BAG-OF-WORDS CLASSIFICATION OF MINIATURE ILLUSTRATIONS

Costantino Grana, Daniele Borghesani, Giovanni Gualdi, Rita Cucchiara

Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905/b - 41100 Modena
name.surname@unimore.it

ABSTRACT

In this paper a system for illuminated manuscripts images analysis is presented. In particular the bag-of-keypoints strategy, commonly adopted for object recognition, image classification and scene recognition, is applied to the classification of automatically extracted miniatures. Pictures are characterized by SURF descriptors, and a classification procedure is performed, comparing the results of Naïve Bayes and histogram intersection distance measures.

1. INTRODUCTION

Artistic or historical documents cannot be made directly available to the public, due to their value and fragility, so museum visitors are usually very limited in their appreciation of this kind of artistic productions. Computer science can help bridging people and these precious libraries: digital versions of the artistic works can be made publicly accessible, both locally at the museums owning the original version and remotely. In this manner, users —either experts, tourists or people keen on art— can explore more comprehensively the document, choosing their own personal way to browse and enjoy it.

Italy, in particular, has a significant collection of illuminated manuscripts, but many of them are not freely accessible. These masterpieces contain thousands of valuable illustrations: different mythological and real animals, biblical episodes, court life illustrations, and some of them even testify the first attempts in exploring perspective for landscapes. Usually manual segmentation and annotation for all of them is dramatically time consuming. For this reason, the accomplishment of the same task with an automatic procedure is very desirable but, at the same time, really challenging due to the visual appearance of these pictures (their arrangement over the page, the decorative parts, etc.).

We propose a solution for automatic manuscript segmentation and pictures extraction, and in particular a modification of the bag-of-keypoints approach to efficiently apply it in the context of automatic categorization of artistic hand-drawn illustrations (i.e. separating illustrations depending on their content, for instance people vs animals). The final goal is to provide automatic content-based functionalities such as similarity search, comparison, recognition of specific elements (people, life scenes, animals, etc...) in artistic manuscripts.

2. RELATED WORK

The problem of image analysis and classification of historical manuscripts is becoming a significant subject of research in recent years, even if the availability of complete systems for the automatic management of illuminated manuscripts digital libraries is quite limited. The AGORA [1] software generates a map of foreground and background and consequently proposes a user interface to assist in the creation of an XML annotation of the page components. The Madonne system [2] is another initiative to use document image analysis techniques for the purpose of preserving and exploiting cultural heritage documents. In [3], Le Bourgeois et al. highlighted some problems due to acquisition and compression, then authors gave a brief subdivision of documents classes, and for each of them provided a proposal of analysis, performing color depth reduction, layout segmentation and main body segmentation using text zones location.

The bag-of-words approach has become increasingly popular and successful in many object recognition and scene categorization tasks. The first proposals constructed a vocabulary of visual words by extracting image patches, sampled from a grid [4]. More advanced approaches used an interest point detector to select the most representative patches within the image [5]. The idea was finally evolved toward the clustering and quantization of local invariant features into visual words as initially proposed by [6] for object matching in videos. Lately the same approach was exploited in [7], which proposed the use of visual words in a bag-of-words representation built from SIFT descriptors [8] and various classifiers for scene categorization. As shown in [9], the bag-of-words approach creates a simple representation but potentially introduces synonymy and polysemy ambiguities, which can be solved using probabilistic latent semantic analysis (PLSA) in order to capture co-occurrence information between elements. In [10] the influence of different strategies for keypoint sampling in the categorization accuracy has been studied: the Laplacian of Gaussian (LoG), the Harris-Laplace de-

tector and random sampling. A recent comparison of vocabulary construction techniques is proposed in [11].

3. AUTOMATIC SEGMENTATION AND RETRIEVAL

In [12] we described a system and the techniques used for text extraction and picture segmentation of illuminated manuscripts. The goal of the automatic segmentation system is to subdivide the document into its main semantic parts, in order to enable the design of new processing modules to manage and analyze each part, relieving the user of the task of manual annotation of the whole book collection. In that work we also introduced a first module for content-based retrieval functionalities by visual similarity with an ad-hoc designed user interface.

The module for *text segmentation* computes the autocorrelation matrix over gray-scale image patches and converts them into a polar representation called *direction histogram*: a statistical framework able to handle angular datasets (i.e. a mixture of Von Mises distributions) generates a compact representation of such histograms that are then the final features used to classify each block through an SVM classifier.

The text-free parts of the image are then passed to a second module that separates plain background, decorations and miniatures. We use here a sliding window approach and represent each window with a descriptor that joins color features (*RGB* and *Enhanced HSV* Histograms) and texture features (*Gradient Spatial Dependency Matrix (GSDM)*. As in [13], we exploited a Lipschitz embedding technique to reduce the dimensionality of the feature space and again used an SVM classifier to obtain the desired classification.

Eventually an interface is proposed to the user, that can perform ranked image retrieval by content similarity, exploiting both visual descriptors and textual information. The next section describes the techniques exploited in a further system module (to be placed before the retrieval by similarity module), that is aimed to categorize the content of the miniatures. As described in Section 5, this module will be also used as an additional rejection step for illustration patches.

4. BAG-OF-KEYPOINTS CLASSIFICATION

One of the most successful strategies to perform object and scene recognition is the bag-of-keypoints approach. The main idea comes from the text categorization (bag-of-words), and it consists in defining, during the training phase, a set of "words" (rich enough to provide a representative description of each and all the classes) and the occurrences of these "words" for each class.

In our context, since we cannot directly extract high level semantic words, we can define "visual words" by clustering accordingly visual descriptors (e.g. keypoint descriptors): the set of centroids of each cluster creates the so called *vocabulary*. After having counted, for each class, the occurrences of each word, the classification can then be easily performed extracting the histogram of the visual words of an example, and then finding the class that has the most similar occurrences distribution.

(a) (b) (c)

Fig. 1. Some representative of the *people*, *animals* and *decorations* classes.

In [7] scene categorization is accomplished following this procedure and making use of Harris affine detector as keypoint detector (mapped to a circular region in order to normalize them for affine transformations) and SIFT as keypoint descriptors. A lot of different keypoint detector and descriptors are available in literature, with different degrees of invariance to distortion or cluttering. Regarding detectors, [14] proved that Hessian-based detectors are better than Harris-based in terms of stability and repeatability, and likewise the use of the determinant of the Hessian matrix is to be preferred to its trace (the Laplacian). Regarding descriptors, SIFT is one of the most widely spread, even if many variations have been proposed, both in terms of distinctiveness and speed: among them SURF [15] is probably among the most successful: it relies on integral images for image convolutions, uses a fast Hessian matrix based interest point detector, performed with *box filters* (an approximation procedure again relying on integral images), and eventually uses a simple Haar wavelet distribution descriptor resulting in a 64 feature vector. These factors make SURF computationally more affordable than SIFT, and very similar in terms of accuracy of the point matching performances. Differently from [14] that proposes a two-fold procedure (Harris-based keypoint detector and SIFT based keypoint descriptor) we propose to use SURF which incorporates both advantages within the same algorithm.

The training set in our system is composed of patches of miniatures belonging to different classes. SURF keypoint descriptors are extracted over all patches and the visual vocabulary V is then made of the k cluster centroids obtained running a k-means clustering procedure: $V = \{v_i\ , i = 1 \dots k\}$, with v_i cluster centroid. Once the vocabulary is computed, each class is characterized by a specific distribution of visual word occurrences: therefore we obtain $p(v_i|C_j)$, for each class j, for each visual word i. In order to avoid later numerical problems, a Laplace smoothing is applied. The number k is a key parameter of the whole process: low k will generate a poorly descriptive vocabulary, while high h will over fit the

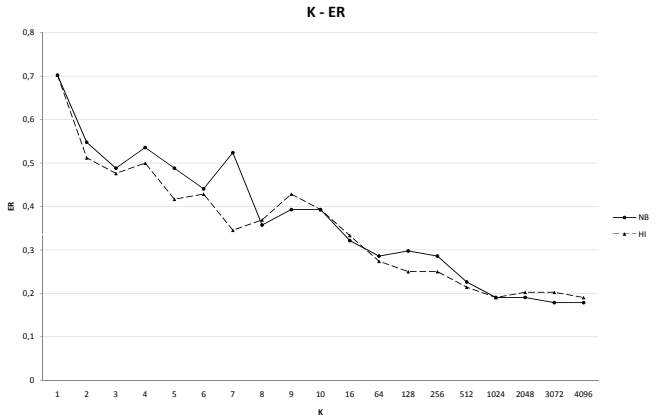

Fig. 2. Classification error rate for *people*, *animals* and *decorations*classes at different k values, with histogram intersection and Naïve Bayes.

training data, therefore the training phase will slide through several k, finding the best value through cross validation.

On any new image patch I to classify, the SURF descriptors are extracted and each casts a vote for the closest cluster centroid; I can be thus described as a histogram of the visual words of the vocabulary: each bin $N(v_i)$ counts the number of times in which a word v_t has occurred in I, constituting the feature vector. The final classification can be accomplished using several methodologies. In this paper, we used histogram intersection and Naïve Bayes.

Histogram intersection [16] is a typical and quite robust metric used to compare histograms. Given two histograms H and G, both with k bins, this metric determines the corresponding similarity S between them by computing:

$$S(H,G) = \sum_{i=1}^{k} \min(H_i, G_i) \tag{1}$$

where H_i and G_i denote the i-th bin of the histogram. The output value is finally normalized.

Naïve Bayes instead is a simple but effective classification technique based on Bayes' rule: given the image I and a prior probability $p(C_j)$ for j-th class, the classifier assigns to I the class with the largest posterior $p(C_i|I)$ according to Eq. 2 (thus assuming the independence of visual words within the image):

$$p(C_j|I) = p(C_j) \prod_{i=1}^{k} p(v_i|C_j)^{N(v_t)} \tag{2}$$

To evaluate the accuracy of the classification, we used the error rate, defined as:

$$ER = 1 - \left(\sum_{j=1}^{N_c} |C_j| M_j \right) / \left(\sum_{j=1}^{N_c} |C_j| \right) \tag{3}$$

	animals	decorations	people
animals	1.00	0.00	0.29
decorations	0.00	1.00	0.14
people	0.00	0.00	0.57

(a)

	animals	decorations	people
animals	0.96	0.00	0.29
decorations	0.00	0.96	0.11
people	0.04	0.04	0.60

(b)

	animals	decorations	people
animals	0.96	-	0.26
decorations	-	-	-
people	0.04	-	0.74

(c)

Fig. 3. The confusion matrix generated by the classification of *animals*, *decorations* and *people* classes, using Naïve Bayes (a) and Histogram Intersection (b) and (c). (c) is obtained with binary classification between *animals* and *people* only.

where N_c is the number of classes to classify, $|C_j|$ is the cardinality of the test samples in the j-th class, and M_j is the percentage of correct classifications within class j.

5. EXPERIMENTAL RESULTS

We performed tests on the digitalized pages of the Holy Bible of Borso d'Este, which is considered one of the best Renaissance illuminated manuscripts. The dataset consists of 320 high resolution digitalized images (3894x2792), a total amount of 640 pages. Each page of the dataset is an illuminated manuscript composed by a two-column layered text in Gothic font, spaced out with some decorated drop caps. The entire surrounding is highly decorated.

The segmentation procedure described in Section 3 was run over a subset of bible pages, providing us a set of valuable illustrations containing mostly miniature illustrations of scenes, symbols, people and animals, rejecting text and most of border decorations: in fact, a small percentage of decorations are wrongly classified at this stage and therefore we consider them as an additional class of miniatures. From this pool of miniatures, we extract a dataset of three semantic classes: *animals*, *people* and *decorations*, and apply the SURF-based bag-of-keypoints classification (Section 4).

Regarding the training set, we used 785 samples for *animals*, 940 for *people* and 600 for *decorations*, while as testing set we used 125 samples for *animals*, 175 for *people* and 120 for decorations. Over all images we extracted SURF keypoints with relative descriptors (Fig. 4), and on the descriptors belonging to the training set we run k-means clustering in order to define the vocabulary. As aforementioned, we train the system at different values of k, to verify the behavior of the error rate w.r.t. the clustering quantization.

Fig. 2 shows that decreasing the clustering quantization (i.e. increasing k) yields higher performances. Best results are obtained with $k = 1024$ for histogram intersection and $k = 3072$ for Naïve Bayes, producing respectively and error rate of 19.05% and 17.86%. These tests show that histogram intersection produces better results with a lower clustering, while the Naïve Bayes approach leads to the best classification accuracy overall, making it our final solution.

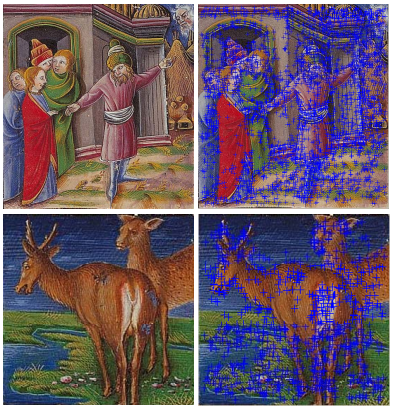

Fig. 4. SURF extraction on *people* and *animals* miniatures.

In Fig. 3(a,b) the confusion matrices related to the multi-class categorization are proposed. The most valuable semantic content is defined within the *people* and *animals* classes: the class of *decoration* was included to apply the proposed approach in a multi-class context. Given the high classification accuracy of decorations w.r.t. the other two classes, we employ this multi-class categorized as a processing stage postponed to the automatic segmentation described in Section 3, in order to remove all decorative pictures erroneously extracted. After this step we apply the bag-of-keypoints approach as a binary classifier, in order to get a more refined *people* vs *animals* categorization: a reduced number of classes allows to have more descriptive vocabularies, thus yielding lower error rates (16.67% of error rate with $k = 512$, exploiting Histogram Intersection; see Fig. 3(c) for confusion matrix).

6. CONCLUSIONS

In this paper, the bag-of-words approach using SURF as keypoint detector and descriptor has been used to classify people and animals widely present in the illuminated manuscript used as dataset. Given a training set automatically extracted and manually selected, some test are reported proving the description capabilities of this approach over these pictures.

7. REFERENCES

[1] J.Y. Ramel, S. Busson, and M.L. Demonet, "AGORA: the interactive document image analysis tool of the BVH project," in *International Conference on Document Image Analysis for Libraries*, 2006, pp. 145–155.

[2] J.M. Ogier and K. Tombre, "Madonne: Document Image Analysis Techniques for Cultural Heritage Documents," in *Digital Cultural Heritage, Proceedings of 1st EVA Conference*, 2006, pp. 107–114.

[3] F. Le Bourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz, "Document Images Analysis Solutions for Digital libraries," in *International Conference on Document Image Analysis for Libraries*, 2004, pp. 2–24.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on PAMI*, vol. 28, no. 4, pp. 594, 2006.

[5] S. Agarwal and A. Awan, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on PAMI*, vol. 26, no. 11, pp. 1475–1490, 2004.

[6] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *ICCV*, 2003, vol. 2, pp. 1470–1477.

[7] Christopher R. Dance, G. Csurka, L. Fan, J. Willamowski, and Cedric Bray, "Visual categorization with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[8] D.G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999, vol. 2, pp. 1150–1157.

[9] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on PAMI*, vol. 29, no. 9, pp. 1575–1589, 2007.

[10] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV*, 2006.

[11] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time bag of words, approximately," in *International Conference on Image and Video Retrieval*, 2009.

[12] C. Grana, D. Borghesani, and R. Cucchiara, "Describing Texture Directions with Von Mises Distributions," in *International Conference on Pattern Recognition*, 2008.

[13] G.R. Hjaltason and H. Samet, "Properties of Embedding Methods for Similarity Searching in Metric Spaces," *IEEE Transactions on PAMI*, vol. 25, no. 5, pp. 530–549, 2003.

[14] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[15] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[16] Michael J. Swain and Dana H. Ballard, "Color indexing," *IJCV*, vol. 7, no. 1, pp. 11–32, 1991.