Object-based Image Retrieval using Local Feature Extraction and Relevance Feedback

Mário H. G. Freitas
Department of Computing
Piim-Lab - CEFET-MG
Belo Horizonte, MG, Brazil

Flávio L. C. Pádua
Department of Computing
Piim-Lab - CEFET-MG
Belo Horizonte, MG, Brazil

Guilherme T. Assis
Department of Computing
Piim-Lab - UFOP
Ouro Preto, MG, Brazil

ABSTRACT

This paper addresses the problem of object-based image retrieval, by using local feature extraction and a relevance feedback mechanism for quickly narrowing down the image search process to the user needs. This approach relies on the hypothesis that semantically similar images are clustered in some feature space and, in this scenario: (i) computes image signatures that are invariant to scale and rotation using SIFT, (ii) calculates the vector of locally aggregated descriptors (VLAD) to make a fixed length descriptor for the images, (iii) reduce the VLAD descriptor dimensionality with Principal Component Analysis (PCA) and (iv) uses the *k*-Means algorithm for grouping images that are semantically similar. The proposed approach has been successfully validated using 33,192 images from the ALOI database, obtaining a mean recall value of 47.4% for searches of images containing objects that are identical to the object query and 20.7% for searches of images containing different objects (albeit visually similar) to the object query.

General Terms

Content-based image retrieval, relevance feedback, feature extraction.

Keywords

Object-based image retrieval, scale invariant feature transform, principal component analysis, vector of locally aggregated descriptors, clustering algorithms.

1. INTRODUCTION

The increasing production of visual information (pictures and videos) in recent years has intensified the demand for multimedia information systems that are able to efficiently store and retrieve files of this nature in large databases [16]. According to [18,9], pictures have to be seen and searched as pictures: by objects, by style, by purpose. In this context, content based image retrieval (CBIR) methods, which use search keys that are extracted automatically from the visual content of images have been developed to improve the performance of visual information management systems [3,20]. This work presents an approach that belongs to this group of methods, which uses local feature extraction, a relevance feedback mechanism and a clustering algorithm to perform object-based image retrieval.

Comprehensive surveys have been developed on the topic of CBIR [3,16,11,18,20], providing information about key theoretical and empirical contributions in the last years. According to [3], CBIR systems are frequently based on three main steps. The first one is related to the extraction of image features describing the visual characteristics of each image in the database (feature extraction step). In the second step, the images and their signatures are stored in order to make easier the search procedure (indexing step). Finally, in the third step, an algorithm performs a search in the database in order to return the most similar images to the input (matching and retrieval step). A relevance feedback approach is commonly applied by visual inspection on the resulting images [16,18].

The feature extraction step of a CBIR system is especially critical and, as described in [3], its corresponding methods can be divided into two main classes: (1) global feature based methods, as those proposed in [4,10,8,6], to cite just a few, and (2) local feature based methods, as for example, the ones proposed in [11,1,8,2]. According to [3], a major shift has been observed from global feature representations, such as color histograms and global shape descriptors, to local features, such as salient points, region-based features and spatial model features [20]. This shift is related to the fact that the image domain is too deep for global features to reduce the semantic gap. Local features, in turn, often correspond with meaningful image components, such as rigid objects, making association of semantics with image portions forthright.

In [12] and [1], two well-known local feature detectors are described: the Scale-Invariant Feature Transform (SIFT) and the Speeded Up Robust Features (SURF), respectively. The SURF detector is partly inspired by the SIFT descriptor, being however several times faster than SIFT. Both works have contributed significantly to advances in the development of novel local features that are robust to scale, rotation and illumination variations.

On the other hand, the authors in [9] evaluate different ways of aggregating local image descriptors into a vector and show that the Fisher kernel achieves better performance than the reference bag-of-visual words approach for any given vector dimension. The experiments demonstrated that the image representation can be reduced to a few dozen bytes while preserving accuracy.

In [2], the authors propose a sparse model for local features, where the geometry of each model part depends on the geometry of its *k* closest neighbors. Moreover, an unsupervised learning algorithm is developed, which is able to form clusters of images with similar appearances, and also estimate the model parameters. The experimental results presented by the authors show that their approach can be applied across a variety of object classes.

In this scenario, where local features demonstrate to be a promising alternative, this work proposes the use of a local feature based method, which combines the techniques SIFT [12], Vector of Locally Aggregated Descriptors (VLAD) [9] and Principal Component Analysis (PCA) [14] to compute robust image signatures. Additionally, the proposed approach

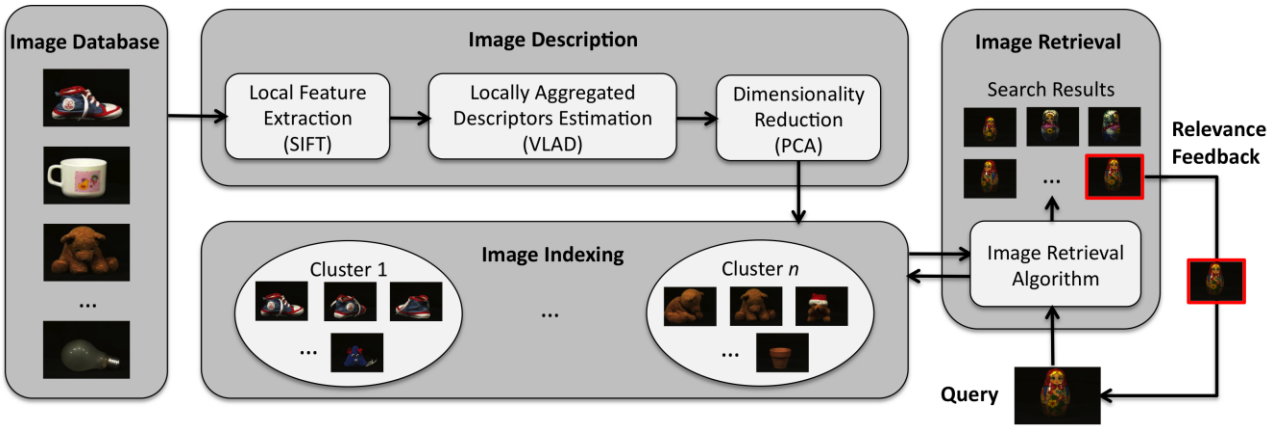

Fig. 1: Overview of the proposed approach for object-based image retrieval.

applies the *k*-Means algorithm [13] for grouping images that are semantically similar, as well as a special image retrieval algorithm that uses a relevance feedback strategy. That image retrieval algorithm is inspired by the solution proposed in [4]. However, unlike the present work, the method developed in [4] is based on global image descriptors (image moments and color signatures).

To evaluate the performance of the proposed approach for object-based image retrieval, the ALOI database [7] is used, which is a color image collection of about 1000 small objects. This database is divided in different subsets depending on the parameter varied at capture time (color, illumination, view-angle, among others).

The remainder of this paper is organized as follows. Section 2 covers the proposed approach for object-base image retrieval. Experimental results are presented in Section 3, followed by the conclusions and discussion in Section 4.

2. THE PROPOSED APPROACH

The proposed approach for object-based image retrieval is divided in three main steps, namely: (1) image description, (2) image indexing and (3) image retrieval. The Fig. 1 presents a conceptual diagram of this approach, illustrating its main steps, which are described in the next sections.

2.1 Image Description

The first step of the proposed approach consists in to describe the visual attributes of the image samples for comparison and retrieval purposes. In this work, the image description is preliminary performed by using local feature extraction, specifically, by applying the well known SIFT technique [12]. In the following, it is used the VLAD method [9] to estimate a fixed length descriptor for each image sample. Finally, it is reduced the VLAD descriptor dimensionality by applying PCA [14]. As a result, the image descriptors produced by the combination of those techniques are robust to scale, rotation and viewpoint variations, making the proposed approach an interesting alternative to object-based image retrieval applications. In the following, those techniques are briefly described.

2.1.1 Scale-Invariant Feature Transform – SIFT

The SIFT technique transforms an image into a large collection of feature vectors, also called keypoints, which are invariant to image translation, scaling, rotation, partially invariant to illumination changes and robust to local geometric distortion [12]. This technique consists of four steps:

1. Scale-space extrema detection: initially, a set of keypoints must be detected. For accomplishing such a task, the image is convolved with Gaussian filters at different scales, and the differences of successive Gaussian-blurred images are taken. Keypoints are searched as maxima/minima of the Difference of Gaussians which occur at multiple scales;
2. Keypoint localization: in this step, the candidate keypoints are localized and the unstable ones (points which are sensible to noise or with low contrast) are eliminated;
3. Orientation assignment: one or more orientations are assigned to each keypoint, based on local image gradient directions. The assigned orientations, scale and location for each keypoint enable SIFT to construct a canonical view for the keypoint, which is invariant to similarity transforms;
4. Keypoint descriptor: finally, keypoints are used for computing descriptor vectors.

Specifically, a keypoint descriptor used by SIFT is created by sampling the magnitudes and orientations of the image gradient in the patch around the keypoint and building orientation histograms to capture the relevant aspects of the patch. Histograms contain 8 bins each, and each descriptor contains a 4×4 array of 16 histograms around the keypoint. This leads to a SIFT feature vector with $4 \times 4 \times 8 = 128$ elements. This 128-element vector is then normalized to unit length to enhance invariance to changes in illumination.

The main intention of SIFT based representations is to avoid problems incurred by boundary effects [12]. Therefore, smooth changes in location, orientation and scale do not cause radical changes in the feature vector. Moreover, it is a compact representation, expressing the patch of pixels using a 128 element vector.

In this work, the SIFT is employed to find the keypoints (and the respective gradient vectors) of the image samples. Besides, it is also used jointly with VLAD and PCA in order to obtain smaller and fixed-length descriptors.

2.1.2 Vector of Locally Aggregated Descriptors

The Vector of Locally Aggregated Descriptors (VLAD) is a technique that is commonly employed for fitting local image descriptors, such as SIFT, into fixed-length descriptors [9]. This method aggregates the image descriptors based on its values and it delivers a fixed-length (often smaller) vector with the most important visual attributes of the input image.

Given an input image I with n descriptors, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, a VLAD can be created as follows:

1. Codebook building: a codebook, with m descriptors (or centroids), $C = [\mathbf{c}_1, \ldots, \mathbf{c}_m]$, is built for the input image. Such as proposed in [9], this task is accomplished by a *k*-Means clustering algorithm [13], using the n original descriptors of the image as the input;

2. Descriptor association: each descriptor $\mathbf{x}_i$ is associated to a centroid $\mathbf{c}_j$, such that:

$$C_j = \left\{ \mathbf{x}_i \in C_j \Leftrightarrow j = \arg\min_{j \in \{1,\ldots,m\}} \left\| \mathbf{x}_i - \mathbf{c}_j \right\|_2 \quad , \quad \forall i \in \{1,\ldots,n\} \right\}, \tag{1}$$

in which C_j is the set of descriptors associated to centroid $\mathbf{c}_j$. In this process, each descriptor is associated to the closest centroid, based on a simple Euclidean distance;

3. Calculating difference vectors: each component of the difference vectors $V = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ is calculated through the following relation:

$$v_{i,j} = \sum_{x_m \in C_i} x_{m,j} - c_{i,j}, \tag{2}$$

in which $v_{i,j}$, $x_{m,j}$ and $c_{i,j}$ are the j^{th} components of vectors $\mathbf{v}_i$, $\mathbf{x}_m$ and $\mathbf{c}_i$, respectively. Importantly, $\mathbf{v}_i$, $\mathbf{x}_m$ and $\mathbf{c}_i$ are $d \times 1$vectors, where d is the number of characteristics of each original descriptor;

4. Finally, the vectors $V = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ are L2-normalized, as shown in Eq. (3):

$$\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i}{\left\| \mathbf{v}_i \right\|_2}. \tag{3}$$

As the main result, this technique delivers a new set of m local descriptors, where the global dimension is $D = m \times d$. In this work, $d = 128$, since it is used the SIFT descriptor. Moreover, according to the authors of [9], the best value for m is located between 16 and 256. In this context and based on the results presented in [9], this work considers $m = 64$, resulting in a VLAD of 8192 elements.

As it is possible to note from the procedure above, a VLAD is created based on the differences between the original descriptors and their respective centroids from the codebook. This procedure can be seen as an adapted and simplified version of the Fisher Kernel [15]. Besides, the employment of a codebook has been inspired from Bag of Features representations [17].

The main advantage of this process is to deliver a fixed-length set of local descriptors. Fixed-length representations can be compared using standard distance metrics, what makes possible to employ robust classification methods, such as Neural Networks and Support Vector Machines or Immune-Inspired algorithms [9].

2.1.3 Principal Component Analysis - PCA

PCA is a tool commonly applied for dimensionality reduction [14], where the eigenvectors with the highest eigenvalues of the empirical vector covariance matrix are used to define a matrix M, mapping a vector $\mathbf{v}_i \in \Re^{128}$, $i = 1, \ldots, m$, to a feature vector $\mathbf{f}_i = \mathrm{M}\mathbf{v}_i$ in a lower-dimensional space.

In this work, PCA projection is used after the application of the VLAD method, in order to highlight even more the image features, creating a final descriptor $\mathbf{f}_i$ with 128 components. The positive impact of applying PCA may be explained by the fact that decorrelated data can be fitted more accurately by a GMM with diagonal covariance matrices [9]. Moreover, GMM estimation is noisy for the smaller components [9,14].

The image description obtained by using the techniques described here contributes to a fast system response, as demonstrated by the experimental results obtained. In the following, it is described the image indexing step, which is based on the *k*-Means algorithm for clustering images that are semantically similar.

2.2 Image Indexing

The image indexing step is responsible for identifying images based on their attributes and for finding natural groups, based on the features extracted from the training samples. Without image indexing, most of the images would remain hidden in the database and, consequently, never seen by the users [21].

Image indexing in this work is accomplished through a content-based approach, in which features of images are automatically identified and extracted during the image description step.

Clustering algorithms, which provide insight into the data by dividing the images into groups that are semantically similar, demonstrate great potential in grouping images for content-based image indexing [21]. In light of this, we propose here the use of the *k*-Means clustering algorithm [13]. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its application here.

The *k*-Means algorithm aims to divide m observations (input descriptors) into k clusters, in such a way that each input descriptor $\mathbf{f}$ is assigned to the cluster with the nearest center. In fact, *k*-Means has an objective function:

$$\min_{\{G_i\}} = \sum_i \sum_{\mathbf{f} \in G_i} w_{\mathbf{f}} \delta(\mathbf{f}, \mathbf{c}_i), \tag{4}$$

where G_i is the i^{th} cluster, $\mathbf{c}_i$ denotes the centroid of G_i (a centroid is the arithmetic mean of the cluster members), $\delta(\cdot)$ is a distance function and $w_{\mathbf{f}} > 0$ is a weight for the input $\mathbf{f}$.

Note that a distance metric is required for the input data and it is necessary to specify, *a priori*, the number of clusters in which the data should be split (k). The evaluation of distances is addressed here by representing the images as vectors embedded in the Euclidean space (the image descriptors), in which the Euclidean distance is defined.

To define the number of clusters (k) in which the data should be divided, it is used the approach presented in [5], which proposes: $k = m^{1/2}$ clusters, where m denotes the number of input descriptors to be clustered. Note that *k*-Means has a single parameter to be set (k), what can make it easier to tune.

The object-based image retrieval approach proposed in this work integrates the *k*-Means method with an image retrieval algorithm that uses a relevance feedback strategy. This fact, in turn, makes it possible to only search in the clusters that are close to the query target, instead of searching in the whole search space. The image retrieval step is described in the next section.

2.3 Image Retrieval

The final step of the proposed approach for image retrieval builds on the algorithm presented in [4], which uses a relevance feedback model.

Basically, this algorithm makes a search space reduction in order to keep searches computationally efficient. This is

accomplished by firstly comparing the descriptors of an image query I_q to the descriptors assigned to centroids of clusters determined in the indexing step. After that, only some clusters are selected, according to the similarity measure established between their centroids and I_q. For each selected cluster, in turn, it is performed the comparison between its elements and the descriptors of the image query I_q. Finally, the most similar images estimated are returned to the user.

The similarity function used to compare the image query I_q to the centroid of the ith cluster, for $i = 1, \ldots, k$, is given by:

$$\phi(F_q, C_i) = \frac{1}{\sqrt{\sum_{j=1}^{m} (\mathbf{f}_j - \mathbf{c}_{i,j})^2}}, \quad (5)$$

where $\phi(\cdot)$ represents the similarity function, $F_q = [\mathbf{f}_1, \ldots, \mathbf{f}_m]$ is the set of descriptors associated to I_q; $C_i = [\mathbf{c}_{i,1}, \ldots, \mathbf{c}_{i,m}]$ is the set of descriptors associated to the ith cluster; $\mathbf{f}_j$ denotes the j^{th} descriptor of I_q; $\mathbf{c}_{i,j}$ represents the j^{th} descriptor assigned to the centroid of the ith cluster and m is the number of descriptors considered.

To support the identification of what the user is looking for, a relevance feedback strategy is applied. In this context, the user is included in the retrieval loop, in such a way that for each iteration the user provides feedback regarding the retrieval results, e.g. by qualifying images returned as either *relevant* or *irrelevant*. From this feedback, the proposed system learns the visual features of the images and returns improved results to the user.

The relevance feedback mechanism implemented was designed to maximize the ratio between the quality of the retrieval results and the amount of interaction between the user and the system. Additionally, since user satisfaction is very subjective and experimenting with users is difficult, the performance measure was defined, relying on the use of a ground truth database for the evaluation of retrieval. To store the feedback information provided by the user, a matrix of counters M is used, as proposed in [4]. The dimensions of M are $k \times p$, where k denotes the number of clusters determined in the indexing step and p is the number of clusters selected according to the similarity measure established. When an image I_s returned by the system is chosen by the user as the most similar to I_q, the corresponding matrix cell M(s,i) is incremented, where s represents the index of image I_s in the dataset and i represents the index of the cluster I_s belongs to.

Two input parameters in special must be provided to the proposed retrieval algorithm, namely, the number of clusters k and the maximum number of images η to be returned to the user, given an image query I_q. This last parameter (η) is estimated by using Eq. (6) and Eq. (7), as suggested in [4]:

$$\eta_i = \frac{\phi(F_q, C_i) \dfrac{\mathrm{M}(s,i)}{\sum_{j=1}^{p} \mathrm{M}(s,j)}}{\sum_{j=1}^{p} \phi(F_q, C_j) \dfrac{\mathrm{M}(s,j)}{\sum_{z=1}^{p} \mathrm{M}(s,z)}} \cdot w, \quad (6)$$

$$\eta = \sum_{i=1}^{k} \eta_i, \quad (7)$$

where η_i and η are, respectively, the number of images returned from the i^{th} cluster and the maximum number of images returned by the system; $\phi(\cdot)$ is the similarity function defined in Eq. (5); M is the matrix of counters used by the relevance feedback mechanism; s is the index of image I_s that is the most similar image to the query I_q in the set of clusters selected; p is the number of clusters selected and w is the number of images returned per iteration.

In the next section, the experimental results obtained by using the proposed approach for image retrieval are presented.

3. EXPERIMENTAL RESULTS

In order to demonstrate the advantages and limitations of the proposed approach, two groups of experiments were performed with image samples of the ALOI database [7]. This database contains 1,000 objects recorded under various imaging circumstances, specifically, under 72 inplane viewing angles, 24 different illumination angles and under 12 illumination colors. A large variety of object shapes, transparencies, albedos and surface covers are considered, making this database quite interesting to evaluate object-based image retrieval approaches. Some image samples of the ALOI database are presented in Fig. 2.

To evaluate the scalability of the proposed approach, the ALOI database was fragmented in four datasets, which are differentiated not only by the number of image samples, but also by the number of objects considered. The size and number of objects of each dataset are listed in Tab. 1. Note that even though ALOI database contains 1,000 objects, only 922 were selected to create the datasets. Some of the remaining objects, however, were used as object queries in the tests performed. For each object in the datasets, 36 image samples were available, illustrating different viewpoints, with rotation angles varying from 0^o to 180^o. Image dimensions in all datasets were 384 × 288 pixels.

On the other hand, two groups of image queries were defined to evaluate the performance of the proposed approach regarding the retrieval of (1) images containing objects that are identical to the object present in the image query (group 1) and (2) images containing objects that are visually similar (albeit not equal) to the object present in the image query (group 2). In the former case, 20 image queries of different objects are used. Those images are present in the four datasets considered and, therefore, for each image query there are 36 images containing the same object. For the latter case, in turn, other 20 image queries of different objects are used. However, the objects of those images are not present in the datasets. A ground truth database was created, considering features as shape, texture and color to determine the similarity between image samples. According to the ground truth created, for each object present in the 20 image queries, there is at least one object considered visually similar in the datasets.

Fig. 2: Image samples of ALOI dataset.

Tab. 1. ALOI datasets used in the experiments.

Dataset	Number of Images	Number of Objects
ALOI-1	5,004	139
ALOI-2	10,008	278
ALOI-3	20,096	556
ALOI-4	33,192	922

To evaluate the relevance feedback mechanism and its impact in the systems recall in a detailed manner, it was considered 10 iterations for each search session of both groups of experiments, even though, such a high number of iterations does not commonly occur in the practice for most users of such a system. The maximum number of images (w) to be returned per iteration was defined as 50 images. Finally, the number of clusters selected (p) by the approach during a specific retrieval session and according to the similarity measure established in Eq. (5), was empirically defined as p = 5, since this value was responsible for the best recall results, while did not impact negatively in the systems computational cost.

3.1.1 Group 1

The first group of experiments aims to evaluate the retrieval capability of images containing objects that are identical to the object present in the image query. In this scenario, as previously described, 20 image queries of different objects have been used. The well-known information retrieval metrics *precision* and *recall* are used to evaluate the approach.

Fig. 3 presents the recall results as a function of the number of relevance feedback iterations performed by the user. The recall values exhibited in Fig. 3 are mean values, thus considering the recall estimated for each image query used. Those results demonstrate the scalability of the proposed approach. Note that, as expected, the greater the dataset, the smaller the recall obtained. Additionally, one may observe from Fig. 3, that the relevance feedback mechanism improves significantly the retrieval capability of the approach until 5 iterations, when the recall values stabilize.

Fig. 4, in turn, shows the precision-recall curves for two different scenarios: (1) when only one relevance feedback iteration is performed (blue curve) and (2) when 10 relevance feedback iterations are performed (green curve). The recall and precision values exhibited are mean values (the results for all image queries are considered). Importantly, those curves are parameterized by the number of images returned in a given retrieval session. Specifically, this number was varied from 5 to w (defined as 50), with increments of 5. Therefore, the first point in any of those curves corresponds to the mean recall and mean precision when only the 5 first returned images are considered. On the other hand, the second point corresponds to mean recall and mean precision for the first 10 returned images, and so on.

Note that, as illustrated in Fig. 4, the values for precision decrease abruptly when the number of returned images is greater than or equal to 15, while the values for recall stabilizes in about 0.3, for the case of only one relevance feedback iteration (blue curve). This fact indicates that, in this case, the relevant images searched by the user appear among the first 15 returned images.

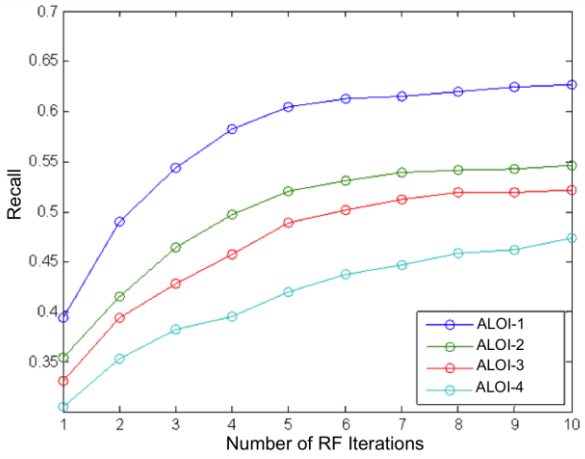

Fig. 3: Recall as a function of the number of Relevance Feedback (RF) iterations in Group 1.

On the other hand, note that for 10 relevance feedback iterations (green curve in Fig. 4), the precision-recall curve is smoother, demonstrating the benefit of using the proposed relevance feedback strategy. In fact, in this case, the recall values have increased significantly more and the relevant images searched by the user appear even among the first 40 returned images.

3.1.2 Group 2

The goal of the second group of experiments consists in to evaluate the retrieval capability of images containing objects that are different of the object present in the image query, but that are considered visually similar (see Fig. 5), according to features, such as, shape and texture, for example. To achieve this goal, a ground truth database was created. As previously described, a new group of 20 image queries of different objects (not present in the four datasets) has been used.

The recall results for this scenario are illustrated in Fig. 6 as a function of the number of relevance feedback iterations. Note that, the behaviors of the curves in Fig. 6 are quite similar to their equivalent curves in Figure 3, that is, the greater the dataset, the smaller the recall. Moreover, the relevance feedback mechanism improves significantly the systems retrieval capability. However, differently from the results for Group 1, the mean recall values are much smaller.

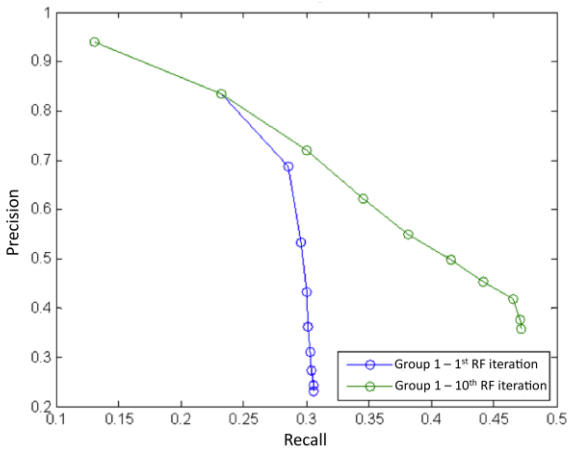

Fig. 4: Precision-recall curves for Group 1.

Fig. 5: Image samples of visually similar objects.

For instance, considering the largest dataset (ALOI-4) and 3 relevance feedback iterations, the recall results for Group 1 and 2 were, respectively, 0.38 and 0.15. This fact was expected, since the scenario of Group 2 is a much more challenging problem for the current state-of-the-art image retrieval algorithms.

Finally, the precision-recall curves for Group 2 are illustrated in Fig. 7, considering two different scenarios: one relevance feedback iteration (blue curve) and 10 relevance feedback iterations (green curve). Those curves were obtained similarly to the ones of Fig. 4 for Group 1. Note that again the values for precision decrease more abruptly when less relevance feedback iterations are performed, demonstrating the importance of this mechanism. The best recall and precision values for Group 2 were, respectively, 0.2 and 0.27, which were obtained when the total number of returned images, specifically 50 images, was analyzed and 10 relevance feedback iterations were performed.

3.1.3 Computational Efficiency

Experiments using the two image query groups and the ALOI-4 dataset presented in Section 3 were performed to evaluate the efficiency of the approach. The experiments were carried out using a workstation with an AMD Turion Dual-Core 2.1 Ghz processor and 2.75 GB RAM, running Windows OS.

The average search times obtained by using the image queries of Group 1 and Group 2 were, respectively, 1.582 ± 0.049s and 1.735 ± 0.169s. Note that the average search time for Group 2 is slightly larger than the one for Group 1, possibly justified by the nature of images of Group 2, which contain objects that are not present in the four datasets considered.

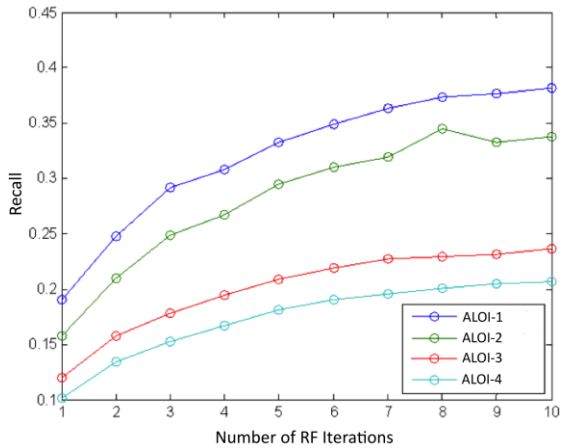

Fig. 6: Recall as a function of the number of Relevance Feedback (RF) iterations in Group 2.

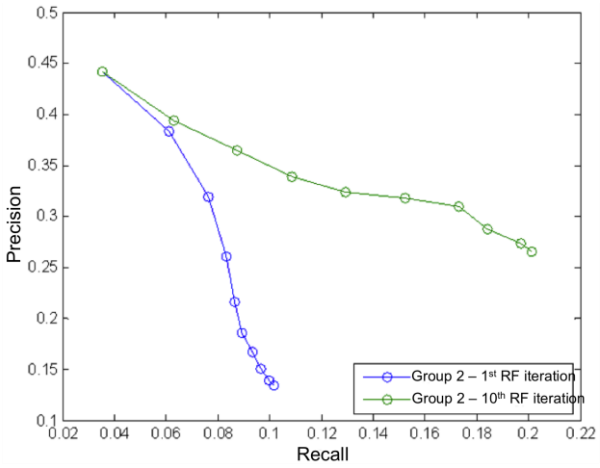

Fig. 7: Precision-recall curves for Group 2.

Considering the computational platform used, the results are quite promising in better computer architectures, demonstrating the potential applicability of the proposed approach in real scenarios.

4. CONCLUDING REMARKS

This work investigates a new approach for object-based image retrieval, which uses local feature extraction to produce image signatures that are invariant to scale and rotation, combining the robust techniques SIFT, VLAD and PCA, as well as a relevance feedback strategy to support the identification of what the user is looking for.

The experiments demonstrate the effectiveness, efficiency and scalability of the proposed approach. From the experimental results, it can be derived that the approach can successfully retrieve not only images containing objects that are equal but also visually similar to the object present in the image query. The results further reveal that the relevance feedback mechanism proposed improves significantly the systems retrieval capability.

The evaluation results give motivation for further investigations on how the approach could benefit from other indexing features and similarity metrics. Besides, the scaling of the proposed relevance feedback mechanism to very large image databases is an important issue that should be more extensively studied. Importantly, features that are shared by some images usually define relevance, but it may concern entire images or parts of images. In this context, the feedback provided by real-world users often contains inaccurate information. Although the proposed approach can tolerate noise to some extent, it should be more exploited how to conduct filtering to remove unreliable feedback before using it for improving the retrieval results. Another important direction for future work consists in to evaluate the approach on databases containing objects in a wide variety of scenes and lighting conditions such as the Corel Stock Photos and Caltech databases.

Finally, although evaluated for image retrieval, the proposed approach is suitable for other types of multimedia retrieval (videos, for instance) applications with only minor changes.

5. ACKNOWLEDGMENTS

The authors thank the support of FAPEMIG-Brazil under Procs. EDT-162/07 and APQ-01180-10; of CEFET-MG under Proc. N° 023-076/09, of CNPq-Brazil and of CAPES-Brazil.

6. REFERENCES

[1] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. 2008. Speeded-up robust features (SURF). Computer vision and image understanding, 110(3), 346-359.

[2] Carneiro, G., and Lowe, D. 2006. Sparse flexible models of local features. In Proceedings of European Conference on Computer Vision (ECCV), pp. 29-43.

[3] Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), 40(2), 5.

[4] Duan, F., Li, X., Liu, J., and Xie, K. 2007. Image Retrieval Model Based on Immune Algorithm. In Proceedings of IEEE Workshop on Intelligent Information Technology Application, pp. 141-144.

[5] Dunn, J. C. 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(1), 32–57.

[6] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., and Yanker, P. 1995. Query by image and video content: the QBIC system. Computer, 28(9), 23-32.

[7] Geusebroek, J., Burghouts, G. J., and Smeulders, A. W. 2005. The Amsterdam library of object images. International Journal of Computer Vision, 61(1), 103-112.

[8] Gevers, T., and Smeulders, A. W. 2000. Pictoseek: Combining color and shape invariant features for image retrieval. Image Processing, IEEE Transactions on, 9(1), 102-119.

[9] Jégou, H., Perronnin, F., Douze, M., and Schmid, C. 2012. Aggregating local image descriptors into compact codes. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(9), 1704-1716.

[10] Jeong, S., Won, C. S., and Gray, R. M. 2004. Image retrieval using color histograms generated by Gauss mixture vector quantization. Computer Vision and Image Understanding, 94(1), 44-66.

[11] Liu, Y., Zhang, D., Lu, G., and Ma, W. Y. 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recognition, 40(1), 262-282.

[12] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[13] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 281-297, p. 14.

[14] Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572.

[15] Perronnin, F., Dance, C. 2007. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8.

[16] Petrelli, D. and Auld, D. 2008. An examination of automatic video retrieval technology on access to the contents of an historical video archive. Program: electronic library and information systems, 42(2), 115-136.

[17] Sivic, J., Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1470-1477.

[18] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content-based image retrieval at the end of the early years. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(12), 1349-1380.

[19] Sukthankar, R., and Ke, Y. 2004. PCA–SIFT: A more distinctive representation for local image descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 506–513.

[20] Suruchi B., and Shahane, N. M. 2013. A survey on textured based CBIR techniques. In IJCA Proceedings on International Conference on Recent Trends in Engineering and Technology, pp. 11-14.

[21] Wang, M., Ye, Z., Wang, Y., and Wang, S. 2008. Dominant sets clustering for image retrieval. Signal Processing, 88 (11) , pp. 2843–2849.