Improving feature tracking by robust points of interest selection

C. Kermad, Christophe Collewet

Improving Feature Tracking by Robust Points of interest Selection

Chafik KERMAD, Christophe COLLEWET

CEMAGREF. 17, Avenue Cucillé, CS 64427, 35044 Rennes, France
e-mail: FirstName.LastName@cemagref.fr

Abstract

This paper deals with robust point features selection for tracking. The aim is to identify unreliable features since the first frame so to track them in all the sequence. We extend a recent version of the well-known Kanade-Lucas-Tomasi tracker [18] by introducing an automatic scheme for rejecting spurious features. We employ a simple and efficient rejection rule based on grey levels co-occurrence entropy and show that its empirically assumptions are satisfied in the scenario of feature tracking. Experiments with real and synthetic images confirm that this approach makes better features tracking. We illustrate quantitatively the benefits introduced by the proposed algorithm.

1 Introduction

As many algorithms rely on the accurate computation of correspondences through a sequence of images, feature tracking has proved to be an essential component of vision-based systems. It remains a difficult problem on which depend many high level tasks as motion estimation, 3D reconstruction, dynamic vision or visual servoing, and, these can concern a variety of applications ranging from medical imaging or continuous inspection to surveillance or road traffic scenes analysis [1] [7] [8] [10] [12] [13].
A usual technique for feature tracking is to minimise the Sum of Squared Differences (SSD) of images intensities. In [19], Tomasi and Kanade introduced a feature tracker based on SSD matching and assuming translational frame-to-frame displacements. Subsequently, in order to take into account more complex displacements, Shi and Tomasi proposed in [17] an affine model which proved adequate for region matching over longer time spans. Their system classified a tracked feature as reliable or unreliable according to the residue of the match between the associated image region in the first and current frames, if the residue exceeded a defined threshold, the feature was rejected. Visual inspection of results demonstrated a well discrimination between good and bad features. Nevertheless, the authors did not specify how to reject bad features automatically. This problem has been solved by Tommasini and *al.* [20] by employing a simple outlier rejection rule. However, this kind of tracking algorithms usually assume that the changes in the scene appearance are only due to geometric deformations. Thus, when changes in illumination are relevant, these approaches perform poorly. In [4], Hager and Belhumeur describe an SSD-based tracker that compensates illumination changes. In their approach, first a target region is defined, then a basis of appearance reference templates for the illumination change is acquired and stored, and finally tracking is performed based on the prior knowledge of the appearance templates. Recently, without intervention of *a priori* knowledges, Soatto and *al.* presented in [18] an extension of Shi-Tomasi tracker which take into account changes in illumination and reflection.

In this paper, following works of Lucas and Kanade [11] which mentioned the importance of the robust selection in the first frame of the features to be tracked, we extend the approach of Soatto and *al.* by introducing an automatic scheme for the selection of significant features in the first frame of the sequence. This approach of processing imposes itself for various components using visual servoing and/or dynamic vision (see [2] for example). We employ a rejection rule based on an entropy criterion, and show that this way to proceed yields to a better behaviour of the tracking process for the

selected points. Experiments with real images confirm that our algorithm makes well features tracking, in the sense that outliers are located reliably. By using the global residual criterion, we present quantitative examples of the benefits introduced by the proposed approach.

This paper is organised as follows: we describe in section 2 the proposed approach for the detection and the selection of the *visual features*. Then, we present in section 3 the method developed by Soatto and *al.* which will allow us to carry out the tracking of the selected features. Experimental results and method evaluation on different sequences are then detailed in section 4. Finally, in section 5, a conclusion recalls the principal results and the limits of the method and proposes some research perspectives.

2 Features Extraction and Selection

Among the visual features, the interest points prove to have the most general character. Indeed, those can be found in the majority of the images and their extraction can be applied to simple objects as well as to complex objects. These points correspond generally to bi-dimensional variations of the intensities of the image [3] [6] [14] [15] [16] [22].

Intuitively one conceives easily that a point lying in an area of homogeneous grey levels has less chances to be well matched that a point belonging to a rich area of information. Moreover, more textures present around a point is rich and varied, more information in the image is discriminating. To evaluate the quantity of visual information present around a point, the information theory offers probabilistic tools based on entropic measures. These measurements have revealed well adapted for the characterisation of the objects of complex and textured nature [5], and, can give an indication on the order of textures in the analysed area of the image, and so, allows to locate the areas where information is richest. Based on these considerations, we propose the use of zones of interest located by measurement of the entropies of co-occurrence of the grey levels of the image in order to select the points candidates to the tracking task. The proposed method is composed of the two following steps:

- A detection of the interest points following the approach of Shi and Tomasi [17]. This detector is characterised by a good repeatability rate [16]. A feature is defined as a region that can be tracked easily from one frame to the other. In practice, since the larger eigenvalue is bounded by the maximum allowable pixel value, the requirement is that the smaller eigenvalue is sufficiently large. Calling λ_1 and λ_2 the eigenvalues of the matrix H with :

$$H = \sum_{W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \qquad (1)$$

  We accept the corresponding feature if $\min(\lambda_1, \lambda_2) > \lambda$, where λ is a user-defined threshold, W is a window of interest in the image I and $[I_x \ I_y]^T = [\partial I/\partial x \ \partial I/\partial y]^T$. Typical values of λ are in [0.1..1.0].

- A selection of the interest points by optimising the criterion of entropy. The most discriminating points according to these values correspond to the points whose illumination variations are large in several different directions. In order not to use all the grey levels of the image and to reduce computational cost, we simplify it by classifying the grey levels automatically in a reduced number using a multithresholding technique [9]. Then, we calculate the entropy of the co-occurrences of the grey levels of the segmented image around the interest points detected in the first step.

  Given a displacement vector $(\Delta_x, \Delta_y)^T$ and the number of grey levels G of the image I, for two grey levels k and l, the value of the co-occurrence matrix noted M, at position (k, l), is given by the number of pixels (x, y) such that $I(x, y) = k$ and $I(x + \Delta_x, y + \Delta_y) = l$. Typical displacements correspond to $(\Delta_x, \Delta_y) \in \{(1, 0), (1, 1), (0, 1), (-1, 1)\}$. The entropy value around an interest points is given by :

$$Ent_P = -\sum_{k \in G} \sum_{l \in G} p_{kl} \log(p_{kl}) \qquad (2)$$

  where $p_{kl} = \dfrac{M[k, l]}{W_h^2}$. Typical value for the size of the window of analysis $W_h \times W_h$ is 16×16.

Only the points having an entropy greater than a threshold, function of the entropies of the interest points, are preserved. In the case of an image coded on a dozen of grey levels, the 2D entropy values

vary in an interval going from 0 to 2.56, and, typically, 70% of the maximal value of entropies represents a good compromise for this threshold value.

3 Features Tracking

This section presents the model of the deformation undergone by the image as a consequence of the relative motion between a rigid scene and a camera, and of illumination changes. Our aim is to develop a model that treads generally with computational efficiency, leading to a real-time implementation.
Let X be a point on a surface S in the scene and $\mathbf{x} = \pi(X)$ the projection of point on the image plane. Let $I(\mathbf{x}, t)$ denotes the intensity value at the location $\mathbf{x}$ of an image acquired at time t. The image deformations of a surface S at time $t+\tau$ can be described as :

$$I(\mathbf{x}, t) = I(\delta(\mathbf{x}), t+\tau) \quad \forall \mathbf{x} \in W \qquad (3)$$

where $\delta(.)$ is generally a non-linear time-varying function which depends on a number of parameters. In the case of an affine model, $\delta(\mathbf{x}) = A\mathbf{x} + d$; where A is a general linear transformation of the plane co-ordinates $\mathbf{x}$ and d is the 2D displacement.
The geometrical model defined by the relation (3) has the advantage of being simple and general enough. Nevertheless, in practice, the conditions of acquisition of the images can not always be controlled. It can yield to discontinuities of illumination and therefore to the violation of (3). Moreover, it generates areas with weak contrast which contribute in the degradation of the images and consequently of the tracking. In order to take into account the variations of lighting of the scene and to operate in more general situations, the combination of the geometrical model with a photometric modelling seems to be necessary. Indeed, between two frames acquisition, the conditions of lighting are different. A pixel in a frame can be clearer than in the other, because of the change of the lighting or the opening of the iris. This change can be approximated by an addition of a grey level. Also, because of the differences of dynamic of the camera, the contrast in a frame can be stronger than in the other. This variation can be approximated by a scale factor [11].
If we consider a light source Ł in the 3-D space and suppose that we observe a smooth surface S. The intensity value of each point on the image plane depends on the portion of incoming light from the source Ł that is reflected by the surface S, and is described by the Bi-directional Reflectance Distribution Function (BRDF). When we assume that for a point p on the smooth surface S, it is possible to consider a neighbourhood U around p such that normal vectors to S do not change within U. Under the above assumptions, and assuming that the surface is Lambertian, the BRDF simplifies considerably and the intensity observed at the point $\mathbf{x}$ can be modelled as [18]:

$$I(\mathbf{x}, 0) = \nu_E E(X) + \xi_E \quad \forall \mathbf{x} \in W_U \qquad (4)$$

where $E(X)$ is the albedo lying on S, $W_U = \pi(U)$, ν_E and ξ_E are constant and can be thought as parameters that represent respectively the contrast and brightness changes of the image. When either the camera or the scene is subject to motion, these parameters will change according to the new geometric setting, and hence ν_E and ξ_E vary along of the time. The model for illumination changes can be written as follows [18]:

$$I(\mathbf{x}, 0) = \nu(t) I(\mathbf{x}, t) + \xi(t) \quad \forall \mathbf{x} \in W_U \qquad (5)$$

where $\nu(t)$ and $\xi(t)$ are recursively defined as: $\nu(t) = \frac{\nu_E(t)}{\nu_E(0)}$, $\xi(t) = \xi_E(t) - \frac{\nu_E(t)}{\nu_E(0)}\xi_E(0)$ for $t > 1$

The combination of the affine and illumination deformations models together gives the following model:

$$I(\mathbf{x}, 0) = \nu(t) I(A\mathbf{x} + d, t) + \xi(t) \quad \forall \mathbf{x} \in W \quad (6)$$

Because of noise and because the affine motion model and the affine illumination model are approximations, equation (6) is generally not satisfied. Therefore, the solution is to move the problem to an optimisation one to find the parameters A, d, ν and ξ, which minimise the following residue:

$$\epsilon = \int_W [I(\mathbf{x}, 0) - \nu I(A\mathbf{x} + d, t) + \xi]^2 w(\mathbf{x}) d\mathbf{x} \quad (7)$$

where $w(.)$ is a weight function depending on the application. In the simplest case, $w(\mathbf{x}) = 1$.
To carry out the minimisation, the model of intensity is approximated by using a first-order Taylor expansion around $A = I_d$, $d = 0$, $\nu = 1$ and $\xi = 0$:

$$\nu I(\mathbf{y}, t) + \xi \simeq \nu I(\mathbf{x}, t) + \xi + \nabla I \frac{\partial \mathbf{y}}{\partial u}(u - u_0) \quad (8)$$

∇I is the gradient of $I(\mathbf{x}, t)$, $\mathbf{y} = A\mathbf{x} + d$, u collects the geometric parameters $A = \{d_{ij}\}$ and $d = [d_1\ d_2]^T$, $u = [d_{11}\ d_{12}\ d_{21}\ d_{22}\ d_1\ d_2]^T$ and $u_0 = [1\ 0\ 0\ 1\ 0\ 0]$. Rewriting (8) in the matrix form, we obtain

$$I(\mathbf{x}, 0) = F(\mathbf{x}, t)^T z \tag{9}$$

with $F(\mathbf{x}, t) = [xI_x\ yI_x\ xI_y\ yI_y\ I_x\ I_y\ I\ 1]^T$, $z = [d_{11}\ d_{12}\ d_{21}\ d_{22}\ d_1\ d_2\ \nu\ \xi]^T$, and, x and y are the co-ordinates of $\mathbf{x}$.

The problem reduces to determining z for each patch. Multiplying (9) by $F(\mathbf{x}, t)^T$ on both sides, and integrating over the whole window W with the weight function $w(.)$, we obtain the following linear 8×8 system $Sz = a$ where :

$$a = \int_W F(\mathbf{x}.t)^T I(\mathbf{x}, 0) w(\mathbf{x}) d\mathbf{x} \tag{10}$$

and

$$S = \int_W F(\mathbf{x}.t)^T F(\mathbf{x}.t) w(\mathbf{x}) d\mathbf{x} \tag{11}$$

When we consider the pixel quantization, S becomes in block-matrix form as:

$$S = \sum_{\mathbf{x} \in W} \begin{bmatrix} T & U \\ U^T & V \end{bmatrix} w(\mathbf{x}) \tag{12}$$

where

$$T = \begin{bmatrix} x^2 I_x^2 & xyI_x^2 & x^2 I_x I_y & xyI_x I_y & xI_x^2 & xI_x I_y \\ xyI_x^2 & y^2 I_x^2 & xyI_x I_y & y^2 I_x I_y & yI_x^2 & yI_x I_y \\ x^2 I_x I_y & xyI_x I_y & x^2 I_y^2 & xyI_y^2 & xI_x I_y & xI_y^2 \\ xyI_x I_y & y^2 I_x I_y & xyI_y^2 & y^2 I_y^2 & yI_x I_y & yI_y^2 \\ xI_x^2 & yI_x I_y & xI_x I_y & yI_x I_y & I_x^2 & I_x I_y \\ xI_x I_y & yI_y^2 & xI_y^2 & yI_y^2 & I_x I_y & I_y^2 \end{bmatrix}$$

$$U^T = \begin{bmatrix} xI_x I & yI_x I & xI_y I & yI_y I & I_x I & I_y I \\ xI_x & yI_x & xI_y & yI_y & I_x & I_y \end{bmatrix}$$

and

$$V = \begin{bmatrix} I^2 & I \\ I & 1 \end{bmatrix}$$

T is the matrix computed in the algorithm of Shi and Tomasi, which is based on geometry only, V comes from the model of photometry and U is the cross terms between geometry and photometry. Finally, when S is invertible, z can be computed as:

$$z = S^{-1} a \tag{13}$$

By solving (13), it is possible to compute all the parameters. However, because of the first-order approximation in (8), it will only give a rough approximation for z even if the model is correct. To achieve a higher accuracy, a Newton-Raphson-style iteration is used. This can be done by approximating (6) around the previous solution, and iterating (13) until the variation in all the parameters is negligible.

4 Experimental Results

In this section, two complementary experiments are presented: the first is applied to a scene of a complex object, and the second one to a scene of textured images.

4.1 Sequence on complex object

The proposed algorithm was tested on several real sequences of complex objects. We present here some results obtained on a sequence resulting from the observation of a food product (sequence FP, figure 1). The 25 frames of the sequence were acquired by a CCD camera fixed on a robot. In this experiment, the camera turns from left to right and during rotation the illumination changes substantially. All the results presented here were obtained under the same experimental conditions: the threshold applied to the eigenvalues (parameter λ defined in section 2) is equal to 0.1 and the threshold of the entropy (also defined in section 2) is fixed at 70% of the value of the maximum entropy. In order to obtain an accurate estimation of the deformations, we have introduced a threshold on the residual values ϵ, only points having an $\epsilon < 0.2$ are tracked.

The figure 2 represents the image of the entropies of co-occurrences of the 12 grey levels obtained by the multi-thresholding of the first image. One can notice that the strongest values (clear grey levels) of this attribute correspond well to representative areas of interest of the image.

The results of extraction and selection of the interest points are respectively presented in figures 3(a) and 3(b). For this example, among the 118 detected points, only 69 are localised in areas with strong entropy and thus considered as good candidates for the tracking.

To evaluate quantitatively the developed approach, we applied the tracking algorithm by using a first version where the extraction of the interest points is made according to the only criterion related to the eigenvalues defined in section 2 (Method EV) while the second version is made according to the criterion of entropy (Method ENT). The figure 4

shows trajectories of the points selected by the two approaches and tracked along the sequence. One can note that for a considerable number of points (generally corresponding to points located on areas with low entropy), the tracking stops before the end of the sequence.
The figure 5 shows that the evolution of the global normalised residues (sum of the residues divided by the number of points tracked after the processing of each frame) according to the number of the frame in the sequence is weaker when one carries out a tracking of the points selected according to the entropy criterion.
In the same way, table 1 recapitulates the results obtained. In particular, it shows that a relative profit, about 15%, on the residues is obtained thanks to the introduction of the criterion of entropy. In addition, in a context of visual servoing, where a small number of points as visual features is used, the experimentation shows that 80% of the best sorted points according to values of entropy are correctly tracked, whereas only 60% of the points are correctly tracked when the sorting is performed according to the eigenvalues.

4.2 Sequence of textured images

The developed algorithm was also tested on a number of sequences of textured images among the well-known sequence "Flower Garden" (Sequence FG, figure 6). The difficulty in this sequence is due to the presence of non homogeneous textures in the images. This can hugely contributes to produce inter-frame false matching. In this experiment, the threshold λ applied to the eigenvalues has been fixed to 0.5 (to limit the number of detected points), the threshold on the entropy remains unchanged.
The results for this experiment are given in figures 7, 8 and 9. They respectively represent: results of detection and selection of the interest points, the trajectory of tracking of the points detected according to two methods (EV and ENT) and the evolution of the normalised residues versus frame numbers. Table 2 quantitatively summarises the obtained results.
Similar remarks to those for the previous section can be made for this sequence. In particular, figure 9 and table 2 show, that in this case also, a better global residue is obtained after selection of the points using the entropies, and, the percentage of tracked points is increasingly more significant when a sorting of points according to the values of entropies is made. The profit obtained on the residues is about 6%, and, on the other hand, the gain obtained on the percentage of the points correctly tracked in all the sequence is about 25%. The tracking of the ten best sorted points according to the two methods is identical (80% of points are correctly tracked). These results show, here again, that our approach contributes to a good selection of the points extracted from the first image, and so, to a good tracking.

5 Conclusion

We have presented an extension of the Soatto and *al.* tracker by introducing a selection of interest points on the initial frame of the sequence, based on grey level co-occurrence entropy attribute. Point features are thus rejected according to their grey level structures. The experimental results on real images of complex objects, textured or not, allowed to illustrate, by using the criterion based on the global residue, the contribution of our method with regard to that of Shi-Tomasi-Kanade classically used. On average, a gain about 10% on residues is obtained thanks to the introduction of the entropy criterion.
Although the discussed algorithm do yield good results in most cases, some points localised on contours are preserved and tracked without being significant. A perspective of this work is thus to improve the proposed method by making it less sensitive to this kind of points. To alleviate in part this problem, we plan to take into account, besides the entropy criterion, information related to the gradient of the image and therefore to eliminate such points.

References

[1] A.A. Amini, R.W. Curwen and J.C. Gore, Snakes and splines for tracking non-rigid heart motion. Proc. ECCV'96, pp. 251-261, 1996.

[2] C. Collewet, F. Chaumette and P. Loisel, Image-based visual servoing on planar objects of unknown shape. Proc. IEEE ICRA'2001, Seoul, May 2001.

[3] W. Förstner, A framework for low level feature extraction, Lecture Notes in Computer Science, Jan-Olof Eklundh (Ed.), Springer-Verlag, 802:383-394, 1994.

[4] G. Hager and P. Belhumeur, Efficient Region Tracking with Parametric Models of Geometry and Illumination, IEEE Trans. on Pattern Analysis and Machine Intell., 20(10):1025-1039, 1998.

[5] R. Haralick, Statistical and structural approaches to textures. Proceedings of the IEEE, 67(5), 1979.

[6] C. Harris and M. Stephens, A Combined Corner and Edge Detector. Proc. 4th Alvey Conference, pp. 147-151, Manchester, Royaume Uni, Aug. 1988.

[7] S. Hutchinson, G.D. Hager, and P.I. Corke, A tutorial on visual servo control. IEEE Trans. on Robotics and Automation, 12(5):651-670, 1996.

[8] M. Isard and A. Blake, Condensation: Conditional Propagation for Visual Tracking. Int'l Journal for Computer Vision, 29(1), p. 5- 28, 1998.

[9] C. Kermad and K. Chehdi, Multi-Spectral Image segmentation : a scalar approach. Proc. IEEE ICIP'00, Vancouver, Sept. 2000.

[10] D. Koller, K. Daniilidis, and H. Nagel, Model-based object tracking in monocular image sequences of road traffic scenes. Int'l Journal of Computer Vision, 10(3):257-281 1993.

[11] B.D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision. Proc. IJCAI'81, pp. 674-679, 1981.

[12] E. Marchand, P. Bouthemy, F. Chaumette and V. Moreau, Robust real-time visual tracking using a 2D-3D model-based approach. Proc. IEEE ICCV'99, pp. 262-268, Sept. 1999.

[13] N.P. Papanikolopoulos, P.K. Khosla, and T. Kanade, Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision. IEEE Trans. on Robotics and Automation, 9:14-35, 1993.

[14] B.J. Robbins, R. Owens, 2D feature detection via local energy, Image and Vision Computing, vol. 15, pp. 353-368, 1997.

[15] K. Rohr, Recognizing corners by fitting parametric models. Int'l Journal of Computer Vision, 9:213-230, 1992.

[16] C. Schmid, R. Mohrand and C. Bauckhage, Evaluation of Interest Point Detectors, Int'l Journal of Computer Vision, 37(2), 151-172, 2000.

[17] J. Shi and C. Tomasi, Good features to track. Proc. IEEE CVPR'94, pp. 593-600, June 1994.

[18] S. Soatto, H. Jin and P. Favaro, Real-time Feature Tracking and Outlier Rejection with Changes in Illumination, Proc. IEEE ICCV'01, Vancouver, July 9-12, 2001.

[19] C. Tomasi and T. Kanade, Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, April 1991.

[20] T. Tommasini, A. Fusiello, E. Trucco, and V. Roberto, Making good features track better. Proc. IEEE CVPR'98, Santa Barbara, California, pp. 178-183, Jun. 1998.

[21] K. Toyama, and G. D. Hager, Incremental Focus of Attention for Robust Visual Tracking. Proc. IEEE CVPR'96, 1996.

[22] Z. Zheng, H. Wang and E.K. Teoh, Analysis of gray level corner detection. Pattern Recognition Letters, Vol. 20, pp. 149-162, 1999.

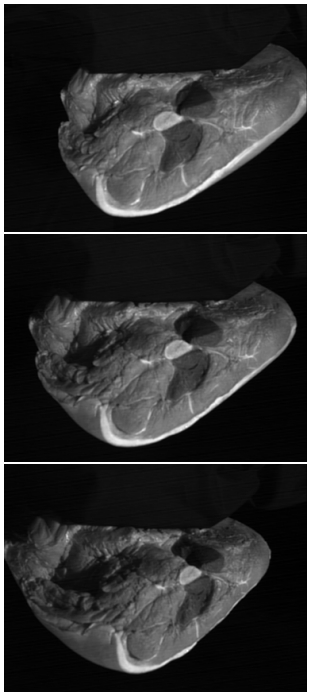

Figure 1: *Examples of frames from the sequence FP.*

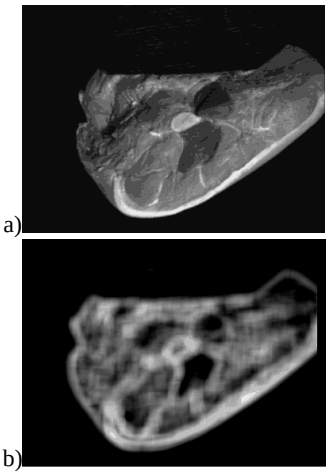

Figure 2: *a) Result of the multi-thresholding of the first image. b) Image of the entropies of the co-occurrences of the grey levels issued from the multi-thresholding.*

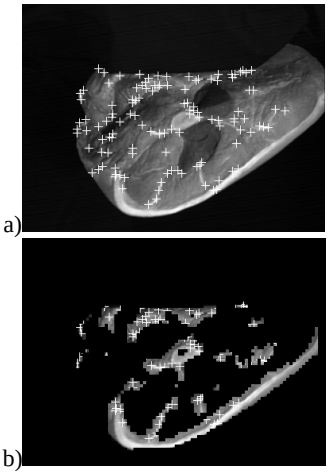

Figure 3: *a) Result of extraction of the interest points using eigenvalues criterion. b) Result of selection of the interest points using entropy criterion (only unmasked points will be tracked). Sequence FP.*

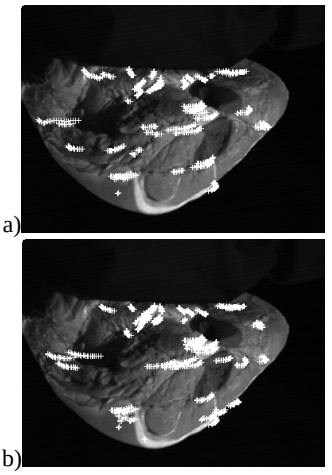

Figure 4: *a) Tracking trajectories of the first classified interest points based on the eigenvalues sorting. b) Selected points tracking trajectories. Sequence FP.*

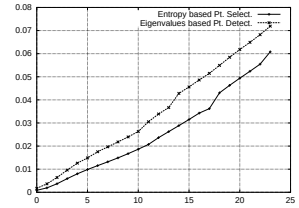

Figure 5: *Evolution of global residues vs. frame number. Sequence FP.*

Meth.	NbIni	NbSel	NbTrk	PTP	SR
EV	118	69	32	46	0.0719
ENT	118	69	36	52	0.0608

Table 1: *NbIni: The number of points detected in the initial image. NbSel: The number of selected points. NbTrk: The number of points tracked along all the sequence. PTP: Percentage of the tracked points correctly. SR: Sum of normalised residues on the final image. Sequence FP.*

Figure 6: *Examples of frames from the sequence FG.*

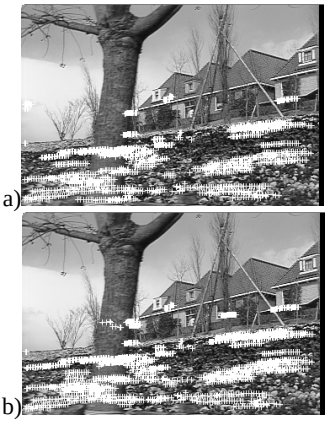

Figure 8: *a) Tracking trajectories of the first classified interest points based on the eigenvalues sorting. b) Selected points tracking trajectories. Sequence FG.*

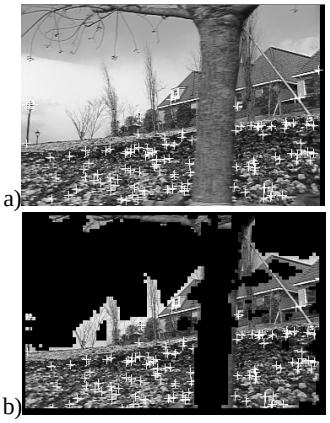

Figure 7: *a) Result of the detection of the interest points using eigenvalues criterion. b) Result of selection of the interest points using entropy criterion. Sequence FG.*

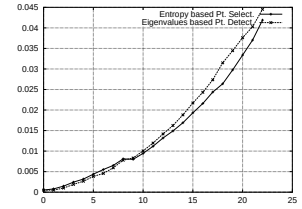

Figure 9: *Evolution of global residues vs. frame number. Sequence FG.*

Meth.	NbIni	NbSel	NbTrk	PTP	SR
EV	116	109	36	33	0.0445
ENT	116	109	48	44	0.0418

Table 2: *NbIni: The number of points detected in the initial image. NbSel: The number of selected points. NbTrk: The number of points tracked along all the sequence. PTP: Percentage of the points correctly tracked. SR: Sum of normalised residues on the final image. Sequence FG.*