Semantic-Based Cross-Media Image Retrieval

Ahmed Id Oumohmed, Max Mignotte, Jian-Yun Nie
DIRO, University of Montréal, CP. 6128, succursale Centre-ville, Montréal, H3C 3J7 Canada
{idoumoha,mignotte,nie}@iro.umontreal.ca

Abstract

In this paper, we propose a new method to cross-media semantic-based information retrieval, which combines classical text-based and content-based image retrieval techniques. This semantic-based approach aims at determining the strong relationships between keywords (in the caption) and types of visual features associated with its typical images. These relationships are then used to retrieve images from a textual query. In particular, the association *keyword/visual feature/characterization* may allow us to retrieve non-annotated but similar images to those retrieved by a classical textual query. It can also be used for automatic images annotation. Our experiments on two different databases show that this approach is promising for cross-media retrieval.

I. Introduction

In general, image retrieval can be either text-based or image-based. In the first case, a user submits a textual query and the system searches for images with similar keyword(s) in its captions. In the latter, the system tries to determine the most similar images to a given query image by using low level visual features such as color, texture or shape. Some recent approaches have tried to go a step further and has proposed a semantic-based approach in order to assign a semantic meaning to the whole image or to its regions. This approach can be also used for automatic images annotation. In this paper, we investigate a similar approach to image retrieval. In our approach, we try to characterize a given (key-)word by the most discriminant and representative visual feature(s) associated with it (i.e., color, texture or shape) and then by one or several descriptions (descriptor or feature vector) of the selected feature.

Among the semantic-based approach, but only image content-based, different kinds of methods have already been investigated. We can cite, for example, the approach used in [1] which consists in grouping images into semantically meaningful categories. This system was applied on 6931 vacation photographs to obtain a classification such indoor/outdoor, city/landscape, etc. This classification is performed by a Bayesian classifier under the constraint that the test image does belong to one of the classes beforehand established by human subjects. We can also cite the approach used in [2] which clusters the image regions into 10 clusters (cloud, grass, etc.) and uses a probabilistic approach to define a semantic codebook of every cluster. Nevertheless, some recent studies [3] have tried to automatically create associations between visual features and keywords. The basic idea is to use a set of annotated images as a set of learning examples, and to extract strong associations between annotation keywords and the visual features of the images. In particular, a segmentation algorithm, such Blobword[4] or Normalized-cuts [5] is used to produce segmented regions, then for each region, feature information (color, texture, position and shape) is computed. The set of computed features are clustered into regions which are

called “blobs” which define the vocabulary for the set of images. Finally images are annotated by the means of a cross-media relevance model.

Instead of using pre-segmented image regions, described by multiple features (color, texture, shape, etc), our approach uses the whole image content and try to find out the most representative clusters of images and the associated visual feature(s). Compared to [3], our approach has the advantage of not being dependant of a specific segmentation and can take into account relationships between regions (e.g., airplane-sky, animal-grass,boat-sea, etc). Besides, some (key)words are best represented by one feature than by considering several features (e.g., airplane with shape, sea with texture, cathedral with contours). Our approach tries to identify such strong associations between words and visual features.

The rest of the paper is organized as follows: In section II, we will present the image processing techniques developed for this retrieval system; i.e., the considered visual features (texture, contours and shape) as well as their corresponding similarity measures. In section III, we will describe the way that relationships between keywords and visual features are extracted by the means of a learning procedure. In section IV, we will present some experimental results on the annotated imageCLEF and Corel© databases and we conclude.

II. Image Processing Retrieval Techniques

Edge, texture and shape (including color) informations are important cues for pattern recognition and retrieval purposes in large image database. In our approach, we have considered these cues as the three fundamental classes of visual characteristics, which we will call features in this paper. For each of the features, we consider a descriptor that will allow to define a discriminant measure for each considered feature.

Edge Descriptor: Wavelet-based measures have often been used in content-based image retrieval (CBIR) systems because of their appealing ability to describe the local texture and the distribution of the edges of a given image at multiple scales. In this study we use the Harr wavelet transform [6] for the gray-level component of the image (because one of the test dataset contains mostly black and white images). The procedure of image decomposition into wavelets involves recursive numeric filtering. It is applied to the set of pixels of the digital image which is decomposed with a family of orthogonal basis functions obtained through translation and dilatation of a special function called *mother* wavelet. At each decomposition scale, we obtain four sub-bands, which we refer to as LL, LH, HL and HH according to their frequency characteristics (L : Low and H : High). The LL sub-band is then decomposed into four sub-bands at the next scale. Three sales of transformation are considered here. For decomposition of each scale, we compute the mean and the standard deviation (μ_n and σ_n) of the energy distribution in each (of the $n = 10$) sub-band. This leads to an edge descriptor $\{\mu_{n=1}, \sigma_{n=1}, \ldots, \mu_{n=10}, \sigma_{n=10}\}$ of 20 components. For this descriptor, the similarity measure between two images i and j is then defined by:

$$d_{i,j} = \sum_{n=1}^{10} \left(\left| \frac{\mu_n^{(i)} - \mu_n^{(j)}}{\sigma(\mu_n)} \right| + \left| \frac{\sigma_n^{(i)} - \sigma_n^{(j)}}{\sigma(\sigma_n)} \right| \right),$$

where $\sigma(\mu_n)$ and $\sigma(\sigma_n)$ are the standard deviations of the components μ_n and σ_n respectively over the entire database.

Texture Descriptor: Tamura *and al.*[7] have proposed to characterize image texture along the dimensions of contrast, directionality, coarseness, line-likeness, regularity and roughness. The directionality is a global texture property which measures the sharpness of the peaks in the oriented edge histogram. Coarseness refers to the average of the best representative sizes of the *textons* (i.e., texture resolution representation). To describe the texture feature, we use the coarseness and directionality histograms. We make two adjustments to the well known coarseness algorithm [7]: 1). We set some predefined texture resolutions $\{2, 8, 14, 20, 26, 32, 38\}$ instead of $2^k \times 2^k$ with $k = 0, 1, \ldots, 6$; 2). We deal with homogeneous regions bigger than the maximum of texture resolutions taken in account. After thresholding, the oriented edges are quantized into an 8-bin histogram. Finally the dimension of the texture descriptor is 15 and histograms are compared by the Jeffrey divergence [8].

Shape and Color Descriptor: Extraction of shapes contained in an image remains a difficult task. One can use a contour detection algorithm (such as the Canny or Sobel edge detector) as a preliminary step in the shape extraction. However, these methods remain highly dependent on some parameters as thresholds. Instead, ollowing [9], we first estimate a segmented image from which we extract the contours of different regions. The segmented image defines a set of connected pixels belonging to a same class. In this procedure, the noise is taken into consideration, edges are always connected, and the only parameter adjustment is the number of regions used in the segmentation procedure. Then, for each edge pixel, we define a direction (horizontal, vertical, first or second diagonal) depending on the disposition of its neighboring edge pixels and compute a 4-bin histogram. We complete this information by computing a 32-bin color histogram by using the HSV color space. The final 36-bin histogram is then exploited for a distance measure, similar to the one used for the wavelets.

III. Associating words with representative images and featres

Given a set of training images with caption, we try to automatically determine one or several clusters of images representative for each word, together with the most discriminative feature(s), i.e. *texture*, *edge* and *shape-color*. The principle is as follows: For each word, we try to group the images associated with it into several clusters (at different scales) According to each feature. Using one cluster as a visual query, if we can find many images annotated with the word among the most similar images According to the associated feature, then the cluster and he feature are considered to be characteristic For the word. In ths way, each word can be ssociated with zero, one or several clusters and features.

More precisely, let us define some notations : let $\mathbf{I}$ and $\mathbf{I}_w$ be respectively the set of all images in the training dataset and the set of all images that are annotated with the keyword w. $|.|$ will designate the cardinal or the number of elements of a considered set: By applying the three visual features characterizations to $\mathbf{I}_w$, we obtain three set of descriptors $\mathbf{D}^{\mathbf{t}_{exture}}_{I_w}$, $\mathbf{D}^{\mathbf{e}_{dge}}_{I_w}$ and $\mathbf{D}^{\mathbf{S}_{hape}}_{I_w}$. We will use the notation $\mathbf{D}^{\mathbf{f}_{eature}}_{I_w}$ to refer to each of these descriptors.

For a fixed number of regions (we consider 1, 2, 4, ?? regions in our case), we use the Generalized LLoyd [10] algorithm to cluster each set $\mathbf{D}^{\mathbf{f}_{eature}}_{I_w}$ in R partitions, thus, we obtain several ${}^{R}_{c}\mathbf{D}^{\mathbf{f}_{eature}}_{I_w}$

clusters, where R denote the number of partitions used in the clustering and c the c^{th} cluster in this R-clustering. The error-distance used in the clustering of $\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}$ is the similarity measure of the feature $feature$. For each value of R, this clustering allows us to approximate the distribution of the set of samples $\mathbf{D}_{I_w}^{feature}$ by R spherical distributions with identical radius. The centers (centroids) of these approximated spherical distributions are then considered as prototype vectors and are denoted by ${}_c^R P_{I_w}^{\mathrm{f}_{feature}}$. Several values of R are used to take in account the fact that a given word may be associated to many images classes. For example, the word BOAT may be associated with images with small shape of boat in sea, or with a closer view of boat, and so on. For each cluster ${}_c^R\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}$, its associated centroid is used as a descriptor vector of a virtual image representative of the word. The virtual image will be used to query the whole training database $\mathbf{I}$ to get the closest descriptors (or images) according to the similarity measure associated to the feature f$eature$. The training process is as follows: • First, in order to associate each (key-)word w with the most discriminant class of visual characteristic $Feature$, we use the following strategy;

For each considered cluster ${}_c^R\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}$, we count the number of images annotated by the word w that are retrieved among the first X ($= 20$ in our case) retrieval result for each $Feature$. Let $topX^{\mathrm{f}_{feature}}$ be this number. We count the sum of the $topX^{\mathrm{f}_{feature}}$ resulting from the query by all corresponding prototype vectors. We then consider the class of visual feature for which this sum is maximal.

• Second, in order to define a set of prototype vectors associated to the pre-estimated class of visual feature, we adopt the following strategy;

We characterize a given cluster ${}_c^R\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}$ by three measures: its proportion ρ within $\mathbf{I}_w$ (simply, $\rho = |{}_c^R\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}|/|\mathbf{I}_w|$), its standard deviation σ (computed according to the similarity measure of f$eature$), and an empirical measure P which represents the number of images, not annotated by the word w, for which the euclidean distance between its descriptor vector and the prototype vector ${}_c^R P_{I_w}^{\mathrm{f}_{feature}}$ is less than the pre-estimated standard deviation σ, namely $P = |\{I \notin \mathbf{I}_w \mid Dist({}_c^R\mathbf{D}_{I_w}^{\mathrm{f}_{feature}}, {}_c^R P_{I_w}^{\mathrm{f}_{feature}}) < \sigma\}|/|I|$.

Once one feature or several weighted features are fixed, we choose representative prototype vectors regarding to P, their proportion and their standard deviation as follows: We use a first criterion to exclude prototype vectors for which $P > 0.05$ and $\rho < 0.05$. If there is no remaining prototype vector, then we ignore this criterion. The second criterion is to retain prototype vectors for which ρ/σ is less than a threshold.

The result of the training process is that a word may be associated with zero, one or several clusters of representative images, together with an associated feature to each cluster.

IV. Experimental Results and Conclusion

The experimental results are based on the historical image database *'St Andrews University Library Photographic Collection'* provided by *ImageCLEF 2004* [11]. This database contains $28,133$ images with caption. The caption text associated to each image contain approximatively ten (key)words. Our goal was to improve textual and multi-words queries by extending words to their associated visual features but our experiments in this context are extremely difficult due to the poor quality of the images

of this database and also the presence of some (key)words used in the request with an abstract concept. (“Scotland”, “north”, “tournament”, etc.).

We have also extended the results of this semantic-based image retrieval system to a set of 20000 images extracted from the Corel© database where each image is annotated by a few concrete and significant keywords. To test the relevance of our approach, we have taken the set of images annotated by a word *w* and we remove hfe word from the caption of 50% of images of this set. We use these images as reference. We try to see how our approach is able to retrieve these images with a query made of the removed word. We will emphasis on two aspects of our results: the retrieved references images and the non-annotated images retrieved.

Table 1 shows some words with the estimated weights for each class of visual feature. Most associations have a significant meaning : animal is associated to shape and texture features, ocean is most described by shape (probably due to the presence of boats), tiger is described by texture and contours, zebra is associated to texture, etc. However, some words have almost the same weights for the three features, for example water, sky, garden and tree. This may be due to the high number of learning vectors. The word texture is strangely associated with shapes and contours.

By choosing clusters with high value of P, we can guess to obtain more images that are not annotated by the word, but which are related to this word. In other hand, low values of this measure may yield to more images that are really annotated by the word; this may be useful in the case of queries with multiple words, so to eventually improve the text retrieval result. 2 shows two query result for the words flower and animal : the algorithm described in III was used to produce this result. Even if the reference images (randomly deannotated images) were not retrieved successfully, we can see that most of images are related to the query word.

REFERENCES

[1] Aditya Vailaya, A. T. Figueiredo, Anil K. Jain, and Hong-Jiang Zhang. Image classifi cation for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130, 2001.

[2] W. Wang, Y. Song, and A. Zhang. Semantic-based image retrieval by region saliency. In *Int’l Conf. on Image and Video Retrieval*, July 2002.

[3] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.

[4] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*, pages 509–516. Springer, 1999.

[5] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[6] S.G. Mallat. A theory for multiresolution signal decomposition : The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[7] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:460–473, 1978.

[8] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann and. Empirical evaluation of dissimilarity measures for color and texture. In *International Conference on Computer Vision*, volume 2, pages 1165–1173, September 1999.

[9] M. Goldberg, P. Boucher, and S. Shlien. Image compression using adaptative vector quantization. *Communications, IEEE Transactions on [legacy, pre - 1988]*, 34:180–187, 1986.

[10] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28:84–95, 1980.

	$\sum$ *top20* over all clusters			associated weights			
word	texture	contours	shape	texture	contours	shape	Number of training vectors
water	61	74	65	0.30	0.37	0.32	2550
sky	65	66	60	0.34	0.34	0.31	2323
tree	85	79	72	0.36	0.33	0.305	2242
people	60	76	51	0.32	0.40	0.27	1908
grass	27	35	28	0.30	0.38	0.31	1061
flower	16	51	61	0.12	0.39	0.47	934
wild	15	17	15	0.31	0.36	0.31	707
bird	24	12	9	0.53	0.26	0.20	595
plant	8	13	10	0.25	0.41	0.32	439
garden	14	14	14	0.33	0.33	0.33	301
sunset	8	15	19	0.19	0.35	0.45	260
ice	6	8	5	0.31	0.42	0.26	240
ocean	15	26	44	0.17	0.30	0.51	231
animal	7	3	11	0.33	0.14	0.52	204
ski	0	4	1	0.00	0.80	0.20	153
texture	8	10	17	0.22	0.28	0.48	126
rural	3	7	3	0.23	0.53	0.23	124
insect	1	10	7	0.05	0.55	0.38	123
tiger	14	10	9	0.42	0.30	0.27	73
zebra	13	9	8	0.43	0.30	0.26	26

Fig. 1. *A list of words with their relative measures of the sum of their associated* top20*; they are classified by the number of the training vectors.*

[11] Carmen Alvarez, Ahmed Id Oumohmed, Max Mignotte, and Jian-Yun Nie. Toward cross-language and cross-media image retrieval. In *Working Notes for the CLEF 2004 Workshop*, volume 1, pages 525–534, September 2004.

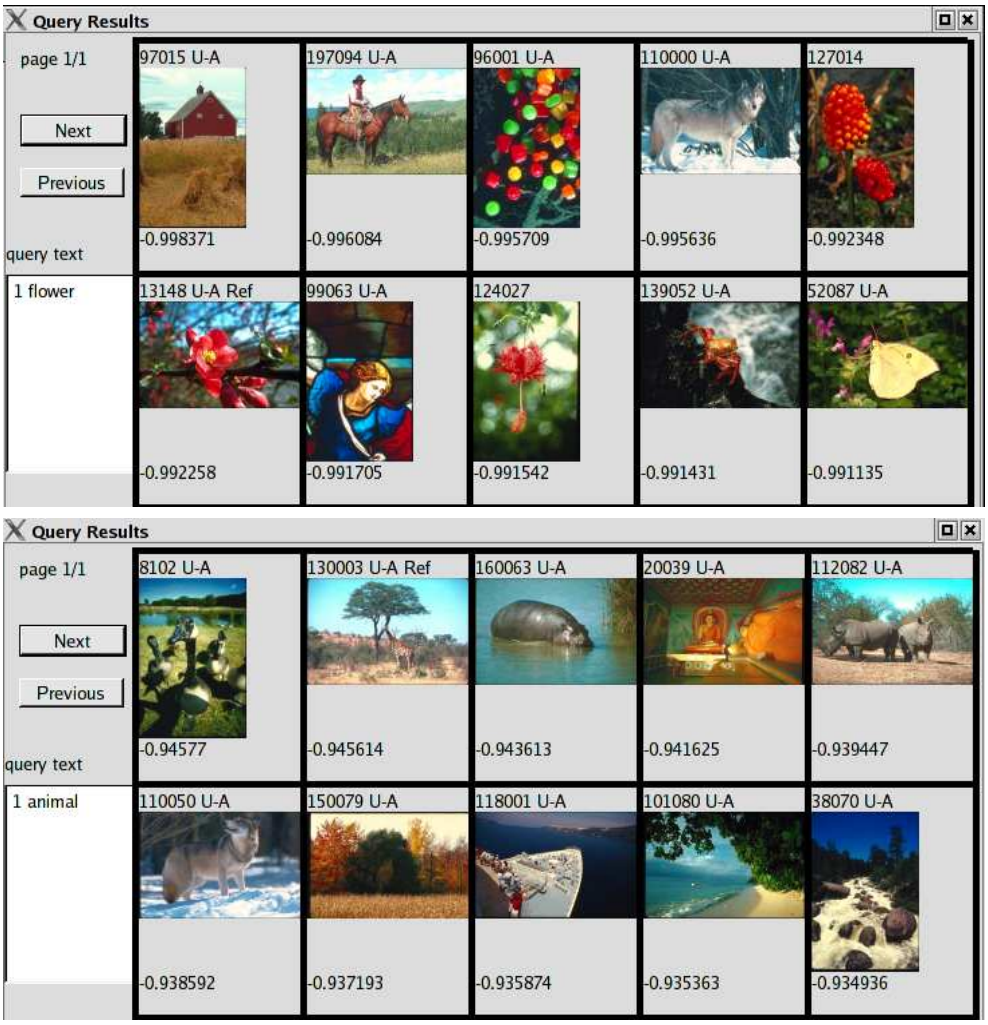

Fig. 2. *A semantic query result for the words flower and animal; the identification number is shown above each image, and the similarity distance is reported below the image. Two others informations are eventually reported above an image : U-A stands foe Un-Annotated image by the word and Ref which stands for a reference image. Negative values of distances are due to the zero mean and unit standard deviation normalization*

3

word	refX				annX				ref sunbj		
	ann10	ann20	ann50	ann100	ref10	ref20	ref50	ref100	subj10	subj 20	subj 30
flower	1	2	3	7	2	3	5	8	9	17	28
animal	1	1	2	3	0	0	0	0	6	9	16
birds	1	1	4	5	1	1	3	5	3	7	9
ice	0	0	0	1	0	0	0	1	0	0	0
grass	0	0	0	5	0	1	1	4	9	15	26

Fig. 3. *Some statistics about the top retrieved un-annotated images which are related to the word (subject judgement) and the top annotated and retrieved images and the top of reference retrieved images. //a reformuler*