A Probabilistic Sensor for the Perception of Activities

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

Abstract

This paper presents a new technique for the perception of activities using statistical description of spatio-temporal properties. With this approach, the probability of an activityin a spatio-temporal image sequence is computed by applying Bayes rule to the joint statistics of the responses of motion energy receptive fields.

A set of motion energy receptive fields are 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 comparing the outputs of 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 a given time is the map of the conditional probabilities that each pixel belongs to an activity of the training set.

The approach is validated with experiments in the perception of activities of walking persons in a visual surveillance scenario. Results are robust to changes in illumination conditions, to occlusions and to changes in texture.

1. Introduction

This paper presents a a technique in which the joint statistics of spatio-temporal filters are used for the perception of classes of activities. This paper presents the design of a family of motion energy receptive fields based on Gabor filters. It shows how triads of such filters can be used to capture motion independent of the texture of the moving object. It then describes the design of a probabilistic classification system for charcterising activity patterns using multi-dimensional histograms of the receptive field responses. The technique is illustrated with results from experiments in detecting the movements of a person in an office environment.

2. The plenoptic function

Adelson and Bergen [2] 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 analysis of the plenoptic function comes within a recognition framework. The use of description techniques and the use of representation models of descriptors responses allow this analysis. 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 thresholded derivatives and the table contents are possible visual elements. Adelson and Bergen use detectors based on low order derivatives 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. Nevertheless the plenopic function provides a powerful framework for the measurement of specific local structures, including spatio-temporal patterns.

This paper concerns the characterization of activity patterns by describing their local visual motion information and modeling the descriptor responses. The result is a software sensor able to discriminate different patterns of activities.

3. Describing spatio-temporal structures

Consider the plenoptic function $I(x, y, t)$ constrained to a gray channel and a fixed view position. The description of $I(x, y, t)$ using spatio-temporal receptive fields enables its analysis. The vector of receptive fields responses describes a subspace in which each dimension is a receptive field. The main problem is to design a minimum number of receptive fields, and to determine an optimal description of appearance for a particular problem. Note that the approach could be extended to more plenoptic dimensions.

3.1. Local signal description

The notion of receptive field in vision is inspired from studies on the description of mammalian visual cortex [6]. Biological systems are observed to discribed visual information in terms of the response of sets of receptive fields. For example Young [11] has shown that biological receptive fields can be described as Gaussian derivatives and Gabor filters.

Classically the description of a signal is obtained by its projection onto a set of basis functions. The two most widely used approaches for signal decomposition are the Taylor expansion (equation 1) and the Fourier transform (equation 2). These two approaches correspond respectively to the projection of the signal onto a basis of function with amplitude modulation and onto a basis of function which are frequency modulated:

$$f(t) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(t_0) \cdot (t - t_0)^n \tag{1}$$

$$f(t) = \sum_{n=-\infty}^{\infty} \hat{f}(n) \cdot e^{i \cdot n \cdot t} \tag{2}$$

These two decomposition methods give an example of local signal description in the spatial domain for the Taylor expansion, and in the frequency domain for the Fourier series. Other local decomposition bases are also possible. A decomposition basis is generally chosen in response to 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. But 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.

3.2. Generic neighborhood operators

The concept of linear neighborhood operators was redefined by Koenderink and Doorn [7] as generic neighborhood operators. Typically operators are required at different scales corresponding to different sizes of estimation support. Koenderink and Doorn have motivated their method by rewriting neighborhood operators as the product of an aperture function $A(\vec{p}, \sigma)$ and a scale equivariant function $\phi(\vec{p}/\sigma)$:

$$G(\vec{p}) = A(\vec{p}, \sigma)\, \phi(\vec{p}/\sigma) \tag{3}$$

The aperture function takes a local estimation at location $\vec{p}$ 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:

$$A(\vec{p}, \sigma) = \frac{e^{-\frac{1}{2}\frac{\vec{p}\cdot\vec{p}}{\sigma^2}}}{\left(\sqrt{2\pi}\sigma^D\right)} \tag{4}$$

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

$$\phi(\vec{p}/\sigma) = (-1)^n\, He_n(\vec{p}/\sigma) \tag{5}$$

and in the case of the Fourier series $\phi(\vec{p}/\sigma)$ are the complex frequency modulation functions tuned to selected frequencies $\vec{\nu}$:

$$\phi(\vec{p}/\sigma) = e^{2\pi j \vec{\nu}\cdot\vec{p}/\sigma} \tag{6}$$

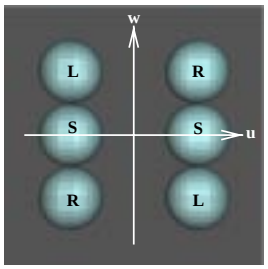
Within the context of spatial, respectively spectral, signal decomposition the generic neighborhood operators are scale normalized Gaussian derivatives [8], and respectively scale normalized Gabor filters.

3.3. Motion energy receptive fields

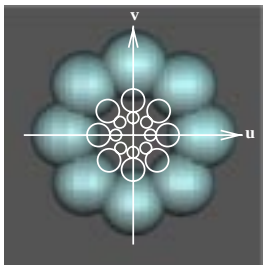
The construction of a framework for the perception of activities involves extraction of local visual motion information. Techniques which reconstruct explicitly the optical flow are often complex and specific to the analyzed scene all the more so since that there are not well suited for describing the motion of moving deformable objects [3]. The extraction of low level motion information involves the use of a decomposition basis sensitive to motion such as a signal decomposition using Gaussian derivatives or Gabor filters.

A measure of motion information rich enough to describe activities is easily obtained in the spectral domain since at a given spatio-temporal frequency an energy measure depends on both the velocity and the contrast of the input signal. In the spatial domain such energy model is hard to design.

A set of Gabor based motion energy receptive fields are used to sample the power spectrum of the moving texture. Their structure relates to the spatio-temporal energy models of Adelson and Bergen [2], and Heeger [5]. Motion energy measures are computed from the sum of the square of



(figure 1)



(figure 2)

Figure 1. 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.

Figure 2. 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.

even (G_{even}) and odd-symmetric (G_{odd}) oriented spatio-temporal Gabor filters tuned for the same orientation in order to be phase independent:

$$H\left(\vec{p}\right) = \left(I\left(\vec{p}\right) * G_{even}\right)^2 + \left(I\left(\vec{p}\right) * G_{odd}\right)^2 \qquad (7)$$

Adelson and Bergen [1] suggest 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 $I\left(\vec{p}\right)$. 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:

$$w\left(\vec{p}\right) = \frac{H_{Right}\left(\vec{p}\right) - H_{Left}\left(\vec{p}\right)}{H_{Static}\left(\vec{p}\right)} \qquad (8)$$

A triad of rightward, leftward and static Gabor energy filters is shown in 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. Figure 2 shows a map of the receptive fields' spatial bandwidths. This set of motion energy receptive fields allows the description of the spatio-temporal appearance of activity.

Whereas Gabor filters are not separable, the implementation of Gabor energy filters (which response is $H\left(\vec{p}\right)$) can be done recursively [9]. Consider the complex notation relative to equation 3 of the response of a 1D Gabor filter:

$$F\left(x\right) = I\left(x\right) * \left[A\left(x,\sigma\right)\phi\left(x/\sigma\right)\right] \qquad (9)$$

The function $\phi\left(x/\sigma\right)$ is a complex exponential involving:

$$F\left(x\right) = \left\{\left[I\left(x\right)\phi\left(-x/\sigma\right)\right] * A\left(x,\sigma\right)\right\}\phi\left(x/\sigma\right) \qquad (10)$$

There are three steps in this computation: the first step is the modulation of the signal $I\left(x\right)$ by the point operator $\phi\left(-x/\sigma\right)$, the second step is a low-pass filter convolution by the kernel $A\left(x,\sigma\right)$, and the third step is a demodulation operation by $\phi\left(x/\sigma\right)$. In the case of Gabor energy filters, the demodulation step is not necessary since:

$$H\left(x\right) = \left\{Re\left[F\left(x\right)\right]\right\}^2 + \left\{Im\left[F\left(x\right)\right]\right\}^2 \qquad (11)$$

$$= \left\{I\left(x\right)Re\left[\phi\right] * A\right\}^2 + \left\{I\left(x\right)Im\left[\phi\right] * A\right\}^2 \qquad (12)$$

For the odd and even Gabor filter only a modulation followed by a low-pass filter convolution are necessary. Since the low-pass filter is a Gaussian kernel, it is separable and can be recursively implemented [10].

4. Probabilistic Analysis of feature space

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 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.

4.1. Measurements probability density

For each class of activity, a multi-dimensional histogram of vectors of measurements is computed. The subspace of receptive fields presents a large number of dimensions which is 12D considering 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.g. number of motion energy receptive fields). A dichotomy tree is designed where each node expects 2^N potential branches corresponding to filled cells. Only filled cells are encoded. A node is a cell of the space of which each dimension is a receptive field. 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.

Such an algorithm allows de computation of high dimensional histograms which are quite sparse. Also Gaussian mixture density models may be used for more dense histograms, but there suffer from the fact that the number of modes must be a priori known or estimated.

4.2. Probabilistic perception of activities

Probabilistic perception 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 subbands.

The probability $p(a_k|\vec{w})$ according to $\vec{w}(\vec{p})$ is computed using the Bayes rule:

$$p(a_k|\vec{w}) = \frac{p(\vec{w}|a_k)\,p(a_k)}{p(\vec{w})} = \frac{p(\vec{w}|a_k)\,p(a_k)}{\sum_l p(\vec{w}|a_l)\,p(a_l)} \quad (13)$$

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 probability $p(a_k)$ of action a_k is estimated according to the context. But without a priori knowledge, it is fixed to the maximum.

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.

5. 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. 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 large visual angle of the camera allows the surveillance of the whole office. The analysed activities are *"coming in"*, *"going out"*, *"sit down"*, *"wake up"* (as stand up), *"dead"* (as somebody fall down), *"first left"*, *"first right"*, *"second left"*, *"second right"* and *"turn left"*, *"turn right"*. Those actions are assumed to take place anywhere in the scene. A view of the scene and an example of the considered activities is shown in figure 3.

Figure 3. A view of the large visual angle camera. Examples of the analyzed activities are shown.

5.1. Assumptions and parameters

The framework for the perception of activities is designed under the assumptions that the camera is fixed, so there are no global motion compensation.

Moreover since there are not a multi-scale strategy to extract motion informations, activities are asumed to be done in the same way for training and for perception. The subfilters 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 using families of filters spaced one octave apart in spatial frequency and with a standard deviation which is twice largest. Those filters are quite large band and are robust to speed changes considering the same activity.

The histograms are computed quantifying the receptive fields responses over 4 bits. Each activity is done 5 times corresponding to 5 people acting anywhere in the scene. The acquisition rate is 10 Hz and the frame size is 192×144 pixels. Each sequence of an activity is between 20 to 35 frames long.

The perception of activities according to Bayes rule (equation 13) is weighted by the a priori probability $p(a_k)$ of action a_k. So without a priori knowledge the probability $p(a_k)$ is fixed to the maximum.

5.2. Results

Because the framework presented in this paper is a sensor sensitive to a set of trained class of activities, it is diffi-

Figure 4. Examples of resulting maps of the local probabilities $p\left(a_k|\vec{w}\right)$.

cult to qualify its sensitivity and its robustness to variations. Regardless of this difficulty, an example of a probabilistic perception of the activity *"second left"* is shown in figure 4. The framework output is a map of the local probabilities $p\left(a_k|\vec{w}\right)$ that each pixel belongs to one of the trained class of activities.

A recognition rule based on the spatial average of $p\left(a_k|\vec{w}\right)$ is used to estimate the robustness of the perception framework. The largest probability average is the decisive one. The figure 5 is an example of the recognition rate of the activities *"coming in"*, *"wake up"*, and *"second right"*. Evolution of the recognition rate is plotted in function of the quantification step (in bits) of the multi-dimensional histograms. Two conclusion emerge from this experiments. Firstly, more the quantification step is little (8 bits for example) and more the histograms are sparse since histograms cells converge to one occurrence. So any probability can be extracted form such measure. It is the case of the activity *"second right"* which recognition rate fall down for a quantification step higher than 5 bits. Secondly the receptive fields are not selective enough to allow the perception of activities like *"coming in"* corresponding to low motion energies.

In the next experiment the training set is reduced to the two simple activities *"sit down"* and *" wake up"*. The sensor robustness to point of view changes and illumination variations is studied. An extract of testing sequences is shown in figure 6. The recognition rates obtained in function of point of view changes and illumination variations are shown in the next table.

%	**view 1**	**view 2**	**view 3**	**view 4**
intensity 1	97.9	73.2	59.5	56.3
intensity 2	99.1	73.5	55.8	53.1
intensity 3	91.4	73.3	58.6	53.2

The robustness to variation of illumination is effective and

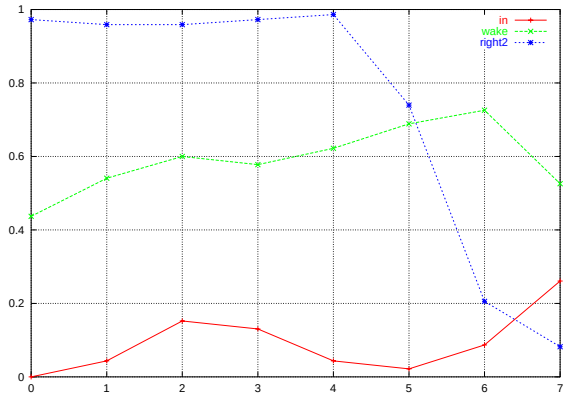


Figure 5. Recognition rate of activities *"coming in"*, *"wake up"*, **and** *"second right"* **in function of the quantification step of the multi-dimensional histograms.**

Figure 6. Extracts of "sit down" and " wake up" sequences under point of view changes and illumination variations.

foreseeable since the framework is sensitive to motion information. On the other hand the perception of activities is quite sensitive to to point of view changes in spite of that the training step was done with activities anywhere in the scene. This loss of robustness is effective for the point of views where the activity takes place far away from the camera, corresponding to low motion energies. In this case the receptive fields need to be more selective.

The main difficulties 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 probability average is not rich enough to do that, and a more complex global decision scheme like Hidden Markov Models could be more efficient.

6. Conclusion and perspectives

The visual recognition of human action has many potential applications in man-machine interaction, inter-personal communication and visual surveillance [4]. 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 spatio-temporal structures. 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 energy filters is that they can be built from separable and recursive components increasing the efficiency of the computation : the probabilistic framework for the perception of activities run at 10 Hz on a standard Pentium II 300 Mhz PC. 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.

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. Heeger [5] and Spinei [9] use a multi-scale strategy with more selective filters in space corresponding to optimal ratio between space and time scales of $\sigma_{x,y} = 4\sigma_t$. 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 (see [8]). On the other hand the framework presented in this paper is sensor able to perceive activities previously learned. To evaluate its sensitivity and its robustness we need to design a recognition scheme taking the map of local probabilities as input. The global decision scheme for recognition is relatively simplistic, corresponding to the average of local probabilities over a frame. A more complex global decision scheme such as a Hidden Markov Model should be more efficient.

Nevertheless, adding a sensor for perception of activities can be an important component of an intelligent environment. 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 spatial context. 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 the computer the probability that the action *"sit down"* occurs is higher than the *"going out"* one. We are currently exploring such a system.

References

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

[2] E. Adelson and J. 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.

[3] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:1, 1994.

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

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

[6] D. Hubel and T. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiol.*, 148:574–591, 1959.

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

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

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

[10] I. Young and L. Vliet. Recursive implementation of the gaussian filter. *Signal processing*, 44(2), 1995.

[11] R. A. Young. The gaussian derivative theory of spatial vision: analysis of cortical cell receptive filed line-weighting profiles. Technical Report GMR-4920, Computer Science Department, General Motors Research Laboratories, Warren, Michigan, US, May 1985.