Probabilistic Recognition of Activity using Local Appearance

Olivier Chomat and James L. Crowley
Project PRIMA - Lab GRAVIR - IMAG
INRIA Rhône-Alpes, 655, avenue de l'Europe
38330 - Montbonnot - FRANCE
Olivier.Chomat@inrialpes.fr

To be published
Computer Vision and Pattern Recognition (CVPR'99)
Fort Collins, Colorado, USA, June 23–25, 1999

April 2, 1999

Abstract

This paper addresses the problem of probabilistic recognition of activities from local spatio-temporal appearance. Joint statistics of space-time filters are employed to define histograms which characterize the activities to be recognized. These histograms provide the joint probability density functions required for recognition using Bayes rule. The result is a technique for recognition of activities which is robust to partial occlusions as well as changes in illumination.

In this paper the framework and background for this approach is first described. Then the family of spatio-temporal receptive fields used for characterizing activities is presented. This is followed by a review of probabilistic recognition of patterns from joint statistics of receptive field responses. The approach is validated with the results of experiments in the discrimination of persons walking in different directions, and the recognition of a simple set of hand gestures in an augmented reality scenario.

1 Introduction

The appearance of an object is the composition of all images of the object observed under different viewing conditions, illuminations, and object deformations. This paper addresses the problem of extending the previous appearance definition to the temporal dimension for the recognition of activity patterns.

Adelson and Bergen [3] define the appearance space of images for a given scene as a 7 dimensional local function $I\,(x, y, \lambda, t, V_x, V_y, V_z)$, whose dimensions are viewing position (V_x, V_y, V_z), time instant (t), position (x, y), and wavelength (λ). They have given this function the name "plenoptic function" from the Latin roots *plenus*, full, and

opticus, to see. The appearance of a scene can be represented as a discrete sampling of the plenoptic function.

Murase and Nayar [8] have demonstrated that the appearance of an object, seen from different viewing angles, can be represented as a continuous surface in a linear subspace obtained by projecting images onto an orthogonal basis determined by principal components analysis. Black [4] has extended this idea to describing deformable objects. However, in both cases, these techniques are applied globally to the entire image and thus suffer from a requirement to segment the region of an image covered by an object from its background and to normalize the size and intensity of the object. Such segmentation is generally unsolvable, and normalization of size can be a source of instability.

Schiele [12] and more recently Colin de Verdière [6] have shown that the problems of segmentation and normalization can be avoided by using sets of local receptive fields. In Schiele's work, joint statistics (multi-dimensional histograms) based on local appearance are used for probabilistic recognition of objects from an image region. Schiele's technique is robust to occlusion and can easily be made independent of viewing position and illumination. Colin de Verdière has shown that the vectors of receptive field responses form a manifold in a hyper-dimensional space, called a "local appearance space". This manifold can be discretely sampled to permit recognition from small neighborhoods by a process which is equivalent to table lookup.

The work described in this paper extends Schiele's result recognition of static objects to the recognition of local spatio-temporal patterns in order to characterize activities.

1.1 Local appearance description

Adelson and Bergen [3] propose to use low order derivatives operators as 2-d receptive fields to analyze the plenoptic function. However, the technique which they describe was restricted to derivatives of order one and two, and does not include measurements involving derivatives along three or more dimensions of the plenoptic function. It appears that the authors did not follow up on their idea and that little or no experimental work was published on this approach.

Schiele [12], and Colin de Verdière [6] use techniques based onto the characterization of the local appearance of static objects for recognition. Those techniques can be related to an efficient description of a plenoptic function $I\left(x, y, \lambda, V_x, V_y, V_z\right)$ taking into account more than two plenoptic dimensions for its description. Colin de Verdière represents appearance as a discrete sampling of a manifold parameterized by object orientation, and viewing position. The extension of such a structural approach for activity analysis poses difficulties because of the complexity of object deformations in space and time. These difficulties can be avoid by using a probabilistic representation of the plenoptic function. Our work is inspired by the techniques developed by Schiele [12] for object recognition using multi-dimensional histograms of filters responses along the spatial dimensions of the plenoptic function. We explore the extension of this technique to the recognition of moving patterns from the statistics of spatio-temporal receptive fields. Joint statistics are used to characterize the spatio-temporal appearance signature of an activity.

1.2 Problem definition

Consider the plenoptic function $I(x,y,t)$ constrained to a gray channel and a fixed view position. Let be $\vec{w}(x,y,t)$ a spatio-temporal neighborhood. $\vec{w}(x,y,t)$ can be viewed as a point in a space where each element of the window $\vec{w}$ is a dimension. This space is called the (local)appearance space. The large number of dimensions does not allow an exhaustive description of object appearance, but quantifying the local appearance of $I(x,y,t)$ using spatio-temporal receptive fields enables its analysis. Receptive fields responses describe an appearance subspace of which each dimension is a receptive field. The main problem is to design a minimum number of receptive fields sensitive to motion, and allowing an optimal description of motion appearance. Note that the approach could be extended to more plenoptic dimensions.

In this paper, a general scheme for the recognition of moving object activities is presented, thus without reconstruction of the motion field. Motion energy models are used as receptive fields to capture the local spatio-temporal appearance of activities. A statistical multi-dimensional analysis is processed to provide recognition using Bayes rule.

2 Motion energy receptive fields

The properties of spatio-temporal filters are studied for recognition of activities in a context of analysis of visual motion information. Consider a space-time image $I(\vec{p})$, and its Fourier Transform $\hat{I}(\vec{q})$ with $\vec{p} = (x,y,t)$ and $\vec{q} = (u,v,w)$. Let r_x and r_y be respectively the speed of horizontal and vertical motion. The Fourier transform of the moving image $I(x - r_x t, y - r_y t, t)$ is $\hat{I}(u,v,w + r_x u + r_y v)$. This means that spatial frequencies are not changed, but all temporal frequencies are shifted by minus the product of the speed and the spatial frequencies. Motion energy receptive fields are designed taking into account that at a given spatio-temporal frequency an energy measure depends on both the velocity and the contrast of the input signal.

While the approach described below is largely inspired by motion estimations techniques based on filters, this approach does not require explicit estimation of the flow field. A set of motion energy receptive fields are designed in order to sample the power spectrum of the moving texture [7]. Their structure relates to the spatio-temporal energy models of Adelson and Bergen [3], and Heeger [7].

2.1 Spatio-temporal energy models

Several authors have proposed physiologically-based models for the analysis of image motion. One popular set of models are spatio-temporal energy models [3] [7] where motion energy measures are computed from the sum of the square of even (G_{even}) and odd-symmetric (G_{odd}) oriented spatio-temporal sub-band filters tuned for the same orientation in order to be phase independent:

$$H(\vec{p}) = (I(\vec{p}) * G_{even})^2 + (I(\vec{p}) * G_{odd})^2 \quad (1)$$

Adelson and Bergen [1] suggested that these energy outputs should be combined in

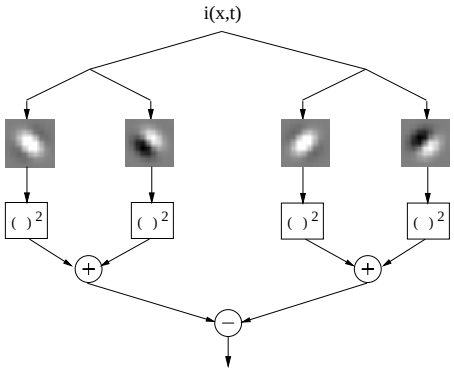


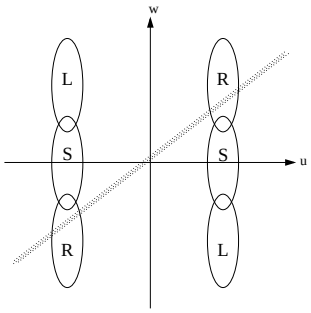
Figure 1: **An example of spatio-temporal energy model applied onto a 2-d signal $I(x,t)$ where x is the spatial dimension, and t the temporal one. By squaring and summing the responses of a quadrature pair of units tuned for the same orientation, the resulting signal gives a phase independent measure of local energy within a given spatio-temporal frequency band. Leftward and rightward motion detector are combined in opponent fashion.**

opponent fashion, subtracting the output of a mechanism tuned for leftward motion from one tuned for rightward motion. An example of spatio-temporal energy model applied to a 2-d signal is shown in figure 1. The output of such filters depends on both the velocity and the local spatial-content of the input signal $I(\vec{p})$. The extraction of velocity information within a spatial frequency band involves normalizing the energy of the filter outputs according to the response of a static energy filter tuned to the same spatial orientation and null temporal orientation:

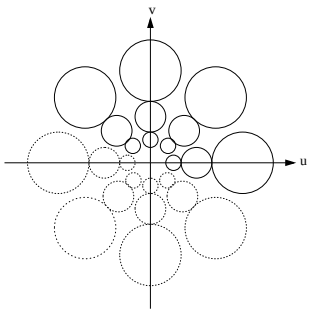
$$w(\vec{p}) = \frac{H_{Right}(\vec{p}) - H_{Left}(\vec{p})}{H_{Static}(\vec{p})} \tag{2}$$

A triad of rightward, leftward and static energy filters is shown in figure 2. Such a spatio-temporal energy model allows the exploitation of low level visual motion information. Using such filters, Adelson and Bergen [2], and, Simoncelli and Adelson [13] provide an interpretation of the standard gradient solution in terms of opponent spatio-temporal energy mechanism for motion estimation. Also Heeger [7] and Spinei and al. [14] propose energy-based techniques for motion estimation. In this paper only the visual motion information is used for action recognition. Extension to a measure of motion field is discuss in the prospective section.

Let us define a motion energy receptive field as a unit composed of 6 filters tuned to the same spatial orientation. Those 6 filters are divided into 3 pairs of quadratic filters. One pair tuned leftward, one rightward and one static. Filters of each pair have the same spatio-temporal frequency sub band.



(figure 2)



(figure 3)

Figure 2: **The spectrum of a moving pattern lies on a plane in the spatio-temporal frequency domain. At a given spatio-temporal frequency, the energy measure depends on both the velocity and the contrast of the input signal $I(x, y, t)$. The responses for rightward (R), leftward (L) and static (S) units are shown for a given spatial band in the frequency domain (u, w) where u are the spatial frequencies and w the temporal ones. Velocity information is extracted comparing the output of a set of spatio-temporal energy filters to a static energy filter in the same spatial frequency band.**

Figure 3: **Map of the spatial bandwidths of a set of 12 motion energy receptive fields in the spatial frequency domain (u, v). There is 4 different orientations and 3 different scales.**

2.2 A family of Gabor filters

Gabor filters with various frequencies and orientations, are organized to sample an image sequence into bandpass energy channels. The sum of the squared output of a sine-phase Gabor filter plus the squared output of a cosine-phase Gabor filter gives a measure of Gabor energy that is invariant to the phase of the signal. The power spectrum of a Gabor energy filter is the sum of a pair of Gaussians centered at $\vec{q_0}$ and $-\vec{q_0}$ in the frequency domain.

A set of 12 motion energy receptive fields are used, corresponding to 4 spatial orientations and 3 range of motions. All filters are tuned for the same temporal frequency $w_0 = \frac{1}{4}$ cycles per frame and the same temporal scale $\sigma_t = 1.49$. All of the results presented in this paper were produced with a spatial frequency tuning of each Gabor filter as $\sqrt{u_0^2 + v_0^2} = \frac{1}{4}$ cycles per pixel and a standard spatial deviation of $\sigma_x = \sigma_y = 1.49$ corresponding to a bandwidth of 0.25. The 4 spatial orientations are 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$ and $\frac{3 \cdot \pi}{4}$. Additional scales are obtained computing a local Gaussian pyramid and convolving with a single family of filter at each level. This is equivalent using families of filters spaced one octave apart in spatial frequency and with a standard deviation which is twice largest. Figure 3 shows a map of the receptive fields' spatial bandwidths. The set of motion energy receptive fields allows the description of the spatio-temporal appearance of activity.

3 Probability density of activities

The outputs from the set of spatio-temporal filters provide a vector of measurements at each pixel. The joint statistics of these vectors allow the probabilistic recognition of activity. A multi-dimensional histogram is computed from the outputs of the filter bank. These histograms can be seen as a form of activity signature and provide an estimate of the probability density function for use with Bayes rule.

Models for the appearance of activities are trained from a large set of training sequences. For each class of activity, a multi-dimensional histogram is computed by applying the filter bank to the image sequences. Probabilistic recognition of action a_k is achieved considering the vector of local measures $\vec{w}(\vec{p})$, which elements i are motion energy measures $w_i(\vec{p})$ tuned for different sub-bands. The probability $p(a_k|\vec{w})$ is computed using the Bayes rule:

$$p(a_k|\vec{w}) = \frac{p(\vec{w}|a_k)\, p(a_k)}{p(\vec{w})} = \frac{p(\vec{w}|a_k)\, p(a_k)}{\sum_l p(\vec{w}|a_l)\, p(a_l)} \tag{3}$$

where $p(a_k)$ is the a priori probability of action a_k, $p(\vec{w})$ is the a priori probability of the vector of local measures $\vec{w}$, and $p(\vec{w}|a_k)$ the probability density of action a_k. The appearance subspace is a 12-d space. The main problem is the computation of an histogram over such a large space. An extension of the quad-tree technique is used to represent the histograms.

The probability $p(a_k|\vec{w})$ allows only a local decision at location $\vec{p} = (x, y, t)$. The final result at a given time (t) is the map of the conditional probabilities that each pixel belongs to an activity of the training set based on its space-time neighborhood. For

the moment the results are presented taking a decision according to the spatial average probability over a given frame. A more reliable recognition scheme could be done using $p\left(a_k|\vec{w}\right)$ as input of Hidden Markov Models.

4 Human actions recognition

This section presents experimental results in the recognition of human actions such as gestures, and full body movements. The context is computer vision understanding of hand and body gestures for wireless interfaces and interactive environments. We demonstrate our technique using a scenario from an augmented reality tool for collaborative work [5], as well as recognition of full body movements in a context of video-surveillance, and results on the recognition of gesture commands.

4.1 Recognition of full body movements

The test sequences are composed of a walking person whose actions are to walk from the background to the foreground (sequence "*Come*"), to walk from the foreground to the background (sequence "*Go*"), to walk from the right to the left (sequence "*Left*") and to walk from the left to the right (sequence "*Right*"). Figure 4 shows extracts of

Figure 4: **Extracts of the training walking man sequences. Related actions are respectively from the left to the right "*Come*", '*Go*", "*Left*" and "*Right*". Images are 192×144 pixels per pixels, and the acquisition rate is 15 Hz.**

the training sequences used for the computation of the probability density $p\left(\vec{w}|a_k\right)$.

As a first experiment the recognition scheme is applied on the training sequence. Examples of the resulting maps of the local probabilities $p\left(a_k|\vec{w}\right)$ computed at a given time (t) are shown in figure 5. In figure 6 the spatial average per frame of $p\left(a_k|\vec{w}\right)$ is plotted for each of the trained actions a_k. The recognition is provided by the maximum of the output probabilities.

As a second experiment, recognition is processed over a new sequence of somebody else performing the same actions as in the training sequences.

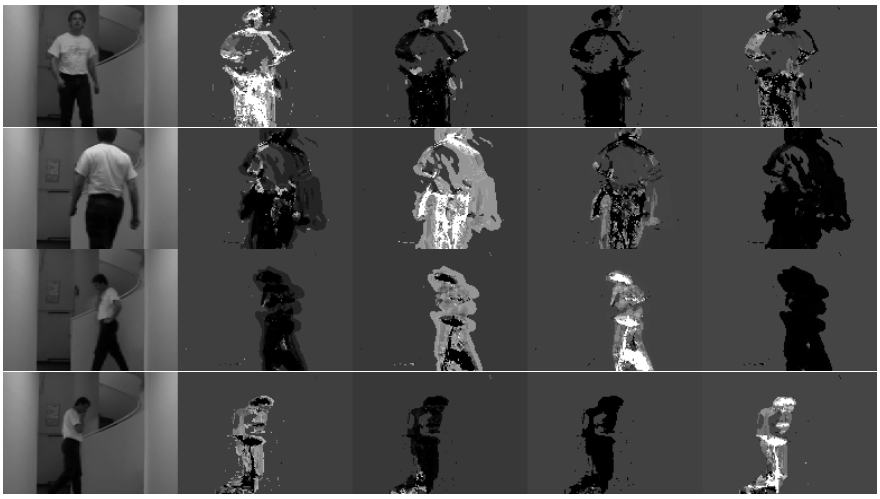


Figure 5: **Examples of resulting maps of the local probabilities $p(a_k|\vec{w})$ computed over extracts of the training sequences. The original images are shown on the first column. Following columns relate respectively on maps of $p(a_k|\vec{w})$ for action a_k equals to "*Come*", "*Go*", "*Left*" and "*Right*".**

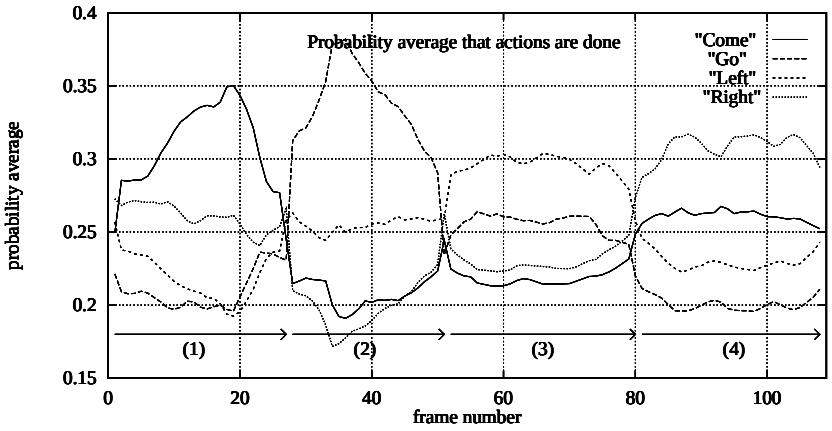


Figure 6: **Spatial average per frame of the local probabilities $p(a_k|\vec{w})$. The analyzed sequences 1, 2, 3 and 4 are the training sequences "*Come*", '*Go*', "*Left*" and "*Right*". The recognition is processed successfully at a given time according to the maximum of the output probabilities.**

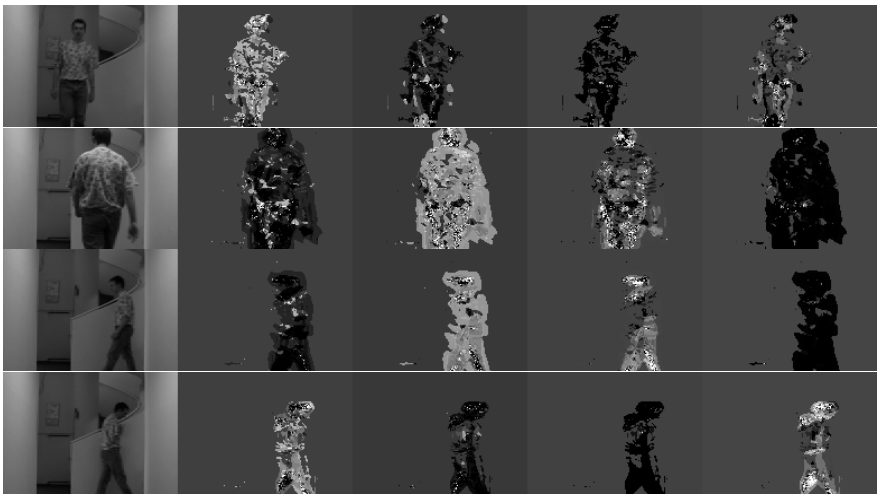


Figure 7: **Examples of resulting maps of the local probabilities $p\left(a_k|\vec{w}\right)$ computed over extracts of new sequences at a given time (t). The original images are shown on the first column. Following columns relates respectively on maps of $p\left(a_k|\vec{w}\right)$ for action a_k equals to "*Come*", "*Go*", "*Left*" and "*Right*".**

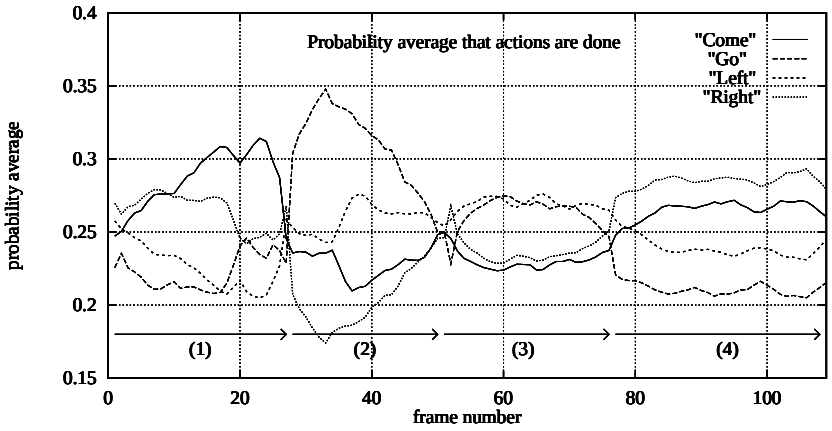


Figure 8: **Spatial average per frame of $p\left(a_k|\vec{w}\right)$. The analyzed sequences 1, 2, 3 and 4 are new sequences performing the actions "*Come*", '*Go*", "*Left*" and "*Right*". Recognition is processed successfully for 1 and 2. The recognition of 3 and 4 is less prominent.**

The resulting probability maps are shown in figure 7, and the spatial average of probabilities is plotted in function of the frame number in figure 8. Recognition is processed successfully for the actions “Come” and “Go”, but the recognition of actions “Left” and “Right” is less prominent due probably to the fact that the two walking figures do not perform exactly the same displacement at the same speed. The problem is all the more difficult since the action “Come”(“Go”) is composed of “Right”(“Left”) action.

Experiments with more selective filters in the temporal dimension are expected to provide improved results. Also the training basis is poor in the sense that the histograms were computed using only one person performing the walking actions. It must be better to learn with several people performing the same actions.

4.2 Gesture recognition

The gesture sequences tested in this experiment are useful for gesture based interactions, such as with a digital-desk [11] where a controlled camera is looking for hand commands and a projector displays feedback onto a desk surface. A set of 4 gesture commands are studied: to “*Rub out*”, to “*Circle*”, to “*Zoom out*” and to “*Zoom in*”. Extracts of the training sequences are shown in figure 9.

Figure 9: **Extracts of the original command gestures sequences. From the left to the right of the figure, the commands “*Rub out*”, “*Circle*”, “*Zoom out*” and “*Zoom in*” are shown. Images are 92×72 pixels per pixels, and the acquisition rate is 10 Hz.**

Only the results on recognition of actions extracted from the training sequences are presented. In figure 10 the resulting maps of the local probabilities are shown. The spatial averages per frame of $p(a_k/\vec{w})$ are plotted in figure 11.

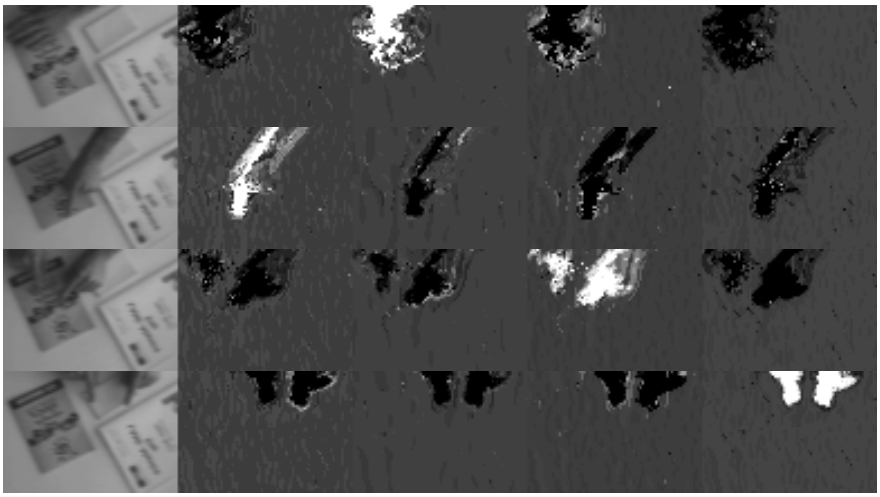


Figure 10: **Examples of resulting maps of the local probabilities $p(a_k|\vec{w})$ computed over extracts of the training sequences at a given time (t). Each row deals, respectively, with the actions "*Rub out*", '*Circle*", "*Zoom out*" and "*Zoom in*" to be analyzed. The original images are shown on the first column. Following columns relates respectively on maps of $p(a_k|\vec{w})$ for action a_k equals to "*Circle*", "*Rub out*", "*Zoom out*" and "*Zoom in*".**

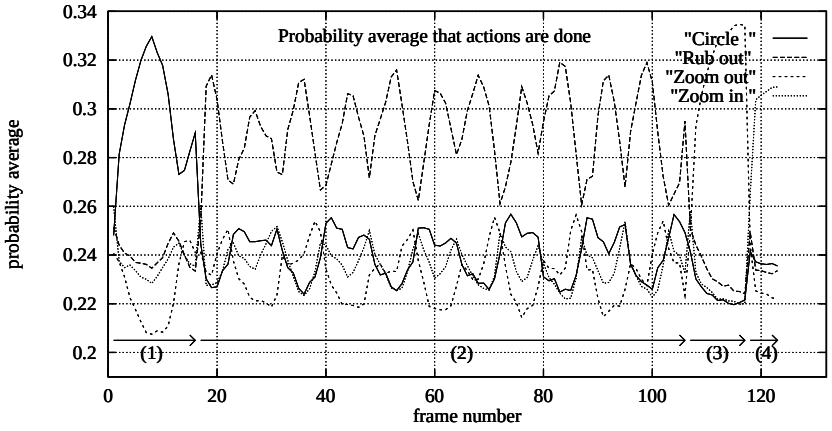


Figure 11: **Spatial average per frame of the local probabilities $p(a_k|\vec{w})$. The analyzed sequences 1, 2, 3 and 4 are the training sequences "*Circle*", '*Rub out*", "*Zoom out*" and "*Zoom in*". The recognition is processed successfully according to the maximum of the output probabilities.**

5 Conclusions and prospectives

The visual recognition of human action has many potential applications in man-machine interaction, inter-personal communication and visual surveillance [5]. A new approach for activity recognition has been presented. Recognition is processed statistically according to the conditional probability that a measure of the local spatio-temporal appearance is occurring for a given action.

The outputs of spatio-temporal Gabor energy filters give measures of the local spatio-temporal appearance. The normalization according to the local static energy leads to a measure of motion information. Multi-dimensional histograms of these measures are used to estimate the probability density of an action. The main advantage of Gabor filters is that they can be built from separable and recursive components increasing the efficiency of the computation. On the other hand Gabor filters are not causal and it may be important for some applications to eliminate delay using filters with a causal temporal response. Alternatively Gaussian derivatives can be used, thus giving an interpretation of the standard gradient equation [9] [13].

This paper describes work in progress and experimental results are limited but encouraging. Further experiments will attempt to quantify the limits of the technique. Also several technical details must be resolved to provided improved results. On one hand the vector of receptive fields responses is sensitive simultaneously to three motion ranges. Space and time scales have been selected to ensure large bandwidth. Heeger [7] and Spinei [14] use more selective filters in space with an optimal ratio between space and time scales of $\sigma_{x,y} = 4\sigma_t$. To make up for it the robustness to scale changes is lesser. A solution is to select automatically local scale parameters according to the maxima over scales of normalized derivatives (see [10]). On the other hand the global decision scheme for recognition is quite simple, corresponding to the average of local probabilities over a frame. A more complex global decision scheme like Hidden Markov Models could be more efficient.

References

[1] E.H. Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Optical Society of America*, 2(2):284–299, 1985.

[2] E.H. Adelson and J.R. Bergen. The extraction of spatio-temporal energy in human and machine vision. In *Workshop on Motion: Representation and Analysis*, pages 151–155, 1986.

[3] E.H. Adelson and J.R. Bergen. *Computational Models of Visual Processing*, chapter The Plenoptic function and the elements of early vision. M.Landy and J.A.Movshons, Cambridge, 1991. MIT Press.

[4] M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *European Conference on Computer Vision*, pages 329–342, 1996.

[5] J. Coutaz, F. Bérard, E. Carraux, and J.L. Crowley. Early experience with the mediaspace comedi. In *IFIP Working Conference on Engineering for Human Computer Interaction*, 1998.

[6] V. Colin de Verdière and J.L. Crowley. Visual recognition using local appearance. In *European Conference on Computer Vision*, pages 640–654, 1998.

[7] D.J. Heeger. Optical flow using spatio-temporal filters. *International Journal of Computer Vision*, pages 279–302, 1988.

[8] H.Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[9] B.K.P. Horn and B.G.Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[10] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[11] W. Newman and P. Wellner. A desk supporting computer-based interaction with paper documents. In *ACM Conference on Human Factors in Computing Systems*, pages 587–592, 1992.

[12] B. Schiele and J.L. Crowley. Probabilistic object recognition using multi-dimensional receptive field histograms. In *International Conference Pattern Recognition*, Vienna, 1996.

[13] E.P. Simoncelli and E.H. Adelson. Computing optical flow distributions using spatio-temporal filters. Technical Report 165, M.I.T. Media Lab Vision and Modeling, 1991.

[14] A. Spinei, D. Pellerin, and J. Herault. Spatio-temporal energy-based method for velocity estimation. *Signal Processing*, 65:347–362, 1998.