IBPRIA - 2nd Iberian Conference on Pattern Recognition and Image Analysis, Estoril, Portugal, June 7-9, 2005

Extracting Motion Features for Visual Human Activity Representation*

Filiberto Pla[1], Pedro Ribeiro[2], José Santos-Victor[2], Alexandre Bernardino[2]

1 Computer Vision Group, Departament de Llenguatges i Sistemes Informàtics,
Universitat Jaume I, 12071 Castellón, Spain
Filiberto.Pla@lsi.uji.es

2 Computer Vision Lab – VisLab, Instituto de Sistemas e Robótica,
Instituto Superior Técnico, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
{pribeiro, jasv, alex}@ isr.ist.utl.pt

Abstract. This paper presents a technique to characterize human actions in visual surveillance scenarios in order to describe, in a qualitative way, basic human movements in general imaging conditions. The representation proposed is based on focus of attention concepts, as part of an active tracking process to describe target movements. The introduced representation, named "focus of attention" representation, FOA, is based on motion information. A segmentation method is also presented to group the FOA in uniform temporal segments. The segmentation will allow providing a higher level description of human actions, by means of further classifying each segment in different types of basic movements.

1. Introduction

Monitorizing human activity is one of the most important visual tasks to be carried out in visual surveillance scenarios. This task includes processes like target tracking, human activity characterization and recognition, etc. Human activity characterization and recognition is a special topic that has been addressed in the literature from different points of views and for different purposes [8] [10] [2] [1] [9].

In the work described here, the objective was to characterize, aimed at building a feature representation for further recognition, the human activity of people in typical visual surveillance scenarios, like airport lounges, public building halls, commercial centers, etc., with a great variety of human action types and ordinary, rather poor, imaging conditions. The main idea of the proposed techniques is to perform a general description of basic human movements, extracting some visual cues that can help to understand the people's actions in higher level recognition tasks.

In order to understand the activity of a person in a given scenario, the human movement can be described as a composition of two different types of movements:

* Work partially supported by grant from the *Spanish Ministry of Science and Education* PR2004-0333, and CAVIAR IST-2001-37540 project from European Union.

1. The movements that a person performs with respect to the environment, that is, the analysis of trajectories and dynamics, performing target tracking. Some of the works are based only on this information [3].
2. The movements that the different parts of the body a person performs during a certain action, with respect to the body point of view.

According to the classification described by [2], human activity recognition approaches can be divided in three different groups. Fist group are *Generic model recovery* approaches, in which, at each time, the person pose is recovered trying to fit it with a 3D body model. These approaches strongly depend on an accurate 3D feature extraction from the image, which usually needs human intervention and controlled environments to facilitate image measurements [6].

Appearance-based models are an alternative to 3D model recovery, appearance based models rely on 2D information extracted from the images, either raw grey level distributions or other processed image features, like region templates, contours, etc. where an action is described as a sequence of 2D poses of the moving target [7].

Finally, *motion-based recognition* techniques try to recognize the human activity by analyzing directly the motion itself, without referring it to any static model of the body. The rational of these approaches lie in the fact that different movements of the body produce defined motion patterns in the image domain [2] [1] [5] [9]. Therefore, some of these works use optical flow measurements as motion features to recognize human activities [10] [8].

The approach presented here is included in the motion-based recognition techniques, aiming at characterizing human activities directly from the motion information. In particular, we will use optical flow information, focusing our attention to the movements of different parts of the body, trying to characterize basic body movements. Therefore, we will assume that a certain target extraction and tracking has already been performed, that is, we will keep our “active” attention to the target only, centering our target in our field of view, the fovea, for further analysis.

2. FOA representation of human motion activity.

As it has already been mentioned, the objective is to characterize, for further recognition, human activities in different scenarios, with variable and realistic conditions that may occur, like low image contrast and resolution, different camera-target relative positions and viewpoints, occlusions of body parts during the movements, and the huge variability of people features and situations.

However, although the human activity recognition task in such conditions may seem unfeasible, it is well known that humans can guess what are the main or basic movements that a target is performing with a non very well defined image structure [2]. Thus, the underlying motion structure of the movement of a target can provide enough visual cues to allow the recognition of basic human body movements.

Keeping this fact in mind, a motion-based structure to characterize basic and general movements of the body is proposed, which has been built on twofold considerations: (a) use of optical flow, and (b) attention centered on the target.

Therefore, the idea is to describe the person movements with respect to some point of the body, assuming the person is being tracked and segmented out from the background. Thus, a previous tracking and target segmentation is performed, which provides us at each time information about the position of the target.

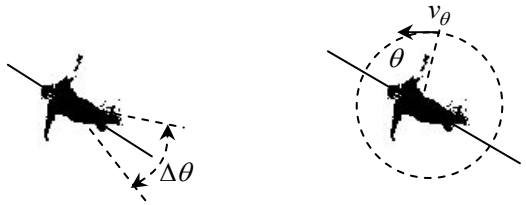

Figure 1. (a) Expected variation of legs movement around the body centroid. (b) Mean optical flow in a given direction

The center of attention, or fovea center, will be situated at the centroid of the region corresponding to the segmented target. In order to refer the motion to the center of the focus of attention $v_c(t)$, the optical flow of the target pixels $v_i(t)$ will be referred to the centroid of the target,

$$v'_i(t) = v_i(t) - v_c(t)$$

Therefore, the target motion with respect to the image coordinates will be compensated, and only the relative motion of the different parts of the target, with respect to the center of attention, will be represented. The objective is to have a qualitative description of the movement, without segmenting or identifying parts of the body, due to the fact that segmenting and tracking each part of the body is a complex and difficult process that cannot be solved in many situations.

Let us have a look to the figure 1. We can assume that the body parts are arranged around the body centroid, and that certain parts of the body usually move around a certain angular range $\Delta\theta$ around the body centroid, for instance, the expected angular variation of the legs movements (figure 1a).

In order we can have a unique reference for all the angular directions with respect to the same origin, they can be referred to the vertical axis of a standing up person. An estimation of the vertical axis of the body can be obtained either by computing the principal axis of the target region, or calibrating the field of view of a static camera, determining the vertical direction with respect to the floor at every image point.

Let us represent the mean optical flow $v_\theta(t)$, at each time *t*, with respect to the centroid at a certain angular direction θ (figure 1b), as:

$$v_\theta(t) = \frac{1}{N_\theta} \sum_{k \in P_\theta} v'_k(t)$$

with P_θ being the set of target pixels (x_k,y_k) that are in the θ direction with respect to the target centroid (x_c,y_c), and $N_\theta = |P_\theta|$ the number of target pixels in such direction.

Decomposing the flow $v_\theta(t)$ in its normal and radial direction with respect to the target centroid, will provide an estimation of the relative motion of that part of the body with respect to the centroid in terms of radial (moving from or towards the centroid) and normal (moving in a perpendicular direction to the radius towards either up/left or down/right, depending on the area of the body where θ is situated).

Representing all angular values of $v_\theta(t)$ along time becomes a 2D signal $foa(t,\theta) = v_\theta(t)$ named *focus of attention representation* (FOA) of the target flow. The FOA provides a description of the evolution of the mean flow of the target pixels at every direction θ with respect to the target centroid. The FOA extracted from a temporal sequence of a tracked person will provide us information about the general movements of the different parts of the body, without having an exact knowledge about the position and motion of each part of the body.

Thus the FOA representation at a given time has the following properties:

- It is a focus of attention representation, inspired in foveal imaging, where the representation is built around a fovea point, in this case, the target centroid.
- It is an active technique based on focusing the attention on the tracked target.
- Provides an angular description of the target with respect to the fovea point.
- The information provided for each angle can be easily interpreted using the normal and radial components of the flow.

Thinking about the discrete form of the FOA representation, it can be further simplified by representing the mean flow of the target along a finite set of orientations θ_i ; $i = 0,...N-1$, where the chosen orientations could integrate the mean flow of nearby directions, that is, at each time *t*, given an orientation θ, the mean flow $foa(t,\theta_i) = v_{\theta_i}(t)$, can be expressed as an integration of a receptive field area around direction θ_i. This receptive field area would cover a certain angular range around the direction. We can define the response of the receptive field area around θ_i direction, as a Gaussian weighted mean of the FOA in the nearby directions, that is

$$foa(t,\theta_i) = \int foa(t,\theta)\, e^{-(\theta-\theta_i)^2/2\sigma_\theta^2}\, d\theta$$

where σ_θ is the typical deviation of the Gaussian receptive field, determining the scope of the receptive field area around each direction. Receptive fields may overlap depending on the scope determined by the standard deviation.

Different types of body movements will activate different receptive fields in different ways, forming defined patterns characterizing basic movements like walking, rising/putting down arms, bending, sitting, etc. The response of the receptive fields forming the FOA representation at each time will provide us a way of identifying such a type of basic movements.

3. Segmenting the FOA representation.

The final aim of the FOA representation is to allow a recognition of human actions. Once we have a representation, in a given feature space, in order to facilitate the

recognition tasks, a temporal segmentation of the body movements would be desirable, in order to decompose a certain human action in simple temporal units containing a unique type of basic body movements.

Other works, like [10], were also aimed at segmenting sequences of human activity to select key pose actions, in order to describe a higher level human activity description. The approach presented here is similar to this basic idea used in [10] about linear prediction, but using other two different concepts.

In order to segment the FOA representation along time, we will look for changes in the FOA representation along time in a similar way changes in video shot sequences are detected. The way changes are detected in the FOA are inspired in the work of [4] for video change detection, which uses the main motion present between two images of a sequence as a way to predict changes in the same video shot.

In a similar way, given the $foa(t-1,\theta_i)$ values of a tracked target for the receptive fields θ_i at a time *t-1*, we can predict the FOA response at a time *t*, $foa^*(t,\theta_i)$. Thus, given the new measured $foa(t,\theta_i)$, we can define the following difference function $Dfoa(t)$ to detect changes:

$$Dfoa(t) = \sum_i \left| foa(t,\theta_i) - foa^*(t,\theta_i) \right|$$

Looking for significant local maxima in the $Dfoa(t)$ function, we can identify the times at which there is a noticeable change in the body movements performed by the target. Bear in mind that the values of $foa(t,\theta) = v_\theta(t)$ are motion vectors of two components, expressed either in the Cartesian components or in the radial-normal components mentioned in the previous section.

To compute the estimate of $foa^*(t,\theta_i)$ from $foa(t-1,\theta_i)$, the following approach is used. Given $v'_k(t-1)$, the vector field referred to the target centroid at time *t-1*, we can estimate the flow field at time *t* of every pixel belonging to the target at time *t-1*. Given the flow vector $v'_k(t-1)$ of pixel $p_k(t-1) = (x_k, y_k)$, we can estimate the new position of pixel p_k in time *t* by

$$p_k^*(t) = (x_k^*, y_k^*) = p_k(t-1) + v'_k(t-1)$$

To the estimated position of the pixel $p_k^*(t)$, the flow vector $v_k^*(t)$ estimated for time *t* at this position will be figured out by applying an uniform movement assumption, that is , $v_k^*(t) = v'_k(t-1)$. Therefore, the estimated mean flow field vector at time *t*, that is, the estimated FOA at time *t*, can be computed as

$$foa^*(t,\theta) = v_\theta^*(t) = \frac{1}{N_\theta} \sum_{k \in P_\theta} v_k^*(t)$$

with P_θ being the set of target pixels (x_k,y_k) that are in the θ direction with respect to the target centroid (x_c,y_c) at time *t*, and $N_\theta = |P_\theta|$ the number of target pixels in such a direction.

4. Experiments and examples

In order to see the effectiveness of the FOA representation and the performance of the FOA segmentation method introduced in section 3, the method has been tested using some sequences of the CAVIAR project [11]. Figure 2 shows some frames of a sequence in a hall of a building entrance, where the tracked person performs a movement combining stepping, turning the upper part of the body and rising arms, afterwards, he stands by for a moment while the arms are up and then he comes back to the initial position.

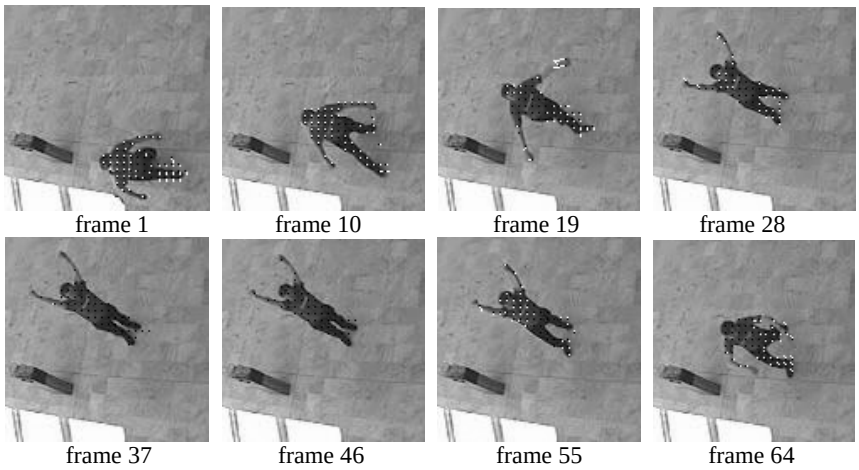

Figure 2. Some frames of a sequence of 70 frames of a target person.

Figure 3 shows the FOA representation of the 70 frames of the sequence in figure 2 using 20 receptive fields. In this case, the fields are placed every 18 degrees from the angle origin, which is placed at the head direction of the principal axis of the target. The principal axis at each frame *t* of the sequence has been estimated from the blob corresponding to the segmented target. The center of the FOA representation has been chosen as the centroid of the blob, which is also placed on the principal axis.

The flow vectors in figure 3 represent, at each time *t*, the mean flow computed by each receptive field $\theta_i ; i = 0,...,N_\theta - 1$, with $N_\theta = 20$. The flow vectors are expressed in terms of the normal and radial components with respect to the direction of the receptive field, that is, $foa(t,\theta) = v_\theta(t) = (v_{R\theta}, v_{N\theta})(t)$. The radial component $v_{R\theta}$ of each vector is represented along the abscissas axis (*t* axis) and the normal component $v_{N\theta}$ is represented along the ordinates axis (θ axis).

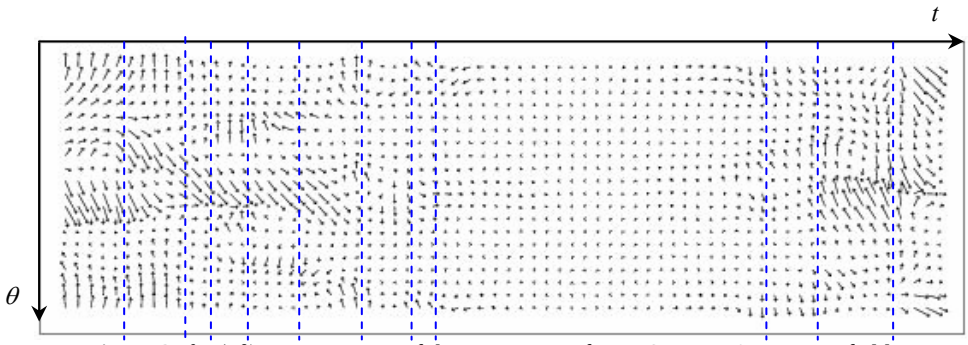

Figure 3. *foa(t,θ)* representation of the sequence in figure 6 using 20 receptive fields.

Looking at figure 3, we can notice how the FOA presents differentiated patterns at different times, corresponding to the different movements of the parts of the body. For instance, the flow field in the first 5 frames corresponds to the activity present at the legs, that is, the receptive fields at the middle, which represents the stepping action of the person. We can even distinguish the movement performed by each leg in opposite direction; all measured with respect to the focus of attention center, that is, the centroid. We can also notice a movement in the upper part of the body, corresponding to the firsts and lasts receptive fields. This movement has a strong normal component that characterizes the turning movement of the upper part of the body the person is performing while stepping, in this case, the person is turning leftwards with respect to the principal axis of the body.

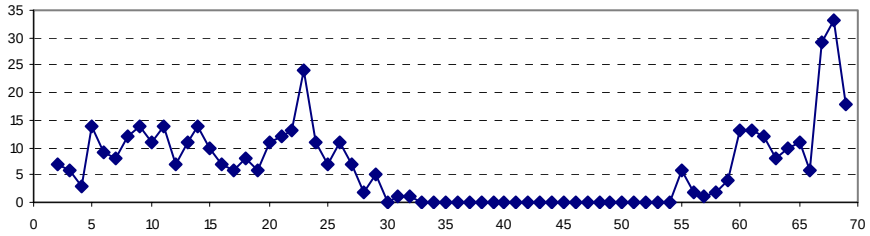

Figure 4. $Dfoa(t)$ of the FOA in figure 3.

Figure 4 shows the computed *Dfoa* of the FOA in figure 3, in order to segment the FOA representation in basic units with uniform motion values of the different parts of the target. The *Dfoa* has been computed using 72 receptive fields, that is, one every 5 degrees, and with a standard deviation of $\sigma_\theta=1$ degree, that is, without no appreciable overlapping between receptive fields. The prediction was approximated by using the *t-1* segmented target instead of the segmented target at *t*, for the sake of computational efficiency. The local maxima of the *Dfoa* in figure 4 have been represented by dashed vertical lines in the corresponding representation of the FOA in figure 3 to illustrate how the segments between these limits show an uniformity in the motion values.

5. Conclusions and further work

This paper has described a technique to characterize human actions in visual surveillance scenarios in order to describe, in a qualitative way, basic human movements in general imaging conditions. The representation proposed is based on the introduced focus of attention approach, the FOA, building the representation from the point of view of the tracked target, thus becoming part of the active vision process to describe target movements. The introduced representation is based on motion information, particularly optical flow from respect to the fovea point.

The representation has been tested in some sequences from the database of the CAVIAR project, and the results obtained show its effectiveness to represent differentiate patters for different types of body moments, which could also be complex or combined movements of the different parts of the body.

The main further work is directed to apply some classification techniques to the FOA segments in order to identify and recognize automatically the sequence of basic movements.

References

1. BenAbdelkader, C.; Cutler, R. and Davis, L.; "Motion-based recognition of people in EigenGait space"; V Int. Conf. on Automatic Face Gesture Recognition, 2002.
2. Bobick, A. F. and Davis, J.W.; "The recognition of human movement using temporal templates", IEEE. Trans. on PAMI, 23(3): 257-267, 2001.
3. Bodor, R.; Jackson, B. and Papanikoloupolos, N.; "Vision-based human tracking and activity recognition", XI Mediterranean Conf. on Control and Automation, 2003.
4. Bouthemy, P.; Gelgon, M. and Ganansia, F.; "A unified approach to shot change detection and camera motion characterization". IEEE Trans. on Circuits and Systems for Video Technology, 9(7):1030-1044, October 1999.
5. Bradski, G.R. and Davis, J.W.; "Motion segmentation and pose recognition with motion history gradients", Machine Vision and Applications, 13: 174-184, 2002.
6. Davis, J.W. and Gao, H.; "An expressive three-mode principal components model of human action style", Image and Vision Computing, 21: 1001-1016, 2003.
7. Davis, J.W. and Tyagi, A.; "A reliable-inference framework for recognition of human actions"; IEEE Conf. on Advance Video and Signal Based Surveillance, 169-176, 2003.
8. Essa, I.A. and Pentland, A.P.; "Coding, analysis, interpretation and recognition of facial expressions", IEEE Trans. on PAMI, 19(7): 757-763, 1997.
9. Masoud, O. and Papanikolopoulos, N.; "Recognizing human activities", IEEE Conf. on Advanced Video and Signal Surveillance, 2003.
10. Rui, Y. and Anandan, P.; "Segmenting visual actions based on spatio-temporal motion patterns", IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2000.
11. CAVIAR Project IST 2001 37540, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.