

Semantic-Based Cross-Media Image Retrieval

Ahmed Id Oumohmed, Max Mignotte, and Jian-Yun Nie

DIRO, University of Montreal, Canada
{idoumoha, mignotte, nie}@iro.umontreal.ca

Abstract. In this paper, we propose a novel method for 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* 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.

1 Introduction

In general, a content-based image retrieval (CBIR) system tries to determine the most similar images to a given query image by using one or a combination of several low-level visual feature(s) such as color, texture or shape. Depending on the content of each image, it's highly difficult to choose the appropriate feature(s) to use and eventually the manner to combine them. While users are mostly interested by the high-level (i.e., abstract) concepts presents within an image query, the most similar images to this latter according to some low-level visual features can be non-relevant in the sens of semantics. This is known as the semantic gap. Usually, an annotated-based image retrieval (ABIR) system is based on a certain model representation of the concepts (words) associated to each image (document). Given a textual query, a such system scores and ranks images according to the importance of each word of text query to images. In this case, the search result is more limited to images that are really annotated by at least one of the words that form the textual query. In this work, we attempt to reach the same objective by finding non-annotated, but similar, images to those retrieved by a classical textual query. To this end, and based on a training set of several images annotated by the same single word, we propose an unsupervised learning procedure which determine the most representative visual feature (visual semantic) of this word. Given an image query and the words of its caption, the user can choose the characterization of a certain word as a new search criterion.

1.1 Related Work

Organizing a set of images into clusters was used by Chen, Wang and Krovetz [1] in their CBIR system (*CLUE*). Instead of sorting images by feature similarities with respect to a query image, the system retrieves image clusters. Especially, the user can navigate between queries according to each defined cluster (semantic *clue*). After the resemblance between the query image and target images are evaluated and sorted, a collection of target images that are “close” to the query image are selected as the neighborhood of the query image. The set of descriptor vectors of this collection is clustered into a dynamically-defined number of regions. This approach offers a different manner to present and visualize the most similar images to a given query image with an interesting interaction with the user.

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 [2] 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 [3] 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 [4] 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 [5] or Normalized-cuts [6] 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.

Among the semantic-based approach trying to model the relationships between image features and associated text, we can cite the interesting work of Barnard et al. [7]. Their approach searches to provide a statistical joint distribution for associated words and features of each region of an image (image segments). After a training step which consists in estimating the parameters of a mixture of (Gaussian) distributions, a query search consists in computing the probability of each candidate image of emitting the query items. This method remains nevertheless highly dependent of the segmentation results and parameters associated to the segmentation (number of classes). Besides it is also highly dependent of the assumption that the cluster-conditional distribution of *index terms* (words or image segments) (i.e., the likelihood of this model) is unimodal and Gaussian. We can also cite the work of Wang et al. in [8] which try to address the challenging and -closely related problem- of automatic linguistic indexing of pictures. Association between an image and textual description of a concept is modeled via a likelihood given by a two-dimensional multi-resolution

hidden markov model (HMM) whose parameters is learned in a training step. Once again, a query search consists in computing the likelihood of each candidate image for each pre-learned concept. As in applications, where this strategy is commonly used (e.g., handwritten text and speech recognition), this method remains highly dependent of the parameter estimation step of the HMM which is then used for the recognition step. In the case of 2D signal (i.e., image) this estimation may not to efficiently model all the diversity of the different concepts and classes of images.

1.2 Our Approach

Instead of using pre-segmented image regions, described by multiple features (color, texture, shape, etc.), our approach uses the whole image content and tries to find out the most representative visual feature(s). Compared to [4], our approach has the advantage of not being dependent 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., sea with texture and cathe-

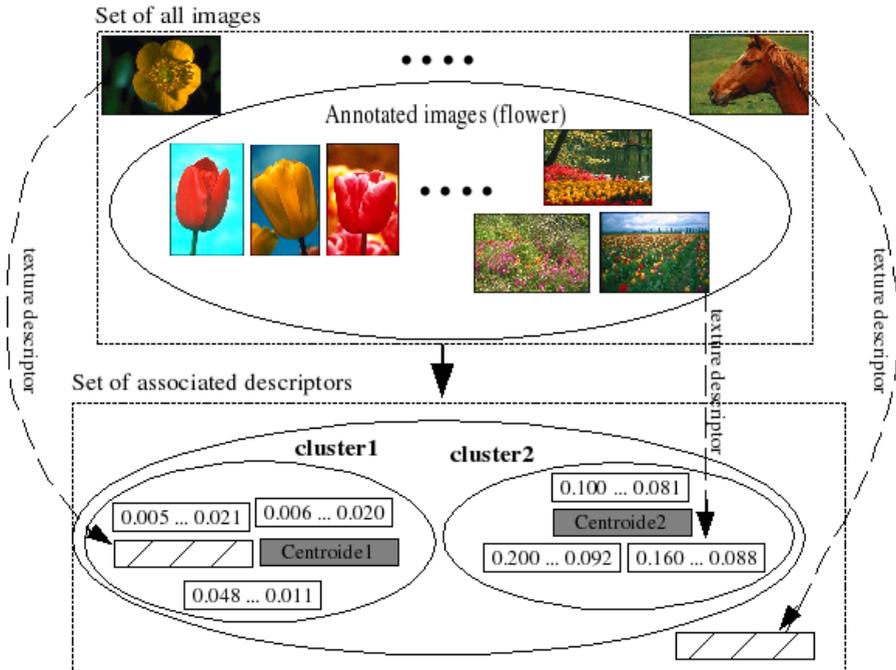


Fig. 1. For each word, the training data is the set of corresponding annotated images which yield to three sets of descriptors (vectors) according to each high-level visual feature. Each set of descriptors is clustered in several regions. The figure shows an example of clustering in 2 regions for the set associated to the texture feature.

dral with contours) which can introduce noise in the automatic retrieval model if they are not relevant. Our approach tries to identify such strong associations between words and visual features.

Our training data for each word is the set of images annotated by this word. This dataset is exploited to obtain several sets of descriptor vectors according to the high-level visual features which will later be associated with the aforementioned (key)word. Each set of descriptors is then clustered by using several number of partitions (cf. Fig. 1 showing an example of clustering associated to a feature with respect to 2 partitions). This clustering allows our system to automatically estimate or capture the optimal number of partitions associated to the number of classes of images in the sens of their visual content (e.g., four types of mountains, six types of cars, etc.). Each cluster is then described by some statistical and spatial characterizations. We also describe the quality and the performance of a query based on the centroid feature (i.e., a model associated to a virtual image) of each clusters. According to some criteria on these descriptions, the key-word is associated with its most representative high-level visual feature, the number of regions used in the clustering and the corresponding cluster centroid.

This unsupervised learning process also allows to propose a new image retrieval method by prompting the user to submit both a query image and a query key-word. To this end, the centroid of the cluster which contains the descriptor of the query image (and which can be viewed as the learned semantic concept of the key-word) can be exploited as a virtual image to perform the query. In particular, this visual semantic allows to retrieve similar images to the image query in the sens of the visual semantic of the given key-word.

1.3 Outline of the Paper

The reminder of the paper is organized as follows: In section 2, we will present the image processing techniques developed for this retrieval system; i.e., the considered visual features (texture, contours and shape/color) as well as their corresponding similarity measures. In section 3, we will describe the way that relationships between keywords and visual features are extracted by the means of a learning procedure. In section 4, we will present some experimental results on the annotated ‘*St Andrews University Library Photographic Collection*’ and *Corel*© databases and we conclude.

2 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 and an associated discriminant measure of similarity $S_{feature}$.

Edge Descriptor: Wavelet-based measures have often been used in content-based image retrieval (CBIR) systems because of their appealing ability to de-

scribe the local texture and the distribution of the edges of a given image at multiple scales. We use the Harr wavelet transform on the gray-level component of the image. 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. Three scales 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 $\{\mu_{n=1}, \sigma_{n=1}, \dots, \mu_{n=10}, \sigma_{n=10}\}$ of 20 components. For this descriptor, the similarity measure (S_{edge}) we use is the weighted-mean-variance distance.

Texture Descriptor: Tamura *et al.* [9] have proposed to characterize image texture along the dimensions of contrast, directionality, coarseness, line-likeness, regularity and roughness. 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 [9]. First, we set some predefined texture resolutions $\{2, 8, 14, 20, 26, 32, 38\}$ instead of $2^k \times 2^k$ with $k = 0, 1, \dots, 6$, then, 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. The similarity measure ($S_{texture}$) used is the Jeffrey divergence [10].

Shape and Color Descriptor: Extraction of shapes contained in an image remains a difficult task. Following [11], 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 similarity measure S_{shape} used for this 36-bin histogram is the weighted-mean-variance distance.

3 Associating Words with Representative Images and Features

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

the feature are considered to be characteristic for the word. In this way, each word can be associated 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 . $|\cdot|$ 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 sets of descriptors $\mathbf{D}_{I_w}^{Texture}$, $\mathbf{D}_{I_w}^{Edge}$ and $\mathbf{D}_{I_w}^{Shape}$. We will use the notation $\mathbf{D}_{I_w}^{feature}$ to refer to each of these descriptors.

For a fixed number of regions (we consider 1, 2, ..., 5 regions in our case), we use the Generalized Lloyd [12] algorithm to cluster each set $\mathbf{D}_{I_w}^{feature}$ in R partitions, thus, we obtain several ${}^R_c\mathbf{D}_{I_w}^{feature}$ 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 $\mathbf{D}_{I_w}^{feature}$ is the similarity measure of the feature $S_{feature}$. 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 ${}^R_cP_{I_w}^{feature}$. 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 ${}^R_c\mathbf{D}_{I_w}^{feature}$, 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 *feature*. 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 ${}^R_c\mathbf{D}_{I_w}^{feature}$, we count the number of images annotated by the word w that are retrieved among the first X ($X = 20$ in our case) retrieved images for each *Feature*. Let $topX^{feature}$ be this number. We count the sum of the $topX^{feature}$ 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 ${}^R_c\mathbf{D}_{I_w}^{feature}$ by three measures: its proportion ρ within \mathbf{I}_w (simply, $\rho = |{}^R_c\mathbf{D}_{I_w}^{feature}|/|\mathbf{I}_w|$), its standard deviation σ (computed according to the similarity measure of *feature*), and an empirical measure P which represents the number of images, not annotated by the word w , for which the distance between its descriptor vector and the prototype vector ${}^R_cP_{I_w}^{feature}$ is less than the pre-estimated standard deviation σ , namely

$$P = |\{I \notin \mathbf{I}_w \mid S_{feature}({}^R_c\mathbf{D}_{I_w}^{feature}, {}^R_cP_{I_w}^{feature}) < \sigma\}|/|I|$$

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 greater 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 (i.e., vectors associated with high peak spherical distribution).

4 Experimental Results and Conclusion

The experimental results are based on the historical image database ‘*St Andrews University Library Photographic Collection*’ provided by *ImageCLEF 2004* [13]. This database contains 28133 images with caption. The caption text associated to each image contains around tens of (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 due to the presence of some (key)words used in the request with an abstract concept. (“Scotland”, “north”, “tournament”, etc.). For our experiments, we have also considered 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 remove each word from the caption of 50% of associated images. We use these images as references and 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 reference images and the non-annotated images retrieved but also related to the word in consideration.

Figure 2 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 or due to the color component included with shape descriptor), 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. Figure 4 shows three semantic query results for the words flower, canal and grass: the algorithm described in 3 was used to produce these results. It shows also a query for word grass according to its second relevant feature. Even if the reference images were not retrieved successfully, we can see that most of images are related to the query word.

database	word	selected feature			Number of training vectors
		Feature 1	Feature 2	Feature 3	
C	water	contours (74)	shape (65)	texture (61)	2550
	sky	contours (66)	texture (65)	shape (60)	2323
	tree	texture (85)	contours (79)	shape (72)	2242
	people	contours (76)	texture (60)	shape (51)	1908
	grass	contours (35)	shape (28)	texture (27)	1061
	flower	shape (61)	contours (51)	texture (16)	934
O	wild	contours (17)	texture (15)	shape (15)	707
	bird	texture (24)	contours (12)	shape (9)	595
R	plant	contours (13)	shape (10)	texture (8)	439
	garden	texture (14)	contours (14)	shape (14)	301
	sunset	shape (19)	contours (15)	texture (8)	260
E	ice	contours (8)	texture (6)	shape (5)	240
	ocean	shape (44)	contours (26)	texture (15)	231
L	animal	shape (11)	texture (7)	contours (3)	204
	ski	contours (4)	shape (1)	texture (0)	153
	texture	shape (17)	contours (10)	texture (8)	126
	rural	contours (7)	texture (3)	shape (3)	124
	insect	contours (10)	shape (7)	texture (1)	123
	tiger	texture (14)	contours (10)	shape (9)	73
	zebra	texture (13)	contours (9)	shape (8)	26
St-	street	contours (119)	shape (101)	texture (96)	2348
	church	contours (57)	texture (48)	shape (48)	2721
	boat	texture (61)	shape (40)	contours (37)	1740
AND-	golfer	texture (18)	shape (14)	contours (10)	309
	canal	texture (3)	shape (3)	contours (2)	178
REW	swing	texture (8)	contours (1)	shape (1)	94

Fig. 2. A list of concepts with their discriminative features ranked by the sum of $top20^{feature}$ over all the clusters of the feature (criterion used to choose the most discriminative feature or eventually to combine several features)

Coref word	top10	top20	top50	top100	ref10	ref20	ref50	ref100	vis20	vis40	vis60
flower (shape)	2	2	3	7	2	3	5	8	9	17	28
animal (shape)	1	1	2	3	0	0	0	0	6	9	16
birds (texture)	1	1	4	5	1	1	3	5	3	7	9
ice (contours)	0	0	0	1	0	0	0	1	0	0	0
grass (contours)	0	0	0	5	0	1	1	4	9	15	26

St-Andrew word	top10	top20	top50	top100	vis20	vis40	vis60
canal (texture)	0	1	1	2	10	17	29
street contours)	1	4	14	26	12	26	37
boat (texture)	1	4	8	10	4	9	12

Fig. 3. Some statistics about the top retrieved images for some words. topX is the number of images annotated by the word among the first X retrieved images. Identically, refX and visX are related respectively to reference images and visually accepted images (a subjective judgment).

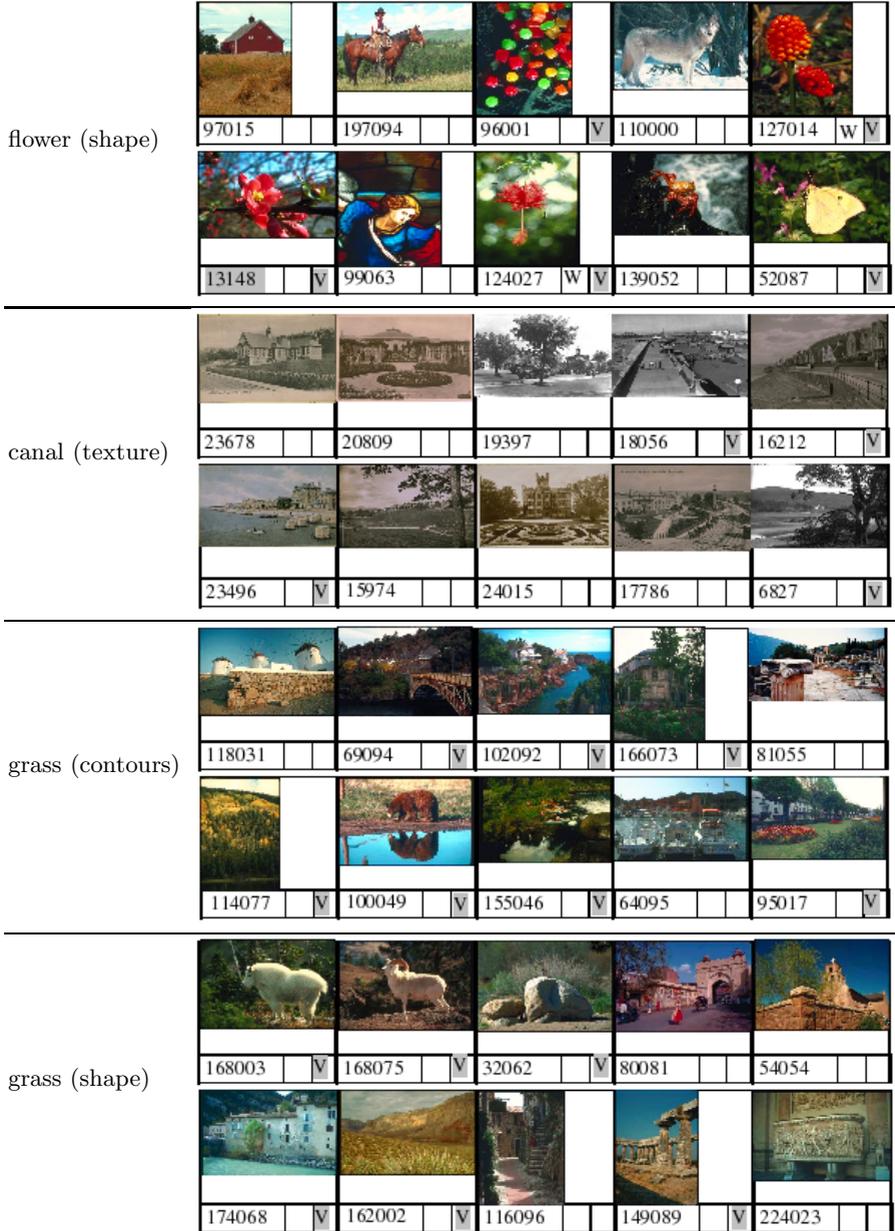


Fig. 4. Semantic query results for concepts flower (shape), canal (texture) and grass (contours). The last query is made according to the best cluster of feature shape. The identification number is shown above each image. Annotated images are marked by a W box. Visually related images to the concept are marked by V box. Reference images have their identification number in a gray box.

References

1. Yixin Chen, James Z. Wang, and Robert Krovetz. Content-based image retrieval by clustering. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, New York, NY, USA, 2003. ACM Press.
2. Aditya Vailaya, A. T. Figueiredo, Anil K. Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130, 2001.
3. 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.
4. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.
5. 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.
6. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.
7. K. Barnard, P. Duygulu, and D. Forsyth. Modeling the statistics of image features and associated text, 2002.
8. Jia Li and James Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
9. H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:460–473, 1978.
10. 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.
11. M. Goldberg, P. Boucher, and S. Shlien. Image compression using adaptative vector quantization. *Communications, IEEE Transactions on [legacy, pre - 1988]*, 34:180–187, 1986.
12. Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28:84–95, 1980.
13. 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.