POSTURE ESTIMATION IN VISUAL SURVEILLANCE OF ARCHAEOLOGICAL SITES

P. Spagnolo, M. Leo, A. Leone, G. Attolico, A. Distante

Istituto di Studi sui Sistemi Intelligenti per l'Automazione - C.N.R.
Via Amendola 166/5, 70126 Bari, ITALY
{spagnolo,leo,attolico,distante}@ba.issia.cnr.it

ABSTRACT

This paper presents a simple and reliable approach to the estimation of body postures in visual surveillance of outdoor environments. The image sequences coming from a still camera are processed by two subsystems to detect motion and recognize objects (humans). Regions corresponding to people are fed to the posture estimation module. The proposed algorithm is based on an unsupervised clustering approach that makes the system substantially independent from any a-priori assumption about the possible output postures. Horizontal and vertical histograms of the binary shapes associated to humans are selected as features for the posture estimation. The Manhattan distance is used for both cluster building and for run-time classification. The BCLS algorithm (Basic Competitive Learning Scheme) has been selected after extensive experimental tests for the construction of clusters. The whole approach has been verified on real sequences acquired while the typical illegal activities involved in stealing were simulated in an archeological site.

1. INTRODUCTION

In the last years, pose estimation has attracted great interest from computer vision researchers due to its promising applications in many areas, firstly visual surveillance. Due to a diffuse request of safety, surveillance systems are being installed in an increasing number of locations such as highways, streets, airports, homes,…. Our application context is the visual surveillance of archeological sites. In this context the main aim is to detect the presence of peoples and to recognize their gestures in order to timely identify illegal actions. In particular, the system should be able to recover the correct posture of human figures, in order to allow to a motion analysis subsystem to detect illegal behaviours.

In this paper we concentrate solely on the pose estimation step, referring previous works about motion detection and object recognition steps.

In literature there are three classes of pose estimation approaches: Model-free, indirect model use and direct model use approaches.

Model free algorithms do not use an a-priori body model; marker points, object shapes and stick figures model based are used to recognize postures. In many works three points relative to centers of mass of head and hands have been used to model posture. In [1] these values have been calculated using color segmentation and blob segmentation techniques. In [2] boundary boxes have been used for modeling human pose; usually these are an intermediate representation during processing, while in the final representation shapes, as ellipses, more similar to human figures, could be used. Stick-figures based approaches include information similar to human skeleton, and are very diffuse where gait studying is necessary. In [3] stick-figures are obtained through axial transformations, while in [4] distance transformation is used; in particular, in this work, it is possible to remove not interesting body parts, reducing computational time. Both axial and distance transformations give an approximation of the human skeleton. A different approach, without using a-priori model, is based on learning a direct mapping between features and posture; low level information are passed to the system as training. In [5] an unsupervised neural network based algorithm is trained with low level joints features.

Indirect model use approaches make use of an a-priori model for pose estimation. Different kinds of model can be used; usually head, hands position, or a generic description of the global human body are used. In [6] a body model given by a simple ratio between the limbs of the figure has been implemented, obtaining good results. In [7] the body model is composed by projections of a silhouette; the algorithm is considerably simple, and gives good results. It searches predefined body parts position, then it uses these information to calculate silhouette projections; these are compared with the same values referring to four principal postures. The main problem of this work is the possibility that figures assumes a posture different from four principal ones. So, authors are working to pass from a supervised posture classification to an

unsupervised one. In [8] real time orientated histograms are calculated and compared with the same relative to a pre-recorded model in different posture: the estimated pose is that with more similarity with the available histograms model.

Direct model use approaches make use of an a-priori model that represents the observed subject. It is continuously updating using the observation of the figure, and gives information about posture at each time instant. About 40% of the methods proposed in literature use a similar approach. Employed models are usually very detailed, they have an autonomous life in the system, and they are very useful in presence of occlusions. The human body model is often represented by means of connected joints; so, it can be represented in a space of states in which every axis represents a joint. To relate these information with postures an *analysis by synthesis* approach is used. This technique is very expansive in terms of computational time, so many constraints has been introduced to reduce dimension of the space of states. Rohr [9] considers only postures parallel to the image plane; Ong e Gong [10] create a subspace by mapping information through PCA. Other interesting works for reducing complexity have been effected by Pavlovic [11] and Moeslund e Granum [12].

For comparing image information with synthetic ones, a lot of abstraction level can be used: edges ([13][14]), silhouettes ([15][16]), contours ([17]), stick ([18][19]), blob ([20]), and texture ([21]).

In the rest of the paper, firstly motion detection and object recognition steps are presented (section 2); then, an reliable and extremely simple approach for pose estimation is proposed (section 3). Finally, the experimental results obtained on real image sequences acquired on an archeological site, are reported (section 4).

2. OBJECT DETECTION AND RECOGNITION

A generic visual surveillance system is composed by a sequence of steps, strictly correlated with the results of each other.

In particular, the posture estimation subsystem is started when the recognition algorithm detects a human figure in the scene. The first step of the global system is the motion detection analysis; if this module detects a foreground object in the scene, an object recognition subsystem is used for recognizing its nature and identity.

Firstly, visual surveillance has to solve the problem of recognizing objects of interest, to be matched with given models, in a complex image. In our contest the objects of interest are both moving and static objects that differ from a background model. For this reason we have implemented an algorithm for objects detection that uses an adaptive background subtraction scheme. It is based on a statistical model of the background: for each pixel a running average and a form of standard deviation is maintained. So, in the test phase, a pixel is labeled as foreground if its intensity value differs from the running average two times more than the standard deviation [22]. Any background subtraction approach is sensitive to variations of the illumination conditions. To solve this problem, a frequent update of the information about the running average and the standard deviation of all background pixel values is necessary. This is done using an exponential filter; the implemented updating equations are described in [23]. After these steps, a reliable model of the background is available at each frame, so it is possible to extract a reliable shape of the objects of interest. But the results of the object detection subsystem described above cannot be used directly by the object recognition system since they present an undesirable source of noise: the shadows. Each foreground object contains its own shadow, because also this area effectively differs from the background. Shadows must be removed, because they radically change object shapes. Starting from the observation that shadows move with their own objects, but that they have not a fixed texture, the removing algorithm proposed in [23] has been implemented. Firstly, the image is segmented calculating the photometric gain for each pixel, then segments that present the same correlation calculated in the background image and in the current image are eliminated. Finally, an additional step is made in order to further simplify the following classification phase. All the moving blob with the area lower than an appropriate threshold are removed. This step allows to concentrate the attention only on the object of interest (cars, people, …). At this point each detected object is represented by a binary shape that can be provided as input to the classifier. In presence of many objects, each of them is considered separately by the moving detection algorithm. Each blob is detached from the other ones and separately analyzed. For each frame the algorithm extracts a number of binary images equals to the number of different objects detected in the scene.

After object detection, their classification must to be done. The features selected for the classification are an appropriate subset of the coefficients of the third level wavelet decomposition.

The wavelet transform decomposes the image in 'subbands' that are localized in frequency and orientation. A wavelet decomposition is performed by passing the image through a series of filter bank stages. In each stage low-pass filtering, high-pass filtering and downsampling are applied in order to obtain a multiresolution representation of the image. After the decomposition a feature reduction method based on the selection of the largest coefficients at the last level of decomposition (level 3) is performed. In this way, it is possible to accomplish object recognition using a few number of coefficients

having both temporal and frequency information. The regions detected as 'human' are the inputs of the pose estimation subsystem. Further information about recognition algorithm can be seen in [24].

In fig. 1 it can be seen the result of these algorithms.

Fig.1: Examples of the images produced by the motion detection and the object recognition subsystems.

3. POSTURE ESTIMATION

The proposed pose estimation subsystem works on binary images: that causes a loss of information but, on the other hand, makes the algorithm very fast and simple. Working on gray level images would enable a more detailed analysis and discrimination of different postures. For example, as proposed in [7], single body parts could be detected and tracked. We have chosen to avoid more complex system in order to design a very simple approach that could be easily integrated in a real time surveillance system without limitation.

At first, it is necessary to select the features that will be used for effectively represent the characteristics of each posture. Then, a similarity measure must be defined that will be used for building the prototypes of each posture and for their run-time classification. An unsupervised clustering approach has been used for building the prototypes on the base of the characteristics of the available data.

3.1. Features selection

The first problem to be solved when designing a pose estimation system is the selection of features allowing the correctly discrimination of different postures. In this way, a pose is represented by a point in a multidimensional feature space: the performance of the system are obviously strictly related to the selected features. A not accurate choice can make postures not separable in the multidimensional space, preventing any useful classification.

Therefore the selected features should have a large discriminatory capability and require the minimum computational time. A good trade-off between these requisites are the histograms of horizontal and vertical projections. Fig. 2 shows histograms relative to three different postures: a "standing" posture (upper image), a "bent" figure (the central one) and a "squatted" pose (lower image).

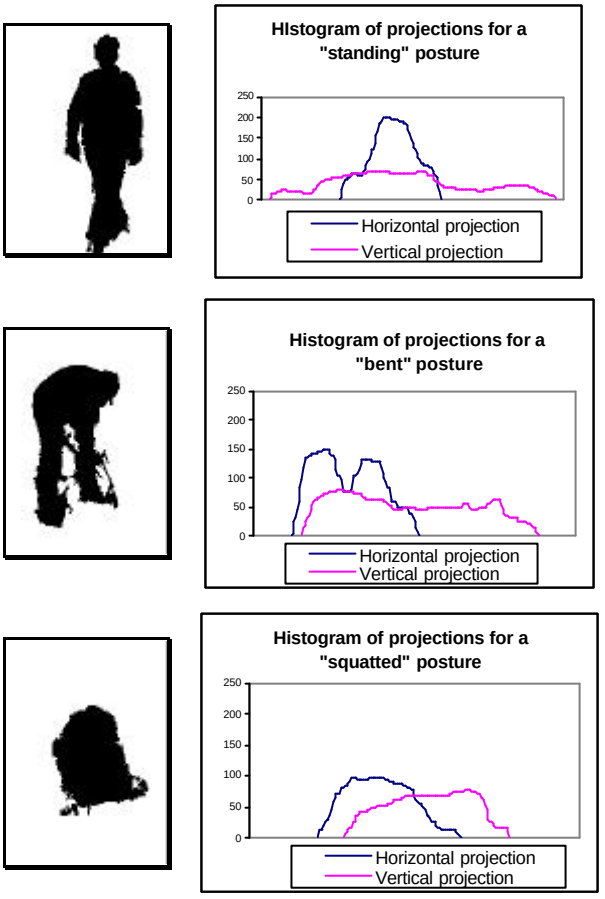

Fig. 2: The histogram projections for three different body postures

The traditional problem of this technique is its sensitiveness to scaling, translation and rotation. In our application this is not a problem: scaling is not present because the images are acquired with a pan-tilt-zoom camera, controlled in order to provide a constant dimension of the observed blob in each frame. This blob is also centered in the analyzed window making translation not relevant. Because people are usually standing rotations are practically negligible: different inclinations of the body correspond to different postures, so what remains of rotation is not a problem but an important information for the posture estimation.

In literature, histogram normalization is often used for coping with the dependence on scaling, to obtain more robust and reliable results. There are two common and extremely simple normalization: an area normalization

(every element is divided for the total area) and a maximum normalization (each element is divided for the maximum) both providing results in the range [0,1]. We have implemented both these techniques, but experimental results show the both of them are not relevant.

A good clustering of the static poses, requires histograms of horizontal and vertical projections to be well defined and with a low level of noise. Small flaws, caused for examples by holes in the binary shape, can change drastically the histograms making the whole process less reliable. In addition, the presence of noise degrades image quality and can affect the pose estimation. To reduce such artefacts, a smoothing operator has been implemented and applied. In particular, a median filter has been used: it is a non linear, spatial filter, and it is ideal to reduce salt and pepper noise. It has been applied to the histograms of projections with a size of 9.

3.2. Proximity measurement

The selected features allow to change the formulation of the original problem. Using the selected features, images become points in a feature space and pose recognition problem can be seen as a clustering problem.

Applying clustering algorithms, supervised or unsupervised, requires the definition of a metric as a way for estimating the "proximity" of two postures according to the selected features.

The selected proximity measurement is a variation of the Manhattan distance: in the original version, it is:

$$d_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{l} w_i \left| x_i - y_i \right| \qquad (1)$$

where **x** and **y** are the histograms, with values x_i, y_i, and w_i are the weight coefficients. Being no difference in significance between coefficient all the weights have been set to 1.

This kind of proximity measurement has the important advantage of being more simple and fast then other more complex distances (as the Euclidean one).

On the other hand, Manhattan distance is very sensitive to the geometrical position of the blob in the image: different position due to translation or mirroring with respect to the vertical axial can change its values up to making impossible to detect the similarity. So, a variation in the traditional Manhattan distance has been introduced to make it more robust to geometric condition. For the vertical projections, the Manhattan distance became:

$$d_1(Y1,Y2) = \min\left(\sum_{j=0}^{DimY-1} \left|Y1(j) - Y2(j+i)\right| \; ; \; \sum_{j=0}^{DimY-1} \left|Y2(j) - Y1(j+i)\right| \right) \qquad (2)$$

where the minimum is evaluated with *i* changing in the interval [0, DimY-1]. Obviously, the only geometrical variation during vertical axis is translation, because mirroring can affect only horizontal projections because mirroring with respect to the horizontal axis is very unlikely!

In fig. 3-4 the resulting histograms in presence of translation (3) and mirroring (4) are shown:

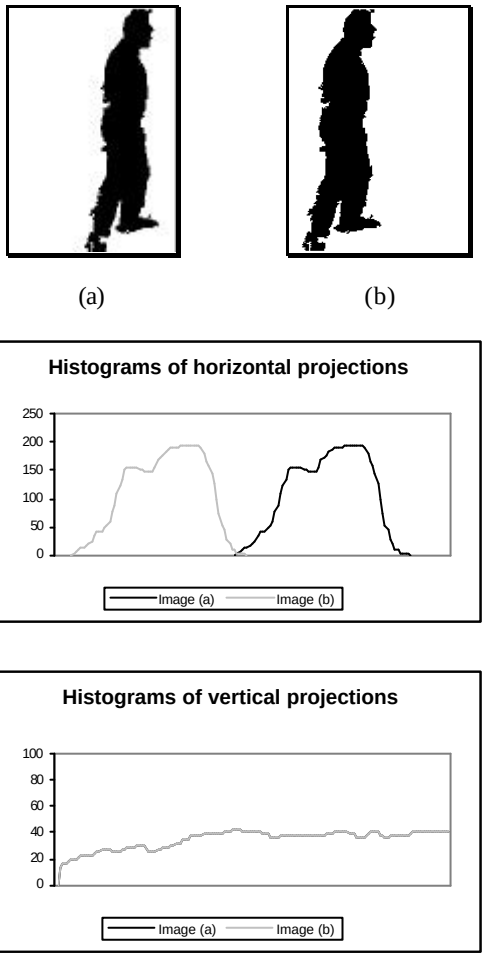

Fig. 3: In images (a) and (b) the same person takes on the same posture, but horizontal projections are different, due to the horizontal translation of the subject in the image. Note that vertical projections do not change.

So, for the horizontal projection, the modified distance will be:

$$d_1(X1,X2) = \min\left(\begin{array}{l} \sum_{j=0}^{DimX1} \left|X1(j) - X2(DimX1-j-i)\right| \; ; \; \sum_{j=0}^{DimX1} \left|X2(j) - X1(DimX1-j-i)\right| \\ \sum_{j=0}^{DimX1} \left|X1(j) - X2(j+i)\right| \; ; \; \sum_{j=0}^{DimX1} \left|X2(j) - X1(j+i)\right| \end{array} \right) \qquad (3)$$

According to equations (2) and (3), overall similarity measurement will be:

$$d_1(Im1,Im2)=a \bullet d_1(X1,X2)+b \bullet d_1(Y1,Y2) \quad (4)$$

where Im1, Im2 are the two images to be compared, and a, b are the coefficients for weighting the horizontal and vertical projections. After several experimental tests they have set to the values a=b=0.5.

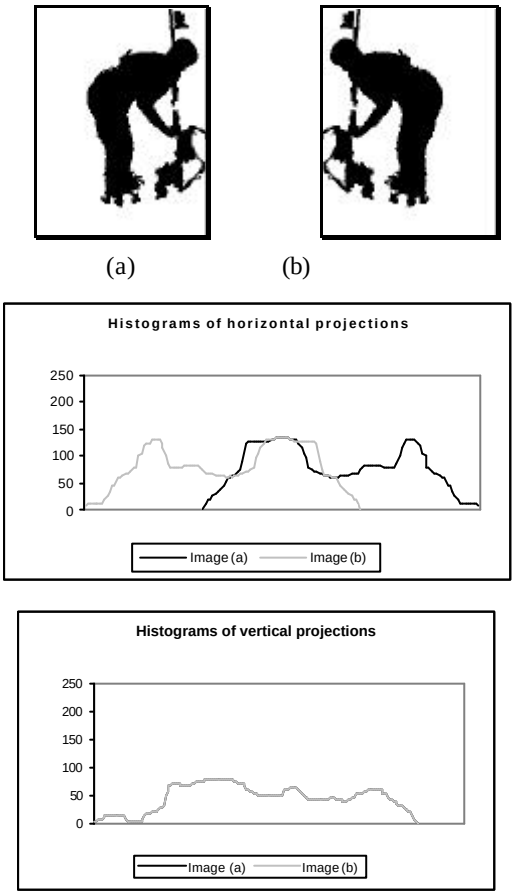

Fig. 4: In images (a) and (b) the same person takes on the same mirrored posture, but horizontal projections are different, due to the mirroring of the subject in the image. Note that vertical projections do not change.

3.3. Clustering

Clustering algorithms, using the proposed distance measurement, permits to group the available training images using the selected features. The effectiveness of resulting clusters depend on the selected algorithm, on the characteristics of the defined distance and on the selected features. There are two main classes of clustering algorithms: supervised and unsupervised. In the supervised algorithms the training data are already labeled while in unsupervised ones these information are not available. The implemented algorithm uses an unsupervised approach: in this way, no information about desired output is provided.

In literature several clustering techniques have been proposed: we have implemented and experimentally compared four different algorithms. Three of them are sequential (BSAS - Basic Sequential Algorithm Scheme, MBSAS - Modified Basic Sequential Algorithm Scheme, TTSAS - Two Threshold Sequential Algorithm Scheme) and one is based on a competitive scheme (BCLS - Basic Competitive Learning Scheme). We have applied all these algorithms to our data. The best results have been obtained using the BCLS. In addition, this algorithm does not require to set thresholds, so it is less sensitive to wrong initial setting than the other ones. During the training period a "representative" for each cluster is found. During the run-time classification the previously defined distance is applied to assign each image to the correct posture. The training phase is decisive to obtain good performance. The most important parameters for the BCLS algorithm are: the number of expected clusters, the quality of input images, the learning velocity during the training period, the number of training images, the "representative" selection step and the order of presentation for the training images.

4. EXPERIMENTAL RESULTS

The experiments have been performed on real image sequences acquired with a static TV camera Dalsa CA-D6; the frame rate selected is 30Hz. The resolution of original images is 528 X 512 pixels. After the Motion Detection and Object Recognition steps, the regions analysed by the module for Posture Analysis have a resolution of 155 X 215 pixels. The test sequences have been acquired in a real archeological site while the movements normally performed by intruders were simulated. The algorithms have been tested on 834 images selected from the acquired sequences. The computational time obtained on a AMD Athlon XP 1600+, 256 Mb SDRAM, HD 41 Gb, 7200 rpm shows that the proposed algorithms are fast enough to be suitable for a real time system.

At a direct visual inspection of images, three different postures can be identified:

Total number of images	Standing pose	Bent pose	Squatted pose
834	501	218	115

The characteristics of the test images emphasize a problem for every posture recognition algorithm: sometimes it can be difficult to univocally assign an image in a specific cluster; even for a human. An example of such ambiguities is provided in fig. 3.

An important input parameter for the BCLS algorithm is the expected number of clusters that must be fixed a-priori. We have done experimental tests looking for two or three output clusters, as it can be seen from tables 1 – 2. Another important parameter for the clustering algorithm is the value of learning velocity during the training period: experimental results have indicated a value of 0.2 as the one providing the best results.

Fig. 5: These images show that in some cases it is not easy to assign images to a specific posture: it is difficult to state if the man is bent or squatted.

Clustering algorithms are very sensitive to the order in which training data are given as input. In order to obtain reliable estimation of the performance of different algorithms and of the same algorithm when changing its parameters, a set of random sequence have been stored and used for feeding the training phases during tests. The result in the following tables show the mean of the corresponding value over several runs during which the input data have been examined in different orders.

% data set images used as training	2 clusters		
	% correct detection (100 run)	Standard Deviation	Computational time
20	**97.3**	2.4	9.88
40	**97.4**	1.5	10.37
100	**94.9**	4.9	12.75

Tab 2: Performance of the BCLS algorithm when 2 output clusters were expected, as a function of the training set dimension

% data set images used as training	3 clusters		
	% correct detection (100 run)	Standard Deviation	Computational time
20	**89.9**	5	15.23
40	**90.0**	4.4	15.87
100	**90.8**	7.1	16.11

Tab 2: Performance of the BCLS algorithm when 3 output clusters were expected, as a function of the training set dimension

As it can be seen, performance of the system does not significantly depend on the size of the training set. Computational time is not a critical problem: in fact with the proposed hardware architecture it is possible to process about 84 frame/sec. So, this subsystem can be integrated in a visual surveillance system working in real-time.

In previous experiments the output postures have been assigned to two or three different classes; but this information cannot be always available for the module. Moreover, it can be useful to have an information about the confidence of the system in its classification. So, we have calculated a "similarity coefficient" whose values provide the information about the similarity of the current posture with respect to the main postures: "standing" and "squatted". The training has been done only with features relative to these two postures. As it can be seen, in this case the results of the classification process are two coefficients, each of which indicates how similar the image is to the couple of reference postures. According with (2) and (3) the similarity coefficients are:

$$i(Im,"standing") = \frac{d_1(Im,"standing"\ r.)}{d_1(Im,"standing"\ r.) + d_1(Im,"squatted"\ r.)} \bullet 100 \qquad (5)$$

$$i(Im,"squatted") = \frac{d_1(Im,"squatted")}{d_1(Im,"standing"\ r.) + d_1(Im,"squatted"\ r.)} \bullet 100 \qquad (6)$$

where i(Im, "standing") is the similarity between current image and the pose "standing", and "standing"r. is the prototype of the pose "standing". The same stand for the pose "squatted".

The figure 6 shows these coefficients for some typical situations.

As it can be seen, in presence of a postures similar to a representative one (first and second rows, respectively "standing" and squatted" poses) the similarity coefficient values clearly identify the correspondent pose class; instead for an image in a bended posture, the similarity coefficient take on an intermediate value. This kind of classification provides a way for estimating the confidence of the classification.

Moreover, it allows a variable number of clusters to be obtained (changing an appropriate set of thresholds on the

output coefficients) without requiring a new and specific training phase.

Im	i(Im, "standing")	i(Im, "squatted")
Standing 	**77.51**	**22.49**
Squatted 	**16.37**	**83.63**
Bent 	**50.89**	**49.11**

Fig. 6: These coefficients express how much each image in the first column is similar to the extreme posture: standing and squatted

5. CONCLUSION AND FUTURE WORKS

In this paper a simple and reliable approach to the estimation of body postures in visual surveillance of archeological site has been implemented.

The selected features for pose estimation are the horizontal and vertical histograms of the binary shapes associated to humans. The Manhattan distance has been used for both cluster building and for run-time classification. After extensive experimental tests, an unsupervised clustering technique approach (the BCLS algorithm: Basic Competitive Learning Scheme) has been selected for the construction of classes. The whole approach has been verified on real sequences acquired while the typical illegal activities involved in stealing were simulated in an archeological site.

The study presented in this work shows that BCLS algorithm is well suited to clustering human poses starting from the binary image of the people in the scene.

The classification performance is very high although no a-priori knowledge of the scene has been used. Moreover the experiments realized on real image sequences acquired in an archaeological site show that the method involves low computational time and allows the system to work in real time.

The good results of this work are the starting point for the recognition of the human behaviours through stochastic models as HMM. For this purpose, algorithms has been implemented and tested to produce in output not only the clustered features, but also a similarity coefficient that suggests the affinity between the examined feature and two main reference poses.

6. REFERENCES

[1] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, "Pfinder: real-time tracking of human body", *IEEE Trans. Pattern Anal. Mach. Intelligence*, 19(7), pp. 780 – 785, July 1997.

[2] T. Darrell, P. Maes, B. Blumberg, A.P. Pentland, "A Novel Environment for Situated Vision and Behavior", in Proc. *IEEE Workshop on Visual Behaviors*, IEEE Computer Society Press, Seattle, 1994.

[3] A.G. Bharatkumar, K.E. Daigle, M.G. Pandy, Q. Cai, J.K. Aggarwal, "Lower limb kinematics of human walking with the medial axis transformation", in *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, pp. 70-76, November 1994.

[4] S. Iwasawa, K. Ebihara, J. Ohya, S. Morishima, "Real-time estimation of human body posture from monocular thermal images", in Proc. of 1999 IEEE Computer Society *Conference on Computer Vision and Pattern Recognition*, pp.15-20, 1997.

[5] R. Rosales, S. Sclaroff, "Learning and synthesizing human body motion and posture", in *The Fourth international Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000.

[6] J. Njastad, S. Grinaker, G.A. Storhaug, "Estimating parameters in a 2½D human model", *11th Scandivian Conference on Image Analysis* , Kangerlussuaq, Greenland, 1999.

[7] I. Haritaoglu, D. Harwood, L.S. Davis, "Ghost: A human body part labeling system using silhouettes", *Fourteenth International Conference on Pattern Recognition*, Brisbane, August 1998.

[8] W.T. Freeman, K. Tanaka, J. Ohta, K. Kyuma, "Computer vision for computers games", in *International Workshop on Automatic Face- and Gestrure-Recognition*, Zurich, Switzerland, 1995.

[9] Rohr, K., *Human Movement Analysis Based on Explicit Motion Models*, Kluwer Academic, Dordrecht/Boston, 1997.

[10] E.J. Ong, S. Gong, "Tracking hybrid 2D-3D human models from multiple views", in *International Workshop on Modeling People at ICCV'99*, Corfu, Greece, September 1999.

[11] V. Pavlovic, J.M. Rehg, T.J. Cham, K.P. Murphy, "A dymanic Bayesian network approach to figure tracking using learned dynamic models", in *International Conference on Computer Vision*, Corfu, Greece, September 1999.

[12] T.B. Moeslund, E. Granum, "3D human pose estimation using 2D-data and an alternative phase space representation", in *Workshop on Human Modeling, Analysis and Synthesis at CVPR*, Hilton Head Island, SC, June 2000.

[13] D. Hogg, "Model-based vision: A program to see a walking person", *Image and Vision Computing*, pp. 5-20, 1(1), February 1983.

[14] D.M. Gravila, L.S. Davis, "3D model-based tracking of humans in action: a multi-view approach", in *Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 1996.

[15] Y. Kameda, M. Minoh, K. Ikeda, "Three dimensional pose estimation of an articulated object from its silhouette image", in *Asian Conference on Computer Vision*, pp. 612-615, November 1993.

[16] C. Hu, Q. Tu, Y. Li, S.Ma, "Extraction of parametric human model for posture recognition using genetic algorithm", in *The fourth International Conference on Automatic face and Gesture Recognition*, Grenoble, France, March 2000.

[17] D. Meyer, J. Denzler, H. Niemann, "Model based extraction of articulated objects in image sequences", in *Fourth International Conference On Image Processing*, 3, pp. 78-81, Santa Barbara, CA, 1997.

[18] Y. Guo, G. Xu, S. Tsuji, "Tracking human body motion based on a stick figure model", *J. Visual Comm. Image Representation*, pp. 1-9, 1994.

[19] C.I. Attwood, G.D. Sullivan, K.D. Baker, "Model-based recognition of human posture using single synthetic images", in *Fifth Alvey Vision Conference*, University of Reading, UK, 1989.

[20] C.R. Wren, A.P. Pentland, "Understanding purposeful human motion", in *International Workshop on Modeling People at ICCV'99*, Corfu, Greece, September 1999.

[21] F. Lerasle, G. Rives, M. Dhome, "Human body limbs tracking by multi-ocular vision", in *Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, 1997.

[22] T.Kanade, T.Collins, A.Lipton, "Advances in Cooperative Multi-Sensor Video Surveillance", *Darpa Image Underst. Work*., Morgan Kaufmann, pp. 3-24, Nov. 1998.

[23] P. Spagnolo, A. Branca, G. Attolico, A. Distante: "Fast Background Modeling and Shadow Removing for Outdoor Surveillance", *Visualization Imaging and Image Processing*, IASTED VIIP, pag. 668-671, Sept. 2002.

[24] M. Leo, G. Attolico, A. Branca, A. Distante, "Object classification with multiresolution wavelet decomposition", in *Proc. of SPIE Aerosense 2002, conference on Wavelet Applications*, 1-5 April, 2002, Orlando, Florida, USA.