IEEE Transactions on Multimedia (2014). 16(1), 169-183.
DOI: http://dx.doi.org/10.1109/TMM.2013.2286083

A Generative Model for Concurrent Image Retrieval and ROI Segmentation

Iván González-Díaz*, *Member, IEEE,* Carlos E. Baz-Hormigos, and Fernando Díaz-de-María, *Member, IEEE* .

***Abstract*—This paper proposes a probabilistic generative model that concurrently tackles the problems of image retrieval and region-of-interest (ROI) segmentation. Specifically, the proposed model takes into account several properties of the matching process between two objects in different images, namely: objects undergoing a geometric transformation, typical spatial location of the region of interest, and visual similarity. In this manner, our approach improves the reliability of detected true matches between any pair of images. Furthermore, by taking advantage of the links to the ROI provided by the true matches, the proposed method is able to perform a suitable ROI segmentation. Finally, the proposed method is able to work when there is more than one ROI in the query image.**

Our experiments on two challenging image retrieval datasets proved that our approach clearly outperforms the most prevalent approach for geometrically constrained matching and compares favorably to most of the state-of-the-art methods. Furthermore, the proposed technique concurrently provided very good segmentations of the ROI.

Furthermore, the capability of the proposed method to take into account several objects-of-interest was also tested on three experiments: two of them concerning image segmentation and object detection in multi-object image retrieval tasks, and another concerning multiview image retrieval. These experiments proved the ability of our approach to handle scenarios in which more than one object of interest is present in the query.

I. Introduction

This paper considers the problem of large-scale query-by-example image retrieval. This problem has been traditionally tackled using the well-known Bag-of-Words (BoW) model [1], [2], a robust and computationally affordable method. This model involves the generation of a visual vocabulary, which allows for associating each local descriptor of an image with one visual word through a quantization process. As a result, each image can be described as a histogram of word occurrences that 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 improve the baseline similarity ranking provided by the BoW model.

The last research directions on this topic can be broadly categorized into four classes: a) those aiming to improve the visual vocabulary; b) those performing a query expansion; c) those optimizing efficiency and memory resources related to image representation; and d) those improving the matching process by taking into account geometric considerations.

Authors are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Madrid. e-mail: {igonzalez,cebaz, fdiaz}@tsc.uc3m.es

Manuscript received August 9, 2012;

Regarding the first direction, several approaches in the literature have proposed the use of very large vocabularies (up to 16.7M words). Using such large vocabularies improves the discrimination capabilities of the model by giving more importance to image details. The most common approach to build vocabularies is the well-known k-means clustering algorithm [2]; however, its results usually scale poorly with the size of the vocabulary. Consequently, more recent works have moved towards either hierarchical approaches (such as trees), or approximate nearest neighbor techniques [3]. Furthermore, other authors have proposed a soft quantization approach for the representation of the local descriptors. In this manner, the actual distances between each descriptor and the closest words in the vocabulary are also taken into account. In [4], a soft quantization process was proposed that provided a notable increase in the system performance. It is also worth mentioning the approach suggested in [5], where a kernel-based density estimation was used to jointly address the quantization and the matching processes.

With respect to the second direction, query expansion techniques, [6] and [4] used top-ranked images as new queries in order to perform several iterations of the matching process. A similar idea is explored in [7], where the authors use a visual query expansion method to enhance the ranking results of an initial text-based search. These methods achieved notable improvements in retrieval performance at the expense of an important increase in the computational time.

The third direction involves obtaining compact image representations, either by reducing the number of detected features per image (see [8] for an example), or by using compact image representations, such as hashes [9], compressed representations of Fisher vectors [10], or binary vectors computed using a Hamming embedding of GIST descriptors [11]. Furthermore, we can also consider in this direction those approaches that use latent topic models to obtain compact higher-level representations, such as that presented in [12].

Finally, with respect to the use of geometric considerations, the most prevalent approach consists of a geometric-based post-processing step [3], [4]. However, other proposals have also been successful in taking geometric constraints into account. In [13], [14], the authors proposed a combined use of Hamming embedding and weak geometric consistency to improve the retrieval process. In [15], bundling features were proposed so that the large regions detected by the MSER detector [16] contained several local patches detected by the SIFT detector [17]. In doing so, these features combine the higher discrimination properties of larger regions with the repeatability and robustness to occlusions of local patches. In

[18], geometry-preserving visual phrases were proposed that capture short- and long-range spatial relations between visual words.

Moreover, the inclusion of geometric considerations in the matching process generates some spatial information that can help to detect the Region of Interest (ROI). In [3], for example, only those matches obeying a specific transformation were considered as true matches. This true/false match classification provided a segmentation mask that allowed for identifying the ROI of the query image. The model in [19] efficiently estimated an affine transformation between every two images by discretizing the transformation space, decomposing it into rotation, scaling and translation. This model was then utilized to generate spatial voting maps in the query, allowing for bounding-box based localization of the object of interest. Furthermore, in [20] a latent topic model named Geometric Latent Dirichlet Allocation was introduced that takes into consideration some geometric constraints. In particular, the topic model was used to unsupervisely model images as mixtures of topics hopefully associated with objects in the scene. Then, they evaluated the performance of this approach for image retrieval applications and, finally, demonstrated that is possible to automatically discover and segment regions-of-interest.

In this paper we propose a geometric-aware matching that relies on a probabilistic mixture model to concurrently solve both image retrieval and ROI segmentation problems. While the model proposed in [3] uses image transformations as the unique constraint in the matching process, our proposal provides a unified framework that takes into account three kind of constraints: spatial coherency between points belonging to the same object, underlying geometric transformations between matched objects, and visual similarity between matched points. As a result, the proposed method naturally provides a segmentation mask identifying the ROI in the query image.

In comparison with [20], which aims to analyze a set of images to discover objects that consistently appear in some of them, our method focuses on the matching process between a query image and a set of reference images. Furthermore, in [20], the geometric transformations are estimated at the local feature level, whereas our method models the transformation between objects appearing in two images.

From our point of view, the proposed method provides three main benefits with respect to traditional retrieval approaches: first, the segmentation of the ROI may be useful in many applications (e.g. video editing); second, it improves the retrieval process by enforcing the matches to fulfil a set of geometric constraints; and, third, using a mixture model to represent the matching process allows us to consider more than one image region being matched in a reference image. As we will show in the experimental section, it successfully addresses several problems of interest in computer vision, such as multi-object retrieval, detection and segmentation, or multiview retrieval.

The model presented in this paper was initially proposed in [21]. In this paper we present an in-depth discussion about the proposed model and a complete development of the formulation. Additionally, we provide an comprehensive assessment of the method in new scenarios of application.

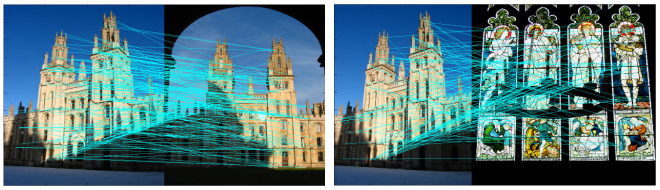


Fig. 1. Results of the matching process between a query image and two images of the reference database. As can be observed, the thresholds used were very conservative so that either correctly (left) or wrongly (right) retrieved images exhibited a relevant number of potential matches. Subsequently, the proposed generative model is in charge of filtering out false matches and providing a refined image ranking.

The remainder of this paper is organized as follows. In Section II we describe the problem to be addressed and present our probabilistic solution. In Section III we assess our proposal in comparison to several state-of-the-art approaches. Finally, in Section IV we discuss the results, draw conclusions, and outline future lines of research.

II. A GENERATIVE MODEL FOR IMAGE RETRIEVAL

Given a query image I^q and a set of R reference images $\{r = 1,...,R\}$, the objective of an image retrieval system is to compute a similarity measure between I^q and each one of the reference images in order to generate a similarity ranking. The computation of this similarity measure involves several steps that are briefly reviewed next. Salient points (keypoints) are detected for every image.

Subsequently, a descriptor is obtained for each keypoint. Each descriptor depicts the appearance of a local region around the corresponding keypoint. Then, a keypoint-level matching process is performed for each pair of images (the query and each one of the reference images). As a result, a set of N_r potential matches are generated between the query image and each reference image I^r.

This step usually relies on several thresholds on the (visual) distance between descriptors, so that non-likely matches are filtered out. Finally, these low-level matching results serve as the basis for computing the similarity measure.

Usually this matching process is prone to false negative and false positives. In this context, we propose to use a generative model of the query image that allows us to incorporate some assumptions that make the matching model more robust.

A. Model assumptions

Our model starts working from a set of preliminary matches. A first subset of potentially false matches are filtered out by means of two thresholds: one on the absolute distance between the descriptors and another on the ratio between the distances to the first and second neighbors. However, the values of these thresholds are conservative enough so that the following steps of the matching process are still responsible for deciding on true and false matches. An illustrative example of the results obtained at this stage are shown in Fig. 1.

In the proposed model, the query image is considered as the result of a composition process that combines several components coming from reference images. The use of probabilistic mixture models is very common in the computer

vision field, e.g.: Gaussian Mixture Models (GMM) [22] or Lantent Topic Models [23]. In our case, the keypoints in the query image and their associated matches are modeled as a mixture of K components, 1 associated with the image *background* (B), $k = 1$, and $K-1$ associated with *foreground* (F) areas $k = 2, ..., K-1$. Each component is defined by a set of keypoints of the query image and their matches in the reference images. The foreground components are then intended to represent objects that also appear (geometrically transformed) in any of the reference images. In contrast, the background component will consist of false matches, i.e., those keypoints in the query image that do not appear in any other image in the dataset.

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 doing so, the proposed model allows the query image to share some specific areas (objects) with a reference image and to be different in others.

More specifically, the proposed model relies on imposing some constraints to the matching process with the aim of identifying (more reliably) the true matches. These constraints are based on what we call 'the model assumptions', which are inspired by observations that generally hold for true matches.

Let us describe a match i between the query and a reference image as a three-dimensional vector $\{\mathbf{x}_i^q, \mathbf{x}_i^r, d_i\}$, where $\mathbf{x}_i^q$ denotes the spatial coordinates of the keypoint in the query, $\mathbf{x}_i^r$ denotes the corresponding coordinates in the reference image, and d_i the matching distance. The following three assumptions generally hold for true matches:

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 $\mathbf{x}_i^r = (x_i^r, y_i^r, 1)$ belongs to a specific object that is also present in the reference image. Therefore, there exists an object-level geometric transformation that maps the object of I^r into I^q. We model this mapping as an Affine transformation:

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

where A_{kr} is a 3x3 matrix that defines the geometric transformation that the object k undergoes from the reference image I_r to the query.
2) The object k tends to appear at a certain location of the image. Consequently, the keypoints belonging to this object should appear in that certain location.
3) True matches tend to produce lower matching distances. Therefore, we suggest to reinforce those matchings whose distances exhibit low values.

Relying on the previous assumptions, we have built a generative probabilistic model of the matches between I^q and I^r.

B. Proposed generative model: basic version

We will describe the model in two phases to make its understanding easier. A basic version is described in this subsection, and two extensions will be explained in the next subsection. With respect to the basic version, we describe first each part of the model (each one related to one of the previous assumptions), and then the whole model resulting from the integration of all its parts.

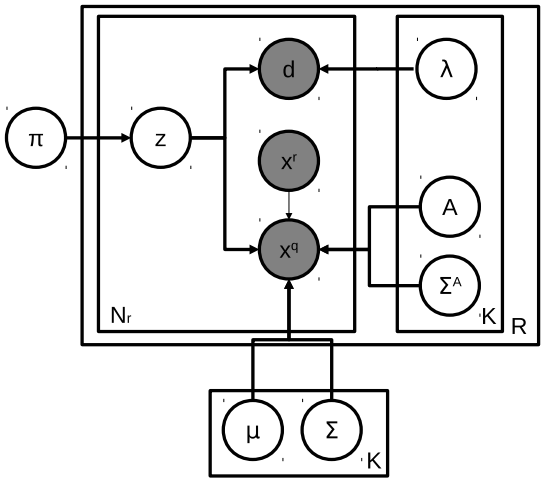


Fig. 2. Proposed graphical model. Nodes represent random variables (observed-shaded, latent-unshaded); edges show dependencies among variables; and boxes refer to different instances of the same variable.

The model assumes that I^q has been generated as the mixture of K components. The first part of the model defines the "a priori" probability of each component or, in other words, the *mixture weights*. The second part describes the location of each keypoint of the query image by means of an Affine transformation that aims to capture the geometric transformation that each object k undergoes to fit the same object in the query image (*transformation-based location*). The third part provides additional insight into the object location, but now according to the expected location of the object in the image (*Spatial consistency-based location*). Finally, the forth part considers the visual similarity itself by taking into account how likely the computed distance is, given each one of the potential objects (*visual similarity*). Let us describe each one of these parts more in-depth.

- *Mixture weights*: Let us define z_i as a simple indicator variable that associates a match i with a specific component of the mixture through a probability $p(z_i = k)$. Specifically, we have modeled these "a priori" probabilities through a multinomial distribution defined by a multinomial parameter π, i.e., $p(z_i = k) = \pi_k$ is the prior probability that the keypoint i belongs to the component k of the mixture.

- *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 associated with the match i has been generated by transforming $\mathbf{x}_i^r$ through the geometric transformation A_{kr}. It is worth noticing that, for the sake of compactness, we have used the index k meaning conditioning on $z_i = k$. For the F components of the mixture, this probability distribution is modeled by a Gaussian distribution of mean $A_{kr}\mathbf{x}_i^r$ and covariance matrix Σ_{kr}^A . The mean $A_{kr}\mathbf{x}_i^r$ represents the expected location given the transformation, while the covariance matrix Σ_{kr}^A models the uncertainty of

the transformation. For the B component, we propose a Uniform distribution over all the possible spatial locations (the MxN pixels of the query image). Integrating both terms (F and B), the complete formulation of the transformation-based location distribution is as follows:

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

- *Spatial consistency-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 takes into account the fact that a particular object (component) tends to appear consistently in a certain area of the image. Accordingly, this probability distribution is used to impose certain spatial consistency over the components. The expected location of each of the F objects 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. In summary, the proposed distribution is as follows:

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

- *Visual similarity*: the distribution $p(d_i|k,\lambda_k)$ models the probability of the computed visual similarity d_i (matching distance between descriptors), given the component k, with λ_k being the distribution parameter. An Exponential distribution is proposed for foreground components and a Uniform distribution for the background component, thus leading to the following definition:

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

Once the parts of the model have been presented separately, we explain how the individual distributions have been integrated into a generative model that describes probabilistically each potential match $\{\mathbf{x}_i^q,\mathbf{x}_i^r,d_i\}$.
Figure 2 shows the graphical model of the proposed algorithm. Following this model, the probability of a match, defined through the variables $\mathbf{x}_i^q$ and d_i, given a potentially matching keypoint $\mathbf{x}_i^r$ in the reference image, is computed as follows:

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

where θ is the parameter vector 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 information coming from considering both the affine transformation and the spatial consistency. Specifically, this final location-based distribution consists of two parts: a Uniform distribution for the background component and the following factorized conditional distribution for the foreground components:

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

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 probability density function (pdf) over $\mathbf{x}_i^q$. Furthermore, given a set of reference keypoints $\mathbf{x}_i^r$ and the parameters of the distributions, this normalizing factor does not depend on the data $\mathbf{x}_i^q$ and can be pre-computed as:

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

Inference: Considering the previous definitions of the variables and the graph shown in Fig. 2, the log-likelihood of a corpus consisting of R reference images can be stated as:

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

which is not directly optimizable due to the sum inside the logarithm.

Applying the Jensen's inequality, a lower bound of the log-likelihood is obtained:

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

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 the expected values of the posterior probabilities ϕ_{ik} are computed as follows:

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

In the M-step, the values of the model parameters that maximize the Likelihood are obtained as:

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

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

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

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

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

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

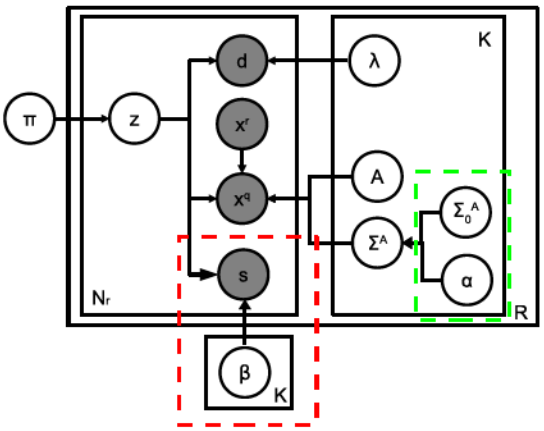


Fig. 3. Proposed extended graphical model. New elements are inside dashed boxes: in red, new segmentation-based localization (section II-C1); in green, hyperparameters of the conjugate prior distribution for the covariance matrix of the transformation (section II-C2).

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

C. Model extensions

Two model extensions are proposed for the generative model. The first aims to provide more precise segmentations of the ROI. The second adds flexibility to the transformation of the model to cope with practical issues found in real databases. Figure 3 shows the graphical representation of the extended model.

1) Improving the ROI segmentation: We have proposed the use of a Gaussian distribution for modeling the spatial location of matched objects in the query image. However, although a Gaussian distribution works properly in terms of location, it obviously provides a coarse approximation of the object shape, what leads to imprecise segmentations of the region-of-interest.

With the aim of providing more precise ROI segmentations, we suggest a new probability distribution that relies on a previous segmentation of the query image. Since the regions resulting from the segmentation have more realistic shapes, a much more precise estimation of the object shape can be provided. In particular, the query image is segmented based on color information [24] and a set of S regions are obtained. Then, the location of the keypoint associated with each match i is indexed by an indicator variable s_i that points to the region that contains the keypoint. And finally, the original distribution $p(\mathbf{x}_i^q|k,\mu_k,\Sigma_k)$ is substituted by a new discrete distribution with parameter β_k:

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

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

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

and substitutes previous equations (12) and (13) of the basic version of the model concerning the Gaussian distribution related to the spatial consistency term.

In addition, in order to obtain a simple analytical solution, we consider the new variable s as conditionally independent of $\mathbf{x}_i^q$ given k. This assumption allows us to factorize their probabilities.

It is also worth commenting that, apart from providing better ROI segmentations, the inclusion of this new distribution produced slight improvements in the overall performance, as will be shown in the experimental section.

2) Managing the flexibility of geometric transformations: Using a Gaussian distribution to model the transformation-based location provides certain degree of flexibility since the covariance matrix Σ_{kr}^A allows us to relax the geometric constraints imposed by transformation when necessary. However, as we found in our experiments, it would be desirable to have control on the flexibility of the model, so that it could fit to either more or less similarity demanding tasks.

With this requirement in mind, we propose to introduce a regularizing prior over the covariance matrix Σ_{kr}^A. In this manner, although small covariance values usually turn out to be more appropriate, the method is endowed with a means that allows us to control on the flexibility of the transformations. In particular, we have considered a Wishart distribution of m_{kr} degrees of freedom, which is the conjugate prior of the inverse of the covariance matrix:

$$p(\Sigma_{kr}^A|\Sigma_{0kr}^A, m_{kr}) \propto |\Sigma_{kr}^A|^{\frac{m_{kr}-D-1}{2}} \exp\left(-\frac{1}{2}Tr\left(\Sigma_{0kr}^A \Sigma_{kr}^{A\,-1}\right)\right) \tag{19}$$

where D is the dimension of the data ($D = 2$, in our case), $Tr(\cdot)$ stands for the Trace operator, and Σ_{0kr}^A and m_{kr} are hyperparameters.

The inclusion of these priors, which only apply to the F components of the model, leads to the following Maximum a Posteriori optimization:

$$\log MAP = \log L + \sum_{k=2}^{K}\sum_{r=1}^{R} \log p(\Sigma_{kr}^A|\Sigma_{0kr}^A, m_{kr}) \tag{20}$$

where $\log L$ has been previously defined in eq. (8).

Consequently, the update expression of the corresponding covariance matrix should change accordingly:

$$\Sigma_{kr}^A = \frac{\sum_{i=1}^{N_r} \phi_{ik}(\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)(\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)^T + \Sigma_{0kr}^A}{\sum_{i=1}^{N_r} \phi_{ik} + m_{kr} - D} \tag{21}$$

Now, following the approach in [25], appropriate values for the hyperparameters have been chosen so that the covariance matrix update equation takes the desired form. In particular, the hyperparameters are chosen as follows:

$$\Sigma_{0kr}^A = \alpha_{kr}\bar{\Sigma}_{0r}^A \tag{22}$$

$$m_{kr} = \alpha_{kr} + D \tag{23}$$

where $\bar{\Sigma}_{0r}^A$ represents the prior of the covariance matrix (the same for all F objects), and α_{kr} manages the balance between

the free term and $\bar{\Sigma}^A_{0kr}$. In addition, α_{kr} is used for setting the value of m_{kr} so that the final update equation for the covariance looks like:

$$\Sigma^A_{kr} = \frac{\sum_{i=1}^{N_r} \phi_{ik}(\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)(\mathbf{x}_i^q - A_{kr}\mathbf{x}_i^r)^T + \alpha_{kr}\bar{\Sigma}^A_{0kr}}{\sum_{i=1}^{N_r} \phi_{ik} + \alpha_{kr}} \quad (24)$$

which substitutes equation (15) of the basic version of the model. The rest of the update equations remain unaltered as described in subsection II-B.

In our experiments, in the absence of information we consider the same value of α_{kr} for all the foreground objects in a scene; in particular:

$$\alpha_{kr} = \alpha_r = C\frac{N_r}{K} \quad (25)$$

where C is a constant that has been set to $C = 10$, and N_r and K had been previously defined as the number of matched points in the reference image r and the number of components in the mixture, respectively. Additionally, the prior value of the covariance has been empirically set to $\bar{\Sigma}^A_{0kr} = 10^{-3}\mathrm{I}$.

D. Generating the ROI

The proposed generative model is also able to unsupervisely discover the ROI in the query image. This region is usually associated with an element (building, object) of special interest in the query that is successfully matched in several reference images. The process followed to obtain the ROI segmentation can be summarized as follows (for simplicity, we describe the procedure for K=2):

1) Generate a binary mask by labeling those points that belong to the F component.
2) Perform an opening morphological operation over the binary mask using a disk-type structuring element (we use radius of 50 pixels in our experiments).
3) After re-labeling the generated connected components, remove those ones whose size is relatively small (smaller than half the size of the largest one, in our experiments).

Some visual examples of the generated ROIs can be found in the experimental section (Figure 7).

E. Automatic selection of the number of model components K

In this section we propose a simple method to automatically determine the value of K based on the query image content. It consists in a splitting approach that iterates adding new components to the mixture when necessary. In the following paragraphs we describe the proposal in detail.

When K is lower than the actual number of objects, each component will represent more than one object. Therefore, we decide to associate each component with the main object (among those represented by that component) and then look for new ones. We start running our model with $K = 2$ and, at the end of each iteration, we look for new components by following this process:

1) For each foreground component k, we select the reference image that best represents it. To that end, for each reference r we compute the accumulated posterior probability χ_{rk} of the matches associated with that component as follows:

$$\chi_{rk} = \sum_{i=1}^{N_r} \phi_{rk} \quad (26)$$

and then select that reference that produces the highest accumulated probability.
2) For each pair *foreground component-best reference image*, we generate the ROI of the main object (as described in previous subsection) and save the corresponding segmentation for future testing of potential new components. As a result of these two first steps, we achieve a segmentation of the main object associated with each foreground component in the mixture. This segmentation, called the 'accumulated segmentation', will be used in the fourth step of the algorithm.
3) For each foreground component, steps 1) and 2) are repeated for every reference image, generating candidate regions that are considered as potential new components.
4) For each potential candidate component, three features are extracted: a) % of overlapping with the segmentation of the main object of this component; b) the relative size of the region with respect to image dimensions; and c) the density of points inside the region (a ratio between the number of matches and the area of the region). The decision whether or not to accept the region as a new component is made using a linear classifier relying on these features as inputs. If accepted, the associated mask is included to the accumulated segmentation.

The linear classifier was trained on some manually generated data.

It is worth noticing that when a reference provides an object that actually coincides with the main object already considered, the degree of segmentation overlapping will be high. Hence, overlapping becomes the main measure to make decisions in this process. Nevertheless, the relative size and density are still useful complementary variables that help to avoid adding new components associated with either very small regions or sparse sets of points.

In addition, it is also worth mentioning how once the algorithm decides that a candidate is is being incorporated to the mixture, its region is added to the accumulated segmentation, thus preventing from the addition of two candidate regions with high degree of overlapping.

A visual example of the process is illustrated in Figure 4.

III. Experiments and results

In this section we describe the assessment of the proposed generative model.

We have used various datasets for our experiments:

- The Oxford Building 5K dataset [26]: a database that contains 5.062 high resolution images (1024x768) showing either one of the Oxford landmarks (the dataset contains 11 landmarks), or other places in Oxford. The database includes 5 queries for each landmark (55 queries in total), each of them including a bounding box that locates the object of interest.

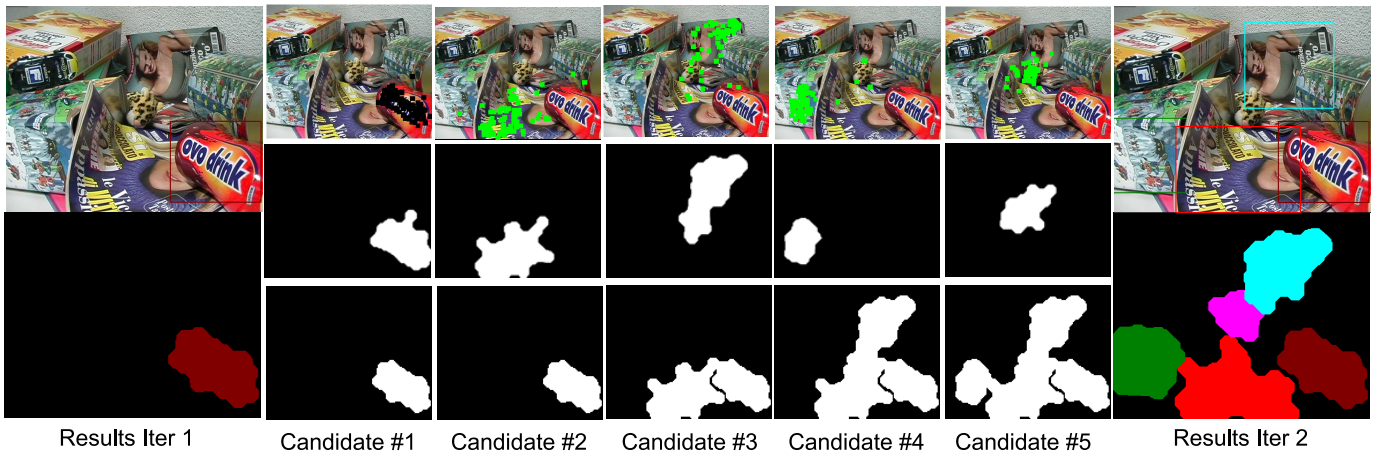


Fig. 4. A visual example of the iterative process to automatically set the value of K. First column: Output of the (K=2) first iteration of our algorithm, with bounding boxes on detected objects (top), and segmentations (bottom). Each color represents a particular object category. Columns 2-6: detected regions candidate to become new components. For each particular case we show: the points belonging to the component (top), the associated candidate mask (middle), and the accumulated segmentations (bottom). If a candidate is accepted, points are shown in green, otherwise in black. Column 7: output of the (K=5) second iteration of the method.

- The INRIA Holidays dataset [13]: a dataset with 1491 personal holiday photos, in which several transformations or artifacts can be evaluated: rotations, viewpoint, illumination changes, etc. This dataset contains 500 image groups or scenes, with the first image of each group being the query (500 queries).
- The ETHZ toys dataset [27]: a dataset for specific object recognition. It contains 40 images modeling the appearance of 9 specific objects, and 23 test images in which several objects appear under different scale, viewpoint, and occlusion conditions. This dataset, although small to properly assess the performance of the proposed system in an image retrieval problem, allows us to demonstrate the model capability of handling more than one object of interest.
- The RGB-D object dataset [28]: a large image dataset of 300 common household objects that was conceived as a 3D dataset and contains 2D images and depth maps of every object (which was placed on a turntable to capture images for one whole rotation). In our experiments we have only used 2D images of the objects viewed from different angles, discarding depth information. We have considered images containing isolated objects as reference dataset, and used the so-called RGB-D Scenes Dataset as test/query dataset. The RGB-D scenes dataset consists of video recorded at different locations containing more than one object of interest per scene. Let us note that many of the objects in this dataset are very homogeneous and therefore, not suitable for a salient feature-based recognition such as the one used in this paper. Hence, we have restricted our experiments to a subset composed of the 10 most-textured object categories and, therefore, more salient features (e.g. cereals box, food box, cap, notebook, etc.). Then, only those frames in the scenes dataset that contained more than one object belonging to the considered categories were selected as queries. This process led to a query set containing 54 images, and a reference dataset of 4314 images, which are figures very similar to those of the Oxford dataset.

Our experiments have been divided into two blocks. First, we assessed our model for image retrieval and automatic ROI segmentation when each image contains only one object of interest, following the conventional experimental protocols for the Oxford Building and Holidays datasets, respectively. In both datasets, since each image contains only one object of interest, we have selected $K = 2$, i.e., one foreground object. Second, we proved the usefulness of the proposed model for $K > 2$ by solving three tasks: a multi-object category-based segmentation in the ETHZ toys dataset, a multi-object detection task in the RGB-D object dataset, and a multiview object retrieval task on the Oxford Building dataset.

In order to establish a meaningful comparison, we followed the feature extraction protocol described in [3]. In particular, we detected salient points using the affine-invariant Hessian detector [29]. Then, we described the local region around these keypoints with a 128-dimensional SIFT descriptor [17]. Subsequently, a Bag-of-Words (BoW) model was used; in particular, we employed the same BoW as in [3] with the 1M-sized hard-assigned vocabulary. Finally, the authors of [3] performed a re-ranking step using RANSAC [30], an efficient geometric-based matching technique, which we substituted by our probabilistic generative model. Furthermore, in order to limit the complexity of the overall process, we used a Fast Nearest Neighbour search [31].

A. Image retrieval and ROI segmentation with one object of interest

As a similarity metric between two images I^q (the query) and I^r (each of the reference images), we propose a new measure χ_r that can be expressed as follows:

$$\chi_r = \sum_{i=1}^{N_r} \phi_{r2}, \tag{27}$$

TABLE I
SUBSYSTEM VALIDATION: PERFORMANCE EVALUATION IN TERMS OF AP OF DIFFERENT VERSIONS OF THE PROPOSED GENERATIVE MODEL FOR THE RE-RAKING OF 300 IMAGES.

Version	AP
BM	0.6640
BM w/o spatial consistency	0.6296
BM w/o visual similarity	0.6612
BM w/o affine transformation	0.6578
EM-improved transformation	0.6810
EM-improved transformation & segmentation	0.6929

i.e., χ_r is the sum of the posterior probabilities ϕ_{rk} of the points belonging to the foreground component ($k = 2$).

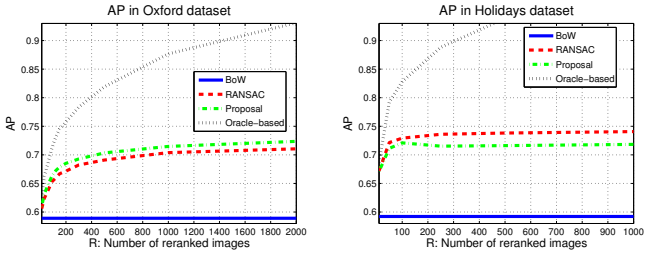
This new measure χ_r allowed us to generate a ranked sequence of images that was then evaluated in terms of Average Precision (AP), a measure that has been extensively used to assess information retrieval systems. AP requires a set of ranked images as system output and combines both recall- and precision-related factors in a single measure, which is also sensitive to the complete ranking. For a detailed description of the AP measure the reader is referred to [32].

1) Validating the elements of the model: With the purpose of validating each one of the elements of the proposed model, we assessed separately their influence on the complete system performance on the Oxford Building dataset. In Table I, we show the results achieved by different versions of the proposed system resulting from disabling the operation of each subsystem separately. The notation used in the table for denoting the resulting systems is described next:

- *BM*: Basic version of our proposal, as described in Section. II-B.
- *BM w/o spatial consistency*: Basic version with the spatial consistency-based location disabled.
- *BM w/o visual similarity*: Basic version with the visual similarity disabled.
- *BM w/o affine transformation*: Basic version with the transformation-based location disabled.
- *EM-improved transformation*: Extended version with improved transformation, as described in subsection II-C2.
- *EM-improved transformation & segmentation*: Complete model, including improved transformation and segmentaion, as described in subsections II-C2 and II-C1, respectively.

The results of this experiment are shown in Table I. In what concerns to the basic model, the *spatial consistency* turns out to be most significant element, showing notably higher relevance than the other two distributions. However, we claim that they are still important and should not be removed from the model. On the one hand, in what concerns the *visual similarity*, it has been included in the final model due to two reasons: a) it actually produces a slight improvement on the performance, and 2) the computational complexity of this element is, by far, much lower than the other two (the update equation is linear and does not involve any complex matrix manipulation such as inversions).

With respect to the *transformation*, although its contribution was minor in the basic model, the version of the system that adds a regularizing prior over the transformation-based



(a) (b)

Fig. 5. An image retrieval performance comparison for different numbers of re-ranked images. a) Oxford dataset, (b) Holidays dataset

location produced a great improvement of the results, what means that the covariance matrix Σ^A was not initially restrictive enough to deal with our task.

In addition, we can see that the combination of the regularization term with the other proposed extension (improvement of the ROI segmentation) shows a notable impact on the system performance. Hence, from now on, the complete version of our model will be used in the rest of experiments.

Finally, it should be mentioned that these results were obtained for the re-ranking of 300 images, and that the performance increments due to every subsystem or extension tend to grow with the number of re-ranked images R (due to the increasing effect of our model on the final ranking).

2) Image retrieval: In order to assess the performance of the proposed generative model, we have compared it to RANSAC [30], a well-known geometric-based technique that robustly computes transformation matrices (between each pair of images) in the presence of outliers. In our implementation, RANSAC re-ranked images according to the number of matches considered as inliers for the corresponding affine transformation, i.e., according to those matches that fitted the estimated transformation.

Results in terms of AP for different numbers of re-ranked images R are shown in Fig. 5(a) and 5(b) for the Oxford 5K and Holidays datasets, respectively. Results of an oracle-based re-ranking process are also included as an upper-bound performance limit of re-ranking for each value of R.

For Oxford 5K dataset, it is worth noticing that performance keeps increasing up to $R = 2000$ images, where the influence of the previous BoW-based ranking might be almost neglected (the oracle-based approach achieves an AP=0.93). Additionally, from these results, it is easy to conclude that our approach outperforms the RANSAC-based re-ranking. As we can see from the figure, a relative improvement of a 2% is achieved by our method with independence on the number of re-ranked images, what proves the robustness of our approach when the proportion of positive images (images showing the landmark of the query) decreases.

In our opinion, this improvement is due to two main reasons: first, the proposed generative model combines several elements, some of which are not considered in the RANSAC-based approach, in particular: *spatial consistency*, *visual similarity*, and the extension concerning *improved segmentation*. Second, our generative model jointly considers all the reference images when performing the ranking. This is an important difference with respect to the RANSAC-based approach,

Fig. 6. 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).

in which the transformation estimation between the query and each reference image is addressed independently. Hence. the outlier detection process should be more accurate when considering the complete reference set and, consequently, the inferred affine transformation should be better.

Similar conclusions can be drawn from the result in Holidays dataset. In this case, our approach consistenly achieves improvements with respect to RANSAC. Futhermore, these relative improvements even grow with the size of the re-ranked set which means that our approach better handle situations in which the number of relevant images decreases. However, we have found that for this dataset the performance saturates for quite a low value of R (R=250). The rationale behind is that, in this dataset, there is just a short number of relevant images (in general, between 1-3) per query. This issue gives very much influence to the quality of the previous ranking achieved by the BoW.

Some visual results including correctly retrieved images and also some errors are provided in Fig. 6 for the Oxford 5k dataset. Images have been selected to show how our model successfully handles geometric transformations and partial occlusions.

Next, in Tables II and III, we show a comparison of our proposal to other state-of-the-art techniques whose performances were reported under this same conditions. Let us note that two methods reported in the literature were not considered in the comparison: a) query expansion (also known as k-nn reranking by several authors) since it is a complementary technique to all the compared methods (including ours), and it would contribute to improve all the results in similar proportions, and b) orientations priors or manually rotation of the images, since they are either completely database dependent or require human manual effort to be run.

TABLE II
A COMPARISON OF OUR PROPOSAL TO OTHER STATE-OF-THE-ART APPROACHES IN OXFORD DATASET.

Algorithm	AP
Hard BoW + RANSAC [3]	0.66
Soft BoW [4]	0.68
Soft BoW + RANSAC [4]	0.73
Kernel Density Estimation [33]	0.61
GVP + RANSAC [18]	0.71
[19]	**0.75**
Proposal	**0.75**

TABLE III
A COMPARISON OF OUR PROPOSAL TO OTHER STATE-OF-THE-ART APPROACHES IN THE HOLIDAYS DATASET

Algorithm	AP
[13]	0.75
[19]	0.76
[14]	**0.78**
Proposal	0.76

For the Oxford 5K dataset, although it is not the main objective of this work (we aim to automatically detect the area of interest in the query image), we present results achieved using the bounding boxes associated with the landmarks (as proposed in the experimental protocol described in [3]). In our model, the location information coming from the bounding box was incorporated on the spatial coherency-based location

TABLE IV
SEGMENTATION ACCURACY (%)

Algorithm	Acc
RANSAC	61.8
Our method	68.2

distribution.

For the Holidays dataset, we have found that, due to the short number of relevant images per query, the influence of the re-ranking methods is not so notable and, in addition, that they perform better with short lists to re-rank. This gives much importance to the initial ranked list which, in our case and due to the low performance of classical BoW, was generated by simply counting the number of matches between images. The number of re-ranked images is $R = 10$ for Hilidays dataset.

The results prove that our approach successfully compares to the main state-of-the-art approaches in both datasets. For the Oxford 5k dataset, our method achieves the best results among all compared techniques, whereas for the Holidays dataset, our performance is very close to the best performing method. As we have already mentioned this dataset only contains between 1-3 positives per query, what gives more importance to the previous ranking.

3) ROI segmentation: The proposed generative model is also able to unsupervisely discover the ROI in the query image. This region is usually associated with an element (building, object) of special interest in the query that is successfully matched in several reference images. The process followed to obtain the ROI segmentation was described in subsection II-D.

Fig. 7 illustrates some ROI segmentation results.

Furthermore, since the same segmentation approach can be also applied to RANSAC-based reference system, we conducted a comparative experiment to assess the segmentation performance. To that end, we have manually segmented the foreground objects of the 55 queries in the database. The resulting binary masks are available online[1]. Specifically, we have computed a segmentation accuracy measurement as the percentage of correctly labeled pixels over the total number of pixels. The results are shown in Table IV and, as it can be seen, our method clearly outperforms the results obtained by RANSAC, the classical geometric-based method for image matching.

B. Handling more than one object of interest

In previous experiments, since there was only one landmark per query image, we fixed to $K = 2$ the number of foreground components in our model. Nevertheless, the proposed model provides the capability of dealing with more than one foreground component. In order to assess this capability we have conducted three different experiments, namely: a) a multi-class category-based segmentation experiment on the ETHZ toys dataset; b) an object detection experiment on the RGB-D object dataset; and c) a multiview object retrieval experiment on the Oxford building dataset.

[1]http://www.tsc.uc3m.es/~igonzalez/maskqueries.zip

TABLE V
MULTI-CLASS SEGMENTATION RESULTS ON ETHZ TOYS DATASET

Algorithm	Acc
RANSAC	72.4
Proposed $K = 2$	73.6
Proposed $K = 3$	72.4
Proposed $K = 4$	70.7
Proposed K_{opt}	77.9
Proposed K_{aut}	76.4

1) Multi-class category-based segmentation experiment on the ETHZ toys dataset: Concerning this first experiment, it is worth mentioning that we are interested in assessing our algorithm in a category-based segmentation problem, rather than in assessing individual detections (as it is more usual for this dataset). In particular, we do not only consider the image partition into regions, but also the correct labeling of each region with the corresponding object. In this manner, the fact of addressing a multi-class problem allows us to evaluate the capability of our model to work with $K > 2$. Hence, we aim not only to detect the presence of an object in an image, but also to properly segment it. Furthermore, to be consistent with our unsupervised approach, our method is not aware of the category represented by each reference image. Just at the end, for evaluation purposes, each component in the mixture is labeled with the object-category of the most likely reference image.

Results of this experiment are shown in Table V in terms of pixel-wise segmentation accuracy. This table provides the results achieved by the RANSAC-based reference method and several versions of the proposed method; in particular:

- RANSAC: to obtain the best possible segmentation, we have properly mixed the individual segmentations provided by RANSAC. Specifically, since RANSAC considers an individual matching problem between a query and each reference image, when two objects were detected in the same pixel, the class corresponding to the image with more 'inlier'-type matches was selected.
- Proposed method with a fixed K: for each case, a maximum of K objects can be detected per image.
- Proposed method with the optimal K value (K_{opt}): in order to evaluate the upper limit of the algorithm, we have run our model for $K = 1...10$ and then, for each particular test image, we have selected the optimal result "a posteriori".
- Proposed method with an automatically selected K value (K_{aut}): the proposed algorithm using the simple iterative method described in section II-E to automatically select K. As mentioned in that section, the proposed method relies on a linear classifier that, in our experiments, have been trained using data generated from the first image and evaluated on the whole dataset.

As can be observed, setting predetermined value of K does not turn to be optimal for this dataset since each image contains a different number of objects. Furthermore, since the test images contain 1-3 objects (except for the first one, that contains 9 objects), lower values of K perform better.

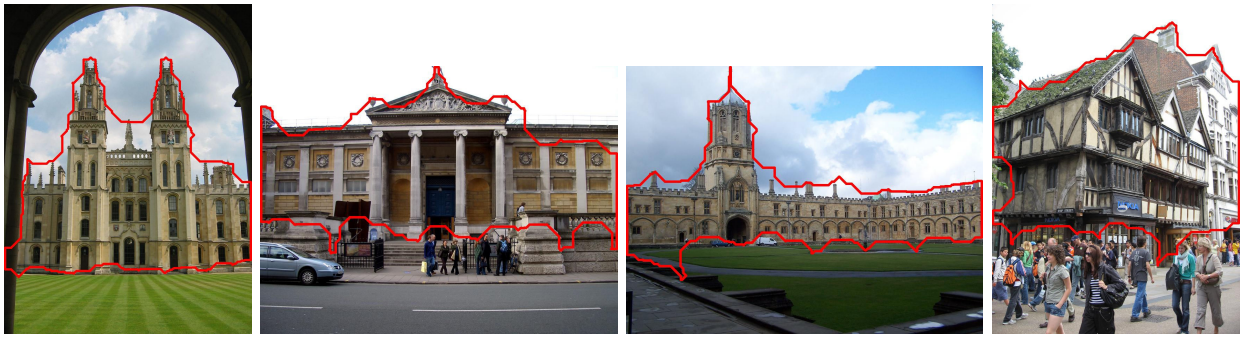


Fig. 7. ROI segmentation examples.

TABLE VI
OBJECT DETECTION RESULTS ON THE RGB-D OBJECT DATASET

Algorithm	F-score
RANSAC	45.9
Proposed $K=2$	47.4
Proposed $K=3$	43.6
Proposed $K=4$	29.8
Proposed K_{opt}	61.1
Proposed K_{aut}	50.0

As expected, the version of the proposed algorithm using the optimal K produces an upper bound of the algorithm performance. However, the results obtained by the automatic method are quite close, what demonstrates that it is possible to automatically select a suitable value for K. Again, our proposal provides better performance than RANSAC due to the fact that it concurrently considers all the foreground objects rather than solving individual matching problems between queries and models.

2) Object detection experiment on the RGB-D object dataset: In this dataset, precise pixel-wise ground-truth object segmentations are not available, but bounding boxes locating the objects are available instead. Therefore, we have relied on these bounding boxes to provide object detection results. To this end, our algorithm aims to provide one bounding box per foreground component corresponding to a detected object. Then, to assess the system performance, we have considered as detections only those for which the relative overlap between ground truth and retrieved bounding boxes exceeds 0.5.

Results of this experiment are shown in Table VI in terms of detection F-score. The same methods considered in the previous experiments have been compared again for this second dataset. The conclusions are very similar to those found in the previous task; in particular, the proposed automatic muti-object retrieval method has shown to overcome RANSAC and to achieve the closest results to the optimal case (K_{opt}).

3) Multiview object retrieval experiment on the Oxford building dataset: A multiview object retrieval task, in which several views of the same object are provided to enhance the system performance, could be another interesting application of our model. With that purpose, we have manually generated 11 new query images that contain five different views of each of the Oxford landmarks. Since the Oxford Building dataset contains 55 queries, 5 corresponding to each landmark, we have concatenated those 5 images to end up with a composite query in which several views of the building are provided to improve the retrieval process. Figure 8 shows some illustrative examples of the generated multiview query images.

For each query image we have run the proposed generative model with $K=6$ components, one of them representing background regions in images, and the others modeling foreground components. Our experiments showed how the inference process led to well defined foreground components, each of them associated with one of the building views and a background component covering those areas that cannot be consistently matched in the reference images (see Figure for an illustrative example). In particular, we have measured that the 96.1% of the points belonging to each FG component are associated with the same building view, what supports our observations.

In order to establish a ranking of the reference images, a combined similarity measure χ_r was computed as the sum of the posterior probabilities ϕ_{ik} for all the foreground components ($k>1$):

$$\chi_r = \sum_{i=1}^{N_r} \sum_{k=2}^{K} \phi_{rk} \tag{28}$$

With the purpose of assessing the quality of this new ranking obtained using $K=6$, we have compared this new result to that of our system using $K=2$. Specifically, we have considered the 5 individual rankings for each of the 5 query images of each landmark and $K=2$ and, then, averaged these results. In this case, the average provided better performance than other simple fusion operators such as *max*.

In table VII we show the obtained results for two different numbers of re-ranked images ($R=200$ and $R=2000$). From these results, we can conclude that the proposed multi-component method attains higher performance by jointly considering all the available views of the landmark. A query-by-query analysis shows that our method is particularly effective when the views included in the composite query are very diverse; for example, in the case of the bottom of Figure 8 our proposal achieved an AP increase of 0.18 with respect to the reference (due to the high variability among query images).

A more in-depth analysis of the results demonstrates that our method tends to assign each reference image to a particular component in the mixture, which is in turn associated with the most similar image in the composite query. This fact

Fig. 8. Three examples of multiview query images.

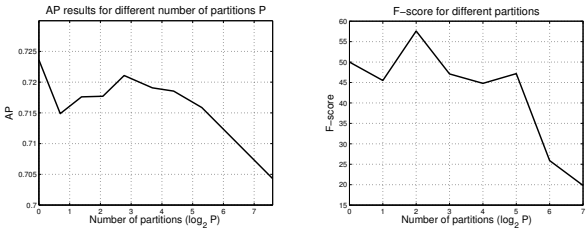
Fig. 9. An illustrative example of the segmentations of the foreground components generated by our algorithm for $K = 6$ (i.e., for 5 foreground components).

TABLE VII
AP RESULTS FOR THE MULTIVIEW OBJECT RETRIEVAL

Algorithm	R=200	R=2000
Proposed - $K = 2$	0.74	0.78
Proposed - $K = 6$	0.78	0.82

TABLE VIII
AVERAGE COMPUTATION TIME AND STANDARD DEVIATION PER QUERY WITH OUR APPROACH AND RANSAC

Algorithm	R=100	R=1000
RANSAC	0.58 ± 0.34 secs	4.71 ± 1.88 secs
Proposed	1.36 ± 1.12 secs	8.81 ± 3.22 secs



(a) (b)

Fig. 10. Illustration of the scalability properties of the proposed approach. The datasets have been divided into several numbers of partitions P for computational purposes. Then, the system performance has been represented as a function of the number of partitions. a) AP for various values of P on the Oxford Building dataset for the $K = 2$ image retrieval task and a re-ranking of a total of $R = 2000$ images; and b) F-score for various values of P on the RGB-D dataset for the $K > 2$ multi-object detection task. In both cases, for better visualization, a log-scale of the P axis has been used

can be interpreted as a nice consequence of the use of just one transformation matrix $A_k r$, given a component k, between the composite query and a reference image. This unique transformation matrix avoids associating points belonging to different views of the landmark with the same foreground component in the mixture.

C. On the computational complexity and the scalability of the model

In what concerns to the computational complexity of our model, we have measured the computation time of our approach and compared it with that of the RANSAC approach. Let us note that both techniques were implemented in a high level programming language (Matlab), they were run with single threading, with the particular setup used in our experiments, and without any specific optimization. In addition, in order to provide a fair comparison, just the time devoted to

the geometric re-ranking was measured (we did not consider other previous tasks such as descriptor computation, keypoints matching process, reading/writing files, etc.). The total time devoted to the re-ranking process per query is included in Table VIII for two different numbers of re-ranked images (N=100,1000).

From these results, we can see that both RANSAC and our method show similar behavior: the execution time shows notable variations (standard deviation) depending on the number of detected points in the query, and approximately grows linearly with the number of re-ranked images.

Furthermore, running our method requires approximately twice the time than the RANSAC approach, due to the extra elements we are including in the generative model that are not taken into account in RANSAC. However, executions times are still comparable and, as we have seen in the experimental section, our method shows many advantages over the RANSAC approach.

Considering now the scalability of the proposed model to deal with very large image data bases, our main concern turned to be the memory consumption. Since our proposal jointly processes all the reference images, the memory consumption increases with the number of references images R. Therefore, managing all the references jointly becomes impractical. Alternatively, a sub-optimal implementation of the method can be made by splitting the reference image dataset into P subsets that can be successfully handled by our model.

Since better performance was expected from a low number P of large subsets, as long as many images are concurrently handled, it was worthy to assess the sensibility of the performance to this parameter and, consequently, the feasibility of the method for working with very large databases. With this purpose, we conducted two series of experiments: a) in the $K = 2$ image retrieval scenario with the Oxford building dataset; and b) in the $K > 2$ multi-object detection scenario with the RGB-D dataset. Our objective was to test sub-optimal implementations using different numbers of partitions P. The results of these experiments are shown in Fig. 10, where the system performance is depicted as a function of the number of partitions into which the dataset was divided.

As can be observed, for the $K = 2$ (Fig. 10(a)) image retrieval task, the performance of the system keeps being high until P goes beyond 32 ($log_2 P = 5$) partitions. This is a nice result since if $P = 16$ is used, which corresponds with a partition size of 125 images, very good performance is achieved at the same time that memory consumption is minimized.

For the $K > 2$ multi-object detection experiment, it should be noticed that the fact dividing the dataset into various subsets entails both pros and cons: on the one hand, it is clear that, as in the previous case, working on small subsets reduces the potential benefits of considering the whole dataset (what normally would lead to a better discrimination between true and false matches). Furthermore, if the subsets became so small that they would not contain at least one relevant sample of each object of interest, the output of the proposed approach would become unstable (this situation is obviously more likely in the multi-object scenario). On the other hand, by doing parallel executions on different subsets, we can take advantage of the various independent detections at the detection fusion stage; in particular, given a number of partitions P, we fuse those detections exhibiting high degree of overlap and remove those ones that do not appear in a predefined number of subsets (we have heuristically set this value to 20% of the subsets). In this manner, many false detections are filtered out according to their lack of consistency along various subsets.

The results shown in Fig. 10(b) support these ideas: we can see how performance improves for partitions with just a few number of subsets (e.g. for $P = 4$, what corresponds to subsets of 1000 images), and can be still considered good enough for larger values in the range $P = 8...32$ (subsets of 500-134 images). However, when P is too large ($P = 64$), and therefore the subsets are too small, the lack of relevant examples in many subsets makes our approach unstable, and thus penalizes its performance.

Hence, in general, we can conclude that, our approach works fine and may work even better when the whole set of data is divided into a small number of subsets. The rationale behind this result is that the fusion stage allows us to filter out many erroneous results when they do not appear consistently in various executions. Just when we bring the scalability to the extreme, i.e., when the number of subsets is too large, the performance of the approach decreases due to the lack of relevant examples in the reference subsets. Since, in our experiments, we have obtained good results with subsets of reasonable sizes (125 images for simple $K = 2$ image retrieval, or 134 images for a $K > 2$ multi-object retrieval), it can be concluded that our proposal would be scalable for very large scale image-retrieval tasks.

IV. Discussion

In this paper we have proposed a generative probabilistic model that concurrently tackles image retrieval and ROI segmentation problems. By jointly modeling several properties of true matches, namely: objects undergoing a geometric transformation, typical spatial location of the region of interest, and visual similarity, our approach improves the reliability of detected true matches between any pair of images. Furthermore, the proposed method associates the true matches with any of the considered foreground components in the image and assigns the rest of the matches to a background region, what allows it to perform a suitable ROI segmentation.

We have conducted a comprehensive assessment of the proposed method. Our results on two well-known databases, Oxford building and Holidays, prove that it is highly competitive in traditional image retrieval tasks, providing favorable results in comparison to most of the state-of-the-art systems. Regarding ROI segmentation, assessed on the Oxford database, the proposed model outperformed RANSAC, the most well-known geometric approach.

In our opinion these results are due to two main reasons: first, our model jointly manages several properties of true matches; and second, by considering the whole set of reference images at once, the proposed method provides a robust method for estimating the actual geometric transformation undergone

by the objects. In particular, by computing the posterior probability that a match is considered as true (e.g. it belongs to any of the considered foreground components), successfully rejects outliers in the estimation of the geometric transformation. This outlier rejection ability notably improves when all the reference images are jointly considered in comparison to traditional techniques where each pair of images (query and reference) are addressed independently.

In addition, our model can also work in scenarios where there is more than one object-of-interest in the query image. To assess the performance of the proposed model, we have conducted three different experiments: a Multi-class category-segmentation experiment on the ETHZ toys dataset; a multi-object detection experiment on the RGB-D dataset; and a multiview object retrieval experiment on the Oxford building dataset. For the first two cases, we developed and tested a method for automatically selecting the number K of objects-of-interest in the query image, with results very close to those ones achieved with the optimal K in each case. In the third experiment, the results showed a significant performance improvement when the number of foreground objects considered by the model fitted the actual number of objects-of-interest. These results allow us to conclude that the performance of the retrieval process can be notable improved when different views of the object-of-interest are available.

V. 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.

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*, 2004, pp. 1–22.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[3] 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*, vol. 0, pp. 1–8, 2007.

[4] 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*, June 2008, pp. 1 –8.

[5] 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*, ser. ICMR '11, New York, NY, USA, 2011.

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007.

[7] L. Yang and A. Hanjalic, "Prototype-based image search reranking," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 871 –882, june 2012.

[8] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, 27 2009-oct. 4 2009, pp. 2109 –2116.

[9] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 17 –24.

[10] F. Perronnin, Y. Liu, J. Sa andnchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 3384 –3391.

[11] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009, pp. 19:1–19:8. [Online]. Available: http://doi.acm.org/10.1145/1646396.1646421

[12] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 17–24. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282283

[13] 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.

[14] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, pp. 316–336, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0285-2

[15] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, Jun. 2009, pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2009.5206566

[16] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of British Machine Vision Conference*, vol. 1, London, 2002, pp. 384–393. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.2484

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

[18] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases." in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 809–816.

[19] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3013–3020.

[20] J. Philbin, J. Sivic, and A. Zisserman, "Geometric latent dirichlet allocation on a matching graph for large-scale image datasets," *Int. J. Comput. Vision*, vol. 95, no. 2, pp. 138–153, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1007/s11263-010-0363-5

[21] I. González-Díaz, C. E. Baz-Hormigos, M. Berdonces, and F. D. de María, "A generative model for concurrent image retrieval and roi segmentation," in *Proceedings of the 7th International Workshop on Content-Based Multimedia Indexing (CBMI), Annecy, France*, 27-29 June 2012.

[22] H. Permuter, J. Francos, and H. Jermyn, "Gaussian mixture models of texture and colour for image database retrieval," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 3, April, pp. III–569–72 vol.3.

[23] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, vol. 20, 2007.

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

[25] J. Goldberger and H. Greenspan, "Context-based segmentation of image sequences," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 463–468, March 2006.

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

[27] V. Ferrari, T. Tuytelaars, and L. J. V. Gool, "Simultaneous object recognition and segmentation by image exploration," in *8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I*, 2004, pp. 40–54.

[28] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1817–1824.

[29] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, pp. 63–86, October 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=990376.990402

[30] 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*, vol. 24, pp. 381–395, June 1981. [Online]. Available: http://doi.acm.org/10.1145/358669.358692

[31] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISSAPP'09)*, 2009.

[32] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009)," http://www.pascal-network.org/challenges/VOC/voc2009/, 2009.

[33] 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*, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 28:1–28:8. [Online]. Available: http://doi.acm.org/10.1145/1991996.1992024

Iván González-Díaz (M'11) received the Telecommunications Engineering degree from Universidad de Valladolid, Valladolid, Spain, in 2005, the M.Sc. and Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007 and 2011, respectively.
His current research interests include image and video object segmentation, computer vision and video coding.

Carlos E. Baz-Hormigos received the Audiovisual System Engineering degree from Universidad Carlos III de Madrid, in 2009. He has worked as engineer in various companies and as a researcher on the Multimedia Processing Group of Universidad Carlos III de Madrid. His main research interests include image and video analysis and processing.

Fernando Díaz-de-María (M'97) received the Telecommunication Engineering degree in 1991 and the Ph.D. degree in 1996 from the Polytechnic University of Madrid, Madrid, Spain. From Oct. 1991 to Oct. 1995, he worked as a Lecturer at Universidad de Cantabria, Santander, Spain. From Nov. 1995 to Sep. 1996, he worked as an Assistant Professor at Universidad Politécnica de Madrid. He reached the Ph.D. Degree in July 1996. His Ph.D. Thesis focused on speech coding. In particular, on the use of Artificial Neural Network-based nonlinear prediction in CELP ("Code-Excited Nonlinear Predictive")-type coders. From Oct. 1996, he is an Associate Professor at the Department of Signal Processing & Communications, Universidad Carlos III de Madrid, Madrid, Spain. From Oct. 97 till nowadays, he has held several offices in both, his Department and his University.
His primary research interests include robust speech recognition, video coding and multimedia indexing. He has led numerous Projects and Contracts in the mentioned fields. He is co-author of several papers in prestigious international journals, two chapters in international books and quite a few papers in revised national and international conferences.