Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 89–96

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Comparing salient point detectors

Nicu Sebe *, Michael S. Lew

Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333CA, Leiden, Netherlands

Received 15 August 2001; received in revised form 20 March 2002

Abstract

The use of salient points in content-based retrieval allows an image index to represent local properties of the image. Classic corner detectors can also be used for this purpose but they have drawbacks when are applied to various natural images mainly because visual features do not need to be corners and corners may gather in small regions. In this paper, we present a salient point detector using wavelet transform and we compare it with two corner detectors using two criteria: repeatability rate and information content. We determine which detector gives the best results and show that it satisfies the criteria well.

Keywords: Wavelet-based salient points; Corner detectors; Information content; Repeatability rate

1. Introduction

Many computer vision tasks rely on low level features. A wide variety of feature detectors exist and results can vary enormously depending on the detector used. An image is "summarized" by a set of features, the image index, to allow fast querying. Local features are of interest since they lead to an index based on local properties of the image. The feature extraction is limited to a subset of the image pixels, the interest points, where the image information is supposed to be the most important (Schmid and Mohr, 1997; Sebe et al., 2000). 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.

Haralick and Shapiro (1993) consider a point in an image *interesting* if it has two main properties: distinctiveness and invariance. This means that a point should be distinguishable from its immediate neighbors and the position as well as the selection of the interesting point should be invariant with respect to the expected geometric and radiometric distortions.

Schmid and Mohr (1997) proposed the use of corners as interest points in image retrieval using the Harris corner detector (Harris and Stephens, 1988). The basic idea is to use the auto-correlation function in order to determine locations where the signal changes in two directions. A matrix related to the auto-correlation function which takes into

* Corresponding author. Tel.: +31-71-527-7050; fax: +31-71-527-6985.
E-mail addresses: nicu@liacs.nl (N. Sebe), mlew@liacs.nl (M.S. Lew).

0167-8655/03/$ - see front matter
PII: S0167-8655(02)00192-7

account the first derivatives of the signal on a window is computed. The eigenvectors of this matrix are the principal curvatures of the auto-correlation function. Two significant values indicate the presence of an interest point.

Different interest point detectors are evaluated and compared in (Schmid et al., 2000). Besides the Harris corner detector and an improved variant of it called PreciseHarris, the authors also consider the detectors proposed by Heitger et al. (1992), Förstner (1994), and Horaud et al. (1990). Heitger et al. (1992) developed an approach inspired by experiments on the biological visual system. They extract 1D directional characteristics by convolving the image with orientation-selective Gabor filters. In order to obtain 2D characteristics, they compute the first and second derivatives of the 1D characteristics. Förstner (1994) classifies image pixels into categories—region, contour, and interest point—by using the auto-correlation function. Local statistics allow a blind estimate of signal-dependent noise variation and thus an automatic selection of thresholds. Horaud et al. (1990) extract line segments from the image contours. These segments are grouped and the intersections of grouped line segments are the interest points. The authors (Schmid et al., 2000) concluded that the best results are provided by the Harris detector (Harris and Stephens, 1988). Zheng et al. (1999) proposed a method derived from the Harris detector (in their paper they call it Plessey corner detector). The most important improvement of their corner detector is that it decreases the complexity (instead of calculating the Gaussians they calculate smoothed gradient-multiple images). They conclude that the performance of their gradient-direction corner detection is slightly inferior to that of the Harris detector but the performance of localization (defined as the closeness to the true location of the corner) is better than that of the Harris detector.

Corner detectors, however, were designated for robotics and shape recognition and they have drawbacks when are applied to natural images. 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 only in corner points, as is 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. Corners also gather in textured regions. 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. However, we do not want to have points in all possible regions: regions where there is nothing interesting (e.g., a region with a constant grey level) should not contain any "interesting" points.

We believe that other points based on image information can be extracted using approaches other than the corner differential framework. Studies on visual attention, more related to human vision, propose different models. The basic information is still the variation in the stimuli. However, this is not longer taken into account in a differential way but mainly from an energy point of view (Itti et al., 1998). Another approach is to integrate a scale space approach into the corner extraction algorithm (Lindeberg, 1998; Mikolajczyk and Schmid, 2001). The idea is to select a characteristic scale by searching for local extreme over scales.

In this context, we aim for a set of interesting points called *salient points* that are related to any visual interesting part of the image whether it is smoothed or corner-like. Moreover, to describe different parts of the image the set of salient points should not be clustered in few regions. We believe multiresolution representation is interesting to detect salient points. Multiresolution representations are usually implemented using image pyramids. This representation has various properties that makes it very popular in image processing and computer vision algorithms: (1) the adaptation of resolution is suitable for coarse-to-fine multigrid iteration strategies; (2) iterative algorithms that proceed by successive refinements usually require less computations and have faster convergence; (3) in the context of iterative algorithms, the smoothing effect of the pyramid reduces the

likelihood of getting trapped in local extrema, which increases robustness; and (4) analogies can be made with the hierarchical organization of the human primary visual cortex. One of the earliest example of a pyramid is due to Burt and Adelson (1986). Their Gaussian filtering, however, produces excessive smoothing which leads to some loss of image details. Higher-quality image reduction can be obtained by designing a filter that is optimum in the least-squares sense (Unser, 1992) or by using the lowpass branch of a wavelet decomposition algorithm (Mallat, 1989).

Taking these into account, we present a salient point extraction algorithm that uses the wavelet transform, which expresses image variations at different resolutions. Our 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. The algorithm presented in this paper is an improved version of our algorithm presented in Loupias et al. (2000), Tian et al. (2001), and Loupias and Bres (2001). There we were interested in using the salient points in a content-based retrieval scenario and we showed that extracting color and texture features in the location given by the salient points provided significantly improved results in terms of retrieval accuracy, computational complexity, and storage space of feature vectors as compared to global features approaches. In a content-based retrieval application the geometric stability of the salient points is not really critical. There, the features stability is more important since image matching is done at feature level. For example, even if a salient point moves along an edge, the matching does not change as long as the feature extracted in that point remains stable. However, if we want to use the salient points in other applications, such as object tracking and recognition or stereo matching, the geometrical stability becomes really critical.

In order to evaluate the "interestingness" of the points (as was introduced by Haralick and Shapiro (1993)) two criteria are considered: repeatability rate and information content. The repeatability rate evaluates the geometric stability of points under different image transformation. Information content measures the distinctiveness of greylevel pattern at an interest point. A local pattern is described using rotationally invariant combinations of derivatives. The entropy of these invariants is measured for a set of interest points.

2. Wavelet-based salient points

The wavelet representation gives information about the variations in the image at different scales. 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) $1/2, 1/4, \ldots, 2^j$, $j \in Z$ and $j \leqslant -1$. 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 consider 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 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:

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

where p is the wavelet regularity and $0 \leqslant n < 2^j N$ 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 (Loupias and Bres, 2001), 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 (Fig. 1). We set its saliency value as the sum of the absolute value of the wavelet coefficients in the track:

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

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 tracking using the highest gradient works well only if one of the $2p$ points clearly has a much higher gradient compared to the other points. However, if two or more points will have close gradient values the tracking using only the maximum may contribute to geometrical instability of the extracted salient point due to the presence of noise. Consider that there are m points out of the possible $2p$ points which have the gradient very close to the maximum gradient. In this case, in order to enhance the robustness to noise of the salient point extraction algorithm, we trace all the m points. In the end, we select the tracking branch that provides maximum saliency according to the Eq. (2).

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 (Daubechies, 1988). Examples of salient points extracted using Daubechies4, Haar, Harris, and Zheng detectors are shown in Fig. 2. Note that while for Harris and Zheng detectors the points lead to an incomplete image representation, for the other two detectors the salient points are detected for smooth edges (as can be seen in the fox image) and are not gathered in texture regions (as can be seen in the girl image). Hence, they lead to a more complete image representation.

Schmid et al. (2000) evaluated and compared different point detectors and concluded that the best results are provided by the Harris detector (Harris and Stephens, 1988). Taking into account that Zheng et al. (1999) showed that the performance of their gradient-direction corner detection is slightly inferior to that of the Harris detector, in our further experiments we consider only Harris

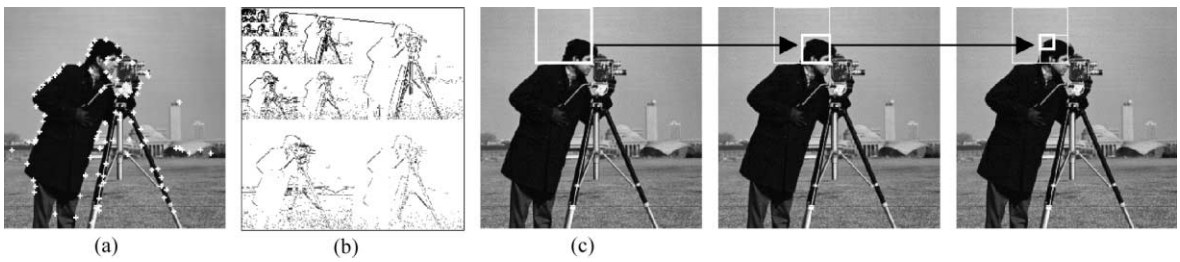

Fig. 1. Salient points extraction: (a) salient points, (b) tracked coefficients, (c) spatial support of tracked coefficients.

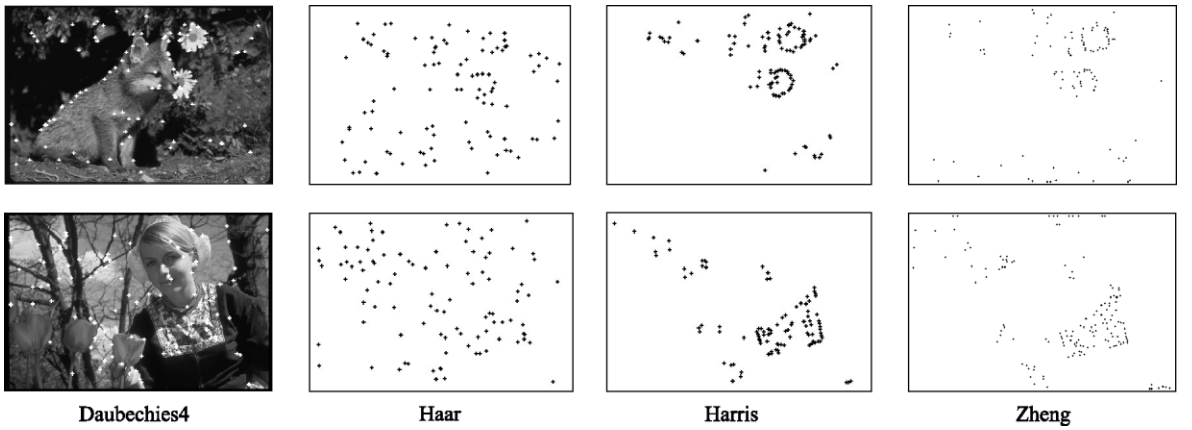

Fig. 2. Salient points examples. For Daubechies4 and Haar salient points are detected for smooth edges (fox image) and are not gathered in textured regions (girl image).

and PreciseHarris (see Schmid et al., 2000) detectors as benchmarks.

3. Repeatability

Repeatability is defined by the image geometry. Given a 3D point P and two projection matrices M_1 and M_2, the projections of P into two images I_1 and I_2 are $p_1 = M_1P$ and $p_2 = M_2P$. The point p_1 detected in image I_1 is repeated in image I_2 if the corresponding point p_2 is detected in image I_2. To measure the repeatability, a unique relation between p_1 and p_2 has to be established. In the case of a planar scene this relation is defined by an homography: $p_2 = H_{21}p_1$.

The percentage of detected points which are repeated is the *repeatability rate*. A repeated point is not in general detected exactly at position p_2, but rather in some neighborhood of p_2. The size of this neighborhood is denoted by ε and repeatability within this neighborhood is called *ε-repeatability*. Moreover, to measure the number of repeated points, we have to take into account that the observed scene parts differ in the presence of changed imaging conditions, such as image rotation or scale change. The salient points which cannot be observed in both images corrupt the repeatability measure and therefore, only the points which are detected in the common scene part should be used to compute the repeatability. Points d_1 and d_2 which are detected in the common part of images I_1 and I_2 are defined by $\{d_1\} = \{p_1|H_{21}p_1 \in I_2\}$ and $\{d_2\} = \{p_2|H_{12}p_2 \in I_1\}$, where $\{p_1\}$ and $\{p_2\}$ are the points detected in images I_1 and I_2, respectively. The set of point pairs (d_2, d_1) which correspond within an ε-neighborhood is $D(\varepsilon) = \{(d_2, d_1)|\mathrm{dist}(d_2, H_{21}d_1) < \varepsilon\}$.

Under these conditions, the repeatability rate for image I_2 is given by:

$$r(\varepsilon) = \frac{|D(\varepsilon)|}{N} \tag{3}$$

where N is the total number of points detected. One can easily verify that $0 \leqslant r(\varepsilon) \leqslant 1$.

4. Information content

Information content is a measure of the distinctiveness of a salient point. Distinctiveness is based on the likelihood of a greyvalue descriptor computed at the point within the population of all observed salient point descriptors. Given one image, a descriptor is computed for each of the detected salient points and the information content will measure the distribution of these descriptors. If all descriptors are spread out, information content is high and matching is likely to succeed. On the other hand, if all descriptors are close to each other, the information content is low and

matching can easily fail as any point can be matched to any other.

Information content of the descriptors is measured using entropy. The more spread out the descriptors are, the larger the entropy is. Entropy measures average information content. In information theory the information content of a message i is inversely related to its probability and is defined as $I = -\log(p_i)$. The average information content per message of a set of messages is $-\sum_i p_i \log(p_i)$ which is the entropy.

In the case of salient points we would like to know how much average information content a salient point "has" as measured by its greylevel pattern. The more distinctive the greylevel patterns are, the larger the entropy is. To measure the distribution of local greyvalue patterns at salient points, we have to describe a measure which describes such a pattern. In order to have rotation invariant descriptors, we chose to characterize salient points by local greyvalue rotation invariants which are combinations of derivatives. We computed the "local jet" (Koenderink and van Doorn, 1987) which consisted of the set of derivatives up to Nth order. These derivatives describe the intensity function locally and are computed stably by convolution with Gaussian derivatives. The local jet of order N at a point $\boldsymbol{x} = (x, y)$ for an image I and a scale σ is defined by: $J^N[I](\boldsymbol{x}, \sigma) = \{L_{i_1,\ldots,i_n}(\boldsymbol{x}, \sigma) | (\boldsymbol{x}, \sigma) \in I \times R^+\}$, where $L_{i_1,\ldots,i_n}(\boldsymbol{x}, \sigma)$ is the convolution of image I with the Gaussian derivatives $G_{i_1,\ldots,i_n}(\boldsymbol{x}, \sigma)$, $i_k \in \{x, y\}$ and $n = 0, \ldots, N$.

In order to obtain invariance under the group SO(2) (2D image rotation), Koenderink and van Doorn (1987) and ter Haar Romeny et al. (1994) computed differential invariants from the local jet:

$$\vec{v}[0\ldots3] = \begin{bmatrix} L_x L_x + L_y L_y \\ L_{xx} L_x L_x + 2L_{xy} L_x L_y + L_{yy} L_y L_y \\ L_{xx} + L_{yy} \\ L_{xx} L_{xx} + 2L_{xy} L_{xy} + L_{yy} L_{yy} \end{bmatrix} \quad (4)$$

The computation of entropy requires a partitioning of the space of $\vec{v}$. Partitioning is dependent on the distance measure between descriptors and we consider the approach described by Schmid et al. (2000). The distance we used is the Mahalanobis distance given by: $d_M(\vec{v}_1, \vec{v}_2) = ((\vec{v}_1 - \vec{v}_2)^T \Lambda^{-1} (\vec{v}_1 - \vec{v}_2))^{1/2}$, where $\vec{v}_1$ and $\vec{v}_2$ are two descriptors and Λ is the covariance of $\vec{v}$. The covariance matrix Λ is symmetric and positive definite. Its inverse can be decomposed into $\Lambda^{-1} = P^T D P$ where D is diagonal and P an orthogonal matrix. Furthermore, we can define the square root of Λ^{-1} as $\Lambda^{-1/2} = D^{1/2} P$ where $D^{1/2}$ is a diagonal matrix whose coefficients are the square roots of the coefficients of D. The Mahalanobis distance can then be rewritten as: $d_M(\vec{v}_1, \vec{v}_2) = \|D^{1/2} P(\vec{v}_1 - \vec{v}_2)\|$. The distance d_M is the norm of difference of the normalized vectors: $\vec{v}_{norm} = D^{1/2} P \vec{v}$. This normalization allows us to use equally sized cells in all dimensions. This is important since the entropy is directly dependent on the partition used. The probability of each cell of this partition is used to compute the entropy of a set of vectors $\vec{v}$.

5. Results

In the experiments we used a set of 1000 images taken from the Corel database and we compared four salient point detectors. In Section 2 we introduced two salient point detectors using wavelets: Haar and Daubechies4. For benchmarking purposes we also considered the Harris corner detector (Harris and Stephens, 1988) and a variant of it called PreciseHarris, introduced by Schmid et al. (2000). The difference between the last two detectors is given by the way the derivatives are computed. Harris computes derivatives by convolving the image with the mask $[-2\ -1\ 0\ 1\ 2]$ whereas PreciseHarris uses derivatives of the Gaussian function instead.

5.1. Results for repeatability

Before we can measure the repeatability of a particular detector we first had to consider typical image alterations such as image rotation and image scaling. In both cases, for each image we extracted the salient points and then we computed the average repeatability rate over the database for each detector. The repeatability rate was computed using Eq. (3).

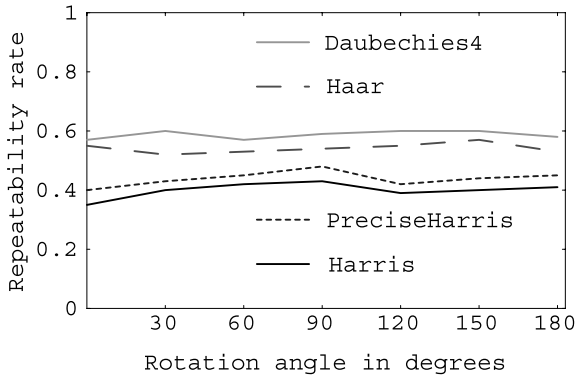

Fig. 3. Repeatability rate for image rotation ($\varepsilon = 1$).

In the case of image rotation the rotation angle varied between 0° and 180°. The repeatability rate in a $\varepsilon = 1$ neighborhood for the rotation sequence is displayed in Fig. 3.

The detectors using wavelet transform (Haar and Daubechies4) give better results compared with the other ones. Note that the results for all detectors are not very dependent on image rotation. The best results are provided by Daubechies4 detector.

In the case of scale changes, for each image we considered a sequence of images obtained from the original image by reducing the image size so that the image was aspect-ratio preserved. The largest scale factor used was 4. The repeatability rate for scale change is presented in Fig. 4.

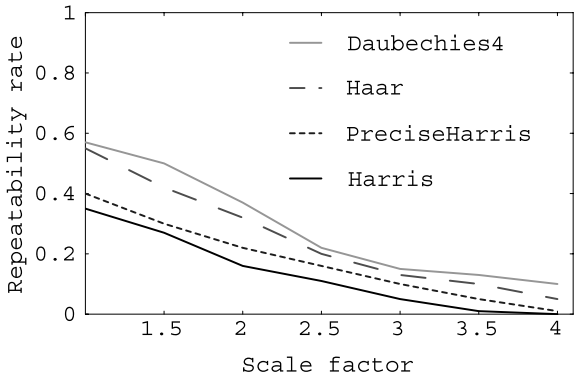

Fig. 4. Repeatability rate for scale change ($\varepsilon = 1$).

All detectors are very sensitive to scale changes. The repeatability is low for a scale factor above 2 especially for Harris and PreciseHarris detectors. The detectors based on wavelet transform provide better results compared with the other ones.

5.2. Results for information content

In these experiments we also considered random points in our comparison. For each image in the database we computed the mean number m of salient points extracted by different detectors and then we selected m random points using a uniform distribution.

For each detector we computed the salient points for the set of images and characterized each point by a vector of local greyvalue invariants (cf. Eq. (4)). The invariants were then normalized and the entropy of the distribution was computed. The cell size in the partitioning was the same in all dimensions and it was set to 20. The σ used for computing the greylevel invariants was 3.

The results are given in Table 1. This table shows that the detector using the Daubechies4 wavelet transform has the highest entropy and thus the salient points obtained are the most distinctive. The results obtained for Haar wavelet transform are almost as good. The results obtained with PreciseHarris detector are better than the ones obtained with Harris but worse than the ones obtained using the wavelet transform. Moreover, the results obtained for all of the salient points detectors are significantly better than those obtained for random points. The difference between the results of Daubechies4 and random points is about a factor of 2.

In summary, the most "interesting" salient points were detected using the Daubechies4

Table 1
The information content for different detectors

Detector	Entropy
Haar	6.0653
Daubechies4	6.1956
Harris	5.4337
PreciseHarris	5.6975
Random	3.124

detector. These points have the highest information content and proved to be the most robust to rotation and scale changes.

6. Conclusion

We presented a salient point detector based on wavelets. The wavelet-based salient points are interesting because they are located in visual focus points without gathering in textured regions. We used the Haar transform for point extraction, which is simple but may lead to bad localization. A better approach is to use Daubechies4 wavelets which avoid these drawbacks.

We also compared our wavelet-based salient point extraction algorithm with two corner detectors using the criteria: repeatability rate and information content. Our points have more information content and better repeatability rate than those of the other detectors. All detectors have significantly more information content than randomly selected points.

Acknowledgement

We would like to thank Etienne Loupias who designed and contributed to the first salient point extraction algorithm.

References

Burt, P.J., Adelson, E.H., 1986. The Laplacian pyramid as a compact code. IEEE Trans. Comm. 15 (1–2), 1–21.

Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets. Comm. Pure Appl. Math. 41, 909–996.

Förstner, W., 1994. A framework for low level feature extraction. ECCV 2, 383–394.

Haralick, R., Shapiro, L., 1993. Computer and Robot Vision II. Addison-Wesley, Reading, MA.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: 4th Alvey Visual Conference, pp. 147–151.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., Kubler, O., 1992. Simulation of neural contour mechanism: From simple to end-stopped cells. Vision Res. 32 (5), 963–981.

Horaud, R., Veillon, F., Skordas, T., 1990. Finding geometric and relational structures in an image. ECCV, 374–384.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Machine Intell. 20 (11), 1254–1259.

Koenderink, J.J., van Doorn, A.J., 1987. Representation of local geometry in the visual system. Biological Cybernet. 55, 367–375.

Lindeberg, T., 1998. Feature detection with automatic scale selection. Internat. J. Comput. Vision 30 (2), 79–116.

Loupias, E., Bres, S., 2001. Key points-based indexing for preattentive similarities: The KIWI system. Pattern Anal. Appl. 4, 200–214.

Loupias, E., Sebe, N., Bres, S., Jolion, J.-M., 2000. Wavelet-based salient points for image retrieval. In: Internat. Conf. on Image Processing, Vol. 2, pp. 518–521.

Mallat, S.G., 1989. A theory of multiresolution signal decomposition: The wavelet transform. IEEE Trans. Pattern Anal. Machine Intell. 11 (7), 674–693.

Mikolajczyk, K., Schmid, C., 2001. Indexing based on scale invariant interest points. In: Internat. Conf. on Computer Vision. pp. 525–531.

Schmid, C., Mohr, R., 1997. Local grayvalue invariants for image retrieval. IEEE Trans. Pattern Anal. Machine Intell. 19 (5), 530–535.

Schmid, C., Mohr, R., Bauckhage, C., 2000. Evaluation of interest point detectors. Internat. J. Comput. Vision 37 (2), 151–172.

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

ter Haar Romeny, B., Florack, L., Salden, A., Viergever, M., 1994. Higher order differential structure of images. Image Vision Comput. 12 (6), 317–325.

Tian, Q., Sebe, N., Loupias, E., Lew, M.S., Huang, T.S., 2001. Image retrieval using wavelet-based salient points. J. Electron. Imag. 10 (4), 1132–1141.

Unser, M., 1992. An improved least squares laplacian pyramid for image compression. Signal Process. 27 (2), 187–203.

Zheng, Z., Wang, W., Teoh, E.K., 1999. Analysis of gray level corner detection. Pattern Recognition Lett. 20, 149–162.