

Contextual-Guided Bag-of-Visual-Words Model for Multi-class Object Categorization

Mehdi Mirza-Mohammadi¹, Sergio Escalera^{1,2}, and Petia Radeva^{1,2}

¹ Dept. Matemàtica Aplicada i Anàlisi, Gran Via 585, 08007, Barcelona, Spain

² Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona
me@memimo.net, {sergio,petia}@maia.ub.es

Abstract. Bag-of-words model (BOW) is inspired by the text classification problem, where a document is represented by an unsorted set of contained words. Analogously, in the object categorization problem, an image is represented by an unsorted set of discrete visual words (BOVW). In these models, relations among visual words are performed after dictionary construction. However, close object regions can have far descriptions in the feature space, being grouped as different visual words. In this paper, we present a method for considering geometrical information of visual words in the dictionary construction step. Object interest regions are obtained by means of the Harris-Affine detector and then described using the SIFT descriptor. Afterward, a contextual-space and a feature-space are defined, and a merging process is used to fuse feature words based on their proximity in the contextual-space. Moreover, we use the Error Correcting Output Codes framework to learn the new dictionary in order to perform multi-class classification. Results show significant classification improvements when spatial information is taken into account in the dictionary construction step.

1 Introduction

Multi-class object categorization is one of the most challenging problems in Computer Vision, which has been applied to a wide variety of applications. Usually, the problem of object categorization is split into two main stages: object description, where discriminative features are extracted from the object to represent, and object classification, where a set of extracted features are labeled as a particular object given the output of a trained classifier.

A general tendency in object recognition to deal with the object description stage is to define a bottom-up procedure where initial features are obtained by means of region detection techniques. These techniques are based on determining relevant image keypoints (i.e. using edge-based information [1]), and then to define a support region around the keypoint (i.e. looking for extrema over scale-space [1]). Several alternatives for region detection have been proposed in the literature [1]. Once a set of regions are defined, they should be described using some kind of descriptor (i.e. SIFT descriptor [2]), and the region-descriptions are related in some way to define a model of the object of interest.

Based on the previous tendency, a recent technique to model visual objects is by means of a bag-of-visual-words. The BOVW model is inspired by the text classification problem using a bag-of-words, where a document is represented by an unsorted set of contained words. Analogously, in object categorization problems, an image is represented by an unsorted set of discrete visual words, which are obtained by the object local descriptions.

Many promising results have been achieved with the BOVW systems in natural language processing, texture recognition, Hierarchical Bayesian models for documents, object classification [3], object retrieval problems [4,5], or natural scene categorization [6], just to mention a few. However, one of the main drawbacks of the BOVW model is that dictionary construction does not take into account the geometrical information among visual instances. Although this issue can be beneficial in natural language analysis, its adaptation to visual word description needs special attention. Note that based on the description strategy used to describe visual words, very close regions can have far descriptors in the feature space, being grouped as different visual words. This effect occurs for most of the state-of-the-art descriptors, even when coping with different invariance, and thus, a grouping based on spatial information of regions could be beneficial for the construction of the visual dictionary.

In this paper, we present a method for considering spatial information of visual words in the dictionary construction step, namely Contextual-Guided Bag-of-Visual-Words model (C-BOVW). Object's interest regions are obtained by means of the Harris-Affine detector and then described using the SIFT descriptor. Afterward, a contextual-space and a feature-space are defined. The first space codifies the contextual properties of regions meanwhile the second space contains the region descriptions. A merging process is then used to fuse feature words based on their proximity in the contextual-space. Moreover, the new dictionary is learned using the Error Correcting Output Codes (ECOC) framework [7] in order to perform multi-class object categorization. We compared our approach to the standard BOVW design and validated over public multi-class categorization data sets, considering different state-of-the-art classifiers in the ECOC multi-classification procedure. Results show significant classification improvements when spatial information is taken into account in the dictionary construction step.

The rest of the paper is organized as follows: section 2 describes the C-BOVW algorithm. Section 3 introduces the multi-class ECOC strategy used to learn the BOVW and C-BOVW dictionaries. Section 4 shows the experimental evaluation, and finally, section 5 concludes the paper.

2 Contextual-Guided Bag-of-Visual-Words

In this section we reformulate the BOVW model so that geometrical information can be taken into account in conjunction with the keypoint descriptions in the dictionary construction step.

The algorithm is split into four main stages: contextual and feature space definition, merging, representant computation, and sentence construction.

Space definition: Given a set of samples for a n -multi-class problem, a set of regions of interest are computed and described for each sample in the training set. Then, K -means is applied over the descriptions and the spatial locations of each region to obtain a K -cluster feature-space and a K -cluster contextual-space, respectively. In our case, we use the Harris-affine region detector and SIFT descriptor. The x and y coordinates of each region normalized by the height and width of the image in conjunction with the ellipse parameters that define the region are considered to design the contextual-space.

Merging: Let define a contextual-feature relational matrix M , where the position (i, j) of this matrix represents the percentage of points from the j th visual word of the feature-space that match with the points of the i th visual word of the contextual-space. Then, from each row of M , the two maximums are selected. These maximums correspond to the two words of the feature-space which share more percentage of elements for a same contextual word. In order to fuse relevant feature words, we select the contextual word which maximizes the minimums of all pairs of selected maximums. It prevents unbalanced feature words to be merged. Finally, the two feature words with maximum percentage in M for that contextual word (which have not been previously considered together) are labeled to be merged at the end of the procedure, and the process is iterated while an evaluation metric is satisfied or a maximum number of merging iterations is reached. Once the merging loop finishes, the pairs of feature words labeled during the previous strategy are merged and define the new visual words.

Representant computation: When the new C-BOVW dictionary is obtained, a set of representant for each final word is computed. In order to obtain an stratified number of representant related to the word densities, only one representant is assigned to the word with the minimum number of elements. Then, a proportional number of representant is computed for the rest of words by applying k -means and computing the mean vector for each of the word sub-clusters. With the final set of representant feature vectors, a normalized sentence of word occurrences is computed for each sample in the training set, defining its probability density function of C-BOVW visual words. The whole C-BOVW procedure is formally described in Algorithm 1. An example of a two-iteration C-BOVW definition for a motorbike sample is shown in Figure 1. At the top of the figure, the initial spaces are shown. In the second row, the shared elements from the two spaces which maximize the percentage of matches for a given contextual word are shown. The contextual-space just considers the x and y coordinates, and the 128 SIFT feature-space is projected into a two-dimensional feature-space using the two principal components. Note that the feature descriptions for the two considered words are very close in the feature space though they belong to different visual words before merging. On the right of the figure the new merged feature

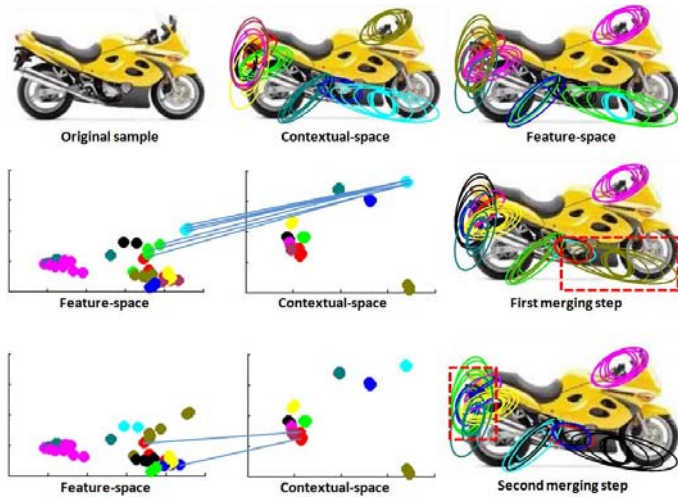


Fig. 1. Two iterations of C-BOVW algorithm over a motorbike sample

cluster is shown within a dashed rectangle. The same procedure is applied for the second iteration of the merging procedure in the bottom row of the figure.

Sentence construction: After the definition of the new dictionary, a new test sample can be simply described using the bag-of-visual-words without the need of including geometrical information since it is implicitly considered in the new visual words. The sentence for the new sample is then computed by matching its descriptors to the visual words with the nearest representant. Finally, the test sentence can be learned and classified using any kind of classification strategy.

3 Multi-class Extension

Error-Correcting Output Codes (ECOC) were defined as a framework to combine binary problems in order to deal with the multi-class case [7]. This framework is based on two main steps. At the first step, namely coding, a set of binary problems (dichotomizers) are defined based on the learning of different sub-partitions of classes by means of a base classifier. Then, each of the partitions is embedded as a column of a coding matrix M . The rows of M correspond to the codewords codifying each class. At the second step, namely decoding, a new data sample that arrives to the system is tested, and a codeword formed as a result of the output of the binary problems is obtained. This test codeword is compared with each class codeword based on a given decoding measure, and a classification prediction is obtained for the new object.

One of the most widely applied ECOC configurations is the one-versus-one design [8]. This strategy codifies the splitting of each possible pair of classes as a dichotomizer, which results in $N(N - 1)/2$ binary problems for an N -class

Algorithm 1. Contextual-Guided Bag-of-Visual-Words algorithm

-
- Require:** $D = \{(\mathbf{x}_1, l_1), \dots, (\mathbf{x}_m, l_m)\}$, where \mathbf{x}_i is an object sample of label $l_i \in [1, \dots, n]$ for a n -class problem, K clusters, and I merging steps.
- Ensure:** Representant $R = \{(r_1, w_1), \dots, (r_v, w_b)\}$, where r_v is a representant for word w_i , $i \in [1, \dots, b]$ for b words. Sentences $S = \{(s_1, l_1), \dots, (s_m, l_m)\}$, where s_i is the sentence of sample \mathbf{x}_i .
- 1: **for** each sample $\mathbf{x}_i \in D$ **do**
 - 2: Detect regions of interest for sample \mathbf{x}_i :
 $X_i = \{(x_1, y_1, \rho_1^1, \rho_1^2, \rho_1^3), \dots, (x_j, y_j, \rho_j^1, \rho_j^2, \rho_j^3)\}$, where x and y are spatial coordinates normalized by the height and width of the image, and ρ^1 , ρ^2 , and ρ^3 are ellipse parameters for affine region detectors.
 - 3: Compute region descriptors: $X_i^r = \{r_1, \dots, r_j\}$, where r_j is the description of the j th detected region of sample \mathbf{x}_i .
 - 4: **end for**
 - 5: Define a contextual-space $C = \{(c_1, w_1^C), \dots, (c_v, w_q^C)\}$ using K -means to define K contextual clusters, where w_i^C is the i th word of the contextual-space.
 - 6: Define a feature-space $F = \{(f_1, w_1^F), \dots, (f_v, w_q^F)\}$ using K -means to define K feature clusters, where w_i^F is the i th word of the feature-space.
 - 7: Initialize a contextual-feature relational matrix $M: M(i, j) = 0, i, j \in [1, \dots, K]$
 - 8: Initialize $W = \emptyset$ the list of feature words to be merged
 - 9: **for** I merging steps **do**
 - 10: update M based on the contextual clusters and new feature clusters so that $M(i, j) = \frac{d(C, F, i, j)}{|w_j^F|}$, where $d(C, F, i, j)$ returns the number of points from contextual-space of word w_i^C that belong to the feature-space j th word w_j^F , and $|w_j^F|$ is the number of regions of the j th feature word.
 - 11: Select the pair of positions with the maximum value for each row of M :
 $\max_{j, k} M(i, \cdot), j \neq k, \forall i$, where \cdot stands for all row positions.
 - 12: $W = W \cup (w_j^F, w_k^F)$: Select the contextual word w_i^C and words w_j^F and w_k^F from the feature-space based on $\max_i (\min(M(i, j), M(i, k)))$, $\forall j, k$
 - 13: **end for**
 - 14: **for** each pair (w_j^F, w_k^F) in W **do**
 - 15: update F so that $w_j^F \leftarrow w_k^F$, and rename feature words so that $w_i^F, i \in [1, \dots, p]$ becomes $w_i^F, i \in [1, \dots, p-1]$
 - 16: **end for**
 - 17: Compute representant $R = \{(r_1, w_1), \dots, (r_v, w_b)\}$ for the new F , where:
 $z_i = \text{round} \left(\frac{w_i}{\min |w_j|_{\forall j}} \right)$
is the number of representant for word w_i , computed using z_i -means, and $\{r_1, \dots, r_{z_i}\}$ representant are computed as the mean value for each sub-cluster of w_i , obtaining an stratified number of representant respect the words densities.
 - 18: Compute sentences $S = \{(s_1, l_1), \dots, (s_m, l_m)\}$ for all training samples of all categories comparing with word representant of R .
-

problem. The one-versus-one ECOC technique is defined in the ternary ECOC framework $M^{N \times M} \in \{-1, 0, +1\}$, being M a coding matrix of N rows (as the number of classes), M the number of columns (dichotomizers to be trained, where $M = N(N-1)/2$ in the case of the one-versus-one design), $\{-1, +1\}$ symbols codify the class membership, and the zero symbol ignores a particular class for a given dichotomizer. Each column of the matrix M corresponds to the i th binary problem h_i , which splits a pair of classes using a given base classifier. Figure 2 codifies a one-versus-one coding matrix M for a 4-class problem. The black positions correspond to the symbol $+1$, the white positions to the symbol -1 , and the gray positions to the zero symbol. Once the set of binary problems $h = \{h_1, \dots, h_M\}$ is trained, a new test sample ρ that arrives to the system is tested applying the set h , and a test codeword $X^{1 \times M} \in \{-1, +1\}$ is obtained.

The decoding step was originally based on error-correcting principles under the assumption that the learning task can be modeled as a communication

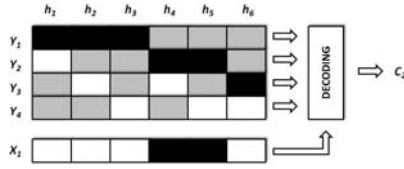


Fig. 2. One-versus-one ECOC coding design for a 4-class problem. A decoding function d for a new input test sample is performed, classifying by class C_2 .

problem, in which class information is transmitted over a channel [7]. The first attempt for ECOC decoding is the Hamming Decoding (*HD*). The Euclidean Decoding (*ED*) is another of the most preferred decoding strategies in the literature. Still very few alternative decoding strategies have been proposed [8]. Then, after the codeword X for the test sample ρ is obtained, a decoding function $d(X, Y_j)$ applying any of the previous decoding strategies is used to compare the test codeword X with each codeword Y_j (j th row from M) codifying class C_j . Finally, the classification prediction corresponds to the class C_j which corresponding codeword Y_j minimizes d (C_2 in the case of the example of Figure 2).

4 Experimental Evaluation

Before the presentation of the results, first, we discuss the data, methods, and validation protocol of the experiments.

Data: The data used in the experiments consists of 15 categories from public Caltech 101 [9] and Caltech 256 [10] repository data sets. One sample for each category is shown in Figure 3. For each category, 50 samples were used, 10 samples to define the BOVW and another 40 images to define new test sentences.

Methods: We compare the C-BOVW with the classical BOVW model. For both methods, the same initial set of regions is considered in order to compare both strategies at the same conditions. About 200 ± 20 object regions are found by image using the Harris-Affine detector [1] and described using the SIFT descriptor [2]. The visual words are obtained using the public open source K -means software from [11]. After computing the final words and representant, multi-class classification is



Fig. 3. Considered categories from the Caltech 101 and Caltech 256 repositories

performed using an one-versus-one ECOC methodology with different base classifiers: Mean Nearest Neighbor (NMC), Fisher Discriminant Analysis with a previous 99% of PCA (FLDA), Gentle Adaboost with 50 iterations of decision stumps (G-ADA), Linear Support Vector Machines with the regularization parameter $C = 1$ (Linear SVM), and Support Vector Machines with RBF Kernel with C and γ parameters set to 1 (RBF SVM)¹. Finally, we use the Linear Loss-weighted decoding to obtain the class label [8].

Validation protocol: We used the sentences obtained by the 50 samples of each category and performed stratified ten-fold cross-validation evaluation.

4.1 Caltech 101 and 256 Classification

In this experiment, we started classifying from three Caltech categories increasing by 2 up to 15. For each step, different number of visual words are computed: 30, 40, and 50. These numbers are obtained by performing ten iterations of the merging procedure (experimentally tested). In order to compare the BOVW and C-BOVW methods at the same conditions, the same detected regions and descriptions are used for all the experiments. The order in which the categories are considered is the following: (1-3) airplane, motor-bike, watch, (4-5) tripod, face, (6-7) ketch, diamond-ring, (8-9) teddy-bear, t-shirt, (10-11) desk-globe, backpack, (12-13) hourglass, teapot, (14-15) cowboy-hat, and umbrella. The obtained results applying ten-fold cross-validation are graphically shown in Figure 4 for the different ECOC base classifiers. Note that the classification error significantly varies depending on the ECOC classifier. In particular, Gentle Adaboost obtains the best results, with a classification error inferior to 0.2 in all the tests when using 30 C-BOVW words. Independently of the ECOC classifier, in most of the experiments the C-BOVW model obtains errors inferior to those obtained by the classical BOVW. BOVW only obtains slightly better results in the case of Gentle Adaboost for eleven classes and 50 visual words.

An important remark of the C-BOVW model is about the selection of the number of merging iterations. This parameter has a decisive impact over the generalization capability of the new visual dictionary. First iterations of the merging procedure use to fuse very close feature-words which belong to different visual words whereas final merging iterations fuse more far regions of the feature-space. Thus, a large number of iterations could be detrimental since the new merged words could be too general for discriminating among sentences of different object categories. Thus, this parameter should be estimated for each particular problem domain (i.e. applying cross-validation over a training and a validation subset). In the previous experiment we checked that ten merging iterations obtains significant performance improvements, though we are aware that this parameter could be not optimal for all the data sets.

¹ We decided to keep the parameter fixed for the sake of simplicity and easiness of replication of the experiments, though we are aware that this parameter might not be optimal for all data sets.

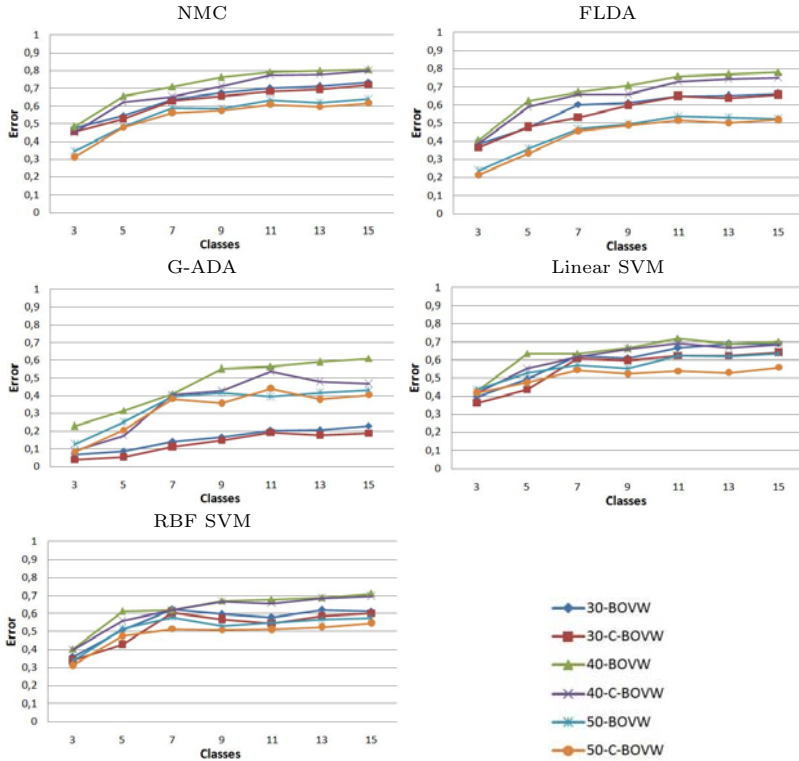


Fig. 4. Classification results for the Caltech categories using BOVW and C-BOVW dictionaries for different number of visual words and ECOC base classifiers

5 Conclusion

In this paper we re-formulated the bag-of-visual-words model so that geometrical information of significant object region descriptions are taken into account in the dictionary construction step. In this sense, regions which have slightly different descriptors because of small displacements in the region detection process can be merged together in a same visual word. The method is based on the definition of a contextual-space and a feature-space. The first space codifies the geometrical properties of regions meanwhile the second space contains the region descriptions. A merging process is then used to fuse feature words based on their proximity in the contextual-space. The new dictionary is learned in an Error-Correcting Output Codes design to perform multi-class object categorization. The results when spatial information is taken into account showed significant performance improvements compared to the classical approach for different number of object categories and visual words.

Acknowledgements

This work has been supported in part by projects TIN2006-15308-C02, FIS PI061290, and CONSOLIDER-INGENIO CSD 2007-00018.

References

1. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Kadir, T., Van Gool, L.: A comparison of affine region detectors. *IJCV* 65(1-2), 43–72 (2005)
2. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2005)
3. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: *ECCV*, pp. 1–22 (2004)
4. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: *CVPR*, pp. 1–8 (2007)
5. Chum, O., Philbin, J., Sivic, J., Zisserman, A.: Automatic query expansion with a generative feature model for object retrieval. In: *ICCV*, pp. 1–8 (2007)
6. Carneiro, G., Jepson, A.: Flexible spatial models for grouping local image features. In: *CVPR*, vol. 2, pp. 747–754 (2004)
7. Dietterich, T., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes 2, 263–282 (1995)
8. Escalera, S., Pujol, O., Radeva, P.: On the decoding process in ternary error-correcting output codes. *Transactions in PAMI* 99 (2009)
9. Caltech 101, http://www.vision.caltech.edu/image_datasets/caltech101/
10. Caltech 256, http://www.vision.caltech.edu/image_datasets/caltech256/
11. Clustering package, <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/>