Probabilistic Spatio-temporal 2D-Model for Pedestrian Motion Analysis in Monocular Sequences[*]

Grégory Rogez[**], Carlos Orrite, Jesús Martínez[***], and J. Elías Herrero

CVLab, Aragon Institute for Engineering Research, University of Zaragoza, Spain
{grogez, corrite, jesmar, jelias}@unizar.es
http://www.cv.i3a.unizar.es

Abstract. This paper addresses the problem of probabilistic modelling of human motion by combining several 2D views. This method takes advantage of 3D information avoiding the use of a complex 3D model. Considering that the main disadvantage of 2D models is their restriction to the camera angle, a solution to this limitation is proposed in this paper. A multi-view Gaussian Mixture Model (GMM) is therefore fitted to a feature space made of Shapes and Stick figures manually labelled. Temporal and spatial constraints are considered to build a probabilistic transition matrix. During the fitting, this matrix limits the feature space only to the most probable models from the GMM. Preliminary results have demonstrated the ability of this approach to adequately estimate postures independently of the direction of motion during the sequence.

1 Introduction

In recent years, human motion analysis has grown to become one of the most active research areas in computer vision [1]. It has a wide spectrum of promising applications in many fields, especially in video-surveillance where the possibility of automatic video understanding and activity recognition would enable a single human operator to monitor wide areas. The most efficient systems are based on the use of a model [2], which is, most of the time, a representation of the human body. The election of an appropriate model is a critical issue. The use of an explicit body model is not simple, given the high number of degrees of freedom of the human body and the self-occlusions, direct consequences of the monocular observation. In previous works, the structure of human body has been represented as 2D or 3D Stick figure [3], 2D (Active) Contour or Shape [4] or 3D volumetric model [5]. The benefits from using a more sophisticated and appropriate model can be reduced or annihilated by poor parameter estimates.

In this paper we present a probabilistic 2D model for pedestrian motion analysis in monocular sequences. The disadvantage of 2D models is their restriction

[*] This work is supported by grant TIC2003-08382-C05-05 from Spanish Ministry of Science and Technology (MCyT) and FEDER.

[**] Funded by FPU grant AP2003-2257 from Spanish Ministry of Education.

[***] Supported by Spanish Ministry of Education under FPI grant BES-2004-3741.

F.J. Perales and R.B. Fisher (Eds.): AMDO 2006, LNCS 4069, pp. 175–184, 2006.

to the camera's angle. We therefore propose to construct 2D dynamical models independent of the orientation of the person with respect to the camera and that can respond robustly to any change of direction during the sequence.

To carry out this goal, we follow the methodology proposed by Bowden [6]. We construct a human model encapsulating within a Point Distribution Model (PDM) the information of the full body silhouette (given by the 2D Shape made of a series of landmarks located along the human contour) and the structural information (given by the corresponding 2D Stick figure). Both training and testing sets comprise of hand-labelled data. The CMU Mobo database [7] has been used for training and real video-surveillance sequences for testing.

The method is based on learning dynamical models. A series of local motion models is learnt by clustering the Stick figure subspace. Using this structure-based partitioning, correspondences between several different views of the same walking sequences are established. This leads to a clustering in the global Shape-Skeleton feature space where all the views considered are projected together. The different clusters correspond in terms of dynamic or view-point. We consider in this work the use of Gaussian Mixture Models (GMM) to cope with the problem of non-linearity of the model as proposed in various papers [8,9]. GMM are fitted to the total Shape-Skeleton training data using the Expectation Maximization (EM) algorithm [8,9]. Temporal and spatial constraints are considered to build a probabilistic transition matrix. This enables a frame to frame prediction of the most probable local models from the GMM that have to be considered.

Once the model has been generated (off-line), it can be applied (on-line) to real sequences. Given an input human blob provided by a motion detection algorithm, the model is fitted for inferring both body shape and posture.

The structure of the paper is as follows: in Sect. 2, we introduce probabilistic modelling. Model construction and fitting are respectively explained in Sect. 3 and Sect. 4. Results are presented in Sect. 5 and conclusions drawn in Sect. 6.

2 Probabilistic Modelling

Our Point Distribution Model (PDM) consists of 2D Shape landmarks concatenated with 2D Skeleton joints. The total space will be clustered following temporal approach (clusters C_j) as well as spatial approach (clusters R_j) as described in Section 3. The first one will partition the dynamic of the motion, and the second one, the direction of motion. The purpose of this probabilistic dynamic model is to obtain a transition matrix combining both constraints.

2.1 Markov Chain for Modelling Temporal Constraint

Following the standard formulation of probabilistic motion model [3], the temporal prior $p(S_t|S_{t-1})$ satisfies a first-order Markov assumption where the choice of the present state S_t is made upon the basis of the previous state S_{t-1}. In the same way, if we partition the state space into N clusters $\mathcal{C} = \{C_1, ..., C_N\}$, the conditional probability mass function defined as $p(C_j^t|C_k^{t-1})$ corresponds to the probability of being in cluster *j* at time *t* conditional on being in cluster *k* at time *t-1* [10]. A *N*x*N* State Transition Matrix (STM) that gives the probabilities density function (pdf) is then constructed, using the procedure described in [11]. Each cluster corresponds to a state in the Markov chain.

2.2 Modelling Spatial Constraint

In this paper, we introduce a novel spatial prior $p(D_t|D_{t-1,t-2,...t-\mathrm{m}})$ for modelling spatial constraint. It expresses the statement that D_t (the present direction of motion of the observed pedestrian in the image) can be predicted given his m previous directions of motion $(D_{t-1}, D_{t-2}, ..., D_{t-\mathrm{m}})$. In this approach, the continuous values of all possible directions of motion in the image plane are discretized. This leads to a discrete set of M particular directions of motion corresponding to M clusters $\mathcal{R} = \{R_1, ..., R_M\}$ in the feature space.

Let $\Delta_t = [R^t_{k_0}, R^{t\text{-}1}_{k_1}, ..., R^{t\text{-}m}_{k_m}]$ be the *m+1*-dimensional vector representing the sequence of the *m+1* cluster labels (denoted by k_i) up to and containing the one at time *t*. Note that some of these k_i labels might be the same. We call $p(R^t_j|\Delta_{t\text{-}1})$ the probability of being in R_j at time *t*, conditional on being in R_{k_1} at time *t-1*, in R_{k_2} at time *t-2*, etc. (i.e. conditional on the m preceding clusters). In this work, we consider a reasonable approach making this probability a normal distribution, with expected value equal to the local mean trajectory angle $\overline{\theta}_t$ and, variance calculated as a function of the sampling rate.

$$p(R^t_j|\Delta_{t\text{-}1}) = p(R^t_j|R^{t\text{-}1}_{k_1}, R^{t\text{-}2}_{k_2}, ..., R^{t\text{-}m}_{k_m}) \sim \mathcal{N}(\overline{\theta}_t, \sigma), \tag{1}$$

where $\overline{\theta}_t = \frac{1}{m+1}\sum_{\mathrm{i}=t}^{t\text{-}m}\theta_i$, being m a function of the sampling frequency.

2.3 Combining Spatial and Temporal Constraints

Let *T* be the *N*x*M* "Toroidal Transition Matrix" (TTM), whose columns represent the *N* temporal clusters and rows correspond to the *M* spatial clusters (See Fig.1). Thus the probability $p(C^t_j \cap R^t_r) = p(T^t_{j,r})$ denotes the unconditional probability of being in C_j and in R_r at time *t*.

The conditional spatio-temporal transition probability is therefore defined as $p(T^t_{j,r}|C^{t-1}_k, \Delta_{t\text{-}1})$, the probability of being in C_j and in R_r at time *t* conditional on being in temporal cluster *k* at time *t-1* and conditional on the m preceding spatial clusters. In this paper, the assumption is made that the two considered events, state and direction changes, are independent, even if it is not strictly true. Some comments about this assumption will be made in Sections 5 and 6. This leads to the following simplified equation:

$$p_{j,r} = p(T^t_{j,r}|C^{t-1}_k, \Delta_{t\text{-}1}) \propto p(C^t_j|C^{t-1}_k).p(R^t_r|\Delta_{t\text{-}1}). \tag{2}$$

The resulting *N*x*M* toroidal matrix is the Probabilistic Transition Matrix (PTM) that gives, at each time instant, the discrete probability density function (pdf). Its content can be visualized by converting it to grey scale image as will be shown in the Section 5.

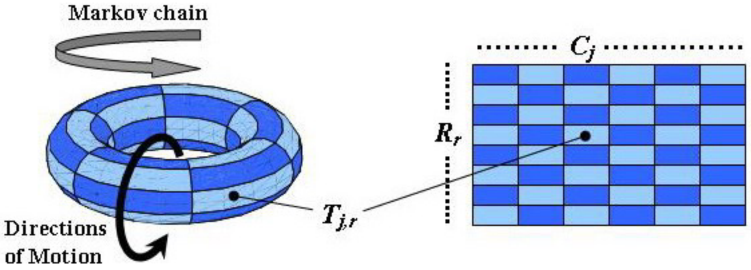


Fig. 1. 3D and 2D representations of the Toroidal Transition Matrix (TTM)

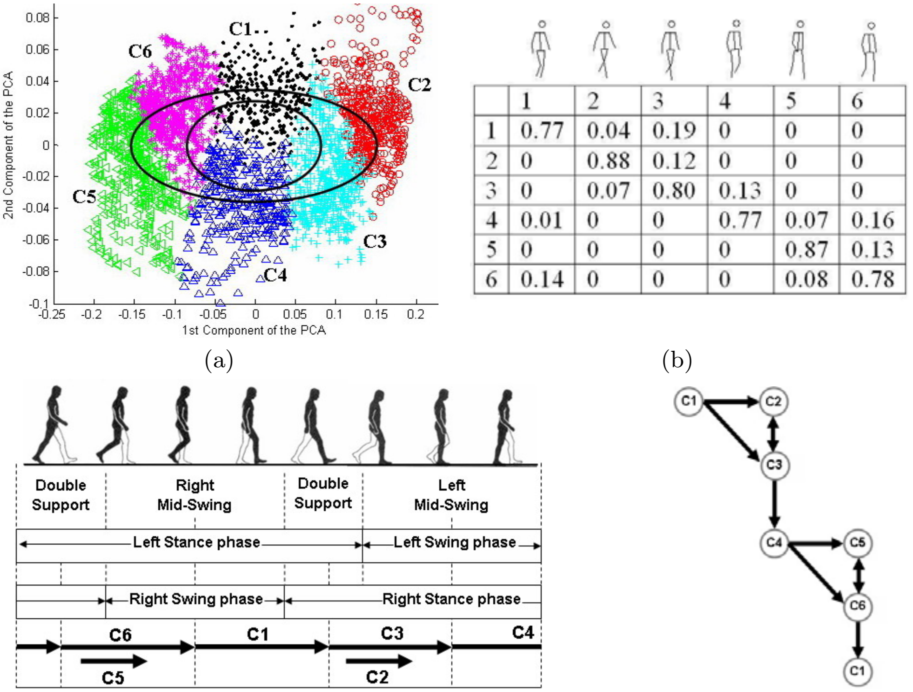
3 Pedestrian Models and Spatio-temporal Transitions

Once we have introduced some theoretical aspects, next we will describe the construction of the model and the transition matrix.

Training Data Base Construction. Precise training human shapes are extracted from Mobo database sequences [7] considering two walking speeds (high and low) and 8 different views (4 manual and 4 interpolated from the previous ones). These views directly provide the spatial clustering. Simultaneously, we labelled 13 fundamental points corresponding to a Stick model. By this process we generated a training database encompassing 21600 Shape-Skeleton vectors, SS-vector (2700 vectors for each different viewpoint).

Training Data Base Normalization. Reliable correspondences between members of the training set have to be established. The case of walking human silhouettes is a very difficult one since pedestrians take a very large number of different poses that affect the contour appearance. We propose to divide the contour into 4 segments (head, right arm, left arm and legs), delimited by a series of Fixed Points (FP), and assign them a fixed number of landmarks equally spaced. The FP are automatically selected with horizontal cutting lines placed at 1/3 and 2/3 of the height. The Shapes are normalized to 100 points. The training set (SS-vectors) is then aligned using Procrustes Analysis to avoid bad effects of position, size and rotation. and PCA is applied for dimension reduction.

Skeleton Clustering. The approach consists in clustering the training set using only the skeleton information that describes more adequately the dynamic of the motion. Thus 2D Skeletons corresponding to the 8 views are concatenated. In this way, the resultant vectors contain the 3D structural information. The set is then pre-processed by PCA and clustered by Kmeans. In this paper, we consider K=6 (when clustering presents the better visual aspect) and leave as suggestion for further research the determination of the optimal K. To make the clustering independent from the initial seeds, we run the K-means algorithm many times and proceed to cluster the results. This leads to the recognition of basic gait cycle phases [12], as illustrated by Fig.2, in an unsupervised way. The patches are ordered according to the logic of the cyclic motion: C1 starts with the Right Mid-Swing and ends with the double support phase, then C3 starts until the



	1	2	3	4	5	6
1	0.77	0.04	0.19	0	0	0
2	0	0.88	0.12	0	0	0
3	0	0.07	0.80	0.13	0	0
4	0.01	0	0	0.77	0.07	0.16
5	0	0	0	0	0.87	0.13
6	0.14	0	0	0	0.08	0.78

(a) (b)

(c) (d)

Fig. 2. (a) Skeleton Clustering in the PCA-plane defined by 1^{st} and 2^{nd} components, typical short and long cycles can be observed. (b) Markov State Transition Matrix. (c) Correspondences between Gait cycle and the 6 clusters obtained, (d) State Diagram.

Left Mid-Swing. C4 follows until the second double support of the cycle which ends with C6. C2 and C5 complete C3 and C6 phases in case of a higher speed gait with larger steps. A Markov State Transition Matrix (STM) [9,11] is then constructed (Fig.2b), associating each sample to one of the 6 patches. This gives the state transition probabilities, valid for the 8 sets (views) of SS-vectors.

Shape-Skeleton Gaussian Mixture Model (GMM). The SS-vectors corresponding to the 8 views are grouped following the cluster labels previously obtained, leading to 8 x 6 = 48 clusters in the global SS-PCA space. Following the procedure of [9] a GMM is fitted to the Data by applying EM. Local PCAs are then applied on each cluster [6] leading to the extraction of local modes of variation, in which both Shape and Skeleton deform (see Fig. 3).

Toroidal Transition Matrix. All the different models are ordered and classified according to the direction of motion and the states. This process leads to the creation of the Toroidal Transition Matrix (TTM) which 2D representation is illustrated in Fig.3: the 6 columns correspond to the 6 temporal clusters C_i while the 8 rows represent the 8 spatial clusters R_i. Spatial and temporal relations can be appreciated between local models from adjacent cells.

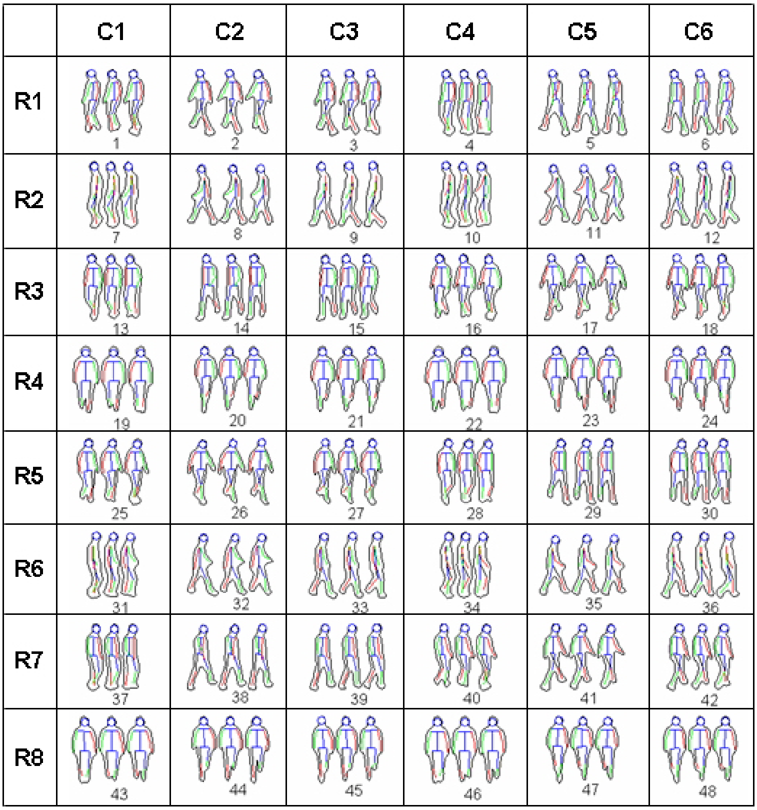


Fig. 3. Toroidal Transition Matrix: 1st Variation Modes of the 48 local Models

4 Model Fitting for Body Pose Inferring

Given an input human blob provided by a motion detection algorithm and the previous m states (poses and trajectory angles), the prediction of the most probable models from the GMM can be estimated by means of the PTM defined in Sect 2.3. It allows a substantial reduction in computational cost since only few models have to be considered. Assuming we have an initial estimate for the Shape parameters the matching process follows these steps:

1. A Shape S is extracted from the blob, looking along straight lines through each model point, following the methodology presented in [8].
2. S and an estimate for the Skeleton (e.g. initially mean Skeleton $\overline{K}$) are concatenated in $V = [S\overline{K}]$ and projected into the SS-PCA obtaining X.
3. Find the nearest cluster by calculating the distance between X and each one of the most probable clusters given by the PTM.
4. Update the parameters to best fit the “local model” defined by its mean $\overline{X}$, eigenvectors Φ and eigenvalues λ_i ,obtaining X^* [8].

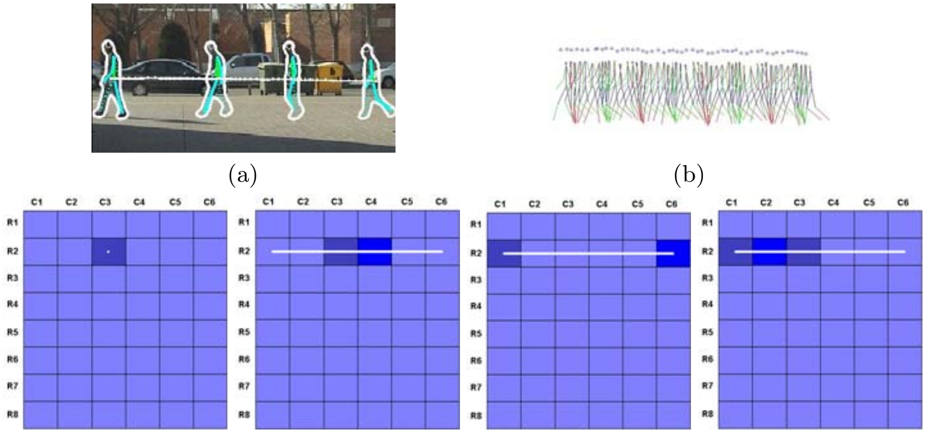
5. We project the vector X^* back to the feature space obtaining V^* which contains a new estimation of both Shape S^* and Skeleton K^*: $V^* = [S^* K^*]$.
6. A new background subtraction with an adaptive threshold inside the Contour S^* is applied, leading to an improved human blob detection.
7. Repeat until convergence and store useful data: θ_t, $T^t_{j,r}$, S^*_t and K^*_t.

This leads to an accurate silhouette segmentation and posture estimation directly obtained from the mapping created between Contours and Stick figures.

5 Results

The model is now evaluated with a series of testing sequences that illustrate different situations which may occur in the analysis of pedestrian motion: straight line walking, changes of direction, of speed, etc. Since we want to test both model fitting and pose estimation, and not the tracking in the image, we provide the system with the bounding-box manually selected avoiding the possible problems due to the tracking. The process begins with a manual initialization: indicating the adequate model in the first frame. In the PTM matrices from Fig.4, 6 and 7, the colored cells represent the probability $p_{j,r}$ from (2). The obscured cell is the "winning one": the local model that best fits the silhouette. For each frame, the row of the "winning" model in the TTM indicates the orientation of the pedestrian with respect to the camera. Additionally, both trajectory and previous states are respectively plotted in the image/matrix with a white line.

As illustrated in Fig.4, the resultant vectors from a pedestrian crossing the scene straight ahead without stopping or turning towards anything all belong to



(a) (b)

(c) (d) (e) (f)

Fig. 4. (a) Outdoor straight line walking sequence at constant speed. (b) Estimated Stick figures. (c, d, e & f) PTM corresponding to the 4 silhouettes depicted in (a).

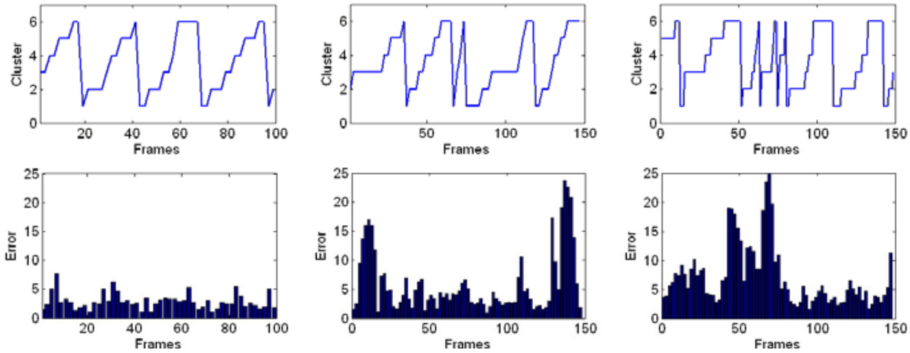


Fig. 5. Feet position error in pixels (bottom) and temporal clusters (top) - given by the column of the TTM corresponding to the "winning" model - of the Straight line walking (left), Indoor (centre) and "Walk-circle" (right) sequences

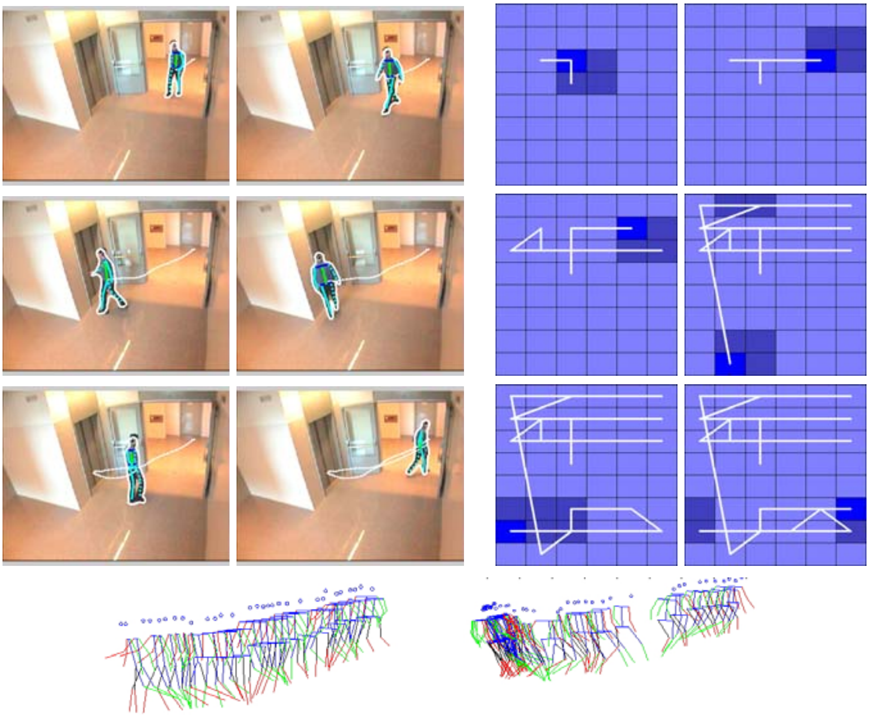


Fig. 6. Indoor sequence with orientation changes and estimated Stick figures

models from the same row of the TTM. Any change of direction is observed as a progressive change of row (See Fig.6 and Fig.7).

Fig.5 shows the pose estimation results for the 3 tested sequences. The mean position error (in pixels) is calculated as the feet-distance between the Skeleton

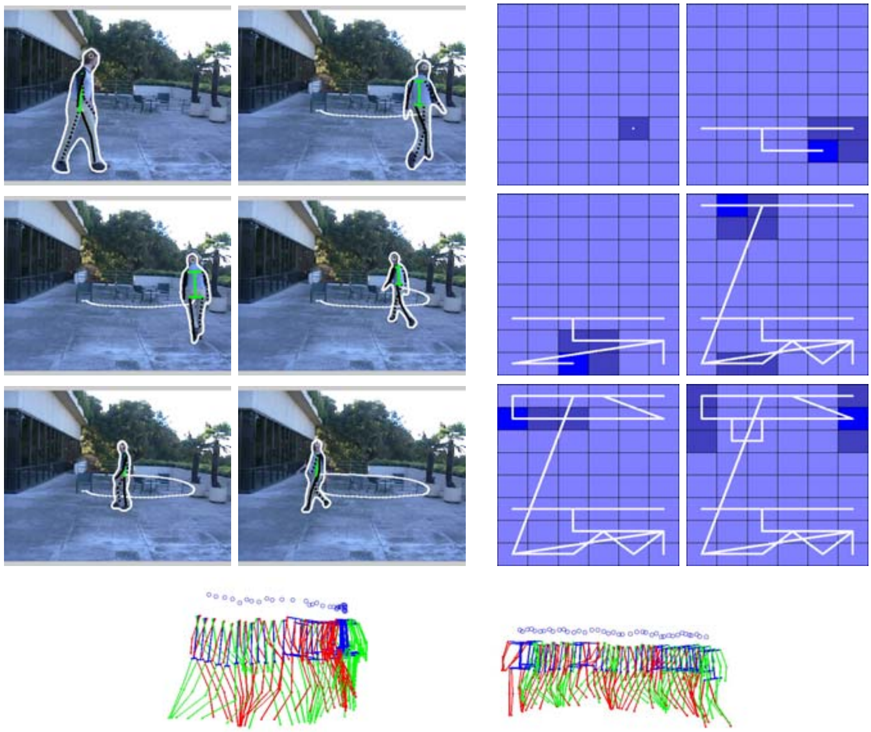


Fig. 7. "Walk-circle" sequence from www.nada.kth.se/∼hedvig/data.html and estimated Stick figures

estimated by the algorithm and the hand-labelled one. Some peaks can be noticed in this figure. For instance, in the indoor sequence (centre) the model failed because of the excessive difference of viewpoint-angle between training and input images, when the subject goes in and out of the scene. In the "walk-circle" sequences (right) the model fails because of the stationary behaviour of the tracking that stays stuck in a cluster during too many frames and then can hardly get out of it. It needs to wait until the next cycle to recuperate the dynamic behaviour of the input motion. For the rest of the frames, the results are globally very satisfactory which means that the model is conveniently tuned to the suitable viewpoints and that the assumption made in Section 2.3 is reasonable.

6 Conclusions and Work-in-Progress

This paper describes new probabilistic spatio-temporal models for human motion analysis. Temporal and spatial constraints are considered to build a Probabilistic Transition Matrix (PTM) that gives a frame to frame prediction of the most probable models from a multi-view GMM.

The proposed fitting algorithm, combined with the new probabilistic models, allows a faster and more reliable estimation of both pedestrian Silhouette and

Stick figure in real monocular sequences. Preliminary results have demonstrated that it works independently of the direction of motion in the image, and that it also responds quite robustly to any change of direction during the sequence. However, further work must be done.

For instance, the fitting process has been initialized providing a good model in the first frame. In order to develop a non-supervised system an automatic initialization has to be considered. Moreover, we have made the assumption that temporal and spatial events are independent. In future research this assumption have to be evaluated in detail since it is not strictly true: a pedestrian can change direction only during the second part of the Swing phases of the gait cycle.

Future work relies on combining this approach with a particle filtering framework in order to obtain a robust human motion tracker in feature space. On the other hand, a perspective correction could be applied to avoid the problem of viewpoint correspondences. Finally, more complicated cases such as various pedestrians with partial occlusions will be considered, and others kinds of motion should be taken into account in more complete models that could be built synthetically from a 3D motion capture system.

References

1. Wang, L., Hu, W., Tan, T.: Recent developments in human motion analysis. Pattern Recognition **36** (2003) 585–601
2. Kakadiaris, I.A., Metaxas, D.N.: Model-based estimation of 3d human motion. IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 1453–1459
3. Sidenbladh, H., Black, M.J., Sigal, L.: Implicit probabilistic models of human motion for synthesis and tracking. In: ECCV (1). (2002) 784–800
4. Baumberg, A., Hogg, D.: Learning flexible models from image sequences. In: ECCV (1). (1994) 299–308
5. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: CVPR (1). (2003) 69–76
6. Bowden, R., Mitchell, T.A., Sarhadi, M.: Reconstructing 3d pose and motion from a single camera view. In: BMVC. (1998)
7. Gross, R., Shi, J.: The cmu motion of body (mobo) database, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2001)
8. Cootes, T., Taylor, C.: A mixture model for representing shape variation. (1997)
9. Ponsa, D., Roca, F.X.: A novel approach to generate multiple shape models for tracking applications. In: AMDO. (2002) 80–91
10. Bowden, R., Sarhadi, M.: Building temporal models for gesture recognition. In: BMVC. (2000)
11. Heap, T., Hogg, D.: Wormholes in shape space: Tracking through discontinuous changes in shape. In: ICCV. (1998) 344–349
12. Inman, V.T., Ralston, H.J., Todd, F.: Human Walking. Williams and Wilkins, Baltimore, USA (1981)