

Joint Visual Vocabulary For Animal Classification

Heydar Maboudi Afkham Alireza Tavakoli Targhi Jan-Olof Eklundh Andrzej Pronobis
KTH-CVAP Stockholm
{heydarma,att,joe,pronobis}@kth.se

Abstract

This paper presents a method for visual object categorization based on encoding the joint textural information in objects and the surrounding background, and requiring no segmentation during recognition. The framework can be used together with various learning techniques and model representations. Here we use this framework with simple probabilistic models and more complex representations obtained using Support Vector Machines. We prove that our approach provides good recognition performance for complex problems for which some of the existing methods have difficulties. Additionally, we introduce a new extensive database containing realistic images of animals in complex natural environments. We assess the database in a set of experiments in which we compare the performance of our approach with a recently proposed method.

1. Introduction

Animals as objects have significance in content-based image and video retrieval as they carry a lot of semantic information about natural scenes. Unfortunately, they are also difficult to recognize since they have deformable bodies that could self occlude and often appear in complex backgrounds. Additionally, as all objects they may appear under different illumination conditions, view points and scales. There are attempts to apply recognition methods on images of animals but the specific problem of animal categorization has attracted limited interest.

Furthermore, existing databases are usually limited to simple settings and contain small numbers of animal classes. The performance of recognition or segmentation algorithms is usually shown on just a few classes e.g. zebra, cheetah and giraffe [4] or cow, sheep, bird, cat and dog (along with 16 non-animal classes) [16].

As we will show, many existing methods showing promising results for recognition can't properly repre-

sent the diversity of animal classes with complex background that occur in natural scenes. Therefore, to better represent the intra-class variability of the animal classes and the different contexts in which the animals can appear, more complex representations are required. As a result, the main contributions of this paper are as follows: (a) A dense method for visual object categorization based on joint textural information of objects; (b) A database of animal images which captures a broad range of natural variations and common technical difficulties; (c) Two different model representations of different complexity within the same classification framework; (d) Experimental evaluation on the proposed database.

2 Database

The images in the database determine how realistic the analysis is. A recent study on available databases for object recognition benchmarking [10] shows that the image classes within these databases lack many necessary features required for making realistic models, such as similar viewpoints and orientations, normalized sizes, position of the object, and little or not background clutter. An example of such a non-realistic database is MSRC [19] which many authors used for evaluation of their segmentation and recognition methods [16, 9, 19, 1, 3, 14]. High recognition rates based on this database (90%) reported, but this rate significantly decreases (40%) when it comes to a realistic animal image database capturing natural variations. Due to non-existence of a comprehensive annotated animal image database which captures all of these variations, a database containing 1239 images in thirteen classes of different animals is created which captures most possible natural conditions and variations.¹ The images were gathered from images of the Corel database [5], combined with images gained from Yahoo Internet image search results and segmented manually into foreground

¹The database is available at <http://nada.kth.se/~heydarma/database/>

and background regions for learning purposes.

3 Related Work

Object recognition: An accurate segmentation can be a great aid for most recognition methods. Since segmentation is not a well-defined problem [7], segmentations produced using different algorithms might not be similar. Furthermore, segmentation algorithms have the same complexity as object recognition algorithms with their own problems. Algorithms such as normalized cuts [6] or graph cuts [2], highly depend on their parameters. These methods were applied on our animal database and failed due to the large complexity in the background.

When dealing with a complex image database, it is important to use a method capable of capturing information from all the different classes. Sparse texture descriptors [8] depend on the region detector used and how complex the structure and texture of object and background are. As a result using such methods may fail, especially for animal classes with smooth skin texture and complex background. Either a segmentation algorithm or a dense method, which uses the information of all the pixels of the image, should be applied to avoid this. One often used dense method is MRF[18]. This method uses the exact intensity of the image or segmented region to extract textural information. It is often used for images containing a single texture and do not perform well on outdoor images without an accurate segmentation. To adopt it to databases with outdoor images with multiple objects, several approaches have been introduced. One of the earliest ones, which showed good performance on the MSRC database [19], was introduced by Winn *et al.*. They divide an image into several sub-regions and each region is classified separately depending on its distribution of visual words. They report a 93.4% classification rate. Later, Savarese *et al.* [14] defined a model for the appearance and the shape of the object in a class by finding correlations between different visual words. The performance of 93.8 classification rate on MSRC database is reported. These approaches focus on recognizing the different regions of the image, which brings the need for using a segmentation algorithms before performing the recognition task. To avoid the use of segmentation algorithms, Schroff *et al.* introduced the single-histogram class models [16]. Their method uses an average histogram of visual word distribution in local neighborhoods for classification. The single-histogram models are much simpler than the models used in the previously mentioned methods, still classification performance reported is comparable to the performance of other meth-

ods with 93.43% classification rate. The segmentation accuracy of this method was reported to be 75.07%. The single-histogram model is not rich for capturing large variations, but sufficient enough for the available databases[10].

Recently Shotton *et al.* [17] introduced the texton-boost method which uses different types of information such as shape, color and texture to classify the pixels of the images. Their approach is based on learning the parameters of conditional random field models from a set of shape features which are considered as weak classifier for the Joint Boosting Algorithm. The method most relevant to this work is the single-histogram class model [16]. This is because to perform the recognition task no extra information such as segmentation is required. For this reason a special attention is given to this method.

Animal Recognition: One of the earliest attempts to perform recognition on an animal database was done by Schmid [15]. They constructed models for content-based image retrieval using Gabor-like filters. The method was tested on only four different classes. All animals used in this work had complex skin texture. Later, Ramanan *et al.* introduced methods to detect textured animals using the shape and texture information in video sequences [12, 13]. In an application for searching images on the Internet Berg and Forsyth [1] used four cues: nearby texts on the web pages, color, texture and shape to re-rank the images retrieved by Google image search. They reported that animals are among the hardest classes of objects for recognition in computer vision.

4 Theory And Method

Recently the use of dense textural visual word dictionaries in image segmentation and object recognition problem has become more popular. Most of the methods designed on the visual word dictionaries neglect information hidden between the different neighboring visual words[16, 9, 19, 1, 3, 14]. The studies done in this paper which uses of such information will result a significant increase in the classification rate. In this section, we introduce our classification framework which is based on a joint visual vocabulary. Furthermore, we apply two different classification methods within this framework. The first classifier is based on a probabilistic approach, as the second one we use Support Vector Machines (SVM). We used the same methodology as [11] for SVMs on the distributions obtained from joint probabilities in local neighborhoods.

In this approach the visual words dictionary, T , is generated using the extension of MRF descriptor, described in the single-histogram models [16].

4.1 Joint Probabilities

The idea of using the joint probabilities is to capture the joint distribution of the visual words, in order to obtain better features for classification. Joint distribution of the textons captures the probability of different visual words appearing in a neighborhood of each other, in different classes. In this application the neighboring visual words were determined using a sliding window technique. It can be easily shown that the maximum likelihood estimate of appearing the visual word t_j in a neighborhood of the visual word t_i with respect to the class c is calculated as

$$\Pr(t_j|t_i, c) = \frac{\sum_{q \in N_{(c, t_i)}} (T(q) == t_j)}{\sum_{q \in N_{(c, t_i)}} 1} \quad (1)$$

where, $N_{(c, t_i)}$ denotes the union of local neighborhoods of the pixels labeled as the visual word t_i in the training region of class c and $T(q)$ returns the visual word positioned at the pixel q .

The model learned for each class is $M_c = [m_{ij}^c]_{K \times K}$ with $m_{i,j}^c = \Pr(t_j|t_i, c)$. The i^{th} row of this matrix is the visual words distribution around the visual word t_i . This vector is denoted by $m_i^c = [m_{i,j}^c]_{1 \times K}$.

With having the visual words dictionary and models for classes, C_1, C_2, \dots, C_n and the background class C_B , we wish to define the probabilities $\Pr(C_i|I_{test})$ for every test image I_{test} . Every visual word in the test region is then classified according to its neighboring visual word distribution. Assume that $n(N|T)$ is the normalized histogram of the visual words within the neighborhood N with the center visual word t_i , positioned at pixel p . Using these information this visual word is classified as

$$c^*(p) = \underset{c \in \{C_1, \dots, C_n, C_B\}}{\operatorname{argmin}} \{d(n(N|T), m_i^c)\}, \quad (2)$$

where, $d(., .)$ denotes the χ^2 distance between the histograms. Finally the probability of occurrence of each class within the test image for $c \in \{C_1, \dots, C_n\}$ is defined as

$$\Pr(c|I_{test}) = \frac{|p \in I_{test} : c^*(p) = c|}{\sum_{j=1}^n |p \in I_{test} : c^*(p) = C_j|}. \quad (3)$$

5 Results

5.1 Experimental Setup and Procedure

For the experiments, the database was randomly split into a test and a training sets. Each set contained 50% of the images. The training set was first used to build the

Method Name	Rate 1	Rate 2	Rate 3
Single-hist. [16]	0.39±0.01	0.60±0.01	0.70±0.01
Joint Prob.	0.65±0.02	0.78±0.01	0.84±0.02
SVM	0.75	0.84	0.88

Table 1: The rates 1, 2, and 3 correspond to classification rates when one, two, or three best hypotheses were taken into account.

dictionary of visual words and then to train the models. In case of the experiments with the single histogram and joint probability methods, the experiments were performed several times for different random splittings and we used visual words dictionary of size 1500. The kernel and training parameters for the experiments with multi-class SVM classifiers [11] were the same for all models ($\gamma = 1$ and $C = 100$) and selected based on a small set of preliminary experiments.

5.2 Experimental Results

The size of the sliding window determines how rich the neighboring distributions are. In our experiment we varied the size of the window from 31×31 pixels to windows with 211×211 pixels and measured the recognition rate for the joint probability approach. The size of most of the images used in the experiments was approximately 384×256 pixels. The results showed that, when the window is too small the classification rate is low since too little information is captured about the object. On the other hand, when the window is too large the classification rate drops, since the models contain more information about the background than the object itself. The best classification rate was achieved for the sliding window of size 121×121 pixels.

Table 1 shows the performance of our method based on two types of models (simple probabilistic model and SVM) and compared to single histogram technique. It is apparent that the model consisting of a single histogram for each class was unable to encode the complex dependencies in the data. As the more sophisticated methods are employed, the classification rate increases by 26% in case of the joint probability model and another 10% in case of the SVMs. When it comes to image search applications, more than one hypothesis can be considered. For such applications we always expect to get the result among several highest ranked images. Table 1 shows the percentages of correct classification among the first two or three hypotheses. It is clear that, the applied recognition framework can output not only a single decision, but is also able to provide a meaningful ranking of hypotheses.

	Single-Hist	Joint	SVM		
	Rate 1	Rate 1	Rate 1	Rate 2	Rate 3
1-bear	0.07	0.42	0.42	0.60	0.68
2-cougar	0.42	0.40	0.52	0.66	0.74
3-coyote	0.38	0.52	0.60	0.72	0.78
4-elephant	0.32	0.64	0.90	0.92	0.94
5-giraffe	0.27	0.33	0.43	0.60	0.74
6-goat	0.18	0.60	0.76	0.86	0.88
7-horse	0.96	0.96	1.00	1.00	1.00
8-leopard	0.35	0.78	0.88	0.94	0.96
9-lion	0.31	0.88	0.86	0.92	0.96
10-panda	0.50	0.67	0.84	0.92	0.96
11-penguin	0.46	0.68	0.62	0.85	0.90
12-tiger	0.18	0.88	0.94	0.98	1.00
13-zebra	0.54	0.85	0.95	0.97	0.97

Table 2: A detailed comparison of performance of the three methods for each of the single classes.

It can be seen from Table 2 that animals such as giraffe, bear or cougar are particularly difficult to recognize using the evaluated methods. Still, it can be observed that the correct classification is usually among the first two or three hypotheses, and the classification rate quickly improves when more than single decision is considered.

6 Conclusions

This paper presented a model for visual object categorization based on joint textural information of objects and their context. We showed that by using the relation between visual words, more information of the object can be captured. Furthermore the information about context and background surrounding the object can be encoded to facilitate categorization and increase the classification accuracy. When applying the joint probabilistic model we have observed 26% improvement in comparison with most of existing methods which is use visual vocabulary and the recent one the single-histogram approach. This performance was improved even more when more sophisticated and complex classifier (SVM) was used. However, the probabilistic method is more efficient in terms of computational complexity and memory requirements. Our initial studies shows also a significant improvement for non animal object classes by using the joint visual vocabulary. In future work we are going to use this method for general object classification. **Acknowledgments:** This project was supported by MOBVIS (EU-FP6-511051-2), MUSCLE (FP6-507752) and 2005-3600-Complex.

References

- [1] T. L. Berg and D. A. Forsyth. Animals on the web. In *CVPR*, 2006.
- [2] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, 2001.
- [3] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition, 2004.
- [4] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 2005.
- [5] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *PAMI*, 2003.
- [6] J. Malik and J. Shi. Normalized cuts and image segmentation. In *CVPR*, 1997.
- [7] D. R. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Technical report, EECS Department, University of California, Berkeley, 2001.
- [8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 2005.
- [9] M. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.
- [10] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition*, 2006.
- [11] A. Pronobis and B. Caputo. Confidence-based cue integration for visual place recognition. In *IROS*, 2007.
- [12] D. Ramanan, D. A. Forsyth, and K. Barnard. Detecting, localizing and recovering kinematics of textured animals. In *CVPR*, 2005.
- [13] M.-D. Ramanan, S. M.-D. A. Forsyth, and M.-K. Barnard. Building models of animals from video. *PAMI*, 2006.
- [14] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlations. In *CVPR*, 2006.
- [15] C. Schmid. Constructing models for content-based image retrieval. In *CVPR*, 2001.
- [16] F. Schroff, A. Criminisi, and A. Zisserman. Single-histogram class models for image segmentation. In *ICCVGIP*, 2006.
- [17] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.
- [18] M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *CVPR*, 2003.
- [19] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.