Improving classification and retrieval of illuminated manuscript with semantic information

Costantino Grana, Daniele Borghesani, Rita Cucchiara

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

Abstract. In this paper we detail a proposal of exploitation of expert-made commentaries in a unified system for illuminated manuscripts images analysis. In particular we will explore the possibility to improve the automatic segmentation of meaningful pictures, as well as the retrieval by similarity search engine, using clusters of keywords extracted from commentaries as semantic information.

1 Introduction

The availability of semantic data is a well known advantage for different image retrieval tasks. In the recent literature, a lot of works has been proposed to create, manage and further exploit semantic information to be used in multimedia systems. The reason is that the information retrieval from textual data is quite successful. Semantic data are typically exploited in web searches in the form of tags (i.e. Google Images), but the process of tagging an image is known to be very boring from a user perspective, and likewise the process of linking correctly textual information about an image to the image itself is very tricky and error-prone. Nevertheless, often the amount of information and details held by a human-made description of an image is very precious, and it cannot be fully extracted using techniques based on vision.

Regarding the system globally, it is known that many artistic or historical documents cannot be made available to the public, due to their value and fragility, so museum visitors are usually very limited in their appreciation of this kind of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. For this reason, the availability of digital versions of the artistic works, made accessible —both locally at the museums owning the original version and remotely— with a suitable software, represents undoubtedly an intriguing possibility of enjoyment (from the tourist perspective) and study (from an expert perspective).

Italy, in particular, has a huge collection of illuminated manuscripts, but many of them are not freely accessible to the public. 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, their various framing into the decorative parts and so on).

In this papers we propose a solution for automatic manuscript segmentation and pictures extraction. A modification of the bag-of-keypoints approach is used to efficiently apply it in the context of automatic categorization of artistic hand-drawn illustrations (i.e. separating illustrations depending on their content, e.g. people vs animals). On the top of this system, we integrated the knowledge of a complete commentary available for the specific manuscript under analysis. A standard keyword clustering approach (usually known as tag-cloud) has been used to find out the most relevant topics within the entire book on in a smaller section, then we explored the correspondence of clusters of words and clusters of extracted pictures to prove that we can use textual data to help and improve object recognition. The final goal is to provide automatic content-based functionalities such as searches for similarity, comparison, recognition of specific elements (people, life scenes, animals, etc...) in artistic manuscripts including also the textual search in the retrieval engine.

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 performs a map of the foreground and the background and consequently propose 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 with acquisition and compression, then authors gave a brief subdivision of documents classes, and for each of them provided a proposal of analysis. They distinguished between medieval manuscripts, early printed documents of the Renaissance, authors manuscripts from 18th to 19th century and, finally, administrative documents of the 18th - 20th century: the authors perform color depth reduction, then a layout segmentation that is followed by the main body segmentation using text zones location. The feature analysis step uses some color, shape and geometrical features, and a PCA is performed in order to reduce the dimensionality. Finally the classification stage implements a K-NN approach.

The bag-of-keypoints 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. SIFT in particular was one of the first algorithms which combined an interest point detectors and a local descriptors to gather a good robustness to background clutter and good accuracy in description. Lately several aspects have been investigated. For example, 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 detector used in [11] and random sampling. A recent comparison of vocabulary construction techniques is proposed in [12].

The idea to exploit text in order to improve or integrate image (and video) retrieval is not new. In fact a lot of complex retrieval systems present a fusion stage in which visual features are somehow fused to audio and/or textual features. This process is usually referred as multimodal fusion. For instance, in [13] and [14] in the context of video retrieval, Snoek *et al.* learned a list of concept-specific keywords, and based on this list they construct a word frequency histogram from shot-based speech transcripts. In [15], Chen *et al.* aimed to improve image web search engines by extracting textual information from the image "environment" (tags, urls, page content, etc...) and users' logs. The text description (which semantically describes the image) is then combined with other low-level features extracted from the image itself to compute a similarity assessment. Textual information are also a very familiar approach for querying the system (since web search engines rely on it), so several works propose the query-by-keyword functionality along with the query-by-example and query-by-concept modalities. For references about this topic, please refer to [16].

3 Automatic segmentation and retrieval

In [17] 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

Fig. 1: Example of picture detection results

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 [18], 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.

Some examples of picture extraction results are shown in Fig. 1.

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. a set of "words" that is rich enough to provide a representative description of each and all the classes;
b. 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.

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. In our system, for performance reasons, we preferred the use of SURF [19]: it is a very successful local descriptor, it relies on integral images for image

(a) (b)

Fig. 2: Some representative of the *people*, *animals* and *decorations* classes.

convolutions, it uses a fast Hessian matrix based interest point detector, performed with *box filters* (an approximation procedure again relying on integral images), and eventually it 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. 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, we apply a Laplace smoothing. 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 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 been extracted from the image, constituting the feature vector. The final classification has been accomplished using Naïve Bayes.

Naïve Bayes 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. 1 (thus assuming the independence of visual words within the image):

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

5 Knowledge summary by tag cloud

Given a text describing a dataset of images, the correlation between frequency (thus importance) of terms and visual content is straightforward. For this reason, a clustering of keywords in such a document will probably lead to highlight the most important concepts (and thus visual objects) included in the dataset. The standard text processing in text retrieval systems is the following:

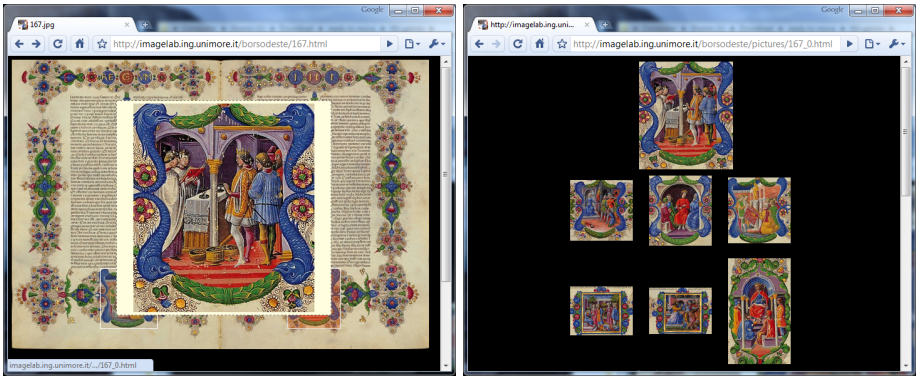
1. Parsing of the document into words.
2. Representing words by their stems. For example: "draw", "drawing" and "drawn" are represented by the stem "draw".
3. Definition of a stop list, i.e. a list of words to reject due to being common (such as "the") or recurring often in most documents and thus not being discriminant for a particular document.
4. Representation of the document as a vector of words and the relative frequency of occurrence within the document (different weighting techniques are possible).

A very rough but effective clustering procedure to extract important keywords from a document is the tag cloud. Tag clouds are a common visual representation of user-generated tags or generally word content of a document, at different size and/or color based on its incidence, typically employed to describe the content of the document itself. We employed this simple procedure to generate and select from the commentary some keywords to use in our tests.

6 Retrieval by similarity

The final application devised for this work is a system aimed at presenting all the details of illuminated manuscripts in a user friendly way. Eventually an interface is proposed to the user, that can perform ranked image retrieval by content similarity: given a query picture, relative histograms are compared using histogram intersection metric, and similarity values are normalized, fused and finally ranked, from the most similar to the query.

As referred in the next section, text information can also be used along with visual descriptors to describe visual content. In fact, the scope of this work is just focused on investigating the relation between textual data (retrieved from



(a) query (b) results

Fig. 3: Example of content-based retrieval example of people. Navigating through the interface, the user can select the query image, and the system proposes the retrieval results ordered by descending appearance similarity.

commentaries) and pictures in order to propose a multimodal search engine with a joint use of both pictorial and textual information.

So far, the similarity results provided as an example in Fig.3 are computed using color histograms (HSV and RGB) as visual descriptors.

7 Results

In this paper, we used the digitalized pages of the Holy Bible of Borso d'Este, which is considered one of the best Renaissance illuminated manuscripts. Tests have been performed on a dataset 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 the bible pages, providing us a set of valuable illustrations within the decoration texture (miniature illustrations of scenes, symbols, people and animals), rejecting border decorations (ornaments) and text. Some samples are shown in Fig. 2. Once the pictures have been processed, a tag cloud has been generated from the commentary (Fig. 4).

The first tag we analyzed was "Geremia", that is the Italian for prophet Jeremiah. A section of the Bible is dedicated to him, so in that section there are a lot of visual references about him. In this section of the bible, a total amount of 68 pictures was extracted from the illustration of the pages. The same features used in the retrieval by similarity module referred in 6 have been extracted from the pictures, and then clustered using the hierarchical Complete Link algorithm

Fig. 4: Tag cloud generated from the Bible commentary

with an automatic clustering level selection. We obtained 4 clusters, and over the most populated one we obtained a recall of 73.68% and a precision of 84.85%. See Fig.5 for some samples coming from this cluster.

The second tag we analyzed was "cervo", that is the Italian for deer. A lot of animals are depicted within this book, and the deer is one of the most frequent, so there are a lot of visual references about it. In the subset of pages where a deer is depicted, a total amount of 99 pictures was extracted from the illustration of the pages. Visual features have been extracted and then clustered. We obtained 3 clusters, and over the most populated one we obtained a recall of 84.31% and a precision of 72.88%. See Fig.6 for some samples coming from this cluster.

8 Conclusions

In this paper, we investigated the use of textual keywords from experts-made commentaries in a system for automatic segmentation and extraction of pictures from illuminated manuscripts. We showed that textual information (if available) can be considered a useful addition not only to help categorization of the visual content, but also to give to the user the possibility to search the visual content with textual keywords, providing a multimodal search engine analogously to common web search engine.

Fig. 5: Samples from the more populated of the clusters generated by clustering visual features of pages containing the word "Geremia"

References

1. Ramel, J., Busson, S., Demonet, M.: AGORA: the interactive document image analysis tool of the BVH project. In: International Conference on Document Image Analysis for Libraries. (2006) 145–155
2. Ogier, J., Tombre, K.: Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. In: Digital Cultural Heritage, Proceedings of 1st EVA Conference, Oesterreichische Computer Gesellschaft (2006) 107–114
3. Le Bourgeois, F., Trinh, E., Allier, B., Eglin, V., Emptoz, H.: Document Images Analysis Solutions for Digital libraries. In: International Conference on Document Image Analysis for Libraries, IEEE Computer Society (2004) 2–24
4. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4) (2006) 594
5. Agarwal, S., Awan, A.: Learning to detect objects in images via a sparse, part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(11) (2004) 1475–1490
6. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: International Conference on Computer Vision. Volume 2. (2003) 1470–1477
7. Dance, C.R., Csurka, G., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision. (2004) 1–22
8. Lowe, D.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision. Volume 2. (1999) 1150–1157

Fig. 6: Samples from the more populated of the clusters generated by clustering visual features of pages containing the word “cervo”

9. Quelhas, P., Monay, F., Odobez, J., Gatica-Perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9) (2007) 1575–1589
10. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision. (2006)
11. Lazebnik, S., Schmid, C., Ponce, J.: Affine-invariant local descriptors and neighborhood statistics for texture recognition. In: ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 649
12. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time bag of words, approximately. In: International Conference on Image and Video Retrieval. (2009)
13. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: MULTIMEDIA ’05: Proceedings of the 13th annual ACM international conference on Multimedia, New York, NY, USA, ACM (2005) 399–402
14. Snoek, C.G.M., Worring, M., Geusebroek, J.M., Koelma, D.C., Seinstra, F.J., Smeulders, A.W.M.: The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10) (2006) 1678–1689
15. Chen, Z., Wenyin, L., Zhang, F., Li, M.: Web mining for web image retrieval. J. Am. Soc. Inf. Sci. Technol. **52**(10) (2001) 831–839
16. Jing, F., Li, M., Zhang, H.J., Zhang, B.: A unified framework for image retrieval using keyword and visual features. Image Processing, IEEE Transactions on **14**(7) (July 2005) 979–989

17. Grana, C., Borghesani, D., Cucchiara, R.: Describing Texture Directions with Von Mises Distributions. In: International Conference on Pattern Recognition. (2008)
18. Hjaltason, G., Samet, H.: Properties of Embedding Methods for Similarity Searching in Metric Spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(5) (2003) 530–549
19. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**(3) (2008) 346–359