González-Díaz, I., Baz-Hormigos, C. E., Berdonces, M. & Díaz-de-María, F. (2012). A generative model for concurrent image retrieval and ROI segmentation. In *2012 10th International Workshop pon Content-Based Multimedia Indexing (CBMI)*. (pp. 90-95). IEEE

DOI: http://dx.doi.org/10.1109/CBMI.2012.6269844

A Generative Model for Concurrent Image Retrieval and ROI Segmentation

Iván González-Díaz Carlos E. Baz-Hormigos Moisés Berdonces Fernando Díaz-de-María
Universidad Carlos III de Madrid
Signal Theory and Communications Department
Avda. de la Universidad, Leganés, Madrid, Spain
{igonzalez, cebaz, mberdonces, fdiaz}@tsc.uc3m.es

Abstract

This paper proposes a probabilistic generative model that concurrently tackles the problems of image retrieval and detection of the region-of-interest (ROI). By introducing a latent variable that classifies the matches as true or false, we specifically focus on the application of geometric constrains to the keypoint matching process and the achievement of robust estimates of the geometric transformation between two images showing the same object. Our experiments in a challenging image retrieval database demonstrate that our approach outperforms the most prevalent approach for geometrically constrained matching, and compares favorably to other state-of-the-art methods. Furthermore, the proposed technique concurrently provides very good segmentations of the region of interest.

1 Introduction

This paper tackles a large-scale query-by-example image retrieval problem. This problem has been traditionally tackled using the well-known Bag-of-Words (BoW) model [1], a robust and computationally affordable procedure. This model involves the generation of a visual vocabulary, so that each local descriptor in a image is associated with the most similar visual word in the vocabulary (quantization process). Then, the resulting histogram of word occurrences is used to compute a similarity measure between every pair of images. Since the BoW model does not take into consideration the spatial distribution of the visual words in the image, several geometry-aware approaches have been proposed to refine the baseline ranking provided by the BoW model.

The last research directions on this topic can be broadly categorized into three classes: a) those aiming to improve the visual vocabulary; b) those performing a query expansion; and c) those improving the matching process with geometric considerations.

Regarding the first direction, one of the first approaches to large-scale image retrieval [7] proposed the use of very large vocabularies (up to 16.7M words) and compared the performance of several clustering techniques in terms of their ability to generate the vocabulary. In a more recent approach [8], a soft quantization in the vocabulary assignment provided a notable increase in the performance. It is also worth mentioning the approach proposed in [10], where a kernel density estimation was used to perform a unified treatment of the descriptor quantization and the matching process.

With respect to the second direction, the query expansion technique [8] used top-ranked images as new queries in order to perform various iterations of the matching process. This procedure achieves notable improvements in retrieval performance at the expense of an important increase in the computational time.

Finally, although a geometric-based verification post-processing step is the most prevalent approach to incorporate geometric information to the matching process [7] [8], there have been other proposals in the literature that efficiently take geometric constraints into account. In [4], the authors proposed a combined use of Hamming embedding and weak geometric consistency to enhance the retrieval process. In [12], geometry-preserving visual phrases were proposed that capture local and long-range spatial relations between visual words.

The inclusion of geometric constraints in the matching process lays the foundations for the concurrent segmentation of the ROI. In [7], for example, only those matches obeying a specific transformation between images will be considered as true matches. This classification between true and false matches provides useful information to generate a segmentation mask associated with the ROI in the query image.

In this paper we propose a geometric-aware matching based on a probabilistic mixture model that concurrently solves the problems of retrieval and ROI segmentation. Specifically, we present a unified framework that models several properties of the matching process between two images, such as spatial coherency, geometric transformations and visual similarity. As a result, the proposed method naturally provides a segmentation mask associated with the ROI in the query image.

From our point of view, our procedure provides several benefits: first, the segmentation of the ROI may be useful in many applications (e.g. video editing); and second, it improves the retrieval process by enforcing matches to fulfill a set of constraints. Furthermore, the obtained segmentation masks might be used as filters in new iterations of the retrieval process, so that only specific regions of the query image are searched in the reference database.

The remainder of this paper is organized as follows. In Section 2 we state the problem and present our probabilistic solution. In Section 3 we assess our proposal in comparison with several state-of-the-art approaches. Finally, in Section 4 we discuss our results and outline future lines of research.

2 A generative model for image retrieval

2.1 Problem statement

Let us assume that we have two images that represent the same concept (place, monument, object): a query image I^q and a reference image I^r. The objective is to compute a similarity measure between the two images. This process involves several steps that are described next.

The first step of the process entails generating local image descriptors, on which the matching algorithm relies. These descriptors represent the appearance of local regions computed around a set of salient points (keypoints) in each image. Since the detection of keypoints depends on the image content, we have a set of local descriptors whose size may differ from one image to another. Second, each descriptor in the query image should match another in the reference image. This step usually relies on several thresholds on the visual distance between the descriptors, so that non-likely matches are filtered out. In particular, we have used two thresholds to discard matches: a threshold on the absolute distance between two descriptors, and a threshold on the ratio between the distances with respect to the first and second neighbor. However, the values of these thresholds are conservative, so that the following steps of the matching process are still responsible for deciding on true and false matches.

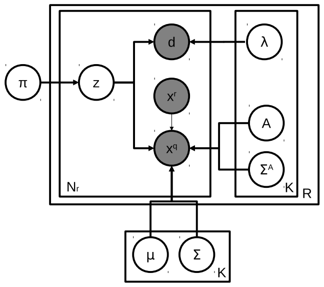

Figure 1. Proposed graphical model. Nodes represent random variables (observed-shaded, latent-unshaded); edges show dependencies among variables; and boxes refer to different instances of the same variable.

Once we have a set of N_r potential matches between the two images, we consider as if the query image had been generated by a composition process involving, on the one hand, geometrically transformed objects from the reference image and, on the other, some background regions. Therefore, we consider the generation of a query image as a mixture of K elements, $K-1$ coming from objects that also appear in any of the reference images and 1 from the background. This approach allows the query image to share specific objects or areas with a reference image while differing in others.

Each match i is then defined by three variables $\{\mathbf{x}_i^q, \mathbf{x}_i^r, d_i\}$, which denote the spatial coordinates in the query and reference images, and the matching distance, respectively. For each true match between the two images, we have made the following three assumptions:

1) A keypoint in the query image $\mathbf{x}_i^q = (x_i^q, y_i^q, 1)$ that has been matched with a keypoint in the reference image $\mathbf{x}_i^r = (x_i, y_i, 1)$ belongs to a specific object that is also present in the reference image. Therefore, there exists an object-level geometric transformation model. In this paper, due to their simplicity and linearity, we propose the use of Affine Transformations:

$$\mathbf{x}_i^q = A_{kr}\mathbf{x}_i^r \tag{1}$$

where A_{kr} is a 3x3 matrix that defines the Affine transformation that the object k undergoes from the reference image r to the query.

2) Keypoints belonging to an object k should appear at certain locations of the query image.

3) True matches tend to show lower matching distances. Therefore, we suggest to reinforce those matchings whose corresponding distances exhibit low values.

In the next subsection we describe the generative model built upon these three assumptions.

2.2 The proposed generative model

Given a query image i and a set of R reference images $\{r = 1, ..., R\}$, a set of N_r potential matches are generated between the query image and the reference image r, as described in section 2.1.

The query image is then represented as a mixture of K components: one *background* (B) component ($k=1$) that is made up of all the false matches (keypoints that cannot be matched in any reference image); and $K-1$ *foreground* (F) components ($k=2,...,K-1$), each one associated with an object in the query image that has been successfully matched in at least one reference image. It is worth noticing that each detected keypoint in the query image might generate up to R matches (one for each reference image), which are treated as independent matches.

In order to generate a probabilistic definition of each match i between the query image and a reference image r, we aim to model the probability distribution $p(\mathbf{x}_i^q, d_i|\mathbf{x}_i^r, \theta)$, where θ is the model paremeter vector. In order to build this probability model, our previous assumptions on true matches have been considered through the corresponding probabilistic distributions, which are described first in an independent manner for the sake of simplicity:

Mixture weights: Let us define z_i as a simple indicator variable that associates a keypoint i in the query image with a specific component of the mixture. Hence, $p(z_i=k)=\pi_k$ is the prior probability of the event that the keypoint i belongs to the component k of the mixture. This distribution is defined by a multinomial parameter π.

Transformation-based location: $p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr}\Sigma_{kr}^A)$ is the probability that the location $\mathbf{x}_i^q$ of the keypoint i is generated by applying the geometric transformation A_{kr} to $\mathbf{x}_i^r$, that stands for the location of the matched keypoint in the reference image. It is worth noticing that, for compactness, the index k means conditioning on $z_i=k$. For the F components in the mixture, this probability is modeled by a Gaussian distribution of mean $A_{kr}\mathbf{x}_i^r$, the expected location given the transformation, and a covariance matrix Σ_{kr}^A, which models the uncertainty of the transformation (for robustness). For the B component, we propose a uniform distribution over the spatial locations (HxW are the dimensions of the query image). As a result, the formulation of this distribution is as follows:

$$p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A)=\begin{cases} U_{x^q}(H,W) & k=1 \\ \mathcal{N}_{x^q}(A_{kr}\mathbf{x}_i^r,\Sigma_{kr}^A) & k>1 \end{cases} \tag{2}$$

Spatial coherency-based location: $p(\mathbf{x}_i^q|k,\mu_k,\Sigma_k)$ models the spatial distribution of the component k in the query image I^q. This term imposes certain spatial coherence over the components so that the keypoints associated with a particular object in the query should be located in certain area of the image. This area is defined by a Gaussian distribution with mean μ_k and covariance matrix Σ_k. For the background component, we propose a uniform distribution over the spatial locations. Hence, we define this distibution as follows:

$$p(\mathbf{x}_i^q|k,\mu_k,\Sigma_k)=\begin{cases} U_{x^q}(H,W) & k=1 \\ \mathcal{N}_{x^q}(\mu_k,\Sigma_k) & k>1 \end{cases} \tag{3}$$

Visual similarity: $p(d_i|k,\lambda_k)$ models the probability of the computed visual similarity d_i between the descriptors (matching distance), given the component. An exponential distribution is proposed for foreground components and, again, a uniform distribution is proposed for the background component, thus leading to the following definition:

$$p(d_i|k,\lambda_k)=\begin{cases} U_d(0,1) & k=1 \\ f_d(\lambda_k)=\lambda_k e^{-\lambda_k d_i}; d\geq 0 & k>1 \end{cases} \tag{4}$$

Integrating all of these distributions, the proposed model probabilistically describes each potential match $\{\mathbf{x}_i^q,\mathbf{x}_i^r,d_i\}$ by means of a finite mixture of hybrid (spatial+transformation+similarity) components. Figure 1 shows the graphical model of the proposed algorithm. Following this model, the probability of a match, defined by the variables $\mathbf{x}_i^q$ and d_i, given the potentially matched keypoint $\mathbf{x}_i^r$ in the reference image, is stated as follows:

$$\begin{aligned} p(\mathbf{x}_i^q,d_i|\mathbf{x}_i^r,\theta)=\sum_{k=1}^{K} p(z_i=k)p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A,\mu_k,\Sigma_k) \\ \cdot\, p(d_i|z_i=k,\lambda_k) \end{aligned} \tag{5}$$

where θ is the set of parameters of the model $\theta=\{\pi,A,\Sigma^A,\mu,\Sigma,\lambda\}$, and $p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A,\mu_k,\Sigma_k)$ is the location-related probability, which fuses the location distributions coming from considering the affine transformation and the spatial coherency. Specifically, this distribution has been formulated as follows: the background component has been represented as a uniform distribution; whereas the foreground components have been defined using the following factorized conditional distribution:

$$p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A,\mu_k,\Sigma_k)=\frac{\mathcal{N}_{x^q}(A_{kr}\mathbf{x}_i^r,\Sigma_{kr}^A)\mathcal{N}_{x^q}(\mu_k,\Sigma_k)}{B(\mathbf{x}_i^r)} \tag{6}$$

where $B(\mathbf{x}_i^r)$ is a normalizing factor that ensures that $p(\mathbf{x}_i^q|z_i=k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A,\mu_k,\Sigma_k)$ is a pdf over $\mathbf{x}_i^q$. Furthermore, given a set of parameters and a set of reference keypoints $\mathbf{x}_i^r$, this normalizing factor does not depend on the data $\mathbf{x}_i^q$ and can be pre-computed as:

$$\begin{aligned} B(\mathbf{x}_i^r)=&|2\pi(\Sigma_{kr}^A+\Sigma_k)|^{-\frac{1}{2}}\cdot \\ &\exp\left[-\frac{1}{2}(A_{kr}\mathbf{x}_i^r-\mu_k)^T(\Sigma_{kr}^A+\Sigma_k)^{-1}(A_{kr}\mathbf{x}_i^r-\mu_k)\right] \end{aligned} \tag{7}$$

2.3 Inference

Considering our definitions of the variables and the graph shown in Fig. 1, the log-likelihood of a corpus of R reference images can be stated as:

$$\log L \propto \sum_{r,i}^{R,N_r}\log\sum_{k=1}^{K}\pi_k p(\mathbf{x}_i^q|k,\mathbf{x}_i^r,A_{kr},\Sigma_{kr}^A,\mu_k,\Sigma_k)p(d_i|k,\lambda_k) \tag{8}$$

which is not directly optimizable due to the sum inside the logarithm. It should be noted that, for compactness, those distributions that differ for background and foreground components have been written in a general form.

Applying the Jensen's inequality we obtain a lower bound of the log-likelihood:

$$\log L \geq \sum_{r,i,k}^{R,N_r,K} \phi_{ik} \Bigg[\log \pi_k + \log p(\mathbf{x}_i^q | k, \mathbf{x}_i^r, A_{kr}, \Sigma_{kr}^A, \mu_k, \Sigma_k) + \log p(d_i | k, \lambda_k) - \log \phi_{ik} \Bigg] \quad (9)$$

where $p(z_i = k | \mathbf{x}_i^q, \theta) = \phi_{ik}$ denotes the posterior (given the data) probability of a keypoint i belonging to the component k of the mixture, and obeys $\sum_{k=1}^{K} \phi_{ik} = 1$.

We propose the use of the Expectation-Maximization algorithm to obtain the values of the parameters that maximize the lower bound of the log-likelihood (Maximum Likelihood or ML values).

EM-Algorithm: Omitting the algebra, in the E-step of the EM algorithm we compute the expected values of the posterior probabilities ϕ_{ik}:

$$\phi_{ik} \propto \begin{cases} \pi_k U_{x^q}(H,W) U_d(0,1) & k = 1 \\ \frac{\pi_k}{B(\mathbf{x}_i^r)} \mathcal{N}_{x^q}(A_{kr}\mathbf{x}_i^r, \Sigma_{kr}^A) \mathcal{N}_{x^q}(\mu_k, \Sigma_k) f_d(\lambda_{kr}) & k > 1 \end{cases} \quad (10)$$

In the M-step we compute the values of the model parameters that maximize the Likelihood:

$$\pi_k = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{N_r} \sum_{i=1}^{N_r} \phi_{ik} \quad (11)$$

$$\mu_k = \frac{\sum_{r=1}^{R} \sum_{i=1}^{N_r} \phi_{ik} \mathbf{x}_i^q}{\sum_{r=1}^{R} \sum_{i=1}^{N_r} \phi_{ik}} ; k > 1 \quad (12)$$

$$\Sigma_k = \frac{\sum_{r=1}^{R} \sum_{i=1}^{N_r} \phi_{ik} (\mathbf{x}_i^q - \mu_k)(\mathbf{x}_i^q - \mu_k)^T}{\sum_{r=1}^{R} \sum_{i=1}^{N_r} \phi_{ik}} ; k > 1 \quad (13)$$

$$A_{kr} = \left(\sum_{i=1}^{N_r} \phi_{ik} \mathbf{x}_i^q \mathbf{x}_i^{rT} \right) \left(\sum_{i=1}^{N_r} \phi_{ik} \mathbf{x}_i^r \mathbf{x}_i^{rT} \right)^{-1} ; k > 1 \quad (14)$$

$$\Sigma_{kr}^A = \frac{\sum_{i=1}^{N_r} \phi_{ik} (\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)(\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)^T}{\sum_{i=1}^{N_r} \phi_{ik}} ; k > 1 \quad (15)$$

$$\lambda_k = \frac{\sum_{r=1}^{R} \sum_{i=1}^{N_r} \phi_{ik}}{\sum_{r=1}^{R} \sum_{i=1}^{N_r} d_i \phi_{ik}} ; k > 1 \quad (16)$$

Let us recall that all the model parameters, with the exception of the mixing weights, exist only for the F components in the mixture.

2.4 Modeling irregular shapes

In section 2.2 we proposed the use of a Gaussian estimate for the spatial location of matched objects in the query image. However, although a Gaussian distribution works properly in terms of location capabilities, it obviously represents a coarse approximation of an object shape, what sometimes leads to poor segmentations of the regions-of-interest.

To overcome this issue we propose a new distribution obtained as follows: first, we perform a previous segmentation of query image based on color information [2] and obtain a set of S regions. Then, the location of each keypoint i in the query is indexed by a new indicator variable s_i that points to the region that contains the keypoint. Therefore the original distribution $p(\mathbf{x}_i^q | k, \mu_k, \Sigma_k)$ can be substituted by a new discrete distribution with parameter β_k:

$$p(s_i | k, \beta_k) = 1[s_i = j] \beta_{jk} \quad (17)$$

where $1[s_i = j]$ means that the keypoint i in the query image lies in the region j of the segmentation; and β_{jk} denotes the probability of a component k locating at a particular region j of the segmentation, and is computed as follows:

$$\beta_{jk} = \frac{\sum_{r,i}^{R,N_r} 1[s_i = j] r_{ik}}{\sum_{m=1}^{S} \sum_{r,i}^{R,N_r} 1[s_i = m] r_{ik}} \quad (18)$$

Since the regions resulting from the segmentation have more realistic shapes, a much more precise estimate of the object shape can be provided. In order to obtain simple analytical solutions, we consider this new variable s as conditionally independent of $\mathbf{x}_i^q$ given the component in the mixture. This assumption allows us to factorize their probabilities. The inclusion of this new distribution has provided slight improvements in the performance (about 2%).

3 Experiments and results

In this section we describe our assessment of the generative model on two tasks: image retrieval and automatic ROI segmentation.

We have used the Oxford Building 5K dataset [9]. This database contains 5.062 high resolution (1024x768) images showing either one of the 11 different Oxford landmarks, or other general places in Oxford. The database includes 5 queries for each landmark, each of them represented by a query image and a bounding box that locates the object of interest.

In order to establish a meaningful comparison, we have followed the setup specified in [7]. In particular, we have detected salient points in images using the affine-invariant Hessian detector [6], and described the local region around the points with a 128-dimensional SIFT descriptor [5]. Then, the authors in [7] propose a serial approach for image retrieval: first, they use a Bag-of-Words (BoW) with large vocabularies; second, they perform a re-ranking step using RANSAC [3], a fast geometric-based matching technique.

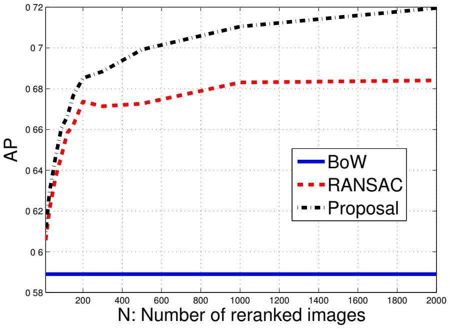

Figure 2. An image retrieval performance comparison for different numbers of reranked images

In our experiments, we have employed the same BoW with the 1M-sized hard-assigned vocabulary provided by the authors but, then, we have substituted the second step of their approach by our probabilistic generative model.

3.1 Image retrieval

For the image retrieval task, we have used the proposed generative model with $K = 2$ (each image contains only one object of interest: the landmark). In order to obtain a similarity metric between two images I^q and I^r, we have computed the ratio between the number of foreground and total (foreground + background) matches between the images. This ratio allows us to generate a ranked sequence of images, that is then evaluated using Average Precision (AP).

In order to assess the performance of the generative model, we have compared it to RANSAC [3], a well-known geometric-based technique that computes affine homographies between each pair of images. RANSAC re-ranks images according to the number of matches considered as inliers by the transformation (those matches that agree with the estimated transformation).

AP results for different numbers N of re-ranked images are shown in Fig. 2. From them, it is easy to conclude that our approach clearly outperforms a RANSAC-based reranking. The improvement is even higher as the number of re-ranked images increases since the RANSAC performance saturates for $N = 1000$ re-ranked images. In contrast, the proposed method performance keeps on improving with N, achieving the best result for $N = 2000$ images, where the influence of the previous BoW-based ranking may be almost neglected.

From our point of view, the rationale behind this improvement is the fact that our generative model jointly consider all the reference images when performing the ranking. This is an important difference with respect to a RANSAC-based approach, in which the transformation process between the query and each reference image is considered independently. We really believe that the selection of outliers considering all the reference set is more accurate than for just one pair of images, so that the quality of the inferred affine transformations is better. In addition, the other elements in the mixture model (spatial coherency and visual similarity) also help to improve the system performance. Some visual results including correctly retrieved images and also some errors are provided in Fig. 3. Images have been selected to show how our model successfully handles geometric transformations and partial occlusions.

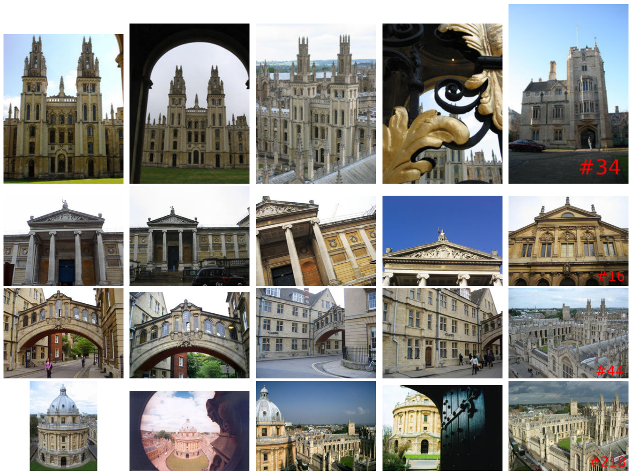

Figure 3. Image retrieval examples. Each row contains: (1) query image, (2-5) correctly ranked images (before first error), (6) first error (position in the ranking is also shown).

Finally, although it is not the main objective of this work (we aim to automatically detect the area of interest in the query image), we also present the results achieved using the bounding boxes associated with the landmarks (provided for the query images in the Oxford 5k dataset). This new procedure follows the one described in [7], and allows us to establish a meaningful comparison to other state-of-the-art techniques whose performances have been reported under the same conditions (e.g. test dataset, vocabulary size, dataset to train the vocabulary, etc.).

In particular, we have encoded the spatial coherency-based location distribution using the bounding box that points to the area of interest in every query image. The results shown in Table 1 prove that our approach successfully compares to the main state-of-the-art approaches. Since query expansion is complementary to any of these methods (including ours), we do not show results for this approach.

3.2 ROI segmentation

The proposed generative model is also able to unsupervisely discover the ROI in the query image. This region is

Table 1. A comparison of our proposal to other state-of-the-art approaches

Algorithm	AP
Hard BoW + RANSAC [7]	0.66
Soft BoW [8]	0.68
Soft BoW + RANSAC [8]	0.73
Kernel Density Estimation [11]	0.61
GVP + RANSAC [12]	0.71
Proposal	**0.74**

usually associated with an element (building, object) of special interest in the query that is successfully matched in several reference images. By labeling those points that belong to an F component, and after some morphology-based post-processing, we obtain the segmentation mask of the ROI (Fig. 4).

In order to evaluate the segmentation performance, we have computed a segmentation accuracy measurement as the ratio between correctly labeled pixels in the segmentation mask and the total number of pixels. The results are as follows: 0.40 for RANSAC and **0.67** our proposal so that, again, our method clearly outperforms the results obtained by RANSAC, the classical geometric-based method for image matching.

4 Discussion

In this paper we have proposed a generative probabilistic model that concurrently tackles the image retrieval and the ROI segmentation problems. By modeling several desired properties of the matching process, our approach successfully discovers 'true' matches between any pair of images and assigns the remainder 'false' matches to a background region. Furthermore, by considering the whole set of reference images at once, it provides a robust estimation method for discovering the actual geometric transformation undergone by the objects. Our assessment has clearly shown that this method is highly competitive with respect to the state-of-the-art image retrieval and segmentation techniques.

Our ongoing research will follow two lines: a) we will apply this method to a scenario with multiple objects ($K > 2$), in which an image shows several regions of interest that can be found in different images of the reference dataset, and b) we will further demonstrate the scalability of the method on very large datasets by splitting them into several subsets and using prior distributions that ensure that similarity values between images belonging to different subsets are comparable.

Acknowledgments This work has been partially supported by the project AFICUS, co-funded by the Spanish Ministry of Industry, Trade and Tourism, and the European Fund for Regional Development, with Ref.: TSI-020110-2009-103, and the National Grant TEC2011-26807 of the Spanish Ministry of Science and Innovation.

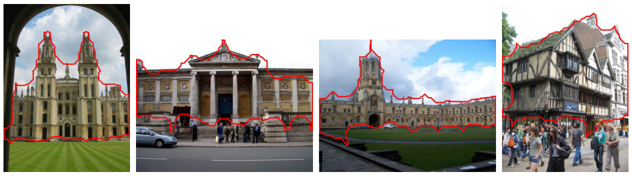

Figure 4. ROI segmentation examples.

References

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

[2] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004.

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, June 1981.

[4] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision*, 2008.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[6] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60:63–86, October 2004.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2007.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] J. Philbin and A. Zisserman. Oxford building dataset. Website. http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/.

[10] W. Tong, F. Li, T. Yang, R. Jin, and A. Jain. A kernel density based approach for large scale image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, New York, NY, USA, 2011.

[11] W. Tong, F. Li, T. Yang, R. Jin, and A. Jain. A kernel density based approach for large scale image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 28:1–28:8, New York, NY, USA, 2011. ACM.

[12] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816. IEEE, 2011.