A Probabilistic Sensor for the Perception and the Recognition of Activities

Olivier Chomat, Jérôme Martin, and James L. Crowley

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

Abstract. This paper presents a new technique for the perception and recognition of activities using statistical descriptions of their spatio-temporal properties. A set of motion energy receptive fields is designed in order to sample the power spectrum of a moving texture. Their structure relates to the spatio-temporal energy models of Adelson and Bergen where measures of local visual motion information are extracted by comparing the outputs of a triad of Gabor energy filters. Then the probability density function required for Bayes rule is estimated for each class of activity by computing multi-dimensional histograms from the outputs from the set of receptive fields. The perception of activities is achieved according to Bayes rule. The result at each instant of time is the map of the conditional probabilities that each pixel belongs to each one of the activities of the training set. Since activities are perceived over a short integration time, a temporal analysis of outputs is done using Hidden Markov Models.
The approach is validated with experiments in the perception and recognition of activities of people walking in visual surveillance scenari. The presented work is in progress and preliminary results are encouraging, since recognition is robust to variations in illumination conditions, to partial occlusions and to changes in texture. It is shown that it constitute a powerful early vision tool for human behaviors analysis for smart-environnements.

1 Introduction

The use of computer vision for recognition of activities has many potential applications in man-machine interaction, inter-personal communication and visual surveillance. Considering several classes of body actions, the machine would be able to react to some command gestures. Such techniques support applications such as video-conferencing, tele-teaching and virtual reality environments, where the user is not confined to the desktop but is able to move around freely. The aim of the research described in this paper is the characterization for recognition of human actions such as gestures, or full body movements.

Analyzing the motion of deformable objects from image sequences is a challenging problem for computer vision. Different approaches have been proposed

D. Vernon (Ed.): ECCV 2000, LNCS 1842, pp. 487–503, 2000.

for this task. With the exception of algorithms whose aim is to determine the 3D motion of objects, two trends emerge: the techniques using 2D geometric model of the objects, and appearance based methods.

In the case of human activity analysis a central problem results from the fact that the human body consists of body parts linked to each other. A possible approach is to model the articulated structure of the human body. For example, in the system Pfinder [WADP96] the human body is modeled as a connected set of blobs, each blob having a spatial and color distribution. An other example is the Cardboard person model of Ju, Black and Yacoob [YB98], where human limbs are represented as a set of connected planar patches. Some of these techniques using human body models assumes that a two-dimensional reconstruction precedes the recognition of action. In either case, the methods require algorithms with relatively high computational costs whose robustness and stability are difficult to analyze.

An alternative to geometric and kinematic modeling is to employ an image based description that captures the appearance of the motion. Davis and Bobick [BD96] have defined a representation of action in terms of Motion History Image (MHI). A MHI is a scalar-valued image where intensity is a function of recency of motion. An appearance-based technique is used to match temporal templates, computing statistical descriptions of MHI with Hu moments. The Motion History Image acts as a local low pass temporal filter, and the distribution over space of the filter output is used for recognition. The MHI of Davis and Bobick is based on local motion appearance. In this case, appearance of motion is defined as the temporal memory of motion occurrences, but no grey level information is used.

Another alternative is the use of object appearance based on contour analysis or active contours. Such an approach has been used by Cootes and al. in [CTL$^+$93] where flexible shape templates are fitted to data according to a statistical model of grey level information around model points. In a more recent work Active Appearance Model (A.A.M.) are used, modeling the object shape and gray level appearance. Baumberg and Hogg [BH95] use a Point Distribution Models of the shapes of walking pedestrians. The main characteristics of the body shape deformations are captured by a Principal Component Analysis of these point sets. This approach is robust to occlusion, but it requires a background segmentation to allow the extraction of the boundary of the pedestrian.

The approach described in this paper is related to the MHI of Davis and Bobick, as a local appearance based method. It is influenced by the work of Murase and Nayar [MN95], where the set of appearances of objects is expressed as a trajectory in a principal component space. It is also inspired by Black and Jepson [BJ96] who extended a global P.C.A. approach to track articulated object in a principal component space. All of these approaches derived a space by performing principal component analysis (P.C.A.) on an entire image. Such global approaches are sensitive to partial occlusions as well as to the intensity and shape of background regions. These problems can be avoided by using methods based on local appearance [Sch97,CC98]. In such an approach, the appearance of neighborhoods is described with receptive fields. Schiele [Sch97] and Colin de Verdière [CC98] define an orthonormal space for expressing local appearance. This space is based on Gaussian derivatives or it is computed via P.C.A. over the set of windows of all the images of the training data. In this space of receptive fields an image is modeled as a manifold. Colin de Verdière achieved the recognition by measuring distance between the vector of receptive fields responses of an observed window and the surface points from a discrete sampling of the manifold. Whereas Schiele has developed a statistical approach using multidimensional histograms of the responses of vectors of receptive fields.

In this paper the appearance of human motion is described using the appearance of small spatio-temporal neighborhoods over a set of sequences, and a statistic approach is used to achieved recognition of activity patterns. The next section of this paper deals with the approach which is used for describing spatio-temporal structures. A synopsis of the local visual motion information is obtained by signal decomposition onto a set of oriented motion energy receptive fields. Section 3 provides the description of a probabilistic framework for analyzing the receptive fields responses. Multi-dimensional histograms are computed to characterize each class of activity. Section 4 shows results from the perception of human activities in the context of computer assisted visual surveillance. Since humans are perceived as deformable moving objects, the challenge is to discriminate different classes of human activities. The output of the probabilistic sensor are maps of the probability that each pixel belong to each one of the trained classes of activities. Recognition of activities elements is done by selecting best local probabilities. Since the temporal aperture window of description is relatively small compared to the temporal duration of activity, Hidden Markov Models (HMM) are employed to recognize the complete activity In a sense, the HMM provides context. That is the purpose of section 5. The last section presents discussions and perspectives.

2 Describing Spatio-Temporal Structures

Adelson and Bergen [AB91] define the appearance space of images for a given scene as a 7 dimensional local function, whose dimensions are viewing position, time instant, position in the image, and wavelength. They have given this function the name "plenoptic function" from the Latin roots *plenus*, full, and *opticus*, to see. Adelson and Bergen propose to detect local changes along one or more plenoptic dimensions and to represent the structure of the visual information in a table of the detectors responses, comparing them two by two. The two dimensions of the table are simple visual detectors such as derivatives and the table contents are possible visual elements. Adelson and Bergen use low order derivatives operators as 2-D receptive fields to analyze the plenoptic function. However, the technique which they describe is 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.

Nevertheless the plenoptic function provides a powerful formalism for the measurement of specific local structures, including spatio-temporal patterns. This paper employs with such framework to describe activity patterns. Activity patterns are characterized by describing their local visual information using a set of spatio-temporal receptive fields, and by statistically modeling the descriptors responses. The result is a software sensor able to discriminate different patterns of activities.

2.1 Using Receptive Fields

The notion of receptive field in vision is stemed from studies on the description of cortex visual cells. Those studies attempt to understand the biological visual system to reach its performance for extracting local information measures.

Classically receptive fields structure relates from signal decomposition techniques. The two most widely used approaches for signal decomposition are the Taylor expansion and the Fourier transform The Taylor series expansion gives a local signal description in the spatial dimension, while the Fourier transform provides a description in the spectral domain. These two methods for signal decomposition correspond respectively to the projection of the signal onto a basis of functions with amplitude modulation and onto a basis of functions which are frequency modulated. Other local decomposition bases are also possible. A decomposition basis is generally chosen to suit the problem to be solved. For example, a frequency-based analysis is more suitable for texture analysis, or a fractal-based description for natural scene analysis. Independently from the basis choice, the description is done over an estimation support relative to the locality of the analysis. The next section formulates the derivative operator of the Taylor expansion and the spectral operator of the Fourier transform as generic operators.

2.2 Generic Neighborhood Operators

The concept of linear neighborhood operators was redefined by Koenderink and Doorn [Kv92] as generic neighborhood operators. Typically operators are required at different scales corresponding to different sizes of estimation support. Authors have motivated their method by rewriting neighborhood operators as the product of an aperture function, $A\left(\boldsymbol{p},\sigma\right)$, and a scale equivariant function, $\phi\left(\boldsymbol{p}/\sigma\right)$:

$$G\left(\boldsymbol{p}\right) = A\left(\boldsymbol{p},\sigma\right)\phi\left(\boldsymbol{p}/\sigma\right) \tag{1}$$

The aperture function takes a local estimation at location $\boldsymbol{p}$ of the plenoptic function which is a weighted average over a support proportional to its scale parameter, σ. An aperture function is the Gaussian kernel as it satisfies the diffusion equation:

$$A\left(\boldsymbol{p},\sigma\right) = \frac{e^{-\frac{1}{2}\frac{\boldsymbol{p}\cdot\boldsymbol{p}}{\sigma^2}}}{\left(\sqrt{2\pi}\sigma^D\right)} \tag{2}$$

The function $\phi\left(\boldsymbol{p}/\sigma\right)$ is a specific point operator relative to the decomposition basis. In the case of the Taylor expansion $\phi\left(\boldsymbol{p}/\sigma\right)$ is the n^{th} Hermite polynomials:

$$\phi\left(\boldsymbol{p}/\sigma\right) = \left(-1\right)^{n} He_{n}\left(\boldsymbol{p}/\sigma\right) \tag{3}$$

In the case of the Fourier series $\phi\left(\boldsymbol{p}/\sigma\right)$ are the complex frequency modulation functions tuned to selected frequencies, $\boldsymbol{\nu}$:

$$\phi\left(\boldsymbol{p}/\sigma\right) = e^{2\pi j \boldsymbol{\nu}\cdot\boldsymbol{p}/\sigma} \tag{4}$$

Within the context of spatial, respectively spectral, signal decomposition the generic neighborhood operators are scale normalized Gaussian derivatives [Lin98], and respectively scale normalized Gabor filters.

2.3 Motion Energy Receptive Fields

The perception of activities involves extraction of local visual motion information. Techniques which reconstruct explicitly the optical flow are often complex and specific to the analyzed scene all the more so since that there are not well suited for describing the motion of moving deformable objects. The extraction of low level motion information involves the use of a decomposition basis sensitive to motion like signal decomposition using combination of Gaussian derivatives or Gabor filters.

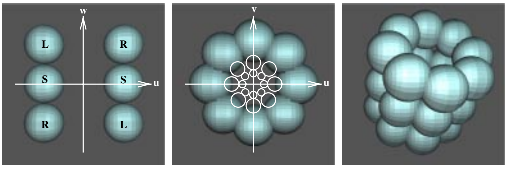
A measure of motion information rich enough to describe activities is easily obtained in the spectral domain, since an energy measure depends on both the velocity and the contrast of the input signal at a given spatio-temporal frequency. Consider a space-time image, $I\left(\boldsymbol{p}\right)$, and its Fourier Transform, $\hat{I}\left(\boldsymbol{q}\right)$, with $\boldsymbol{p} = (x, y, t)$ and $\boldsymbol{q} = (u, v, w)$. Let r_x and r_y be respectively the speed of horizontal and vertical motion. The Fourier transform of the moving image, $I\left(x - r_x t, y - r_y t, t\right)$, is $\hat{I}\left(u, v, w + r_x u + r_y v\right)$. This means that spatial frequencies are not changed, but all the temporal frequencies are shifted by minus the product of the speed and the spatial frequencies. A set of Gabor based motion energy receptive fields is used to sample the power spectrum of the moving texture. Their structure relates to the spatio-temporal energy models of Adelson and Bergen [AB91], and Heeler [Hee88]. Motion energy measures are computed from the sum of the square of even (G_{even}) and odd-symmetric (G_{odd}) oriented spatio-temporal Gabor filters which have been tuned for the same orientation, thus in order to be phase independent:

$$H\left(\boldsymbol{p}\right) = \left(I\left(\boldsymbol{p}\right) * G_{even}\right)^{2} + \left(I\left(\boldsymbol{p}\right) * G_{odd}\right)^{2} \tag{5}$$

Adelson and Bergen [AB85] suggested that these energy outputs should be combined in opponent fashion, subtracting the output of a mechanism tuned for leftward motion from one tuned for rightward motion. The output of such filters depends on both the velocity and the local spatial-content of the input signal, $I\left(\boldsymbol{p}\right)$. 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:

$$w\left(\boldsymbol{p}\right)=\frac{H_{Right}\left(\boldsymbol{p}\right)-H_{Left}\left(\boldsymbol{p}\right)}{H_{Static}\left(\boldsymbol{p}\right)} \tag{6}$$

A triad of rightward, leftward and static Gabor energy filters is shown in part (a) of figure 1. Such a spatio-temporal energy model allows the measurement of low level visual motion information. A set of 12 motion energy receptive fields are used, corresponding to 4 spatial orientations and 3 ranges of motions. This set of motion energy receptive fields allows the description of the spatio-temporal appearance of activity.



(a) (b) (c)

Fig. 1. *Bandwidths of spatio-temporal receptive field triads. Figure (a) represents responses for rightward (R), leftward (L) and static (S) units for a given spatial band in the frequency domain (u, w) where u are the spatial frequencies and w the temporal ones. Figure (b) is a 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. And figure (c) is a 3D view of a set 4 motion energy receptive fields corresponding to 4 orientations and 1 scale.*

Note that the optical flow is not reconstructed explicitly but a vector of measures, $\boldsymbol{w}\left(\boldsymbol{p}\right)$, is obtained, where the elements $w_i\left(\boldsymbol{p}\right)$ of $\boldsymbol{w}\left(\boldsymbol{p}\right)$ are motion energy measures tuned for different sub-bands. The combination of the 12 motion energy receptive fields can lead to a motion estimate. Heeger [Hee88] use a numerical optimization procedure to find the plane that best accounted for the measurements (the error criterion is least-squares regression on the filter energies). Spinei and al. [SPH98] make the response of a triad of Gabor energy filters $w\left(\boldsymbol{p}\right)$ proportional to motion using a non-linear combination of the response of the Gabor filters. Than he merges the estimated motion components corresponding to different orientations and scales. But we insist on that optical flow estimation is not the purpose of the proposed approach since we are motivated by signal decomposition. Low level motion information is extracted using a set of motion energy receptive fields based on Gabor energy filters. The outputs from the set of receptive fields provide a vector of measurements, $\boldsymbol{w}\left(\boldsymbol{p}\right)$ giving a synopsis of the local visual motion information.

3 Probabilistic Analysis of Feature Space

The outputs from the set of motion energy receptive fields provide a vector of measurements, $\boldsymbol{w}\left(\boldsymbol{p}\right)$, at each pixel. The joint statistics of these vectors allow the probabilistic perception of activity. A multi-dimensional histogram is computed from the outputs of the filter bank for each class of activity. 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.

3.1 Measurements Probability Density

For each class of activity a_k, a multi-dimensional histogram of vectors of measurements is computed. The histogram is an estimate of the density probability $p\left(\boldsymbol{w}|a_k\right)$ of action a_k. The subspace of receptive fields presents a large number of dimensions, which is 12D in the case of the basis of motion energy receptive fields defined previously. 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. Let be N the number of dimensions (e.d. number of motion energy receptive fields). A dichotomic tree is designed where each node expects 2^N potential branches corresponding to filled cells. Cells are sub-divided by 2 along each dimension. Among the 2^N resulting new cells, the filled cells are sub-divided themselves until the final resolution.

This algorithm allows the computation and the storage of high dimensional histograms which are quite sparse.

3.2 Probabilistic Perception of Activities

The probabilistic perception of action, a_k, is achieved considering the vector of local measures, $\boldsymbol{w}\left(\boldsymbol{p}\right)$, whose elements i are motion energy measures, $w_i\left(\boldsymbol{p}\right)$, tuned for different sub-bands. The probability, $p\left(a_k|\boldsymbol{w}\right)$, that the pixel $\boldsymbol{p}$ belongs to action a_k according to $\boldsymbol{w}\left(\boldsymbol{p}\right)$ is computed using Bayes rule:

$$p\left(a_k|\boldsymbol{w}\right) = \frac{p\left(\boldsymbol{w}|a_k\right)p\left(a_k\right)}{p\left(\boldsymbol{w}\right)} = \frac{p\left(\boldsymbol{w}|a_k\right)p\left(a_k\right)}{\sum_l p\left(\boldsymbol{w}|a_l\right)p\left(a_l\right)} \tag{7}$$

where $p\left(a_k\right)$ is the a priori probability of action a_k, $p\left(\boldsymbol{w}\right)$ is the a priori probability of the vector of local measures $\boldsymbol{w}$, and $p\left(\boldsymbol{w}|a_k\right)$ the probability density of action a_k. The probability $p\left(a_k\right)$ of action a_k is estimated according to the context. But without a priori knowledge, it is fixed to the maximum.

The probability, $p\left(a_k|\boldsymbol{w}\right)$, allows only a local decision at location $\boldsymbol{p} = (x, y, t)$. 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 appearance.

4 Application to the Perception of Human Activities

The vast amount of raw data generated by digital video units and their poor capacities to filter out useless information lead us to develop a framework for highlighting specific relevant events according to scene activities. Some examples of applications are assisted video-surveillance helping users concentrate their attention, or intelligent office environments understanding and reacting to the configuration of the scene. In this context the probabilistic framework was trained for the perception of human activities of an office fitted out with a camera for visual surveillance.

The wide angle camera allows the surveillance of the whole office. The analyzed activities are *"coming in"*, *"going out"*, *"sit down"*, *"wake up"*, *"dead"* (when somebody fall down), *"first left"*, *"first right"*, *"second left"*, *"second right"* and *"turn left"*, *"turn right"*. Those actions can take place anywhere in the scene and under any illumination conditions. A view of the scene and an example of the considered activities is shown in figure 2.

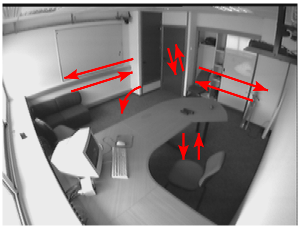


Fig. 2. *A view of the large visual angle camera. Examples of the analyzed activities are shown. Images are* 192×144 *pixels a per pixels and the acquisition rate is 10 Hz.*

4.1 Assumptions and Parameters

This section deals with the conditions of application to evaluate the probabilistic sensor ability to perceive the class of activities defined previously.

Global conditions: It is assumed that the camera is fixed, therefore there is no global motion to compensate. The changes in the scene illumination are uncontrolled and the static objects can move location. Images are 192×144 pixels a per pixels and the acquisition rate is 10 Hz.

Receptive fields parameters: All of the results presented in this paper were produced with a spatial frequency tuning for 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 using families of filters, which are spaced one octave apart in spatial frequency and with a standard spatial deviation which is twice largest. The sub-band-filters are tuned for the same temporal frequency $w_0 = \frac{1}{4}$ cycles per frame and the same temporal scale $\sigma_t = 1.49$.

Histograms computation: The histograms are computed by quantifying the receptive fields responses from 1 to 8 bits. Each class of activity is done between 5 times to 18 times corresponding to 5 people acting anywhere in the scene and depending from the activity rate in a scenario. Each sequence of an activity is between 11 to 44 frames long. The left hand side of table 1 resumes information for each class of activity histograms computation.

Table 1. *Informations on the sequences of each class of activity. The left hand part of the table deals with sequences used for histogram computation and right hand part of the table deals with test sequences.*

class of activity	number of sequences	number of frames per sequence	total number of frames	number of sequences	number of frames per sequence	total number of frames
in	5	26-35	151	18	21-26	423
out	5	30-36	166	18	18-26	394
sit	18	14-27	378	24	13-28	474
wake	18	13-25	320	36	11-33	759
dead	5	20-23	105	12	10-14	143
left1	4	31-41	146	15	10-35	370
right1	4	26-41	140	12	21-32	320
left2	4	29-44	143	28	6-32	513
right2	4	25-40	132	36	6-34	785
turn right	6	11-18	93	15	7-28	182
turn left	6	12-19	82	25	6-33	299

Perception: The perception of activities according to Bayes rule (equation 7) is weighted by the a priori probability $p(a_k)$ of action a_k. Without a priori knowledge the probability $p(a_k)$ is fixed to the maximum.

Test sequences: A set of test sequences are used to evaluate the sensitivity of the probabilistic sensor. Those test sequences are different from ones used to compute the multi-dimensional histograms. Information on each activity test sequences are summarize in the right hand side of table 1.

4.2 Results

The method presented in this paper is a sensor able to perceive elements of trained class of activities. Since the receptive fields integrate temporal information over 9 frames and each of the sequences of activities are typically 20 frames long, the sensor outputs a sequence of elements, rather than a single response element for each trained activity. So it is difficult to qualify its sensitivity and its robustness to variations. Regardless, an example of a probabilistic perception of the activity *"second left"* is shown in figure 3. The framework output is a map

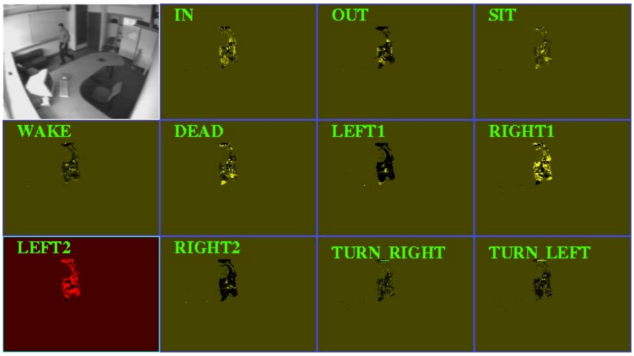


Fig. 3. *Examples of resulting maps of the local probabilities* $p\,(a_k|\boldsymbol{w})$. *The original image is in the upper left corner. Maps of the probability that each pixel belongs to one of the trained class of activity are shown. White pixels correspond to high probabilities and dark pixels to low probabilities. The occurring activity* "second left" *which has been recognized is highlighted in red.*

of the local probabilities $p\,(a_k|\boldsymbol{w})$ that each pixel belongs to one of the trained class of activities.

To evaluate the sensor ability to perceive different classes of activity a global decision rule is first designed with input the map of the local probabilities. Then the recognition rate is evaluated in function of the number of bits used to estimate the density probability of each class of activity, and in function of the number of receptive fields.

Decision rule: A global decision is taken by selecting the largest, $p\,(a_k|\boldsymbol{w})$ among the K classes. The class of activity which has the largest number of largest probabilities is selected for recognition. An example of activity recognition using such a rule is shown in figure 3 where the class of activity *"second left"* is highlighted in red.

Histograms quantification: The subspace of receptive fields responses presents a large number of dimensions and histograms are quite sparse. To bring to the fore the sparseness of histograms, the recognition rates are studied as a function of the quantification rate of the histograms and as a function of the number of dimensions in the subspace of receptive fields. The graphs of figure 4 deal with the evolution of recognition rates as a function of the number of bits per dimensions used to represent histograms. The left hand side of the figure 4 summarizes results in a subspace using only one range of Gabor filters (corresponding to one standard spatial deviation $\sigma_s = 1.49$). In this case the number of dimensions is 4, corresponding to the 4 orientations of the receptive fields. Over an histogram quantification rate of 5 bits the histograms cells are empty and Bases rule is unusable. But below a quantification rate of 4 bits histograms overlaps and activities are confused. The right hand side of figure 4 relates result with the three scale ranges of filters, corresponding to a subspace of receptive fields of 12 dimensions. The graphs show that histograms are too sparse and a quantification rate of 2 or 3 bits is the limit. It appears clearly that the training set of sequences of activities is not large enough for such a large appearance space.

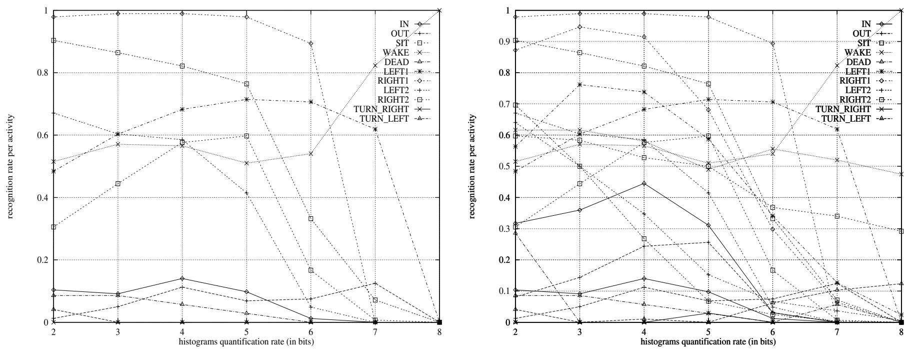


Fig. 4. *Recognition rates per class of activity as a function of the quantification rate of histograms (in bits). The left hand side figure deals with results using a 4D subspace and the right hand side figure with results for a 12D subspace.*

Recognition rates: Recognition rates are compared using histograms computed over one range of receptive fields (4D subspace) and coded with 4 bits per dimension, and with histograms computed over a 12D subspace and coded with 2 bits per dimension. Table 2 summarizes recognition rates for each class of activities. Notes that activities *"turn left"* and *"turn right"* are not recognized. The activities *"in"*, *"out"* and *"dead"* are not well perceived.

Table 2. *Recognition rates for each class of activity of the test sequences. The first row deals with results using a 4D subspace with histograms computed with 4 bits per dimension. The second row show results for a 12D subspace with histograms computed with 2 bits per dimension. The activities* "turn left" *and* "turn right" *are not recognized.*

%	**in**	**out**	**sit**	**wake**	**dead**	**left1**	**right1**	**left2**	**right2**
4D - 4 bits	11.1	5.4	57.6	56.5	5.7	64.5	92.5	58.5	82.1
12D - 2 bits	35.0	12.2	64.5	68.5	10.0	65.9	90.9	65.5	78.7

There are two reasons why those activities can not be discriminated. The first reason is that the acquisition rate is only 10 Hz, and it isn't enough to catch the motion information of short time activities. The second reason comes from the decision rule which is not rich enough to take into account the temporal

Table 3. *Confusion matrix of class of activities for the test sequences. The first left column deals with input activities an the first upper row are outputs. Each cell is the number of output labels for the corresponding input. The last right column is the total number of inputs per class of activity.*

	in	**out**	**sit**	**wake**	**dead**	**left1**	**right1**	**left2**	**right2**	**turn right**	**turn left**	**total**
in	**148**	22	8	4	0	43	39	85	8	0	0	423
out	28	**48**	3	25	0	71	107	8	27	0	2	394
sit	9	2	**306**	29	2	30	46	32	13	1	1	474
wake	23	11	22	**520**	2	66	62	11	41	0	1	759
dead	3	0	88	4	**14**	0	8	24	0	0	0	143
left1	1	1	1	1	0	**244**	9	53	60	0	0	370
right1	1	0	5	7	0	0	**291**	16	0	0	0	320
left2	16	1	120	0	1	8	31	**336**	0	0	0	513
right2	0	11	1	103	0	40	9	0	**618**	0	0	785
turn right	4	1	25	21	0	44	25	40	21	0	1	182
turn left	22	19	26	36	1	66	42	27	51	0	0	299

complexity of activities. Table 3 deals with the confusion matrix of activities. It appears that the activity *"in"* is composed of *"in"* and *"right1"*, the activity *"out"* is composed of *"out"*, *"left1"* and *"right1"*, and the activity *"dead"* is composed of *"sit"* and *"left2"*. And so on. The figure 5 shows examples of sequences of the probabilistic activity sensor outputs for the inputs activities *"dead"*, *"coming in"*, *"wake up"* and *"sit down"*. For example the activity *"sit down"* is perceive with two mains components which are effectively the activity element *"sit down"* followed by the activity element *"first right"*. This time decomposition is natural since the end of the action sit down is a pure horizontal translation corresponding to the phase when the person leans back onto the chair back.

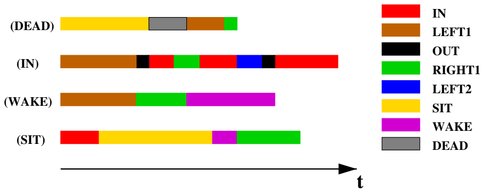


Fig. 5. *Examples of sequences of the probabilistic activity sensor outputs. The inputs are respectively sequences of the activities* "dead"*,* "coming in"*,* "wake up" *and* "sit down"*.*

4.3 Conclusion

The probabilistic sensor allows the discrimination of several class of complex elements of activities. Results are encouraging and make clear several points:

- A large subspace of receptive fields is necessary to perceive and discriminate complex activities. Only one range of receptive fields corresponding to a 4D subspace is not enough to catch signal disparity. Three ranges of receptive fields (12D subspace) have given interesting recognition rates.

- A difficulty is to make up a training basis of classes of activities large enough to allow multi-dimensional histograms computation. It has been shown that histograms computed over such a large subspace (12D) are quite sparse. But sparseness can be limited by enlarging the training set, and results can be improved.
- Improvements can be obtained by increasing the acquisition rate, in order to catch finer temporal motion information, like opening the door in the case of the activity *"coming in"*.

The main difficulty is still in the definition of a recognition framework allowing the evaluation of the robustness of the activity sensor, and to evaluate its sensivity to the histograms computation and to the receptive fields selectivity. The decision rule used previously is not rich enough to take into account that activities elements are complex. The sensor output is a sequence of the activity elements detected over a temporal window which is short relatively to the duration of the input activity. A good cue is to use temporal sequences of decisions (see figure 5) as input of a more global decision scheme. The next section define a complex global decision scheme based on Hidden Markov Models.

5 A Hidden Markov Model Based Recognition Scheme

The output of the probabilistic sensor is the temporal decomposition of complex activities into the most probable class of short activities elements. Since this temporal decomposition is difficult to predict for a given class of activity the use of Hidden Markov Model seems appropriate for recognition. Hidden Markov Model, HMM, are doubly stochastic models, because they income an underlying stochastic process that is not observable. HMMs are appropriate for modeling and recognizing time–warping dynamic patterns. HMMs have been popularized in the application area of speech recognition. Recently, HMMs have also been employed for gesture recognition and activities recognition.

5.1 Discrete Hidden Markov Model

A discrete Hidden Markov Model can be view as a nondeterministic finite automaton. Each state, s_i, is characterized by a transition probability, a_{ij}, (the transition probability to reach state s_j from state s_i), an initial state probability π_i and a discrete output probability distribution, $b_i(O_k)$, which defines the conditional probability of emitting observation symbol, O_k, from state s_i. HMM is denoted by $\lambda = (A, B, \pi)$ where A is the transition matrix, $A = \{a_{ij}\}_{ij}$, B is the observation probability vector, $B = \{b_i(O_k)\}_{ik}$, and Π is the initial probability vector, $\Pi = \{\pi_i\}_i$.

The transition matrix, A, defines the topology of the automation. In the general case, all values of a_{ij} are defined and the HMM is called *ergodic*. If A is band diagonal, HMM is *left–right* A left–right HMM is appropriate when a temporal order appear.

5.2 HMM–Based Activities Recognition Scheme

This section detail the different steps to design our approach for a discrete HMM–based activities recognition thus using one HMM per class of activity for classifying.

1. **Describing an HMM for each activity:**
   A HMM is employed to model each activity, a, which is characterized by $\lambda_a = (A_a, B_a, \Pi_a)$. Even though values of element in A_a, B_a and Π_a will be estimated in the training process, the structure of matrix, A_a, have to be determined. The structure considers in the same time the topology of the model (ergodic or left–right) and the number of states. The number of states can be determined using different methods.
   exhaustive *(a priori)* : testing all possible number of states between one and an arbitrary selected number and selecting the number which maximize the probability of recognition.
   heuristic *(a posteriori)* : the number is selected by studying the problem and the observation sequences.
   automatic : the maximization of *Bayesian Information Criterion* (BIC) [BCG98] provides an automatic method to find the number of states. Given a training sample, $\mathcal{S}_a$, for activity a, the criterion is defined as:

   $$BIC(\lambda_a, N_a) = \log P(\mathcal{S}_a|\lambda_a, N_a, \hat{\phi}) - \frac{\nu_{\lambda_a,N_a}}{2}\log(card(\mathcal{S}_a)) \quad (8)$$

   In the above equation, ν_{λ_a,N_a} is the number of independent parameters in the HMM λ_a composed of N_a states and $\hat{\phi}$ is an estimator of maximum likelihood. The equation can be viewed as the difference between a term measuring the appropriateness of data to the model, and a penalty term which penalizes models with a great number of independent parameters.
2. **Training the HMMs:**
   For each class of activity (i.e. HMM), the model parameters $\lambda_a = (A_a, B_a, \Pi_a)$ are adjusted in order to maximize the likelihood $P(\mathcal{S}_a|\lambda_a)$, the probability of observing a training sample, $\mathcal{S}_a$, given the model parameters, λ_a. Baum–Welch's re–estimation formulas is used to to reestimate model parameters to achieve a local maximum.
3. **Classifying new activity:**
   Given an observations sequence of an unknown activity, O, the classification process estimate the class, a^*, such that:

   $$a^* = \arg\max_{1 \leq a \leq \mathcal{N}} P(\lambda_a|O) \quad (9)$$

   In many cases, only $P(O|\lambda_a)$ is known. Bayes rule allows computation of $P(\lambda_a|O) = kP(O|\lambda_a)$ where k is a constant depending of the probability of each activity. The activities are considered with equal probability, $k = \frac{1}{\mathcal{N}}$. The Baum's *forward–backward* procedure is used to compute efficiently the probability $P(O|\lambda_a)$.

HMMs are composed of several parameters: observation sequences, HMMs topology and number of states. Next paragraphs deal with those parameters and experimental results useful to select there values.

Sequences of observable symbols The vocabulary used as input to the HMMs are the different classes of elements of activities from the probabilistic sensor (see section 4). But only the activities (composed of several elements of activities) *"sit down"*, *"wake up"*, *"first left"*, *"first right"*, *"second left"* and *"second right"* are studied. The reason why is that activities *"coming in"*, *"going out"*, *"dead"* (as somebody fall down), *"turn left"* and *"turn right"* are not considered because of a large number of confusion (see table 3).

HMMs Topology The nature of the activities and the outputs provided by the probabilistic sensor (see section 4 and table 5) result in a succession of different elements of activity in a complete activity. This tendency fits with the left–right topology.

Number of states The number of states can be estimated using methods presented in section 5.2. The number of states fixed a priori (heuristic method) or estimated by the *Bayesian Information Criterion* converges to the same value which 2 states per activity.

Training sets HMMs are trained with 130 sequences divided in $\mathcal{N} = 11$ classes as shown in table 4. The training set is too small to estimate efficiently the HMM. This set allows to have preliminary results and to estimate the feasibility of such recognition. If we consider left–right HMMs with two states, the number of parameters to estimate is 26: 2 for the transition matrix, 2 for the initial probability vector, 2×11 for the observations probability vector (one for each states). Considering between 10 and 20 example per parameters, for future experiments we will have to compose a training set between 260 and 520 example per activity.

Table 4. *Number of training sequences for each class of activity.*

Classes of activity	*sit*	*wake*	*left1*	*right1*	*left2*	*right2*	***Total***
Number of sequences	23	34	12	12	22	27	130

5.3 Recognition of Activities

This section presents preliminary results on the recognition of complete activities. In this experiment, we have used a cross validation on the training set presented in section 5.2. From the 130 sequences, one is extracted for recognition, all the remaining 129 sequences are used to train the 6 HMMs per activity to be recognize. Table 5 shows recognition rates.

Table 5. *Recognition rates of activities. A cross validation is used.*

Classes of activity	*sit*	*wake*	*left1*	*right1*	*left2*	*right2*	***Total***
Recognition rates (%)	91%	88%	83%	92%	82%	85%	87%
Number of misclassified activities	2	4	2	1	4	5	18

Table 5 shows promising results. We obtain a global recognition rate of 87%, corresponding to 18 misclassified activities on 130 ones. Those misclassified activities are due to the small number of training examples which imply impossibility to compute the probability of some observations sequences or the misclassification.

6 Conclusion and Perspectives

A new approach for activity recognition has been presented. Recognition of activity elements is processed statistically according to the conditional probability that a measure of the local spatio-temporal appearance is occurring for a given action. Then a temporal regularisation of perceived activity elements is done to recognize complex activities.

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 provide improved results. On one hand the vector of receptive fields responses is sensitive simultaneously to three motion ranges. The space and time scales have been selected to ensure large bandwidth. Since multi-scale strategies are redundant, a solution will be to select automatically local scale parameters according to the maxima over scales of normalized derivatives [Lin98]. On the other hand the framework presented in this paper is sensor able to perceive activities previously learned. Enlarging the training basis of each class of activities will certainly improve results since instabilities comes from the histograms sparseness.

The output of the probabilistic sensor is the temporal decomposition of complex activities (about 20 frames) into the most probable class of short activities elements (9 frames). Since the temporal aperture window of description is relatively small compared to the temporal duration of activity, Hidden Markov Models are employed to regularize the recognition. In a sense, the H.M.M. provides context. The temporal sequences of decisions are used as input of H.M.M. for the recognition of complex activities. It has been shown that some misclassification are due to the lack of training examples. Further experiments using larger training set will be done soon.

Nevertheless, plugging the perception of activities framework in an intelligent office environment controlled by a supervisor is highly considered. If the intelligent environment knows where people are in the scene, the a priori probability of each class of activities could be estimated according to the context (context cells). Introducing this a priori knowledge into the Bayes rule will improve the sensitivity of activities. For example if the tracked person comes in front of a computer the probability that the action *"sit down"* occurs is higher than the *"going out"* one.

Note that the probabilist framework for the perception of activities runs at 10 Hz on a standard bi-Pentium III 600 MHz PC.

References

[AB85] 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.

[AB91] 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.

[BCG98] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering eith integrated classification likelihood. Technical report, October 1998.

[BD96] A. Bobick and J. Davis. An appearence based representation of action. In *Proc. Int. Conference on Pattern Recognition*, pages 307–312, 1996.

[BH95] A.M. Baumberg and D.C. Hogg. Learning spatiotemporal models from training examples. Technical Report 95.9, School of Computer Studies, Division of Artificial Intelligence, University of Leeds, March 1995.

[BJ96] 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.

[CC98] V. Colin de Verdière and J.L. Crowley. Visual recognition using local appearance. In *European Conference on Computer Vision*, pages 640–654, 1998.

[CTL$^+$93] T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper, and J.Graham. Building and using flexible models incorporating gray-level information. In *International Conference on Computer Vision*, pages 242–246, May 1993.

[Hee88] D.J. Heeger. Optical flow using spatio-temporal filters. *International Journal of Computer Vision*, pages 279–302, 1988.

[Kv92] J.J. Koenderink and A.J. van Doorn. Generic neighborhood operators. *Pattern Analysis and Machine Intelligence*, 14(6):597–605, june 1992.

[Lin98] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[MN95] 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.

[Sch97] B. Schiele. *Object Recognition Using Multidimentional Receptive Field Histograms.* PhD thesis, Institut National Polytechnique de Grenoble, july 1997.

[SPH98] A. Spinei, D. Pellerin, and J. Herault. Spatio-temporal energy-based method for velocity estimation. *Signal Processing*, 65:347–362, 1998.

[WADP96] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. In *International Conference on Automatic face and Gesture Recognition*, pages 51–56, 1996.

[YB98] Y. Yacoob and M.J. Black. Parameterized modeling and recognition of activities. In *International Conference on Computer Vision*, 1998.