Appeared in Conference on Intelligent Autonomous Systems, IAS '95, Karlsruhe, March 95. (best paper prize).

Comparison of Correlation Techniques

Jerome Martin and James L Crowley

IMAG-LIFIA, 46 Ave Félix Viallet, 38031 Grenoble, France

Abstract. This paper presents a theoretical and experimental comparison of different forms of SSD and normalised cross-correlation of image neighborhoods. Signal detection theory is used as a framework for analysis of correlation techniques. A sum of squared difference (SSD) of two image neighborhoods is shown to provide an optimal matching measure for tracking and registration in the case fo additive Gaussian noise. Correlation of the image, its gradient magnitude or its Laplacian are discussed. The relations between SSD and Cross Correlation are demonstrated, and different normalisation techniques are described. An experimental comparison is made of SSD, Normalized Cross Correlation, and Zero-Mean normalised cross correlation, in the presence of changes in light level, additive Gaussian noise, and salt-and-pepper noise.

Keywords. Matching, Tracking, Correlation, Normalisation.

1. Introduction

Many computer vision techniques require matching parts of images. Examples include registering images to determine shift and deformation for reconstruction using stereo and motion, matching to a reference for recognition, verification and event detection and extraction of information to index into an image database. Matching is fundamental to computer vision.

For many years, vision researchers have believed that the central problem of matching was one of computational cost. To reduce the cost of matching, researchers have sought to perform matching at the highest possible level of abstraction. Most techniques involved constructing a hierarchy of increasingly abstract image descriptions, matching at the most abstract level, and then propagating correspondences back down the hierarchy to determine association of parts. A typical image hierarchy is composed of descriptions of the image contents at following abstraction levels:

Level 0: The raw image or its derivatives.
Level 1: Edges (represented as segments or contours).
Level 2: Groupings of edges, (a network of structures).
Level 3: Recognised objects and object parts.

Unfortunately, the processes for deriving each level of description from the immediately lower level are typically prone to errors. At each level, unstable reactions to noise in the lower level leads to both extraneous components and to missing parts. Efforts to overcome such errors lead to processes whose computational complexities are of relatively high order. In particular, recognition by matching networks of structures in the presence of missing and extraneous symbols has an exponential complexity. The complexity and unreliability of ad hoc "higher level" matching processes has lead some non-specialists to assert that computer vision has failed. Fortunately, this opinion is incorrect.

As available computing power has increased (by a factor of 2 each 1.5 years) it has become apparent that mastering computational cost requires mastering computational complexity. Signal processing theory (and the related areas of information theory and estimation theory) provide the mathematical tools to design robust low-level matching procedures, complete with estimates of the probability of error. Furthermore, these techniques have constant computational complexities and can easily be computed at video rates using relatively simple hardware. Image signals can be matched robustly and in real time by pixel based operations derived using signal processing tools.

Optimal techniques for comparing signals generally use a form of "Sum of Squared Difference" of the signals expressed in a discrete basis set (Wozencraft and Jacobs 1965). For image neighborhoods, this similarity measure can be computed using the individual pixels as the basis space, or using more exotic spaces such as multi-resolution pyramids (Burt and Adelson 83; Crowley and Stern 84), tensor spaces based on Gabor filters (Granlund 78;

Knutssson 89), or spaces derived by principle components analysis (Turk and Pentland 91). Sum of Squared Difference (or SSD), (Anandan 87), provides a general expression which can yield a variety of specialised forms of cross-correlation, depending on how normalisation is performed.

The correlation of a template neighbourhood with a search region of an image is a direct extension of operations which are basic tools in signal processing and communications theory. Although cross-correlation was used in the early years of computer vision (Moravec 77), it has been neglected over the last few decades. The common wisdom was unreliable and computationally expensive. Vision researchers sought to reduce computational cost by operating on more abstract descriptions such as groupings of edge segments (Binford 82). Unfortunately, the reliable detection of such abstract descriptions has proven to be very difficult. In addition, rapid increases in available computing power have made video rate correlation possible on inexpensive hardware (Inoue et al 92).

The basic formula for cross correlation can be derived directly from an inner product of two vectors, or equally from the sum of squared differences of two neighbourhoods. Unfortunately analytical considerations leave open a number of questions about the normalisation of the signals to be correlated. As a result, one finds a variety of normalisation techniques, each with its own characteristics. Thus the first part of this paper examines the derivation of cross-correlation and the motivations for the different forms of normalisation. Different normalisation techniques are compared from both an analytic and an experimental point of view.

A comparison is also made of correlation of the raw signal, the gradient magnitude, and the Laplacian. Experimental results show that correlation of the Laplacian is more precise while correlation with the raw signal is more robust with respect to image noise.The Gradient provides a good trade-off between robustness and precision.

This paper presents an experimental comparisons of different forms of SSD and normalised cross correlation of image neighborhoods. The experiments are performed using pixels as the basis set, but but the results can be extended to comparisons in other spaces.

2. Signal Processing Background

The optimal receiver principal from communications theory (Wozencraft and Jacobs 65) provides a mathematical framework for designing systems which communicate messages over noisy channels. This framework permits a designer to estimate the probability of error for a communication and thus to design systems for which this probability of error is minimized. In this framework, minimizing the probability of error for matching leads directly to minimizing a sum of squared difference. When signals are suitably normalized, minimizing a sum of squared difference is equivalent to maximizing a correlation (Duda and Hart 73). This is interesting because of the existence of image coding hardware for the MPEG standard which can beprogrammed to image correlation at video rates (Inoue et al. 92). It is also interesting because correlation is a form of inner product in a vector space defined by the image pixels. The key question is how to normalize the signals.

The optimum receiver principle was developed at about the same time as the first digital computers, and thus pre-dates both computer vision and pattern recognition. This principle has not become a part of the science of computer vision because the basic assumption of additive Gaussian noise does not apply. However, techniques such as principal components analysis and invariance theory provide a means to factor out the effects of some non-Gaussian perturbations.

2.1 The Optimal Receiver Principal

An optimum receiver is designed to determine the identity of a message given a received signal which has been corrupted by noise. The message to be detected, denoted m_i, is known to have come from a finite set $m_i \in \{M\}$. The message is coded as a time varying signal vector, s(t), which is then transmitted across a channel where it is corrupted by a noise signal n(t) to produce a received signal r(t). The receiver uses the received message to produce an estimate of the most probable message, $\hat{m}_i$, as shown in figure 1.

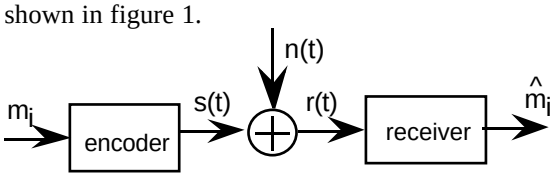

figure 1. A receiver estimates a message from a received signal corrupted with noise.

Having a model of the noise is fundamental to the design of an optimal receiver. Because electromagnetic phenomena obey superposition, the optimum receiver model assumes that the noise source is additive. In computer vision, the signals are discretely sampled vectors, and many of the noise

sources are not additive, but it is possible to design vector spaces which are relatively equi-variant to many sources of noise.

To estimate the identity of the communicated message, the receiver transforms the received message into an new representation by convolving with a set of N basis signals $\varphi_n(t)$, to produce a set of channels $r_n(t)$ for n = 1, ..., N. This basis set can be expressed as a vector basis set $\vec{\varphi}(t) = \{\varphi_n(t)\}$. For continuous time varying signals, convolution is realized by a multiplication followed by integration. A timing device is used to signal the instant T where the integration has completed and the signal may be detected. In computer vision, the simplest such basis set are the individual pixels which compose an image. Basis sets with additional invariance properties can be added using the gradient and Laplacian, or other techniques such as principal components analysis.

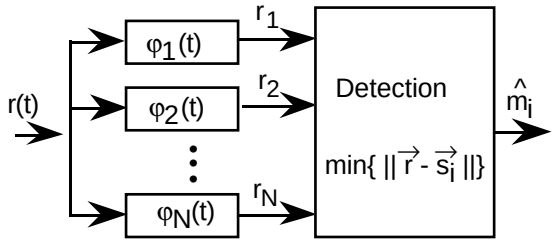

Figure 2. An optimum receiver expresses the received signal with a set of basis functions and then applies a detection rule.

Convolution with the basis set $\vec{\varphi}(t)$ transforms the received signal to a point $\vec{r} = (r_1, r_2, .., r_N)$. Each message, m_i, can be represented as a point, $\vec{s}_i$, in this space obtained by expressing the transmitted signal with the basis set. In a computer vision system, the simplest basis space for the signal vector, r_n, is the image of pixels, R(i, j). A template for a signal to be detected corresponds to a mask of pixels, S(i, j). The pixel vector space can be transformed to a space which has lower dimensionality and better noise invariance by multiplication with another basis set of filters.

In communication theory, the detection process is typically performed by a simple "nearest neighbour" classifier. This is justified by the assumption of additive Gaussian noise, n(t). For such noise, the error probabilities are independent for each channel, and can be described as a noise energy, W. The noise energy specifies the variance for a Gaussian probability distribution, and permits the estimationof the probability of receiving signal vector $\vec{r}$ given that the message was m_i.

$P(\vec{r} \mid m_i) = P(\vec{r} - \vec{s}_i)$.

Applying Bayes rule, and eliminating $P(\vec{r})$ give a probability of a message as a function of the received vector.

$P(m_i \mid \vec{r}) \propto P(\vec{r} \mid m_i)\ P(m_i)$.

Writing the expression for this probability as a Gaussian function and applying a logarithm gives:

$$P(m_i \mid \vec{r}) \propto \| \vec{r} - \vec{s}_i \|^2 + W \, \mathrm{Ln}\{P(m_i)\}$$

Thus, the detector need only determine the message m_i for which distance from $\vec{s}_i$ to $\vec{r}$ is minimum in the signal space. This process is equivalent to a normalized cross correlation (Duda and Hart 72), as can be see by simply computing:

$$\| \vec{r} - \vec{s}_i \|^2 = \| \vec{r} \|^2 - 2 \| \vec{r}\, \vec{s}_i \| + \| \vec{s}_i \|^2.$$

When $\vec{r}$ and $\vec{s}_i$ are normalised to a unit length, then minimizing $\| \vec{r} - \vec{s}_i \|^2$ is equivalent to maximizing the cross correlation $\| \vec{r}\, \vec{s}_i \|$.

2.2 SSD and Correlation.

To apply the optimum receiver principle to the image matching problem, $\vec{r}$ and $\vec{s}_i$ are replaced by arrays of pixels. Let us define the "received" image as R (say of size 512 by 512) and the message signal to be detected as a small image mask S of size M by N. In the absence of noise, S is detected in R at position (i, j) if the sum of squared differences is a local minimum below a threshold. Thus, the matching measure SSD is defined at location (i, j) as shown in equation 1. As seen above, SSD minimizes the probability of error for white Gaussian additive noise.

$$SSD(i,j) = \sum_{m=0}^{M} \sum_{n=0}^{N} \big(R(i+m, j+n) - S(m,n) \big)^2 \qquad (1)$$

In computer vision the goal is often to determine the position at which a reference template best matches the image. Thus the SSD(i, j) must be computed over a search region of possible image locations. For example, in tracking, the search region is centered on the predicted position and its size depends on uncertainty in accelerations of the target. Limiting the size of the search region can speed the search and permit even faster tracking (Berard 94).

As with the optimum receiver (eq. 1), completing the squares of SSD gives two times the inner product of R(i, j) and S subtracted from the sum of

squares of the template and neighborhood. Cross-correlation is an inner product at each pixel. A square root of sum of squares is a measure of the energy of a signal. Thus we can write:

$$SSD(i, j) = E_R^2(i, j) + E_S^2 - 2\, SR(i, j) \qquad (2)$$

where :

$$SR(i, j) = \sum_{m=0}^{M} \sum_{n=0}^{N} S(m, n)\, R(i+m, j+n) \qquad (3)$$

$$E_R^2(i, j) = \sum_{m=0}^{M} \sum_{n=0}^{N} R(i+m, j+n)^2 \qquad (4)$$

and

$$E_S^2 = \sum_{m=0}^{M} \sum_{n=0}^{N} S(m, n)^2 \qquad (5)$$

The term E_S^2 is the energy of the message template and $E_R^2(i, j)$ is the energy of the N x M image neighborhood located at position (i, j).

2.3 The Signal Basis Space

Patterns in images are corrupted by many non-Gaussian and non-additive phenomena. The sensitivity to non-Gaussian noise can be minimized by transforming the received signal to a signal space which is (relatively) invariant to the noise. As an example, consider the case of a change in the direction of light source. Beyond the additive and multiplicative effects of gray-level, which can be corrected by normalising the energy, light source direction also changes the appearance (or gray level pattern) in the image. To overcome this effect, many computer vision techniques use discrete approximations of derivatives of the image.

For a discrete signal, a true derivative can only be computed by convolution with an infinite length filter. However, it is quite adequate to approximate derivatives with discrete differences, provided that the image has been suitably smoothed. The first-order difference for a row of an image is equivalent to a small filter of the form [1 0 –1] or the form [1 –1]. The form [1 0 –1] is less sensitive to high frequency noise. It also makes it possible to define a basis space with is orthogonal to both a second derivative computed by [1 –2 1] a 3 point average [1 1 1]. Thus we define our first difference filters as :

$$\Delta i = \boxed{1\ 0\ -1} \qquad \Delta j = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

For a suitably smoothed image, p(i, j), these filters provide a discrete approximation to the gradient (Chehikian and Crowley 91) of the form:

$$\nabla p(i, j) = \begin{bmatrix} \Delta i * p(i,j) \\ \Delta j * p(i,j) \end{bmatrix}$$

The Gradient magnitude, $\|\nabla p(i, j)\|$ computed as the square root of the sum of the squares, provides an approximation to the first derivative which is relatively invariant to rotation.

A number of researchers have asserted that the Laplacian provides a basis space in which signals may be represented with greater precision and has a relative invariance to changes in lighting which is superior to the derivative. The Laplacian is a sum of second derivatives and thus requires only one filter.

A discrete second difference which is orthogonal to the first order difference filter shown above is given by the filter:

$\Delta^2 i = \boxed{1\ -2\ 1}$.

Approximations to the discrete Laplacian are as old as computer vision. Chehikian (Chehikian and Crowley 91) has shown analytically that weighted sums of the second order difference at different angles provides excellent rotation invariance.

$$\nabla^2 = \boxed{\begin{array}{ccc} -1 & -2 & -1 \\ -2 & 12 & -2 \\ -1 & -2 & -1 \end{array}}$$

So, should one match the image, the gradient magnitude or the Laplacian? The answer depends on the nature of the signal and the requirements of the task. In the frequency domain, an ideal first derivative grows linearly with higher frequency, while a second derivative grows quadratically. Thus a correlation of first derivatives should have a more precise peak than a correlation of raw images, but be more sensitive to high frequency noise, such as the noise caused by aliasing in the digitizer. A second derivative doubles the effect. The gradient magnitude approximation presented above is in fact a band-pass filter, that tapers monotonically to zero at the Nyquist frequency. As a result, it is very well suited to suppress aliasing noise. The Laplacian filter shown above is a high pass filter. The choice of basis depends on the noise characteristics of the signals and the precision and stability requirements

of the task.

3. Normalisation Techniques

The existence of low cost correlation hardware makes cross-correlation a very attractive operation. Section 2 showed that the SSD of a template and an image neighborhood could be replaced by cross-correlation, provided that both the template and the image neighborhood are suitably normalized. This section compares common normalisation techniques.

3.1 Normalized Cross Correlation

The most direct normalisation is to divide each inner product by the square root of the energy of the template and the neighborhood. Mathematically, this can be expressed as

$$NCC(i, j) = \frac{SR(i, j)}{\sqrt{E_R^2(i, j) \cdot E_S^2}} \qquad (6)$$

The template can be normalised before starting the correlation. A simple algorithm exists to compute the energy within an N by M neighborhood at each pixel, based on a fast algorithm for computing the sum of a rectangular neighborhood. The cost of this algorithm is independent of the size of the template, and requires 1 multiply, 4 additions and 1 square root per pixel of the search neighborhood. However, most experimenters consider that individually normalising each image neighborhood is too costly, and design an approximation. For example, it is not unusual to see the energy of a neighborhood replaced by the global energy of the image.

3.2 Zero Mean Normalized Cross Correlation

With zero mean cross correlation, the mean is subtracted from both the template and the image neighborhood before computing either the inner product or the neighborhood energy. The mean of the template, μ_S, and of the image neighborhood $\mu_R(i, j)$ are given by:

$$\mu_S = \frac{1}{M \cdot N} \sum_{m=0}^{M} \sum_{n=0}^{N} S(m,n)$$

$$\mu_R(i, j) = \frac{1}{M \cdot N} \sum_{m=0}^{M} \sum_{n=0}^{N} R(m+i,n+j)$$

This gives the formula, ZNCC, shown in equation 6 (Shown below because of the two column layout required by the conference).

$$ZNCC(i, j) = \frac{\sum_{m=0}^{M} \sum_{n=0}^{N} (S(i, j)-\mu_S)\,(R(i+m,j+n)-\mu_R(i, j))}{\sqrt{\sum_{m=0}^{M} \sum_{n=0}^{N} (R(i+m,j+n)-\mu_R(i, j))^2 \cdot \sum_{m=0}^{M} \sum_{n=0}^{N} (S(m, n)-\mu_S)^2}} \qquad (6)$$

$$MOR(i, j) = \frac{\sum_{m=0}^{M} \sum_{n=0}^{N} (R(i+m,j+n)-\mu_R(i, j))\,(S(m, n)-\mu_S)}{\sum_{m=0}^{M} \sum_{n=0}^{N} (R(i+m,j+n)-\mu_R(i, j))^2 + \sum_{m=0}^{M} \sum_{n=0}^{N} (S(m, n)-\mu_S)^2} \qquad (7)$$

$$FUA(i, j) = \frac{\sum_{m=0}^{M} \sum_{n=0}^{N} \left((R(i+m, j+n)-\mu_R(i, j)) - (S(m,n)-\mu_S)\right)^2}{\sum_{m=0}^{M} \sum_{n=0}^{N} (R(i+m,j+n)-\mu_R(i, j))^2 + \sum_{m=0}^{M} \sum_{n=0}^{N} (S(m, n)-\mu_S)^2} \qquad (8)$$

4. Experimental Comparisons

The choice of basis set and normalisation depends on the sharpness of the peak required, the stability of the peak in the presence of perturbations, and the kinds of noise which are embedded in the signal. This section compares different forms of correlations in the presence of changes in lighting, additive Gaussian noise, and salt and pepper noise. For each noise source the response of the correlation of the raw image, the gradient magnitude and the Laplacian are presented as a function of the parameter of the noise.

4.1 Change of Light Level.

Accommodating changes in ambient light is one of the most fundamental requirements for any computer vision system. To test the sensitivity to light changes, a sequence of 7 images of size 256 by 256 were acquired of a person standing in the laboratory with a background of desk, tables and chairs while the light level was adjusted with the rheostat of a halogen lamp. The average gray level of the images were (229, 209, 183, 154, 124, 94, 65). The 4th image (avg 154) was selected as a template and compared to the other images using SSD. The results are summarized in table 1.

No.	SSD of Image	SSD of Gradient	SSD of Laplacian
1	17231548	1209662	2464934
2	8502748	379890	805602
3	2155414	94139	184864
4	0	82918	150133
5	2160000	82918	150133
6	8640000	82918	150133
7	19360516	86654	174786

Table 1. SSD of image, gradient and Laplacian as a function of change in gray level.

The results shown in Table 1 are displayed in a more synthetic manner by the graph in figure 3. As one would expect, the SSD of the gradient and Laplacian are relatively immune to changes in lighting, while the correlation of the raw image signal is quite sensitive.

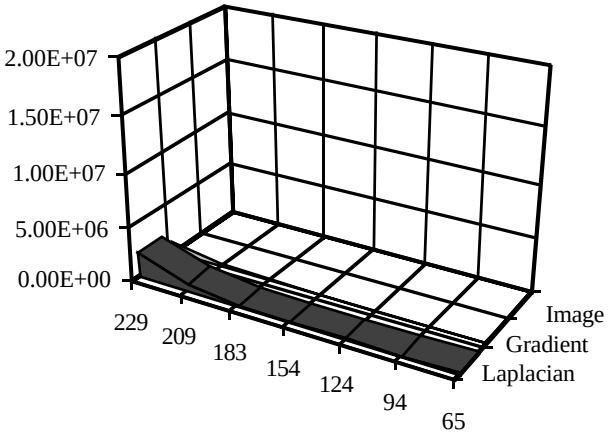

Figure 3. SSD of Image, Gradient and Laplacian as a function of light level.

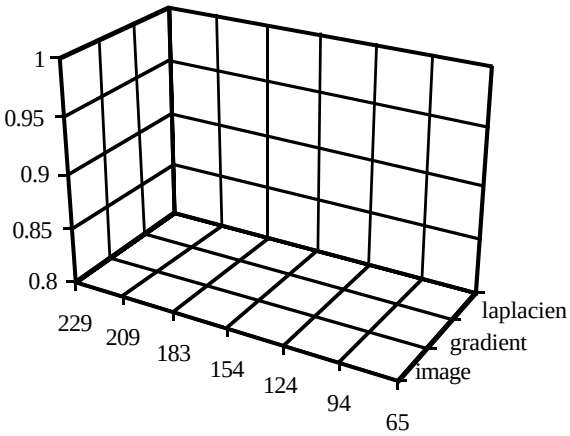

Figure 4. Results of Normalized Cross-Correlation (NCC) as a function of light level.

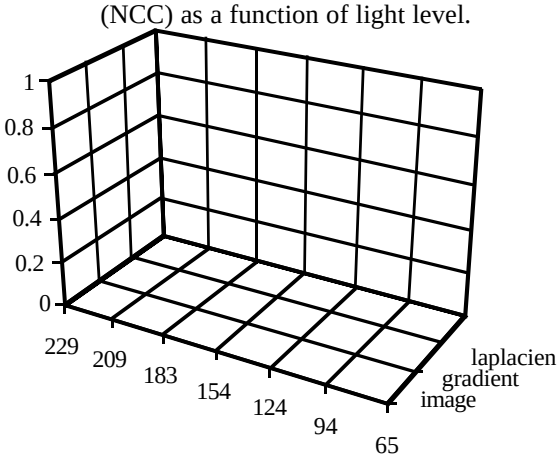

Figure 5. Results of ZNCC as a function of light intensity.

The same images are used to compare NCC in figure 4, and ZNCC in figure 5. All measures except the SSD of the raw image appear to be relatively stable with regards to a change in ambient light intensity. The SSD of the Gradient and Laplacian, and the ZNCC of the image appeared to exhibit the most stability.

4.2 Additive Gaussian Noise.

This second experiment tests the sensitivity of the different match measures to additive Gausian noise. An image was successively corrupted with additive Gaussian noise with a standard deviation varying from 0 to 0.7 in steps of 0.1. The SSD of original image and its corrupted copy were correlated using SSD (figure 6), NCC (Figure 7) and ZNCC (figure 8).

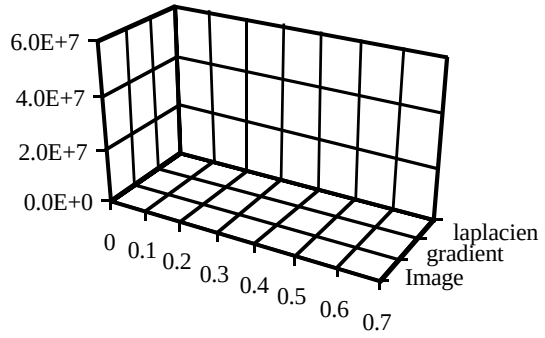

Figure 6. Results of SSD as a function of standard deviation of additive Gaussian Noise.

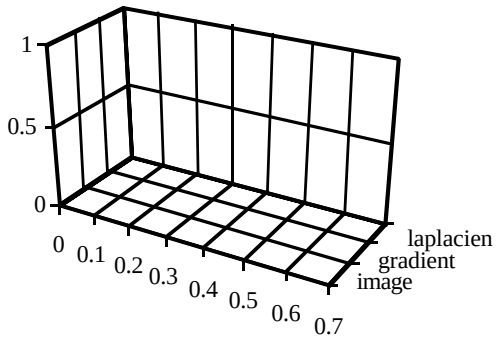

Figure 7. Results of NCC as a function of standard deviation of additive Gaussian Noise.

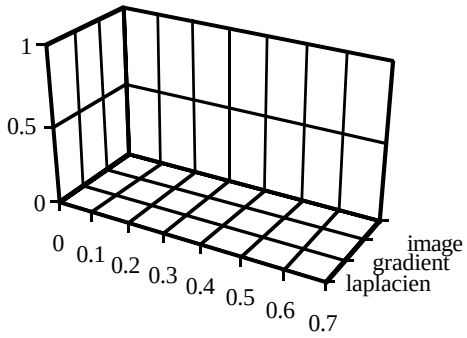

Figure 8. Results of ZNCC as a function of standard deviation of additive Gaussian Noise.

4.3 Salt and Pepper Noise.

Images are sometimes corrupted by "replacement" noise. In such a case, a random value replaces the value of a pixel. A common model for such noise is "Salt and Pepper" noise, in which some percentage of the pixels are randomly replaced by white (255) or black (0) pixels. Figures 9, 10 and 11 show SSD, NCC and ZNCC as a function of number of pixels modified by "Salt and Pepper" noise.

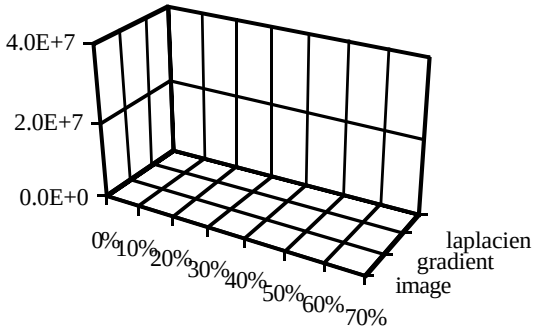

Figure 9. SSD with Salt and Pepper noise as a function of percent pixels.

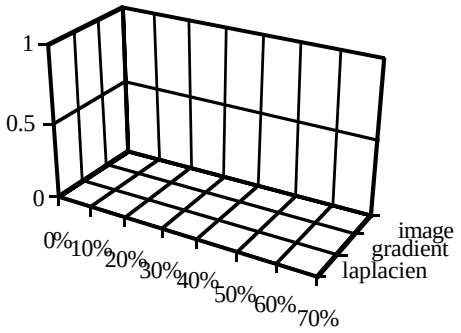

Figure 10. NCC with Salt and Pepper noise as a function of percent pixels.

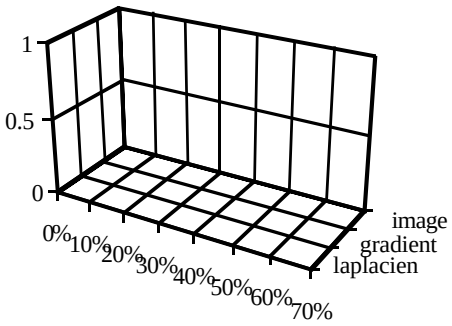

Figure 11. ZNCC with Salt and Pepper noise as a function of percent of pixels changed.

5 Conclusions

The appropriate choice of basis set and normalisation are determined by the task. None-the-less, the analysis and experiments permit us to make the following observations. In general, SSD provides a more stable result than NCC or ZNCC. For the sources of noise in these tests, the SSD of the gradient usually provides the most stable result. When SSD must be replaced by a correlation, it is generally preferable to use the NCC of the gradient. In general, NCC seems to provide a more stable detection than ZNCC (and is also less costly) although, we feel that this last conclusion must be the subject of further experiments. Furthermore, the gradient provides a more stable result than the Laplacian.

Bibliography

Anandan, P. (1987), “Measuring Visual Motion from Image Sequences”, Doctoral Dissertation, Computer Science Department, University of Massachusetts, May 1987.

Bérard, F. (1994), “Vision par Ordinateur pour la Réalité Augmentée : Application au Bureau Numérique.”, DEA INPG, June 1994.

Binford, T., (1982) "Survey of Model Based Image Analysis Systems", International Journal of Robotics Research, 1(18), 1982.

Burt P. J. and E. H. Adelson, (1983), "The Laplacian Pyramid as a Compact Image Code", IEEE Transactions on Communications, Vol 31, No. 4, 1983.

Chehikian A. and Crowley J. L., (1991), "Fast Computation of Optimal Semi-Octave Pyramids. ", 7th S.C.I.A., Aalborg, August 1991.

Crowley, J. L. and Stern, R. M. (1984), "Fast Computation of the Difference of Low-Pass Transform", IEEE Transactions on PAMI, PAMI 6(2), March 1984.

Duda, R. O. and Hart, P. E., (1973), Pattern Classification and Scene Analysis, Wiley, N. Y. 1973.

Fua. P. , (1994), “A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Images Features.” Submitted to the Journal of Machine Vision and Applications.

Granlund, G. H. (1978) "In Search of a General Picture Processing Operator", Computer Graphics and Image Processing, 8(2), pp 155-178, 1978.

Inoue, H. , Tashikawa, T. and Inaba M., (1992), "Robot vision system with a correlation chip for real time tracking, optical flow, and depth map generation", 1992 IEEE Conference on Robotics and Automation, Nice, April 1992.

Knutsson H. (1989), "Representing Local Structure Using Tensors", In the 6th S.C.I.A., Oulou Finland, June 1989.

Martin, J. (1994), "Techniques Visuelles de Détection et de Suivi de Mouvements", Rapport de Magistère, LIFIA-IMAG, 1994.

Moravec, H.P., (1977), “Towards Automatic Visual Obstacle Avoidance”, 5th IJCAI, Cambridge, August 1977.

Murase, H. and Nayar S. K. (1993), “Learning and Recognition of 3D Objects from Appearance”, IEEE Workshop on Qualitative Vision, New York 1993.

Turk, M. and Pentland, A. P., (1991), “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, 3(1):71-86, 1991.

Wozencraft J. M. and Jacobs I. M., (1965), Principles of Communication Engineering, John Wiley and Sons, 1965.