Salient Points for Content-based Retrieval

Nicu Sebe Michael S. Lew
Leiden Institute of Advanced Computer Science
Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands
{nicu mlew}@liacs.nl

Abstract

In image retrieval, global features related to color or texture are commonly used to describe the image content. The use of interest points in content-based image retrieval allows an image index to represent local properties of images. In this paper, we present a wavelet-based salient point extraction algorithm and 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.

1 Introduction

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 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 general, the features are often computed from the entire image. The problem with this approach is that these global features cannot handle all parts of the image having different characteristics. Therefore, local computation of image information is necessary. Local features can be computed at different image scales to obtain an image index based on local properties of the image and they need to be sufficiently discriminative to "summarize" the local image information. These features are too time-consuming to be computed for each pixel in the image and therefore, the feature extraction should be limited to a subset of the image pixels, the interest points [9, 11], 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 discriminative index because they are related to the visually most important parts of the image.

Schmid and Mohr [9] introduced the notion of interest point in image retrieval. To detect these points, they use the Harris' corner detector [5]. This detector, as many others [10], was initially designed for robotics, and it is based on a mathematical model for corners. The original goal was to match corners from a pair of stereo images, in order to obtain a representation of the 3D scene. Since the corners detectors were not designed to

BMVC 2001 doi:10.5244/C.15.42

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 do not need 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, a smoothed edge may have 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 textured regions. Many corners are detected in these regions by a corner detector. The problem is that due to efficiency reasons only a preset number of points per image can be used in the indexing process. Since in this case most of the detected points will be in a small region, the other parts of the image may 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 believe multi-resolution representation is interesting to detect salient points. We 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.

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 easy to understand that using a small amount of such points instead of all image pixels 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 is presented in Section 2. Color and texture features used in this paper are discussed in Section 3 and 4, respectively. The experimental setup is presented in Section 5, followed by our experiments (Section 6). A final discussion is 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 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 attenuated function with zero integral. We study the image f at the scales (or resolutions) 1/2, 1/4, ..., 2^j, $j \in Z$ and $j \leq -1$. 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, 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

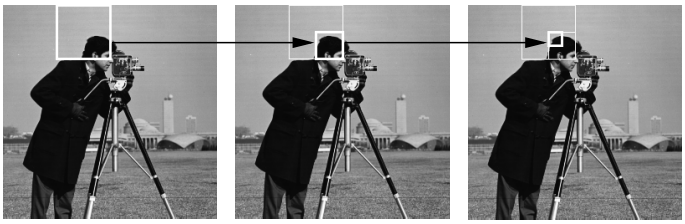

Figure 1: Salient points extraction: spatial support of tracked coefficients

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:

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

where p is the wavelet regularity and $0 \leq n < 2^j N$ 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. 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 (Figure 1). We set its saliency value as the sum of the absolute value of the wavelet coefficients in the track:

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

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 wavelet function, 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 can be avoided with the simplest overlapping wavelet, Daubechies4. However, this drawback is not likely in natural images and therefore, we used Haar transform in our experiments.

3 Color Features

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. Color histograms

are often used [14]. 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. [13] 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 process 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, as well [13]. Furthermore, because most of the information is concentrated on the low-order moments, only the first moment (mean), the second moment (variance), and the third central moment (skewness) were used. If the value of the i^{th} color channel at the j^{th} image pixel is p_{ij} and the number of pixels in the image is N, then the index entries related to this color channel are:

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

We were working with the HSV color space so, for each image in the database a 9-dimensional 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 the identity of an object. Therefore, texture or geometric properties are needed to identify objects. Consequently, color indexing methods are bound to retrieve false positives, i.e., images which have a similar color composition as the query image but have a completely different content. Therefore, in practice, it is necessary to combine color indexing with texture and/or shape indexing techniques.

The method of texture analysis chosen for feature extraction is crucial to the success of texture classification and many methods have been proposed to extract texture features [7]. Recently there was a strong push to develop multi-scale approaches to the texture problem. Smith and Chang [12] used the statistics (mean and variance) extracted from the wavelet subbands as texture features. To explore the middle-band characteristics, tree-structured wavelet transform was studied by Chang and Kuo [2]. Ma and Manjunath [6] evaluated the texture image annotations by various wavelet transform representations and found out that Gabor wavelet transform was the best among the tested candidates, which matched the human vision study results [1].

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 $\Delta x \Delta u \geq \frac{1}{4\pi}$ and $\Delta y \Delta v \geq \frac{1}{4\pi}$, where $[\Delta x, \Delta y]$ gives the resolution in space and $[\Delta u, \Delta v]$ gives the resolution in spatial-frequency. In addition to good performances in texture discrimination and segmentation, the justification for Gabor filter is also supported through psychophysical experiments. Texture analyzers implemented using 2-D Gabor functions produce a strong correlation

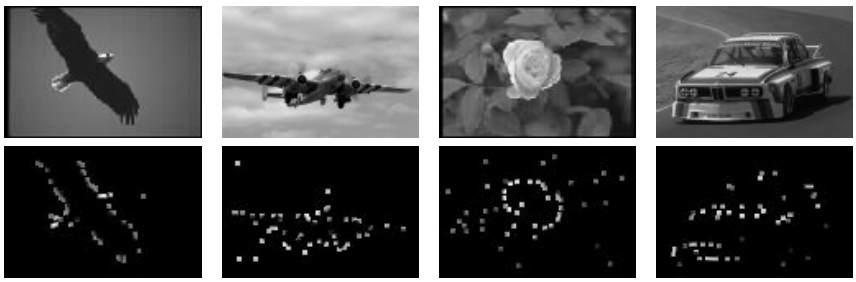

Figure 2: Examples of images from the database and their salient point map

with actual human segmentation [8]. Furthermore, the receptive visual field profiles are adequately modeled by 2-D Gabor filters [3].

5 Experimental Setup

The setup of our experiments was the following. First we extracted salient points for each image in the database using Haar wavelet transform and the algorithm described in Section 2. The number of salient points extracted will clearly influence the retrieval results. We performed experiments (not presented here due to space limitation) in which the number of salient points varied from 10 to several hundreds and found out that when using more than 50 points, the improvement in accuracy we obtained did not justify the computational effort involved. Therefore, in the experiments, 50 salient points were extracted for each image. Figure 2 shows some salient points examples. The original images are shown in the upper row and the salient points together with a small neighborhood around them are shown in the lower row. Clearly, the local information of the objects, i.e., bird, airplane, flower, and car, are captured by the salient points.

For feature extraction, we considered the pixels in a small neighborhood around each salient point. These pixels form the image signature. For each image signature in the database we computed the color moments (in a 3×3 neighborhood) and the Gabor moments (in a 9×9 neighborhood). For convenience, this approach is denoted as the salient approach. For benchmarking purposes we also considered the results obtained using the color moments and the wavelet moments [12] extracted over the entire image (denoted as Global CW approach) and the results obtained using the color moments and the Gabor moments [6] extracted over the entire image (denoted as Global CG approach).

The overall similarity distance D_j for the j^{th} image in the database is obtained by linearly combining the similarity distance of each individual feature, $S_j(f_i)$:

$$D_j = \sum_i W_i S_j(f_i) \qquad j = 1, \ldots, N \tag{4}$$

where N is the total number of images in the database and $S_j(f_i)$ is defined as:

$$S_j(f_i) = (\mathbf{x}_i - \mathbf{q}_i)^T (\mathbf{x}_i - \mathbf{q}_i) \tag{5}$$

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. (4) are set to be equal.

6 Results

In the first experiment we considered a small database consisting of 600 images from the Corel database. As ground truth we used 146 images divided in 7 classes: airplane (25 images), bird (27 images), car (18 images), tiger (18 images), flower (19 images), mountain (19 images), and church paintings (20 images).

In order to test the retrieval results for each individual class, we randomly picked 15 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.

Class	**Salient**	**Global**
Airplane	100	93.33
Bird	93.33	86.66
Tiger	86.66	80
Car	73.33	60
Flower	73.33	60
Church Painting	93.33	93.33
Mountain	93.33	93.33

Table 1: Retrieval accuracy (%) for each individual class using 15 randomly chosen images from each class as queries

From this experiment we can see that for some classes that are mainly composed of a single object on a simple background (e.g., Bird or Airplane where the background represented the blue sky), both the salient and the global approach have similar performances. For the global approach, the color moments were extracted from the entire image and therefore, the color information is mainly determined by the dominant background, e.g., blue sky. In this sense, the birds were found to be similar because of the background, not the objects themselves. In the salient approach, the salient points were mostly found on the boundaries of the objects. The local color moments around the neighborhood of the salient points were extracted and they represented the object information instead of the background. Here, 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 more accurately the user's concept than the global approach in terms of object finding. When the classes have the complex background (e.g., Tiger, Car, Airplane) the retrieval conditions are more difficult and the salient approach performs much better than the global approach. When the images show more global variations, (e.g., Church Painting, Mountain), both approaches perform very well showing that the salient approach can still capture global image information (background) as well.

In our second experiment we considered a database of 479 images (256×256 pixels in size) of color objects such as domestic objects, tools, toys, food cans, etc [4]. As ground truth we used 48 images of 8 objects taken from different camera viewpoints (6 images for a single object).

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. 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, we extracted Gabor texture features from the 9×9 neighborhood of each salient point. The dimension of the Gabor filter was 7×7 and we used 2 scales and 6 orientations/scale. The first 12 features represented the averages over the filter outputs obtained in order for: scale 1 and orientation $1, \ldots,$ scale 1 and orientation 6, scale 2 and orientation $1, \ldots,$ 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 orientations/scale. The first 18 features were the averages over the filters outputs and the last 18 features were the corresponding variances.

We expect 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 3 we show an example of a query image and the similar images from the database retrieved with various ranks.

Query

Salient	1	2	4	15	22
Global CW	2	7	12	25	33
Global CG	2	5	9	20	27

Figure 3: Example of images of one object taken from different camera viewpoints. The ranks of individual images were obtained using salient points information (Salient), global color moments and wavelet moments (Global CW), and global color moments and Gabor moments (Global CG)

The salient point approach outperforms both the global CW approach and the global CG approach. 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 better performances than the global CW approach. This fact demonstrates that Gabor feature is a very good feature for texture characterization. Moreover, it should also be noted that: (1) the salient point approach only uses the information from a very small part of the image, but still achieves a good representation of the image. For example, in our object database $9 \times 9 \times 50$ pixels were used to represent the image. Compared to the global approach (all 256×256 pixels were used), it only uses less than 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 retrieval accuracy for the object database. Each of the 6 images

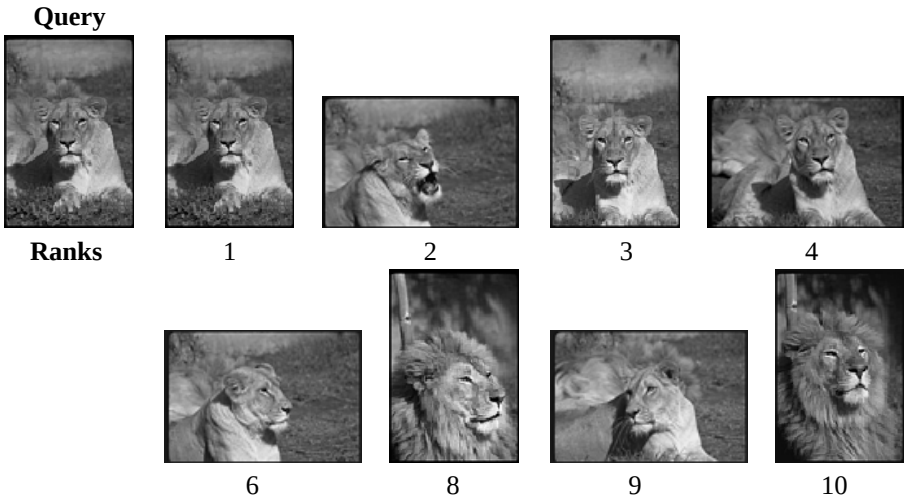

Figure 4: Retrieved images from a query using salient points approach. Match quality decreases from the top left to the bottom right

from the 8 classes was considered as query image and the average retrieval accuracy was calculated.

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

Table 2: Retrieval accuracy (%) using 48 images from 8 classes for object database

Results in Table 2 show that using the salient point information the retrieval results are significantly improved (>10%) compared to the global CW approach. When compared to the global CG approach, the retrieval accuracy of salient approach is with 1.9%, 0.4%, and 1.5% higher in the top 6, 10, and 20 images, respectively. The salient approach has therefore, similar performance comparing with the Global CG approach but much lower computational complexity 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 we considered a database consisted of 4013 various images covering a wide range of natural scenes such as animals, buildings, construction sites, textures, paintings, mountains, lakes, and roads. In order to perform quantitative analysis, we randomly chose 15 images from a few categories, e.g., building, flower, tiger, lion, road, forest, mountain, sunset and use each of them as queries. For each category, we measured how many hits, i.e. how many similar images to the query were returned in the top 20 retrieved images.

Figure 4 shows an example of the retrieved images from a query using the salient approach. Match quality decreases from the top left to the bottom right. Figure 5 shows

the average number of hits for each category using the global CW approach, the global CG approach, and the salient approach. Clearly the salient approach has similar performance comparing with the global CG approach and outperforms the global CW approach when the first five categories are considered. 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 because the images exhibit now more global characteristics and therefore, the global approaches can capture better the image content.

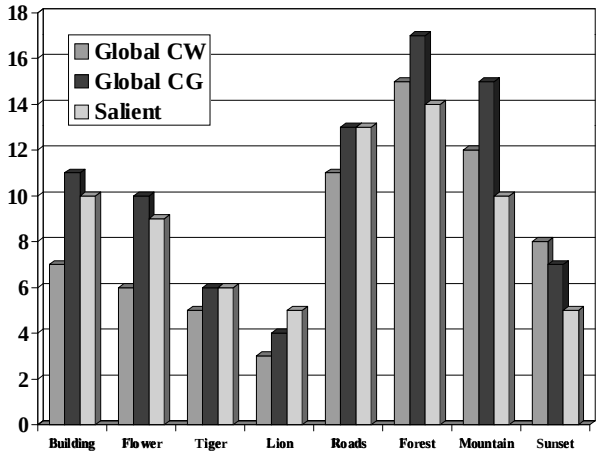

Figure 5: The average number of hits for each category using the global color and wavelet moments (Global CW), the global color and Gabor moments (Global CG) and the salient point approach (Salient)

7 Discussion

In this paper, we presented a wavelet-based salient point extraction algorithm and applied it in 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 (50 points in this paper) 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 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 similar performance com-

pared 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 have very important influence for very large image databases.

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 proving that Gabor features are very good candidates for texture characterization [1].

In conclusion, 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 shape information in the location of the salient points making the retrieval more accurate. We also intend to automatically determine the optimal number of the salient points needed to be extracted for each image.

References

[1] J. Beck, A. Sutter, and A. Ivry. Spatial frequency channels and perceptual grouping in texture segregation. *Comp Vis Graph Imag Process*, 37:299–325, 1987.

[2] T. Chang and C. Kuo. Texture analysis and classificatrion with tree-structured wavelet transform. *IEEE Trans on Imag Process*, 2(4):429–441, 1993.

[3] J. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Trans on Biomed Eng*, 36(1):107–114, 1989.

[4] T. Gevers and A. Smeulders. PicToSeek: Combining color and shape invariant features for image retrieval. *IEEE Trans Imag Process*, 20(1):102–119, 2000.

[5] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vis Conf*, pages 147–151, 1999.

[6] W.Y. Ma and B.S. Manjunath. A comparison of wavelet transform features for texture image annotation. *Int Conf on Imag Process*, 2:256–259, 1995.

[7] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distribution. *Patt Recogn*, 29:51–59, 1996.

[8] T.R. Reed and H. Wechsler. Segmentation of textured images and gestalt organization using spatial/spatial-frequency representations. *IEEE Trans on Patt Anal and Mach Intell*, 12(1):1–12, 1990.

[9] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans on Patt Anal and Mach Intell*, 19(5):530–535, 1997.

[10] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interst point detectors. *I J Comp Vis*, 37(2):151–172, 2000.

[11] N. Sebe, Q. Tian, E. Loupias, M.S. Lew, and T.S. Huang. Color indexing using wavelet-based salient points. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 15–19, 2000.

[12] J.R. Smith and S.-F. Chang. Transform features for texture classification and discrimination in large image databases. *Int Conf on Imag Process*, 3:407–411, 1994.

[13] A. Stricker and M. Orengo. Similarity of color images. *SPIE - Storage and Retrieval for Image and Video Databases III*, 2420:381–392, 1995.

[14] M.J. Swain and D.H. Ballard. Color indexing. *I J Comp Vis*, 7(1):11–32, 1991.