurrence probabilities of contexts) are combined at decision level to take advantage of the complementary information embedded in them.

2. Related works

Bag-of-Words model. Numerous works have recently demonstrated the effectiveness of BoW model on image classification tasks. We focus here on those related to visual word disambiguation.

To deal with synonymy and polysemy, one choice is eliminating the most and least frequent words which are supposed to be the most ambiguous [22]. Another choice is to utilize task-specific information to obtain less ambiguous vocabulary [18]. In addition, the ambiguity of visual words can be reduced by considering their co-occurrences [34].

The hard assignment used in the standard BoW model leads to large loss of information if some visual words have close representations. To address this problem, soft assignment in which a local feature is assigned to different number (including zero) of visual words was proposed [26] and can also help to address the synonymy.

Polysemy of visual words is partly caused by the discard of spatial information. Hence, the use of spatial information can also help to disambiguate visual words. A typical example is the well-known spatial pyramid matching [15].

Topic model, such as *Probabilistic Latent Semantic Analysis (pLSA)* [11], also has the effect to address polysemy [21]. For example, both *bird* and *equipment* topics can give high probability to the word *crane*, but the occurrence probabilities of different topics reduce this uncertainty. In contrast to topic model, our method uses semantic contexts rather than topics learnt from data collection. Please refer to section 3.2 for more details.

In another related work [13], Khan *et al.* proposed to use some category-specific color attention maps to weight local shape features and then concatenate multiple histograms. Our method also uses the idea of weighting local features. However, we adopt semantic contexts (rather than color) to generate attention maps and reserve only the most discriminative context for each visual word (rather than concatenation).

Context. Contextual information is often extracted by modeling interactions between pixels, regions and objects. Conditional random field [10, 20] and co-occurrence [7, 12] are two commonly used modeling methods.

In contrast, our method does not model interactions but adopts different local contexts to enrich the representation of whole image. Similar idea is also presented in [3, 25], in which images or videos are first decomposed into regions and then multiple region-specific BoW histograms are computed and combined. The differences between our method and them are twofold. First, in our method, BoW histograms are context-specific rather than region-specific. Second, our method compresses multiple histograms rather than computing multiple kernels for them [3] or concatenating them [25], therefore resulting in a more compact image representation.

Semantic attributes. The recent literature abounds in approaches making interesting use of semantic concepts and giving proofs-of-concept. Farhadi *et al.* [6] used a set of semantic attributes such as 'hairy' and 'four-legged' to identity familiar objects, and to describe unfamiliar objects when images and bounding box annotations are provided. Lampert *et al.* [14] showed that high-level descriptions in terms of semantic attributes can be used to recognize object classes without any training image, once semantic attribute classifiers are trained from other classes of data.

In addition to describing objects semantically, there also exist some methods which aim to describe the whole image by semantic features. Vogel and Schiele [27] used attributes describing scene to characterize image regions and combined these local semantics into a global image description for natural scene retrieval. Wang *et al.* [28] proposed to represent an image by its similarities to Flickr image groups which have explicit semantic meanings. Li *et al.* [17] built a semantically meaningful image hierarchy by using both visual and semantic information, and represent images by the estimated distributions of concepts over the entire hierarchy. Torresani *et al.* [24] used the outputs of a large number of object category classifiers to represent images.

Our approach bears similarity with [24] and [28], as we also use semantic classifiers to describe images. But different from them, we propose to use the semantic features to disambiguate the visual words in BoW framework and show it outperforms the existing approaches.

3. Approach

In this section, we first explain how to define, learn and predict semantic contexts from training images, and then explain how we describe test images with them.

3.1. Semantic contexts

Following the procedure given in [23], we define 110 semantic contexts by hand with the intention of providing abundant semantic information for image description. (see Fig. 2). Two types of contexts are distinguished: *global* contexts including the contexts of global scene and *local* contexts including the contexts of local scene, color, shape, material and object.

For each semantic context, we learn a classifier by SVM 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. As to the training images, there are two cases. For the semantic contexts that appear in PASCAL 2007 (20 objects e.g., *motorbike*) and Scene-15 databases (15 global scenes e.g., *bedroom*), the training images as well as the annotations are directly obtained from the databases. For other semantic contexts, 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. The context classifiers as well as the training images are publicly available at http://users.info.unicaen.fr/~ysu/semantic.

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

3.2. Context-embedded image representation

In this subsection, we first formulate the process of embedding contexts into BoW model, and then elaborate how to construct the context-embedded image representation by using the previously learned context classifiers.

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).

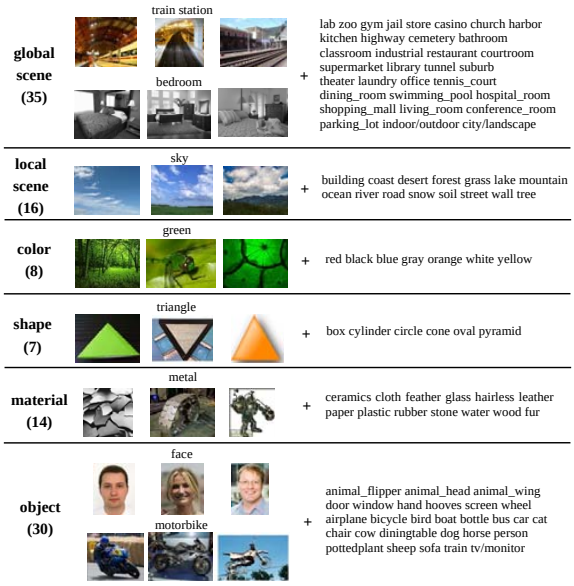


Figure 2. Grouped semantic contexts and some illustrative training images. The values in parentheses are the number of semantic contexts within corresponding groups. In this paper, the contexts of global scene are refereed as *global* contexts, while the context of local scene, color, shape, material and object are referred as *local* contexts.

As mentioned in section 1, a visual word could have different meanings in different contexts. Marginalizing $p(v_j|I)$ over different 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 contexts, $p(v_j|c_k, I)$ 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.

Eq. 3 bears similarities to that in *Probabilistic Latent Semantic Analysis (pLSA)* [11]. But different from pLSA, we do not assume the conditional independence that conditioned on the context c_i visual words v_i are generated independently from the specific image I, i.e., $p(v_j|c_k, I) \neq p(v_j|c_k)$. Instead, we believe that the words generated by a given context constitute some characteristic signatures of the image. As an illustration, if for a particular image, *window* like visual word occurs simultaneously with the *blue* context, it could be a good cue for hypothesizing the presence of a plane in the image. Another difference from pLSA is that we do not consider contexts as latent variables, which we believe would be hard to estimate, but define them offline and predict them for every image by context classifiers (see previous section).

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

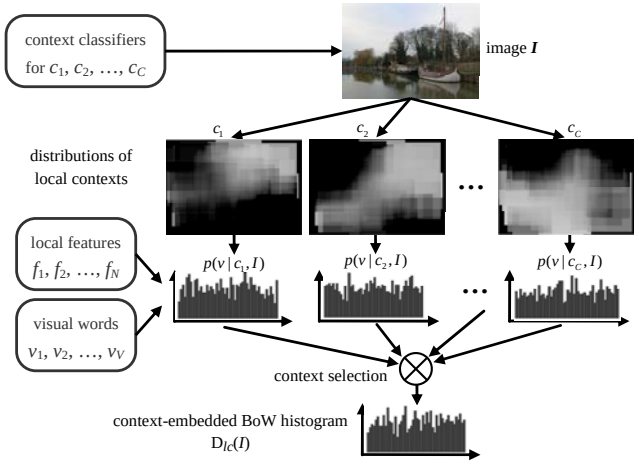


Figure 3. Construction of context-embedded BoW histogram. For an image, probabilistic distributions for *local* contexts are generated by the corresponding context classifiers. Then, a BoW histogram is constructed for each context by weighting local features according to its distribution. Finally, for a specific classification task, a context selection process is used to choose the most discriminative context for each visual word.

[27]. 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 a context-embedded BoW histogram, i.e., $p(v_j|c_k, I)$ and a vector of context-occurring probabilities, i.e., $p(c_k|I)$, which are then combined at decision level (see section 3.3).

3.2.1 Context-embedded BoW histogram

In this work, $p(v_j|c_k, I)$ 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:

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

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 $I_p(f_i)$.

Keeping $p(v_j|c_k, I)$ for all visual words and all contexts would lead to a $V \times C$-dimensional descriptor. In this work C is 75 since only *local* contexts are used to construct $p(v_j|c_k, I)$ and V is usually from hundreds to thousands. If we use this $V \times C$-dimensional descriptor to train a classifier, the number of parameters to be learned would be too large with respect to the number of training images, producing a high risk of over-fitting. Our intuition is that, for a given classification task, a visual word usually appears in a limited set of contexts rather than all contexts. For example, as in Fig. 1, the visual word denoted by 'square' almost only appears in the context *sky* and *river*. In practice, we show in section 4 that using only one context per visual word already gives very good results. By doing that, for a given classification task, an image is finally represented by

$$D_{lc}(I) = (p(v_1|c_{k_1}, I), \ldots, p(v_j|c_{k_j}, I), \ldots, p(v_V|c_{k_V}, I)),$$

where c_{k_j} is the selected context for visual word v_j and the given classification task. We call this representation as context-embedded BoW histogram which has the same dimensionality as the standard BoW histogram. The whole process described above is illustrated in Fig. 3.

Up to now, the only remaining problem is how to choose context for each visual word. This is a feature selection problem and in theory any criterion can be used for that, e.g. max-likelihood. Although more consistent with the proposed probabilistic framework, the max-likelihood criterion does not allow to use category labels of images and therefore performs worse than some supervised ones in practice. In this work, we adopt a supervised *t-test* based criterion for context selection. Specifically, for each visual word v_j and each context c_k, we assume that the value of $p(v_j|c_k, I)$ follows the Gaussian distribution $\mathcal{N}(\mu^+_{j,k}, \sigma^+_{j,k})$ on positive images and $\mathcal{N}(\mu^-_{j,k}, \sigma^-_{j,k})$ on negative images. For a given visual word, we compute the *t-test* statistic between these two distributions for every possible context and take the context giving the highest value. It therefore selects the context for which the representation of positive images is as different as possible from that of negative images, i.e., the most discriminative context. As this context selection process is supervised, the selected contexts depend on the classification task to be addressed. That is to say, the selected contexts for *aeroplane* classification and *person* classification will be very different.

3.2.2 Context-occurring probability

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.,

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

where C' is the number of all contexts (110 in our case) and C is the number of *local* contexts (75 in our case).

3.3. Combination of both representations

Up to now, we have constructed two image representations, i.e., $D_{lc}(I)$ and $D_{gc}(I)$, which encode local and global contextual information respectively. After that, $D_{lc}(I)$ and $D_{gc}(I)$ are combined at the decision level. Specifically, we train classifiers on $D_{lc}(I)$ and $D_{gc}(I)$ separately. When classifying an image, the outputs of two classifiers are combined by a linear combination model. The optimal weights are learned on a validation set.

4. Experiments

4.1. Experimental setup

Local features. Four types of local features, such as described in [6], are used in our experiments: SIFT, Texton filterbank (36 Gabor filters at different scales and orientations), LAB and Canny edge detection. Specifically, SIFT features are computed for 2000 image patches with randomly selected positions and scales (with scales from 16 to 64 pixels), and are quantized to 1024 k-means centers. Texton and LAB features are computed for each pixel, and quantized to 256 and 128 k-means centers respectively, while Canny edge features are quantized to 8 orientation bins. Combining these features gives a 1416-dimensional BoW feature vector.

Context classifiers. The context classifiers are learned by SVM with linear kernel (here we use the implementation of LIBSVM [2]), the inputs to which are BoW feature vectors constructed by pooling local features within image regions (for region-level classifiers) or whole images (for image-level classifiers). The SVM parameter C is set to 10, which is determined by fivefold cross-validation. As to the image regions, on each training image we sampled 100 regions with random positions and scales (with scales from 20% to 40% of the image size).

Databases. Three publicly available image databases are used for evaluation: PASCAL VOC 2007 [4], Scene-15 [15] and MSRCv2 [29].

PASCAL VOC 2007 is the last challenge for which the test data annotations are publicly available. The data set contains 9963 images of 20 object classes which were collected from users uploads to the Flickr website. For the challenge's classification task, the goal is to determine whether or not each test image contains at least one instance of each object class of interest. Performance is measured by calculating the average precision (AP) for each class, and the mean average precision over the 20 categories (mAP), following the protocols given in [4].

Scene-15 database contains 15 scene categories, each of which has 200 to 400 gray-level images. These images come from the COREL collection, personal photographs, and Google image search. Following the experimental setup used in [15], 100 images per category are randomly sampled as training samples (remaining as testing samples). One-versus-all strategy is used for multiclass classification and the performance is reported as the average classification rate on 15 categories.

MSRCv2 is an object category database. We follow the experimental setup used in [35] which chose 9 categories out of 15: cow, airplane, face, car, bike, book, sign, sheep and chair in order to make objects from different categories do not appear in the same image. In experiments, 15 training images and 15 testing images are randomly sampled for each category. One-versus-all strategy is used for multiclass classification and the performance is reported as the average classification rate on 9 categories.

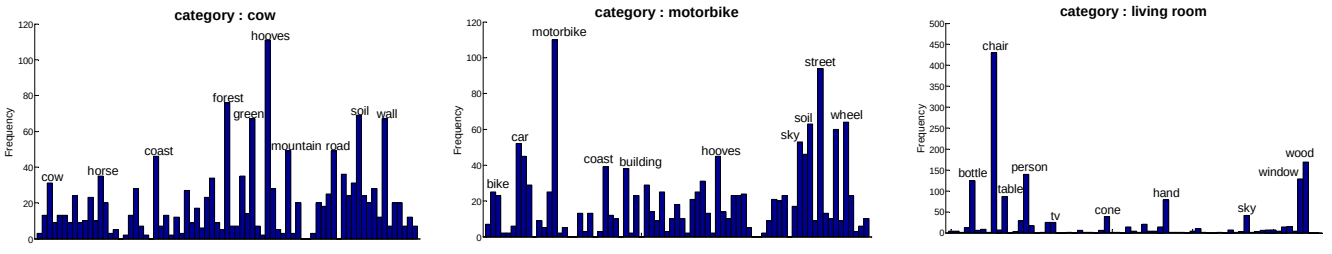
Image classification. For each category, two SVM classifiers with chi-square kernel is learned for D_{lc} and D_{gc} respectively. The value of SVM parameter C and the normalization factor γ of chi-square kernel are determined by fivefold cross-validation. The optimal weights for classifier combination is learned on the validation set of PASCAL 2007 database and adopted directly for Scene-15 and MSRCv2 databases.

To enhance the performance of BoW histogram and D_{lc}, we additionally use spatial pyramid matching (SPM), as proposed in [15]. Using a three level pyramid, 1×1, 2×2, 3×1 (totally 8 channels), gives final image representation with a dimensionality 8×1416.

4.2. Qualitative results

In this subsection, we give some examples to show the effect of context selection. As explained in section 3.2, we choose only a single context for each visual word, depending on the category to be classified. Hence, for each category, we can count the frequency that each context is selected, and higher frequency means higher importance for classifying this category. By doing so, multiple category-specific frequency histograms can be generated. Fig. 4 gives the frequency histogram for category *cow*, *motorbike* and *living room*. It can be seen that the contexts which are related to the category to be classified tend to have high relative importance (frequency). Take Fig. 4(b) as example, besides *motorbike*, the context *street* and *wheel* also play an important role in *motorbike* classification.

As explained before, the context selection depends on the classification task to be addressed. It means an image will be described differently in different classification tasks. For example, in Fig. 5, for *motorbike* classification, the two most important contexts are *motorbike* and *street*. This choice can be easily explained. For *person* classification, the contexts *black* and *sky* dominate the image description. These two contexts seem to have no relation with *person*, whereas one possible explanation is that in daily life people often wears dark and blue clothes.



(a) (b) (c)

Figure 4. Selection frequencies of different contexts for three categories: *cow*, *motorbike* and *living room*. The contexts with high frequency are marked by their names.

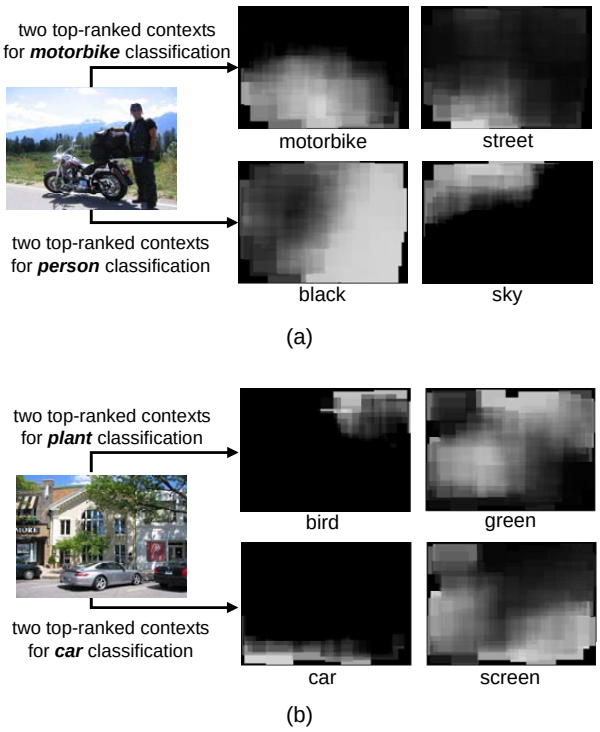


Figure 5. Saliency maps of the two top-ranked contexts for different classification tasks. The value of each pixel on the saliency map is computed by averaging the outputs of corresponding context classifier on the image regions covering this pixel.

4.3. Comparison with related methods

In this subsection, we first compare our methods with the standard BoW model. Table. 1 summarizes the performances of BoW model, context-embedded BoW histogram (D_{lc}), context-occurring probability (D_{gc}) and their combination ($D_{lc} + D_{gc}$) on three databases. By embedding local contexts, the performance of BoW model is improved by 2.8% on PASCAL 2007, 2.1% on Scene-15 and 2.3% on MSRCv2. Although D_{gc} (only 110-dimension) does not give better performance than BoW model, combining it with D_{lc} leads to additional improvement, demonstrating that they are complementary with each other. Finally, the improvement of our method ($D_{lc} + D_{gc}$) to BoW model is 5.3% on PASCAL 2007, 4.5% on Scene-15 and 4.5% on MSRCv2.

For more detailed comparison, Fig. 6 gives the performance improvement for each category in PASCAL 2007 database. It can be seen that D_{lc} performs better than BoW model on 18 of 20 categories (except for *bus* and *cat*), whereas $D_{lc} + D_{gc}$ performs better than BoW model on all categories. In particular, for category *pottedplant*, the improvement of average precision is more than 10%. We believe the reason of this large improvement is that pottedplants are very diverse in appearance and usually in small scales therefore their classification mainly depends on the contextual information.

In [1], images are represented by the mixing coefficients of topics which are learned from visual words via pLSA. This representation is similar to the proposed context-occurring probability (D_{gc}). Thus, we re-implemented the method in [1] and compare it with D_{gc}. To be fair, the number of topics is set to the dimensionality of D_{gc}. The performances of this pLSA-based method are 52.8% on PASCAL 2007, 77.0% on Scene-15 and 78.3% on MSRCv2 respectively, which are worse than those of D_{gc} (refer to Table. 1). In addition to pLSA, we compare our method with another attribute-based methods [28]. In [28], an image is represented by a descriptor of 103 dimensions, each of which corresponds to the similarity of this image to a Flickr image group. Although the dimensionality is a little higher, D_{gc} gives much better performance (55.1%) on PASCAL 2007 than this 103-D similarity-based descriptor (44.9%, cited directly from [28]).

4.4. Influence of local context regions

In the computation of D_{lc}, the number of randomly sampled image regions (i.e., the size of I_p) is a key parameter. Hence, we do several experiments on PASCAL 2007 to evaluate the effect of region number as well as their locations (random sampling vs. regular grid) using only D_{lc}. From

	PASCAL 2007	Scene-15	MSRCv2
BoW + SPM	59.2	83.3 ± 0.7	86.2 ± 2.3
D_{lc}	62.0	85.4 ± 0.5	88.5 ± 2.4
D_{gc}	55.1	79.1 ± 0.9	82.8 ± 2.8
$D_{lc} + D_{gc}$	**64.5**	$\mathbf{87.8 \pm 0.5}$	$\mathbf{90.7 \pm 1.8}$
result from dataset creator	59.4 [4]	81.4 [15]	80.4 ± 2.5 [35]

Table 1. Performance comparison with the standard BoW+SPM model.

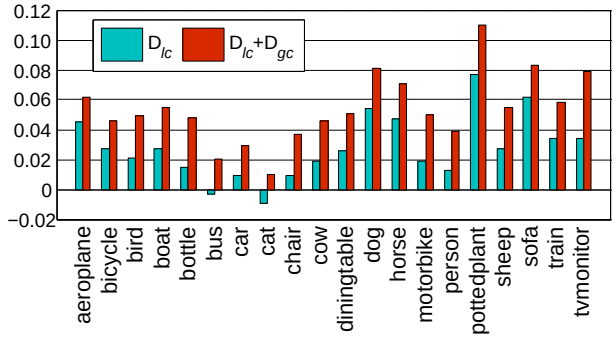


Figure 6. Performance (mAP) improvement of our methods to the standard BoW+SPM model on PASCAL 2007 database.

these experiments we conclude that sampling regions on a regular grid does not give better results than sampling them randomly. However, random sampling raises questions about the stability of results and how many regions to use. If we sample 10, 50 or 100 regions per image, the mAP are respectively 60.8%, 61.5% and 62.0%. Taking more than 100 regions does not improve the results significantly. Regarding stability, the standard deviations observed over 5 runs, if we sample 10, 50 or 100 regions per image, are respectively 0.5%, 0.3% and 0.2%. Hence, if 100 regions are randomly sampled, the choice for these regions does not have a great effect on the performance of D_{lc}.

4.5. Influence of dimension reduction

As mentioned in section 3.2, we rank contexts for each visual word and select only the most discriminative one, resulting in the V-dimensional descriptor D_{lc}. Although it is also possible to reserve more contexts (e.g., top 2, 3 or 5) for each visual word with the cost of higher dimensionality of D_{lc}, Fig. 7 shows that it does not result in significant performance improvement (at most 0.2%). Instead of context selection, we can use other dimensionality reduction methods, such as *Principal Component Analysis (PCA)* or *Linear Discriminant Analysis (LDA)*, to obtain a low dimensional image descriptor. To validate their effects, we use PCA and LDA to project the C-dimensional descriptor $(p(v|c_1, I), p(v|c_2, I), ..., p(v|c_C, I))$ for each visual word into a low dimensional subspace. Fig. 7 gives the performance of PCA (up to 5-D) and LDA (only 1-D due to the binary classification task on PASCAL 2007 database), which are worse than that of context selection. In sum, selecting only one context for each visual word gives the best tradeoff between performance and dimensionality.

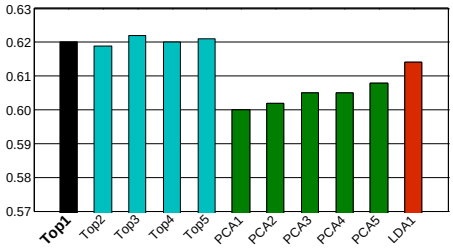


Figure 7. Performance (mAP) comparison of different dimension reduction methods on PASCAL 2007 database. Top$N$ means that the top-ranked N contexts are reserved. The numbers after PCA and LDA denote the dimensions of subspace.

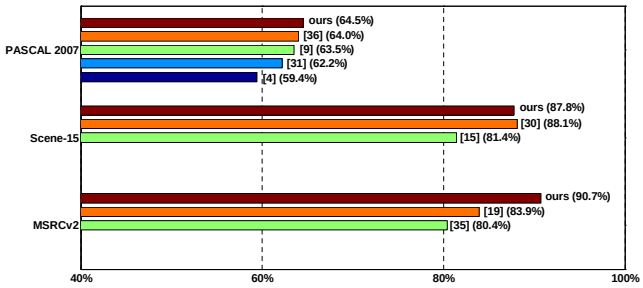


Figure 8. Comparison with the state-of-the-art results on PASCAL 2007 [4, 9, 31, 36], Scene-15 [15, 30] and MSRCv2 [19, 35].

4.6. Comparison with state-of-the-art results

The results of our method on PASCAL 2007, Scene-15 and MSRCv2 databases are 64.5%, 87.8% and 90.7% respectively (refer to Table. 1), which are comparable to or better than the state-of-of-art results on these databases. Please see Fig. 8 for details.

5. Conclusion and discussion

In this paper, we presented a novel method to disambiguate visual words with the help of local and global semantic contexts. Extensive experimental results demonstrated that, by embedding contextual information, our method improves the performance of the standard bag-of-words model by a large margin, say 5.3% on PASCAL VOC 2007, 4.5% on Scene-15 and 4.5% on MSRCv2. Furthermore, our method achieves comparable or better performances compared with the recent state-of-the-art approaches on these challenging image classification tasks.

Finally, it is worthwhile to discuss the practicality of our method. Indeed, it takes some time to collect images and train classifiers for all the semantic contexts. However, this is an offline training phase and the context classifiers are generic therefore they can be used in any image classification task. In the testing phase, since the context classifiers are linear SVMs, the construction of the probabilistic distri-

bution of contexts is quite efficient. Thus, the computation time of context-embedded BoW histogram is comparable to that of traditional bag-of-words histogram.

6. Acknowledgement

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

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV*, 2006. 6

[2] C.-C. Chang and C.-J. Lin. 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/~cjlin/libsvm. 3, 5

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 2

[4] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/. 1, 5, 7

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/. 1

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2, 5

[7] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 1, 2

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 1

[9] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009. 7

[10] X. He, R. Zemel, and A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *CVPR*, 2004. 1, 2

[11] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, 1999. 2, 3

[12] S. Ito and S. Kubota. Object classification using heterogeneous co-occurrence features. In *ECCV*, 2010. 1, 2

[13] F. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *ICCV*, 2009. 2

[14] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 5, 7

[16] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998. 1

[17] L.-J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *CVPR*, 2010. 2

[18] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, 2007. 2

[19] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV*, 2010. 7

[20] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1, 2

[21] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 2

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1, 2

[23] Y. Su, M. Allan, and F. Jurie. Improving object classification using semantic attributes. In *BMVC*, 2010. 3

[24] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 2

[25] M. Ullah, S. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010. 2

[26] J. van Gemert, C. Veenman, A. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010. 2

[27] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007. 2, 3

[28] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *ICCV*, 2009. 2, 6

[29] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005. 5

[30] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 7

[31] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *ICCV*, 2009. 7

[32] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1

[33] D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, 1992. 1

[34] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007. 2

[35] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009. 5, 7

[36] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010. 7