Video-Based Human Motion Estimation System

Mariofanna Milanova and Leonardo Bocchi

Computer Science Department, University of Arkansas at Little Rock
2801 S. University Ave., Little Rock, Arkansas 72204, USA
Dept. of Electronics and Telecommunications, University of Florence,
V. S. Marta 3, Florence, Italy
mgmilanova@ualr.edu, leonardo.bocchi@unifi.it

Abstract. This paper presents the system designed to estimate body silhouette representation from sequences of images. The accuracy of human motion estimation can be improved by increasing the complexity of any of the three fundamental building blocks: the measured data, the prior model, or the optimization method. The vast majority of existing literature on human motion estimation has focused on just one of these building blocks: improving the methods for optimization, also called inference. In contrast, our approach seeks to explore the hypothesis that the other two building blocks are critical components, using extremely high accuracy measured data and shape of body motion priors, so that the objective function is more precise and less noisy, resulting in an easier solution. Our main goal is to develop a new module for extracting accuracy measured data from video imagery.

1 Introduction

Much work has been done to create different automated visual surveillance systems that are based on computer vision technology and designed for security purposes facilitating the detection and tracking of human motion and intrusion. A survey by Wang reviewed the techniques for human motion analysis dealing with detecting, tracking and recognizing [1]. Wang's paper provides a comprehensive survey of research on computer-vision-based human motion analysis. The emphasis is on three major issues involved in a general human motion analysis system, namely human detection, tracking and activity understanding.

Currently the main algorithms for locating and tracking people can be divided into the following four categories: region –based tracking, active contour- based tracking, feature- based tracking and model-based tracking [2].

Techniques for tracking human motion from video present us with still open problems, such as: self occlusion of the legs during motion, occlusions from other objects, different illumination and background.

The task of tracking human motion becomes even more complicated in the case of tracking human activities. A requirement for achieving automatic surveillance of human activity is attaining a reliable tracking of human body parts. A visual surveillance system could identify human behavior that is considered abnormal by simply representing or interpreting the human body movements. For example, in the

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2009, LNCS 5620, pp. 132–139, 2009.
© Springer-Verlag Berlin Heidelberg 2009

case of detecting a distinct and unusual body shape or different body parts with abnormal proportions or people starting a fight, the visual surveillance system will interpret the action.

Keeping a detailed description of the human figures being tracked, such as segmenting into meaningful body parts, allows for a more comprehensive analysis of the human activities being tracked.

The paper is organized as follows: section 2 describes the overall architecture of the system. In sections 2-1 – 2.5 we determine a range of methods implemented to estimate body silhouette representation. Experimental results are presented in section 3. Conclusions are presented in section 4.

2 Materials and Methods

The overall architecture of the system is shown in Fig. 1. In the first step, each frame of the video sequence is processed to extract the Region of Interest (ROI) from the background. The extracted ROI is then filtered to evaluate a vector of feature maps which is used to represent the spatial distribution of features in the frame. The detected blobs are refined to produce a human silhouette. Next, at the body silhouette representation step, we implement body –part detection referred to as a set of local descriptors. For the shape –based analysis, we define a global descriptor. Both the local descriptors and global descriptor are combined to implement pose/shape estimation. In the last stage, an elastic network composed of a set of specialized feature detectors is used to find the optimal match between the image and feature maps. Our generative model predicts silhouettes in each video camera view given the pose /shape parameters of the model.

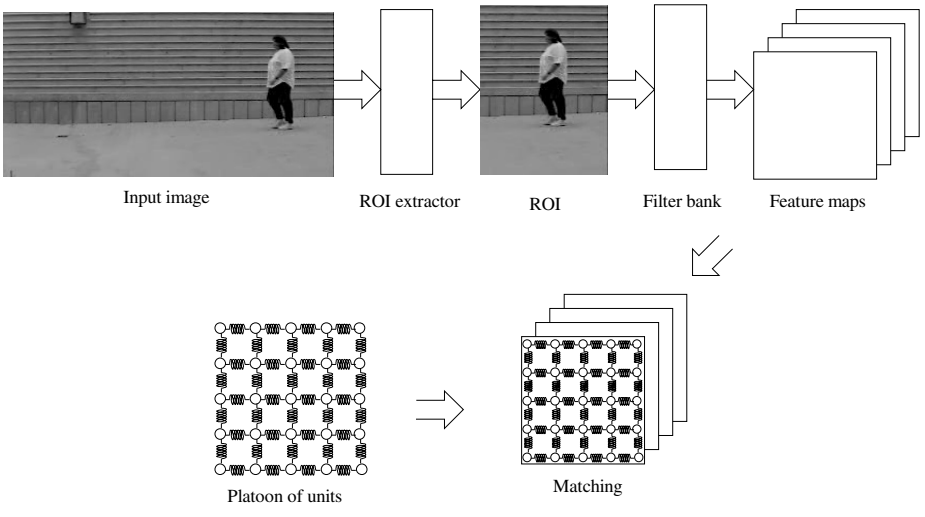

Fig. 1. Block diagram of the system

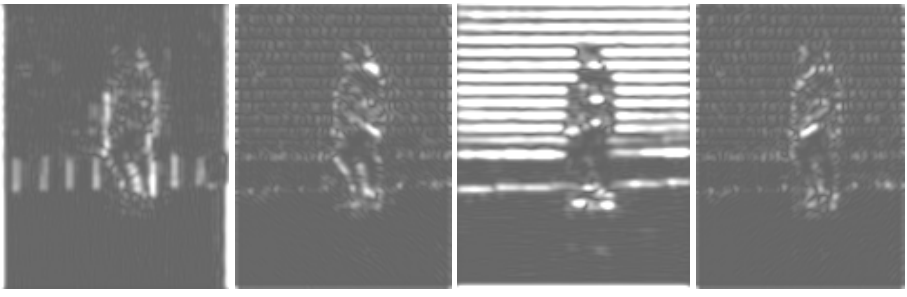

Fig. 2. Sample feature maps obtained from the ROI shown in fig. 1, with scale k=1 and four different orientations (horizontal, left diagonal, vertical, right diagonal, respectively)

2.1 Region of Interest (ROI) Extraction

For ROI extraction we used optical flow algorithm presented by Little and Boyd [3]. We compute the optical flow of the motion sequence to get *n* images (frames) of (*u,v*) data, where *u* is the x-direction flow and *v* it the y-direction flow. The dense optical flow is generated by minimizing the sum of absolute differences between image patches. The result is a set of moving points. For each frame of the flow we compute a set of scalars that characterizes the shape of the flow in that frame. We use all the points in the flow and analyze their spatial distribution. The shape of motion is the distribution of flow, characterized by several sets of measures of the flow. For example, we compute the following scalars: x and y coordinates of the centroid of moving region, aspect ratio of moving region. The system is represented in Milanova [4].

2.2 Feature Extraction

Each ROI has been processed to extract a feature map which describes the spatial position and features of the subject.

A proper selection of the set of feature maps is of primary importance to obtain good performances of the system, and it is mandatory that the features represent the local properties of the image (presence and directions of edges, corners, and similar characteristics of the image). Moreover, a set of optimal features needs to provide a compact description of these local properties of the image to have a feature vector of reasonable size. This suggests that a feature set be selected having both a limited spatial support in order to capture local information and a limited frequency response allowing to reduce noise. These properties, as shown by several researchers, are best exploited by Gabor functions.

The proposed feature map is based on a multiscale Gabor representation, based on the Morlet wavelet. The Morlet wavelet is defined by rotation and scaling of a mother wavelet function which, for a given scale *k* and a given orientation θ, is expressed as:

$$\psi_{k,\theta} = \beta_{k,\theta} \exp\left(-\frac{\mathbf{v}_{k,\theta}^2 \mathbf{x}^2}{2\sigma^2} \right) \exp\left(i \mathbf{v}_k \mathbf{x}\right) \tag{1}$$

where $\boldsymbol{x} = (x,y)$ is the coordinate vector in the image plane, $\beta_{k,\theta}$ is a normalization constant, and the parameter $\boldsymbol{v}$ has direction θ and module $v = 2^k$. Starting from (1), the (discrete) Morlet transform is defined as:

$$M_{k,\mathbf{x}_0} = \sum_{\mathbf{x}} \psi_{k,\theta}(\mathbf{x} - \mathbf{x}_0) \cdot I(\mathbf{x}) \tag{2}$$

For each value of scale k and orientation θ, the Morlet transform represents a map of the features with the corresponding scale and orientation present in the image.

In this work, we used a set of three different scales, and eight equally spaced directions. The resulting transforms are complex-valued images. As the phase information is mainly related to the spatial position of the objects into the image, we converted each feature map to a real valued image, taking it modulus. Therefore, we obtained a total of 24 feature maps from each ROI. Each pixel in the image is associated to a feature vector having 24 components, which describe the local properties of the image in a neighborhood of the pixel.

2.3 Representation of the Silhouette

The proposed representation is based on a self –organizing system designed to learn to recognize both the characteristic features of the image and their spatial relationship.

To this end, we build an architecture composed of a set of specialized feature detectors which are coupled together, by elastic forces, to form a sort of network. Each feature detector acts as an independent unit, which is free to move on the target image to detect a matching feature. The unit is identified by a target vector with 24 components, which is compared to features present in the image in order to find the best match.

However, the elastic coupling between units force the units to act in a coordinated way, and to find the optimal matching between detectors and image, taking into account both the matching between feature detectors and actual features, and the spatial relationships among features, both in the image and in the network units.

When a new image is fed as input to the network a relaxation process takes place allowing units to move on the image and to reach the optimal minimum. During the training phase, at the end of the relaxation phase, the target vector is updated to match more closely the feature vector in the final location of the unit.

In the following, we outline the basic relations which describe the network dynamics. For a more detailed description, see [5].

2.4 Relaxation

The approach used to find the optimal configuration of the network, which is the best position of the feature detectors on the input image, is based on an energy minimization strategy. The network is associated to energy composed of two components: the first part is associated to the elastic stretching of the connection between the units, while the second component relates to the discrepancy between target vectors and the features present in the image.

The elastic energy is evaluated by assuming an ideal spring is connected between each couple of neighboring units in the grid. Assuming each unit *(i,j)* is connected to the 4-neighborhood S_{ij}, the resulting elastic energy E_i can be expressed as:

$$E_i = \sum_{(i,j)} \sum_{(m,n) \in S_{ij}} c \left| P_{ij} - P_{mn} \right|^2 \tag{3}$$

where P_{ij} is the position, on the image, where the unit *(i,j)* is located and *c* represents the elastic constant of the springs. The second energy term has been selected in order to achieve the minimal energy when the best matching occurs between the target vector and the feature maps. Among the several possible rules, we selected the most straightforward, based on the scalar product between the two vectors:

$$E_e = -\sum_{(i,j)} \mathbf{w}_{ij} \times \mathbf{I}(P_{ij}) \tag{4}$$

where *(i,j)* identifies the network units, w_{ij} is the target vector for the unit, and $\mathbf{I}(P_{ij})$ is the vector of the feature maps at the position P_{ij}.

Minimization of the total energy $E_t = E_i + E_e$ is achieved by an iterative procedure. In the first phase, an attention point is randomly selected on the image. Selection procedure is performed using a roulette-wheel procedure, which gives higher probability of selection to points having a larger feature vector. Once the attention point has been selected, all units in the neighborhood of the attention point are tested by moving, in turn, each of them in the attention point, and evaluation the network energy before and after the move. The unit which is associated to the largest energy loss is then selected as winner, and it is moved toward the attention point. Experimental results indicate that the convergence speed can be improved by moving altogether all units in the neighborhood of the winning units.

The process is then repeated for a given number of iterations, slowly decreasing both the speed of the movement toward the attention point and the radius of the neighborhood.

2.5 Adaptation

Once the relaxation phase has been completed the network can be trained to improve the matching between the target vector and the actual features present in the image. To this end, the target vector of each unit is changed in order to reduce the difference between the target vector and the features present in the image in the final position of the unit. The adaptation is performed following a rule similar to the one used in the self-organized neural maps proposed by Kohonen [6]:

$$\Delta \mathbf{w}_{ij} = \varepsilon \left[\mathbf{I}(P_{ij}) - \mathbf{w}_{ij} \right] \tag{5}$$

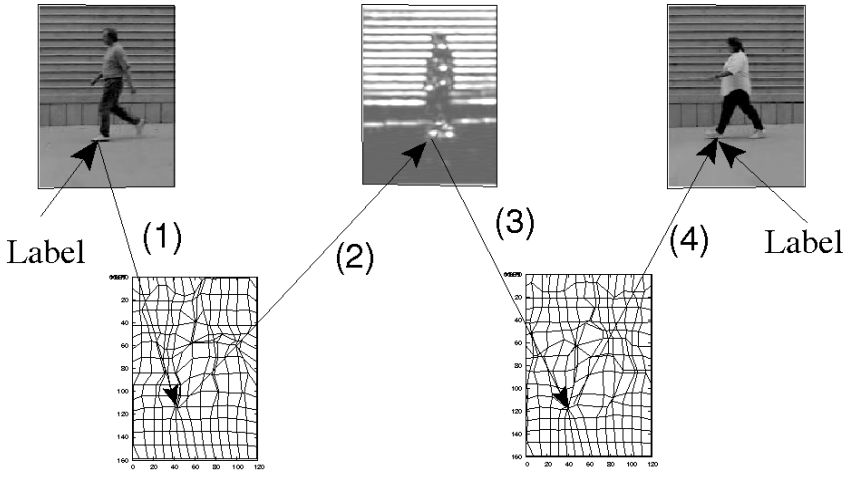

Fig. 3. Results evaluation: a label marks a significant point of the image (walking man) in the training set. At first (1) the match point which is closer to the label is identified. This allows to mark (2) a unit in the grid. A second image is presented to the network, and the unit identifies a new match point (3). The distance between the match point and a label (4) placed on the test image (walking woman) gives an estimate of the results.

where ε is the learning rate. As it occurs with Kohonen maps, best convergence results are obtained using a slowly decreasing value of ε.

3 Experimental Results

The system has been developed and tested on a set of video sequences obtained from surveillance cameras. The video sequences represent, on a simple background, different individuals walking over a straight line. Each sequence lasts approx 5-10s. The set of sequences has been split in two distinct parts, to produce a training set and a test set. A first set of frames has been extracted from the training set and used to design the system and to train the units. After the training phase has been completed, the system is tested on video frames extracted by the test set.

A semi-quantitative evaluation can be obtained by labeling some of the units in the grid according to their location on the images in the training set, as shown in Fig. 3. The procedure is designed to transfer the labels from an image to network units and from an image to a different image. The procedure can be described as follows: each test image is applied to the network input and the mesh completes the relaxation process. At the end of the relaxation process it is possible to identify the unit whose match point is located closest to the label placed on the image. That unit is assumed to represent the tag point in the first image. When a test image is presented to the network and the relaxation process has been completed, the unit is located in a new point. This point is assumed to represent the label in the test image.

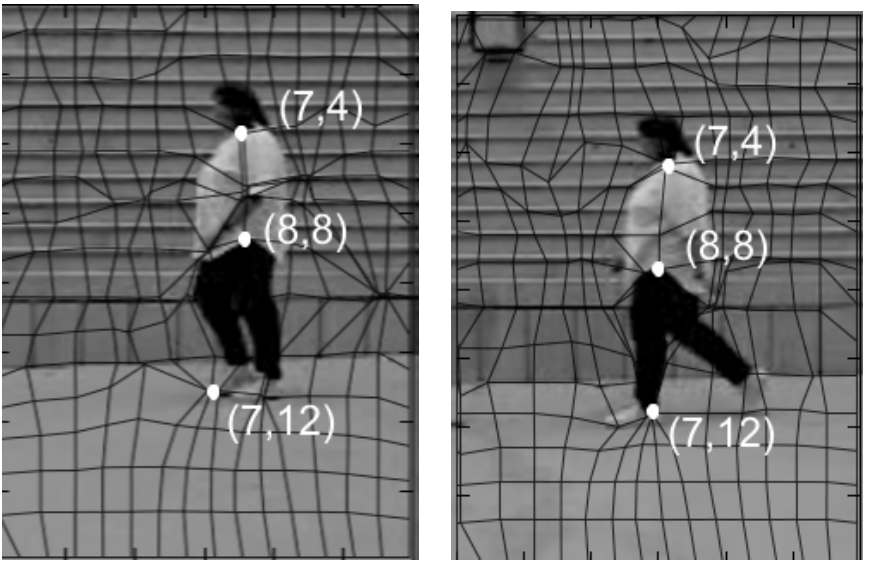

Fig. 4. Results of the matching step. The grid indicates the position of the units, at the equilibrium, on two sample images from the test sequence. Three units have been marked in white on both images.

An evaluation of the performance of the network can be therefore achieved by measuring the distance from the transferred label and the manually applied label. A perfect matching between the images would produce a null relative displacement where the same unit should be located on the same marker in all images. Experimental results indicate that the average absolute displacement between the marks is about 0.9 units.

A visual evaluation of the results is shown in Fig. 4 for two different frames of the sequence. It can be noted that units concentrate on boundaries of the moving figure, where the maximum information is present. As an example, Fig. 4 shows three units which are marked with white labels, together with their position in the grid. As it can be seen in the figures, those units are in a consistent position on both frames, although their spatial position is quite different in the two frames. More specifically, unit (7,4) is positioned on the neck of the figure, unit (8,8) is positioned on the belt, and unit (7,12) is located on the front foot. In the latter case, it can be noted how the grid identifies different points of the feet, due to its different orientation.

4 Conclusions

The proposed system is planned to be part of an automated self-training system designed to perform video surveillance. Positions of labeled units allow extracting information about the position and the dynamics of the observed figure. The parameters which describe the relative distance of the units are, therefore, associated to the position and to the physical dimension of the different body parts.

Our aim is to interpret the deformation parameters of the grid to detect anomalies in the shape or position of the figure, as well as in the dynamics of the movement. Any discrepancy between the learned behavior of the grid and the actual behavior can be used to trigger an alarm stating that something anomalous is occurring.

Acknowledgment

This paper is supported by NSF grant 0619069 Development of Interdisciplinary Arkansas Emulation Laboratory and by funding from the U.S. Defense Threat Reduction Agency (DTRA-BA08MSB008).

References

1. Wang, L., Hu, W.: Recent developments in human motion analysis. Pattern Recognition 36(3), 585–601 (2003)
2. Anderson, P., Corlin, R.: Tracking of Interacting Peoples and Their Body Parts for Outdoor Surveillance, Master Thesis (2005)
3. Little, J., Boud, J.: Recognizing People by their Gait: The Shape of Motion. J of Computer Vision Research 1(2), 2–32 (1998)
4. Milanova, M.: Object Recognition in Image Sequences with Cellular Neural Networks. Neurocomputing 31(1-4), 125–141 (2000)
5. Bocchi, L.: Evolution of an abstract image representation by a population of feature detectors. In: Cagnoni, S., Lutton, E., Olagu, G. (eds.) Genetic and evolutionary computation for image processing, pp. 157–176. Hindawi Publishing Corporatio (2008)
6. Kohonen, T.: Self-Organization and Associative Memory, 3rd edn. Springer, New York (1989)