Content-Based Image Retrieval Using Wavelet-based Salient Points[†]

Q. Tian[1], N. Sebe[2], M.S. Lew[2], E. Loupias[3], T. S. Huang[1]

[1]Beckman Institute, 405 N. Mathews, Urbana, IL 61801 {qitian, huang}@ifp.uiuc.edu
[2]Leiden Institute of Advanced Computer Science, Leiden, The Netherlands {nicu, mlew}@liacs.nl
[3]Laboratories Reconnaissance, de Formes et Vision, INSA-Lyon, France loupias@rfv.insa-lyon.fr

ABSTRACT

Content-based Image Retrieval (CBIR) has become one of the most active research areas in the past few years. Most of the attention from the research has been focused on indexing techniques based on global feature distributions. However, these global distributions have limited discriminating power because they are unable to capture local image information. Applying global Gabor texture features greatly improve the retrieval accuracy. But they are computationally complex. In this paper, we present a wavelet-based salient point extraction algorithm. We show that extracting the color and texture information in the locations given by these points provides significantly improved results in terms of retrieval accuracy, computational complexity and storage space of feature vectors as compared to the global feature approaches.

Keywords: CBIR, Haar wavelet, wavelet-based salient points, Gabor filter

1. INTRODUCTION

Recent years have witnessed a rapid increase of the volume of digital image collections, which motivates the research of image retrieval. [1,2,3] Early research in image retrieval proposed manually annotated images for their retrieval. However, these text-based techniques are impractical for two reasons: large size of image databases and subjective meaning of images. To avoid manual annotation, an alternative approach is content-based image retrieval (CBIR), by which images would be indexed by their visual contents such as color, texture, shape, etc. Many research efforts have been made to extract these low-level image features, [4,5] evaluate distance metrics, and look for efficient searching schemes. [8,9]

In a typical content-based image database retrieval application, the user has an image he or she is interested in and wants to find similar images from the entire database. A two-step approach to search the image database is adopted. First, for each image in the database, a feature vector characterizing some image properties is computed and stored in a feature database. Second, given a query image, its feature vector is computed, compared to the feature vectors in the feature database, and images most similar to the query images are returned to the user. The features and the similarity measure used to compare two feature vectors should be efficient enough to match similar images as well as being able to discriminate dissimilar ones.
In this context, an image index is a set of features, often computed from the entire image. However natural images are mainly heterogeneous, with different parts of the image with different characteristics, which cannot be handled by these *global* features.

Local features can be computed to obtain an image index based on local properties of the image. These local features, which need to be discriminant enough to "summarize" the local image information, are mainly based on filtering, sometimes at different image scales. These kinds of features are too time-consuming to be computed for each pixel of the image. Therefore the feature extraction is limited to a subset of the image pixels, the *interest points*, [10,11,12,13] where the image information is supposed to be the most important.

Besides saving time in the indexing process, these points may lead to a more discriminant index because they are related to the visually most important parts of the image. Schmid and Mohr introduced the notion of *interest point* in image retrieval. [10] To detect these points, they compute local invariants. They use the Harris' detector, one of the most popular corner detectors. This detector, as many others, was initially designed for robotics, and it is based on a mathematical model for corners. The original goal was to match same corners from a pair of stereo images, to obtain a representation of the 3D scene. Since corner

[†] Appear in the proceedings of *SPIE Photonics West, Electronic Imaging 2001*, *Storage and Retrieval for Media Databases,* January 21-26, 2001, San Jose, California

detectors were not designed to give a "summary" as comprehensive as possible of an image, they have drawbacks when applied to various natural images for image retrieval:

1. **Visual focus points need not to be corners**: when looking at a picture, we are attracted by some parts of the image, which are the most meaningful for us. We cannot assume them to be located in corner points, as mathematically defined in most corner detectors. For instance, smoothed edges can also be visual focus points, and they are usually not detected by a corner detector. The image index we want to compute should describe them as well.
2. **Corners may gather in small regions**: in various natural images, regions may well contain textures (trees, shirt patterns, etc). Many gathered corners are detected in these regions by a corner detector. However, a preset number of points per image are used in the indexing process, to limit the indexing computation time. With this kind of detectors, most of the points are in a small region, and the local features are computed from the same texture region, while other parts of the image will not be described in the index at all.

For these reasons, corner points, as designed in robotics, may not represent the most interesting subset of pixels for image indexing. Indexing points should be related to any visual "interesting" part of the image, whether it is smooth or corner-like. To describe different parts of the image, the set of interesting points should not be clustered in few regions. From now on, we will refer to these points as *salient points*, which are not necessarily corners. We will avoid the term *interest points*, which is ambiguous, since it was previously used in the literature as *corner*.

We believe multi-resolution representation is interesting to detect this kind of points. We will present a salient point extraction algorithm using the wavelet transform, which expresses image variations at different resolutions. Wavelet-based salient points are detected for smoothed edges and are not gathered in texture regions. Hence, they lead to a more complete image representation than corner detectors. [14]

In this paper, our idea is first to extract salient points in the image and then in their location to extract local color and texture features. It is quite easy to understand that using a small amount of such points instead of all images reduces the amount of data to be processed. Moreover, local information extracted in the neighborhood of these particular points is assumed to be more robust to classic transformations (additive noise, affine transformation including translation, rotation and scale effects, partial visibility, etc.).

The rest of paper is organized as follows. A wavelet-based salient point extraction algorithm will be presented in Section 2. Color features and texture features adopted in this paper will be discussed in Section 3 and 4, respectively. Similarity measurement will be described in Section 5 and the experimental results will be given in Section 6. Finally, discussion will be given in Section 7.

2. WAVELET-BASED SALIENT POINTS

The wavelet representation gives information about the variations in the image at different scales. In our retrieval context, we would like to extract salient points from any part of the image where "something" happens in the image at any resolution. A high wavelet coefficient (in absolute value) at a coarse resolution corresponds to a region with high global variations. The idea is to find a relevant point to represent this global variation by looking at wavelet coefficients at finer resolutions.

A wavelet is an oscillating and attenuating function with zero integral. We study the image f at the scales (or resolutions) $1/2,\ 1/4,\ldots,2^j$, $j \in \mathbf{Z}$ and $j \le -1$. The wavelet detail image $W_{2^j} f$ is obtained as the convolution of the image with the wavelet function dilated at different scales. We considered orthogonal wavelets with compact support. First, this assures that we have a complete and non-redundant representation of the image. Second, since the wavelets have a complete support, we know from which signal points each wavelet coefficient at the scale 2^j was computed. We can further study the wavelet coefficients for the same points at the finer scale 2^{j+1}. There is a set of coefficients at the scale 2^{j+1} computed with the same points as a coefficient $W_{2^j} f(n)$ at the scale 2^j. We call this set of coefficients the children $C(W_{2^j} f(n))$ of the coefficient $W_{2^j} f(n)$. The children set in one dimension is:

$$C(W_{2^j} f(n)) = \{W_{2^{j+1}} f(k), 2n \le k \le 2n + 2p - 1\} \quad (1)$$

where p is the wavelet regularity and $0 \le n \le 2^j N$ with N the length of the signal.

Each wavelet coefficient $W_{2^j} f(n)$ is computed with $2^{-j} p$ signal points. It represents their variation at the scale 2^j. Its children coefficients give the variations of some particular subsets of these points (with the number of subsets depending on the wavelet). The most salient subset is the one with the highest wavelet coefficient at the scale 2^{j+1}, that is the maximum in absolute value of $C(W_{2^j} f(n))$. In our salient point extraction algorithm, we consider this maximum, and look at his highest child. Applying recursively this process, we select a coefficient $W_{2^{-1}} f(n)$ at the finer resolution $1/2$ (Figure 1(b) and (c)). Hence, this coefficient represents *2p* signal points. To select a salient point from this tracking, we choose among these *2p* points the one with the highest gradient. We set its saliency value as the sum of the absolute value of the wavelet coefficients in the track:

$$saliency = \sum_{k=1}^{-j} | C^{(k)}(W_{2^j} f(n)) |, \quad -\log_2 N \le j \le -1 \qquad (2)$$

The tracked point and its saliency value are computed for every wavelet coefficient. A point related to a global variation has a high saliency value, since the coarse wavelet coefficients contribute to it. A finer variation also leads to an extracted point, but with a lower saliency value. We then need to threshold the saliency value, in relation to the desired number of salient points. We first obtain the points related to global variations; local variations also appear if enough salient points are requested.

The salient points extracted by this process depend on the wavelet we use. *Haar* is the simplest orthogonal wavelet with compact support, so is the fastest for execution. The larger the spatial support of the wavelet, the more the number of computations. Nevertheless, some localization drawbacks can appear with *Haar* due to its non-overlapping wavelets at a given scale. This drawback can be avoided with the simplest overlapping wavelet, *Daubechies 4*. However, this kind of drawback is not likely in natural images and therefore, we used *Haar* transform in our experiments. In Figure 1(a), we present the salient points using the *Haar* transform. Note that our method extracts salient points not only in the foreground but also in the background where some smooth details are present.

(a) Salient points **(b) Tracked coefficients** **(c) Spatial support of tracked coefficients**

Figure 1: Salient points extraction

3. COLOR FEATURES

Of the visual media retrieval methods, color indexing is one of the dominant methods because it has been shown to be effective in both the academic and commercial arenas. In color indexing, given a query image, the goal is to retrieve all the images whose color compositions are similar to the color composition of the query image. In color indexing, color histograms are often used because they are sufficient accuracy. [15] While histograms are useful because they are relatively insensitive to position and orientation changes, they do not capture spatial relationship of color regions and thus, they have limited discriminating power. Stricker, et al., [16] showed that characterizing one dimensional color distributions with the first three moments is more robust and more efficient than working with color histograms.

The idea of using color distribution features for color indexing is simple. In the index we store dominant features of the color distributions. The retrieval progress is based on similarity function of color distributions. The mathematical foundation of this approach is that any probability distribution is uniquely characterized by its moments. Thus, if we interpret the color distribution of an image as a probability distribution, then the color distribution can be characterized by its moments. [16]

Furthermore, because most of the information is concentrated on the low-order moments, only the first moment (mean), the second and the third central moments (variance and skewness) were used. If the value of the *i*-th color channel at the *j*-th image pixel is I_{ij} and the number of image pixels is *N*, then the index entries related to this color channel are:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} I_{ij}\,, \qquad \sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(I_{ij}-\mu_i)^2\right)^{\frac{1}{2}}, \qquad s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(I_{ij}-\mu_i)^3\right)^{\frac{1}{3}} \tag{3}$$

We were working with the HSV color space so, for each image in the database a 9-dimensional color feature vector was considered.

4. TEXTURE FEATURES

Color indexing is based on the observation that often color is used to encode functionality (sky is blue, forests are green) and in general will not allow us to determine an object's identity. [17] Therefore, texture or geometric properties are needed to identify objects. [18] Consequently, color indexing methods are bound to retrieve false positives, i.e., images, which have a similar color composition as the query image but with a completely different content. Therefore, in practice, it is necessary to combine color indexing with texture and/or shape indexing techniques.

Texture analysis is important in many applications of computer image analysis for classification, detection or segmentation of images based on local spatial patterns of intensity or color. Textures are replications, symmetries and combinations of various basic patterns or local functions, usually with some random variation. Textures have the implicit strength that they are based on intuitive notions of visual similarity. This means that they are particularly useful for searching visual databases and other human computer interaction applications. However, since the notion of texture is tied to the human semantic meaning, computational descriptions have been broad, vague and something conflicting.

The method of texture analysis chosen for feature extraction is critical to the success of texture classification. Many methods have been proposed to extract texture features either directly from the image statistics, e.g., co-occurrence matrix, or from the spatial frequency domain. [19] Ohanian and Dubes [20] studied the performance of four types of features: Markov Random Fields parameters, Gabor multi-channel features, fractal-based features and co-occurrence features. Comparative studies to evaluate the performance of some texture measures were made. [21, 22]

Recently there was a strong push to develop multi-scale approaches to the texture problem. Smith and Chang [23] used the statistics (mean and variance) extracted from the wavelet subbands as the texture representation. To explore the middle-band characteristics, tree-structured wavelet transform was studied by Chang and Kuo. [24] Ma and Manjunath [25] evaluated the texture image annotations by various wavelet transform representations, including orthogonal and bi-orthogonal, tree-structured wavelet transform, and Gabor wavelet transform (GWT). They found out that Gabor transform was the best among the tested candidates, which matched the human vision study results. [26]

Gabor filters produce spatial-frequency decompositions that achieve the theoretical lower bound of the uncertainty principle. They attain maximum joint resolution in space and spatial-frequency bounded by the relations $\Delta_x^2 \cdot \Delta_u^2 \geq \frac{1}{4\pi}$ and $\Delta_y^2 \cdot \Delta_v^2 \geq \frac{1}{4\pi}$, where $[\Delta_x^2, \Delta_y^2]$ gives resolution in space and $[\Delta_u^2, \Delta_v^2]$ gives resolution in spatial-frequency. In addition to good performances in texture discrimination and segmentation, the justification for Gabor filters is also supported through psychophysical experiments. Texture analyzers implemented using 2-D Gabor functions produce a strong correlation with actual human segmentation. [27] Furthermore, the receptive visual field profiles are adequately modeled by 2-D Gabor filters. [28] Gabor functions are Gaussian modulated by complex sinusoids. In two dimensions they take the form: [28]

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right) \tag{4}$$

The Gabor filter masks can be considered as orientation and scale tunable edge and line detectors. The statistics of these microfeatures in a given region can be used to characterize the underlying texture information. A class of such self-similar

functions referred to as Gabor wavelets is discussed. [29] This self-similar filter dictionary can be obtained by appropriate dilations and rotations of $g(x, y)$ through the generating function,

$$g_{mn}(x, y) = a^{-m} g(x', y') \tag{5}$$
$$x' = a^{-m}(x\cos\theta + y\sin\theta),\ \ y' = a^{-m}(-x\sin\theta + y\cos\theta)$$

where $\theta = 2\pi / K$, K the number of orientations, $m = 0,1,\cdots, S-1$, S the number of scales in the multi-resolution decomposition, and $a = (U_h / U_l)^{-1/(S-1)}$ with U_l and U_h the lower and the upper center frequencies of interest.

5. SIMILARITY MEASUREMENT

The image similarity is a fuzzy concept, which must be clarified. For user, the implicit image similarity is usually based on the human perceptual similarity. However, this kind of descriptions cannot be extracted automatically from the image without specific knowledge. Image similarity is therefore mainly based on low-level features, such as color, texture and shape.

In MARS [30] system, the overall similarity distance D_j for the *j*-th image in the database is obtained by linearly combining individual feature's similarity distance of the *i*-th feature for the *j*-th image in the database, $S_j(f_i)$:

$$D_j = \sum_i W_i S_j(f_i) \qquad j = 1,...,N \tag{6}$$

where *N* is the total number of the images in the database.

Case 1: If the number of the query image is one, $S_j(f_i)$ is defined as:

$$S_j(f_i) = (\mathbf{x}_i - \mathbf{q}_i)^T (\mathbf{x}_i - \mathbf{q}_i) \qquad j = 1,...,N \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{q}_i$ are the *i*-th feature (*e.g. i=1* for color and *i=2* for texture) vector of the *j*-th image in the database and the query, respectively.

The low-level feature weights for color and texture in Eq. (6) is set to be equal weight in this case.

Case 2: If the number of the query images is greater than one and a preference weight for each query image is assigned, e.g. how much the user likes each query image, $S_j(f_i)$ is a *Mahalanobis* distance defined as

$$S_j(f_i) = (\mathbf{x}_i - \mathbf{q}_i)^T C_i^{-1} (\mathbf{x}_i - \mathbf{q}_i) \quad j = 1,...,N \tag{8}$$

where $\mathbf{x}_i$ is the *i*-th feature vector of the *j*-th image in the database and $\mathbf{q}_i$ is the weighted mean feature vector of the query images, respectively. C_i is the covariance matrix of the *i*-th feature components of the query images. The element of C_i is defined as:

$$C_i(m,n) = \frac{\sum_{k=1}^{L} V(k)[r_i(k,m) - q_i(m)][r_i(k,n) - q_i(n)]}{\sum_{k=1}^{L} V(k)} \tag{9}$$

where *V(k)* is the preference weight for the *k*-th query image, $r_i(k,m)$ and $r_i(k,n)$ are the *m*-th and *n*-th component values of the *i*-th feature of the *k*-th query image, respectively. $q_i(m)$ and $q_i(n)$ are the *m*-th and *n*-th component values of the *i*-th feature of the query, respectively. *L* is the total number of the query images and *L>1*.

In case 2, low-level feature weight W_i in Eq. (6) is defined as:

$$W_i = \frac{1}{d_i} \quad \text{where } d_i = \frac{\sum_{k=1}^{L} V(k) S_k(f_i)}{\sum_{k=1}^{L} V(k)} \tag{10}$$

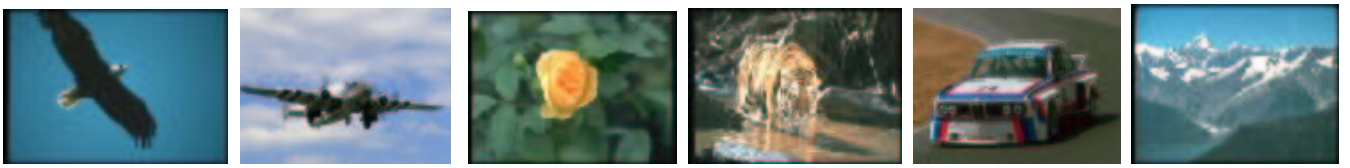
The higher weight is given to the feature that has the smaller average distance d_i, based on the query images. This is because the query images are more similar, i.e., have smaller distance in this feature than other features.

Clearly, case 1 is a special case of case 2 by setting the covariance matrix C_i to be the identity matrix and $L=1$.

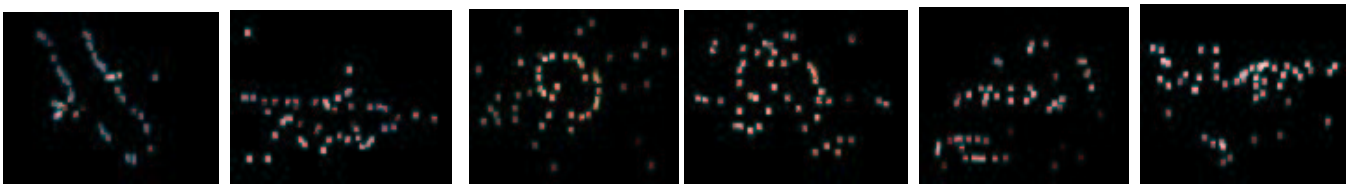
In this paper, we are mainly considering the case 1 where we submit one query image and want to evaluate performance of the different algorithms. The relevance feedback, [5, 30] i.e., submitting more than one query images that are considered to be relevant to each other, is not our main interest in this work.

6. RESULTS

The setup of our experiments was the following. First we extracted a fixed number of salient points for each image in the database using *Haar* wavelet transform and the algorithm described in Section 2. The number of salient points cannot be too small or too large. According to our experiments, 50 ~ 100 is a reasonable range for the number of salient points. Figure 2 shows some salient points examples. The original images are shown in Figure 2(a). Their salient point maps are shown in Figure 2(b). Clearly, the local information of the objects, i.e., bird, airplane, flower, tiger, car and mountain, is captured by the salient points.



(a) Original images



(b) Salient points (50 points) and their 3×3 neighborhood

Figure 2 Examples of salient points (a) Original images (b) Salient point map (c) Salient points and their 3×3 neighborhood

(a) Original images

(b) Salient points together with their 3×3 neighborhood

Figure 3. An experimental result using the color moments extracted from the 3×3 neighborhood of the salient points (rank from the left to right and from top to bottom, the top left is the query image)

For feature extraction, we considered the pixels in a small neighborhood around each salient point that form the image signature. For each image signature in the database we computed the color moments for color and the Gabor moments for texture. In this paper, 3×3 neighborhood for color moment extraction and 9×9 neighborhood for Gabor moment extraction were used. For convenience, this approach is denoted as *salient approach*.

When the user selects one query image, the system computes the corresponding feature vector from the query image signature and compares it with the feature vector of every image in the database. In MARS [30] system, the color moments [16] and wavelet moments [23] were extracted from the entire images to form the feature vectors. In this paper, this approach is denoted as *Global CW approach*. For benchmarking purposes we also considered the results obtained using the color moments and Gabor texture features obtained over the entire image. This approach is denoted as *Global CG approach*. The above-mentioned three approaches will be compared in the following experiments.

In the first experiment we considered a small database consisting of 142 images of 7 classes such as airplane (21 images), bird (27 images), car (18 images), tiger (18 images), flower (19 images), mountain (19 images) and church paintings (20 images). All images in the database have been labeled as one of these classes and this serves as the ground truth.

Figure 3 shows an experimental result using the color moments extracted from the 3×3 neighborhood of the salient points. The top 20 retrieved images are shown in Figure 3(a). The top left image is the query image. The corresponding salient points together with their 3×3 neighborhood are shown in Figure 3(b).

In order to test the retrieval results for each individual class, we randomly picked 5 images from each class and used them as queries. For each individual class we computed the retrieval accuracy as the average percentage of images from the same class as the query image that were retrieved in the top 15 images. Only the color feature was used. Thus the comparison was between the salient approach and the global approach. The results are given in Table 1. The retrieval accuracy was given by the average percentage of correct images from the same class as the query that was retrieved in top *n* images.

Table 1: Retrieval accuracy (%) for each individual class using 5 randomly chosen images from each class as queries

Class	Salient	Global
Airplane	88	86
Bird	97	97
Tiger	89	81
Car	63	49
Flower	58	60
Church painting	93	98
Mountain	97	100

From this experiment we can see that for some classes that were mainly composed of a single object on a simple background (e.g., Airplane, Bird where the background represented the blue sky), both the salient and the global approach have the similar performance. However, they are very different in the way the similarity is found. For the global approach, the color moments were extracted from the entire image and therefore they were determined by the dominant background, e.g., blue sky. In this sense, these airplanes or birds were found to be similar because of the background, not the object themselves. In the salient approach, the salient points were mostly found on the boundaries of the airplanes and birds. The local color moments around the neighborhood of the salient points were extracted and they represented the object information instead of the background. So in this sense, the images were found to be similar in terms of the object, not the background. Therefore, although both approaches give the similar retrieval results, the salient approach captures the user's concept more accurately than the global approach in terms of object finding. When the classes have the complex background (e.g., Tiger, Car) that makes the retrieval difficult the salient approach performs better than the global approach. For the Flower, the global approach performs a little better than the salient approach. This is possibly due to the fact that these images are found to be more similar because of the dominant green background. When the image shows more global variations, (e.g., Church Painting, Mountain), both approaches perform very well and the global approach is slightly better than the salient approach. However this fact still shows that the salient approach can capture global image information (background) as well.

In our second experiment we considered a database of 479 images of color objects such as domestic objects, tools, toys, food cans, etc. The size of the image is 256×256. As ground truth we used 48 images of 8 objects taken from different camera viewpoints (6 images for a single object). The problem is formulated as follows:

Let $Q_1, \cdots, Q_n$ be the query images and for the *i*-th query Q_i, $I_1^{(i)},\cdots,I_m^{(i)}$ be the images similar with Q_i according to the ground truth. The retrieval method will return this set of answers with various ranks.

In this experiment both color and texture information was used. Three approaches, the salient approach, the Global CW approach and the Global CG approach were compared. Color moments were extracted either globally (the Global CW and Global CG) or locally (the salient approach). For wavelet texture representation of the Global CW approach, each input image was first fed into a wavelet filter bank and was decomposed into three wavelet levels, thus 10 de-correlated subbands. Each subband captured the feature of some scale and orientation of the original images. For each subband, the mean and standard deviation of the wavelet coefficients were extracted. The total number of wavelet texture features was 20. For the salient approach, the wavelet moments used in MARS cannot be applied because they are unable to extract meaningful texture information from an image of very small size, e.g., 9×9 neighborhood of the salient point. Since the Gabor texture was the best candidate for texture classification and can be applied to a small image, [25] we extracted Gabor texture feature from the 9×9 neighborhood of each salient point. The dimension of the Gabor filter was 7×7. We extracted from each neighborhood of the salient points 24 Gabor features using 2 scales and 6 orientations. The first 12 features represented the averages over the filter outputs obtained in order for: scale 1 and orientation 1, $\cdots$, scale 1 and orientation 6, scale 2 and orientation 1, $\cdots$, scale 2 and orientation 6. The last 12 features were the corresponding variances. Note that these features were independent so that they had different ranges. Therefore each feature was then Gaussian normalized over the entire image database.

For the global CG approach, the global Gabor texture features were extracted. The dimension of the global Gabor filter was 61×61. We extracted 36 Gabor features using 3 scales and 6 orientation. The first 18 features were the averages over the filters outputs and the last 18 features were the corresponding variances.

We expect that the salient point method to be more robust to the viewpoint change because the salient points are located around the object boundary and capture the details inside the object, neglecting the noisy background. In Figure 4 we represented an example of a query image and the similar images from the database.

Salient	**1**	**2**	**6**	**12**	**18**
Global CW	**1**	**2**	**4**	**42**	**121**
Global CG	**1**	**2**	**5**	**9**	**21**

Figure 4 Example of images of one object taken from different camera viewpoints. The rank of individual image were obtained using salient point information (Salient), global color moments and wavelet moments (Global CW) and global color moments and Gabor moments (Global CG)

The salient point approach outperforms the global CW approach in capturing the last two images. Even when the image was taken from a very different viewpoint, the salient points captured the object details enough so the similar image was retrieved with a good rank. The global CG approach shows slightly better performance than the salient point approach (except for the last image) and much better performance than the global CW approach. This fact demonstrates that Gabor feature could be a powerful feature for texture classification. However, it should also be noted that: (1) the salient point approach only utilizes the information from a very small part of the image, but still achieves a good representation of the image. This shows that the salient points extracted from the whole image have the representing power. For example, in our object database at most $9\times9\times50$ pixels were used to represent the image. Compared to the global approach (all 256×256 pixels were used), it only utilizes *1/16* of the whole image pixels. (2) Compared to the global CG approach, the salient approach has much less computational complexity. Table 2 shows the computation complexity for the three image databases using the salient point

approach and global approach for extracting Gabor texture features. The computation is done on the same SGI O2 R10000 workstation. The total computational cost for the salient approach comes from the two sources. The first source is the time spent on the salient point extraction. The second source is the Gabor texture feature extraction from the neighborhood of the salient points. From Table 2, the average computational complexity ratio of the global approach to the salient approach is about 6.58 (average of 4.67, 8.06 and 7.02) for the listed three image databases. It can be inferred that the computational complexity difference will be very huge for the very large database, e.g., millions of images. (3) When the computational complexity of the color feature extraction is compared, the salient points approach is much faster than the global approach. Color moments were extracted from 3×3 neighborhood of the salient points and 50 salient points were used. If we consider that the computational cost for the salient points extraction has already been counted in the Gabor texture feature extraction step, then it will not be counted for color feature extraction. Table 3 summarizes the results for the same three databases used in Table 2. As one can see, the computational complexity difference is very large. (4) The number of Gabor texture features used in the salient approach and the global approach were 24 and 36, respectively. This won't have a big effect for small database. However, for very large image database, the storage space used for these texture features will surely make big difference. As to the color features, both approaches use the same number of features.

Table 2: The computational complexity comparison between the salient approach and the global approach for extracting texture feature using Gabor filters

Database	1	2	3
Description	Object	Natural images	Scenery images
Number of Images	479	1505	4013
Resolution	256×256	384×256	360×360
Salient Points Extraction (minutes)	23.9	77.5	225
Salient Gabor feature extraction (minutes)	7.98	37.6	108
Salient total time (minutes)	31.88	115.1	333
Global Gabor feature extraction (minutes)	149	928	2340
Ratio (Global/Salient)	4.67	8.06	7.02

Table 3: The computation complexity of color features for the salient point approach and the global approach

Database	1	2	3
Computation Cost (Global/Salient)	145	218	288

Table 4: Retrieval accuracy (%) using 48 images from 8 classes for object database

Top	6	10	20
Global CW	47.3	62.4	71.7
Global CG	61.2	74.2	84.7
Salient	59.3	73.8	83.2

Table 4 shows the retrieval accuracy for this object database. Each of the 6 images from the 8 classes was considered as query image and the average retrieval accuracy was calculated.

Results in Table 4 show that using the salient point information the retrieval results are significantly improved (>10%) compared to the global CW approach implemented in the MARS [30] system. When compared to the global CG approach, the retrieval accuracy of salient approach are 1.9%, 0.4% and 1.5% lower in the top 6, 10 and 20 images, respectively. Although the salient approach is not the best in terms of the retrieval accuracy among the three approaches, it has the similar performance with the Global CG approach but much lower computational complexity (See Table 2 for texture and Table 3 for color) and 33.3% less storage space of feature vectors than the global CG approach. Although the global wavelet texture features are fast to compute, their retrieval performance is much worse than the other two methods. Therefore, in terms of overall retrieval accuracy, computational complexity and storage space of feature vectors, the salient approach is the best among the three approaches.

In our third experiment, two databases were evaluated. The first database consists of 1505 various natural images. They cover a wide range of natural scenes, animals, buildings, construction sites, textures and paintings. The second database consists of

4013 various scenery pictures. Most of them are outdoor images like mountains, lakes, buildings and roads, etc. Since we don't have the ground truth for these two image databases, it is difficult to obtain the precision-recall curve. Instead, in order to perform some quantitative analysis, we randomly choose 5 images from a few categories, e.g., building, flower, tiger, road, mountains, forest, sunset, and use each of the 5 randomly chosen images as query. We measure how many hits, i.e., how many similar images to the query, are returned in the top 20 retrieved images. The average number of hits is calculated for each category. The query images are randomly chosen from the two testing databases.

Figure 5 shows an example of the retrieved images from a query using the salient point approach. Match quality decreases from the top left to the bottom right. Figure 6 shows the average number of hits for each category using the global CW approach, the global CG approach and the salient point approach. Clearly the salient approach has the similar performance as the global CG approach and outperforms the global CW approach for the first five categories, which are building, flower, tiger, lion, and road. For the last three categories, which are forest, mountain and sunset, the global approaches (both global CW and global CG) perform better than the salient approach. This is reasonable because that the image contents show more global property in the last three categories than the first five categories. Therefore the global approach will result in good performance for these categories.

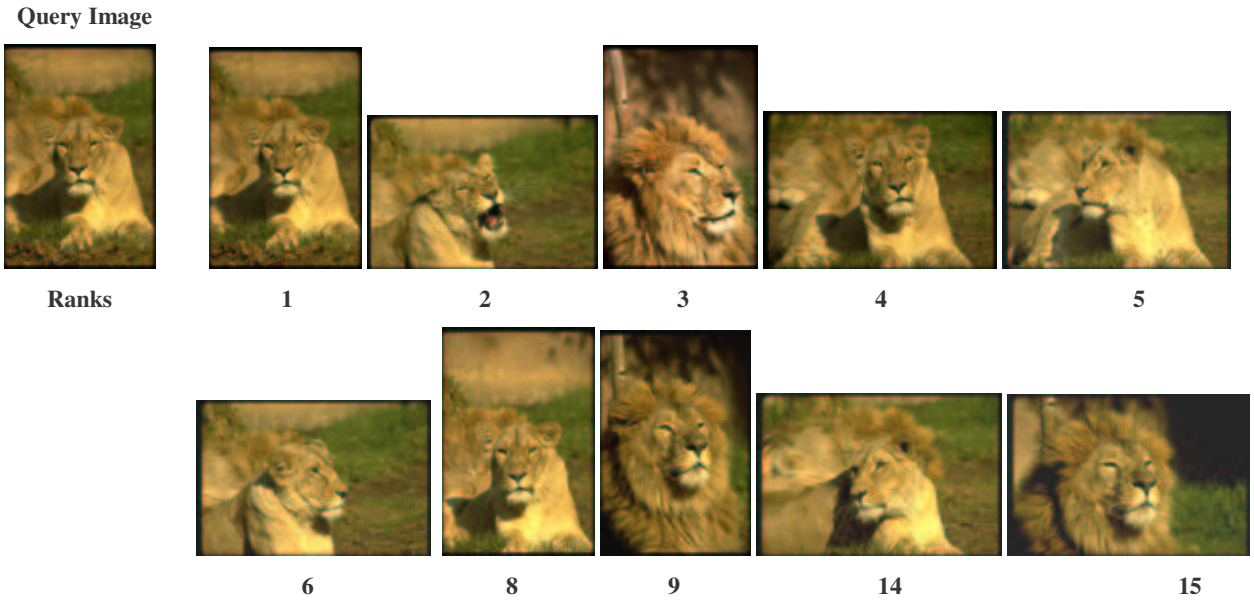


Figure 5 Retrieved images in top 20 returned images from a query using the salient point approach. Match quality decreases from the top left to the bottom right.

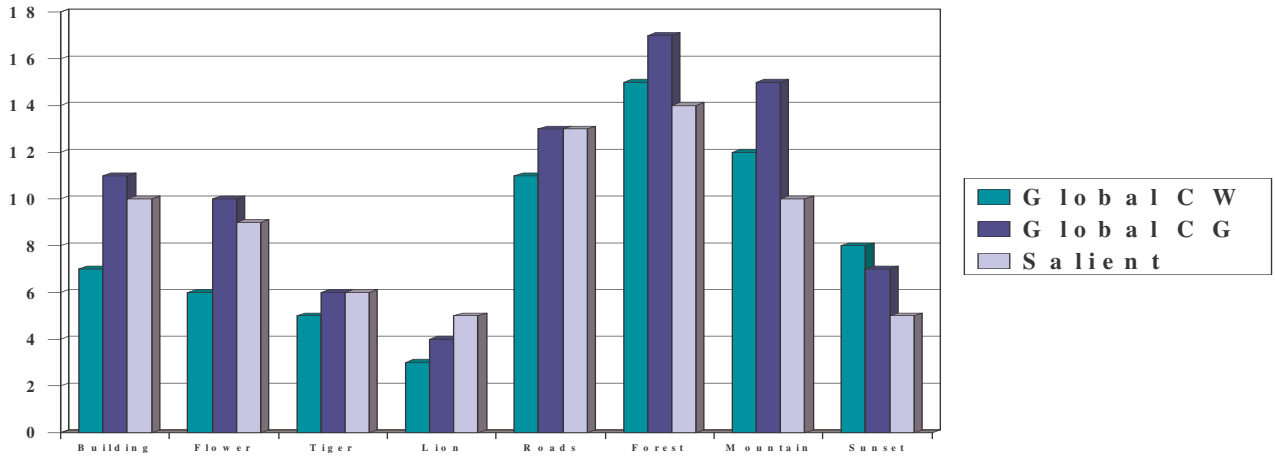


Figure 6. The average number of hits for each category using the global color and wavelet moments (Global CW), the global color and Gabor moments (Global CG) and the salient point approach (Salient)

7. DISCUSSIONS

In this paper, we presented a wavelet-based salient point extraction algorithm and applied it in the content-based image retrieval. The salient points are interesting for image retrieval because they are located in visual focus points and therefore they can capture the local image information. Two demands were imposed for the salient points extraction algorithm. First, the salient points should be located in any visually interesting part of the image. Second, they should not be clustered in few regions.

To accomplish these demands we used a *Haar-based* wavelet salient point extraction algorithm that is fast and captures the image details at different resolutions. A fixed number of salient points were extracted for each image. Color moments for color and Gabor moments for texture were extracted from the 3×3 and the 9×9 neighborhood of the salient points, respectively. For benchmark purpose, the salient point approach was compared to the global color and wavelet moment (Global CW) approach [28] and the global color and Gabor moments (Global CG) approach.

Three experiments were conducted and the results show that (1) the salient point approach has better performance than the global CW approach. The salient point approach proved to be robust to the viewpoint change because the salient points were located around the object boundaries and captured the details inside the objects, neglecting the background influence. (2) The salient point approach has the similar (a little worse) performance compared to the global CG approach in terms of the retrieval accuracy. However, the salient point approach achieves the best performance in the overall considerations of retrieval accuracy, computational complexity and storage space of feature vectors. The last two factors will make very important impact for the very large image database. This is the advantage of the salient approach over the global approach.

Our experimental results also show that the global Gabor features perform much better than the global wavelet features. This fact is consistent with the results of the other researchers in the field. This again proves that Gabor feature is a very powerful candidate for texture classification. [26]

In conclusions, the content-based image retrieval can be improved by using the local information provided by the wavelet-based salient points. The salient points are able to capture the local feature information and therefore, they can provide a better characterization for object recognition.

In our future work, we plan to extract the shape information in the locations of the salient points making the retrieval more accurate and evaluate the optimal number of the salient points needed to be extracted for each image.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundations Grants CDA-96-24396 and EIA-99-75019.

REFERENCES

1. S. Chang, J. Smith, M. Beigi and A. Benitez, “Visual information retrieval from large distributed online repositories”, *Communications of ACM*, Dec. pp. 12-20, 1997.
2. A. Pentland, R. Picard and S. Sclaroff, “Photobook: Content-based manipulation of image database”, Intl. Journal of computer vision, 1996.
3. Y. Rui, T. Huang and S. Chang, “Image retrieval: Current techniques, promising directions and open issues”, *Journal of visual communications and image representation*, Vol. 10, pp. 1-23, 1999.
4. B. Manjunath and W. Ma, “Texture features for browsing and retrieval of image data”, *IEEE PAMI*, Nov. 1996.
5. Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval”, *IEEE Circuits and Systems for Video Technology*, Vol. 8, No.5, pp. 6440655, 1998.
6. M. Popescu and P. Gader, “Image content retrieval from image database using feature integration by Choquet integral”, *SPIE Storage and retrieval for image and video databases*, VII, 1998.
7. S. Santini and R. Jain, “ Similarity measures”, *IEEE PAMI*, Vol. 21, No.9, 1999.
8. D. Swets, J. Weng, “Hierarchical discriminant analysis for image retrieval”, *IEEE PAMI*, Vol. 21, No. 5, pp. 386-400, 1999.
9. H. Zhang, D. Zhong, “A scheme for visual feature based image retrieval”, *Proc. SPIE storage and retrieval for image and video databases*, 1995.

10. C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval", IEEE Trans on PAMI, Vol. 19, No. 5, pp. 530-535, May 1997.
11. S. Bres and J. M. Jolion, "Detection of interest points for image indexation", *3rd Int'l Conf. on Visual Information Systems, Visaul99,* pp.427-434, Amsterdam, The Netherlands, June 2-4,1999.
12. T. Tuytelaars and L. Van Gool, "Content-based image retrieval based on local affinely invariant regions", *3rd Int'l Conf. on Visual Information Systems, Visaul99,* pp.493-500, Amsterdam, The Netherlands, June 2-4,1999.
13. S. Bhattacharjee and T. Ebrahimi, "Image retrieval based on structural content", *Workshop on Image Analysis for Multimedia Interface Services*, Heinrich-Hertz-Institui (HHI), Berlin, German, May 31 – June 1, 1999.
14. N. Sebe, Q.Tian, E. Loupias, M.S. Lew, T.S.Huang, "Color indexing using wavelet-based salient points", will appear in *IEEE workshop on content-based access of image and video libraries*, in conjunction with *IEEE CVPR'2000*, Hilton Head Island, South Carolina, June 13-16, 2000.
15. M. J. Swain and D. H. Ballard, "Color indexing", IJCV 7(1): 11-32, 1991.
16. M. Stricker and M. Prengo, "Similarity of color images", *SPIE – Storage and Retrieval for Image and Video Databases*, 1995.
17. M. Stricker and A. Dimai, "Spectral covariance and fuzzy regions for image indexing", *SPIE –Storage and Retrieval for Image and Video Databases*, 1997.
18. M. Flicker, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system", *IEEE Computer*, 28(9): 23-32, 1995.
19. L. Van Gool, P. Dewaele, and A. Oosterlinck, "Texture Analysis", CVGIP, 29(3): 336-357, 1985.
20. P. Ohanian and R. Dubes, "Performances evaluation for four classes of textural features", *Pattern Recognition*, 25:819-833, 1992.
21. T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distribution", *Pattern Recognition*, 29:51-59, 1996.
22. T. Reed and H. Wechsler, "A review of recent texture segmentation and feature extraction techniques", *CVGIP*, 57(3): 359-373, 1993.
23. J. Smith and S.-F. Chang, "Transform features for texture classification and discrimination in large image database", *IEEE Intl. Conf. on Image Proc*., 1994.
24. T. Chang and C. Kuo, "Texture analysis and classification with tree-structured wavelet transform", *IEEE Trans. Image Proc*., 2(4): 429-441, 1993.
25. Y. Ma and B. Manjunath, "Texture features of wavelet transform features for texture image annotation", *IEEE Intl. Conf. on Image Proc*., 1995.
26. J. Beck, A. Sutter, and R. Ivry, "Spatial frequency channels and perceptual grouping in texture segregation", *Computer Vision, Graphics, and Image Processing*, 37, 1987.
27. T. Reed and H. Wechsler, "Segmentation of textured images and gestalt organization using spatial/spatial-frequency representations", *IEEE Trans. Pattern. Anal. Mach. Intell*., 12(1):1-12, 1990.
28. J. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive field", *IEEE Trans. on Biomedical Engineering*, 36(1), 1989.
29. W. Ma and B. Manjunath, "Texture features and learning similarity", *IEEE CVPR*, 1996.
30. Y. Rui, T. S. Huang, S. Mehrotra, "Content-based image retrieval with relevance feedback in MARS", *Proc. of ICIP, Santa Barbara*, 1997.