ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

PATTERN RECOGNITION

THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/pr

A spatio-temporal 2D-models framework for human pose recovery in monocular sequences

Grégory Rogez*, Carlos Orrite-Uruñuela, Jesús Martínez-del-Rincón

CVLab, Aragon Institute for Engineering Research, University of Zaragoza, Spain

Received 6 July 2007; received in revised form 23 November 2007; accepted 25 February 2008

Abstract

This paper addresses the pose recovery problem of a particular articulated object: the human body. In this model-based approach, the 2D-shape is associated to the corresponding stick figure allowing the joint segmentation and pose recovery of the subject observed in the scene. The main disadvantage of 2D-models is their restriction to the viewpoint. To cope with this limitation, local spatio-temporal 2D-models corresponding to many views of the same sequences are trained, concatenated and sorted in a global framework. Temporal and spatial constraints are then considered to build the probabilistic transition matrix (PTM) that gives a frame to frame estimation of the most probable local models to use during the fitting procedure, thus limiting the feature space. This approach takes advantage of 3D information avoiding the use of a complex 3D human model. The experiments carried out on both indoor and outdoor sequences have demonstrated the ability of this approach to adequately segment pedestrians and estimate their poses independently of the direction of motion during the sequence.

Keywords: Human motion analysis; Human shape modelling; Pose inference

1. Introduction

Human motion capture and analysis has grown to become one of the most active research topics in computer vision over the past decade [1]. This is mainly motivated by the wide spectrum of promising applications in many fields such as video-surveillance, human–machine interfaces, medical diagnosis, sports performance analysis or biometrics.

The human motion analysis divides into three main interacting levels as described in Ref. [2]: human detection, human tracking and human behavior understanding. The detection stage that aims at segmenting people from the rest of the image is a significant issue since the performance of the other two processes highly depends on it. Human activity understanding relies on accurate detection and tracking, but a good prior knowledge of pose can also improve considerably both detection and tracking.

Many efficient systems are based on the use of a model which is, most of the time, a representation of the human body. In previous works, the structure and appearance of the human body have been represented as 2D or 3D stick figure [3], 2D (active) contour or shape [4–6], binary silhouette [7] or 3D volumetric model [8,9]. The selection of the appropriate model is a critical issue and 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 inherent to the monocular observation.

People are able to deduce the pose of a known articulated object (e.g. a person) from a simple binary silhouette. The possible ambiguities can be solved from dynamics when the object is moving. Following this statement, the first step of this work consists in constructing a human model that encapsulates within a point distribution model (PDM) [10] both body silhouette information provided by the 2D-shape and structural information given by the 2D skeleton joints. In that way, the 2D pose could be inferred from the silhouette and vice versa. Due to the high non-linearity of the resulting feature space, mainly caused by the rotational deformations inherent to the articulated structure of the human body, the use of non-linear statistical models will be considered in this work. This approach will be compared to

* Corresponding author. Tel.: +34 635 983 614.
E-mail addresses: grogez@unizar.es (G. Rogez), corrite@unizar.es (C. Orrite-Uruñuela).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.02.012

other two methods previously tested for solving non-linearity issue. Such non-linear statistical models have been previously proposed by Bowden [11] that demonstrated how the 3D structure of an object can be reconstructed from a single view of its outline. While Bowden only considered the upper human body and the frontal view, in this work the complete body will be modelled and viewpoint changes will be taken into account.

One of the difficulties when employing 2D-models relies on dealing with this viewpoint issue. Most of the previous related works are based on the fundamental assumption of "in-plane" motion or only present results obtained from data satisfying such condition [12]. Few consider motion-in-depth and out-of-plane rotation of the tracked people. Freeing algorithms from the view dependency appears as a critical issue for practical applications. Therefore, the goal of this work is to construct 2D dynamical models that can perform independently of the orientation of the person with respect to the camera and that can respond robustly to any change of direction during the sequence.

1.1. Related work

There are basically two main schools of thought on human pose recovery: model-based *top-down* approaches and model-free *bottom-up* strategies. Model-based approaches presuppose the use of an explicit model of a person's kinematics [9,13]. The number of degrees-of-freedom and the high dimensionality of the state space make the tracking problem computationally difficult. Recent research has investigated the use of learnt models of human motion to constraint the search in state space by providing strong priors on motion [12,14,15]. In *bottom-up* strategies, the individual body parts can be detected and then probabilistically assembled to estimate the 2D pose as in Ref. [16] or an example-based method can be followed. This last method consists in comparing the observed image with a database of samples as in Refs. [17–19] to cite a few. In some cases, a mapping from 2D image space to 3D pose space is learnt for directly estimating the 3D pose [20–22]. Instead of storing and searching for similar examples, Agarwal and Triggs [20] use non-linear regression of joint angles against shape descriptor vectors to distill a large training database into a compact model. Grauman et al. [22] inferred the 3D structure from multi-view contour using a probabilistic "shape+structure" model. As mentioned before, this idea was first introduced by Bowden [11].

Shape-models have appeared as powerful tools for human motion analysis. Baumberg and Hogg [4] used active shape models to track pedestrians from a fixed camera. The same active shape tracker was considered by Siebel and Maybank [6] that extended it by a head detector and a region tracker, all integrated in the visual surveillance system ADVISOR. Fan et al. presented in Ref. [23] a compound structural and textural image model for pedestrian registration. In Ref. [24], the authors exploit the shape deformations of a person's silhouette as a discriminative feature for gait recognition, indicating that methods based on shape perform better than methods based on kinematics alone. Giebel et al. [25] proposed a Bayesian framework for tracking pedestrians from a moving vehicle: a method for learning spatio-temporal shape representations from examples was outlined, involving a set of distinct linear subspace models. Recently, Zhang et al. [26] introduced a statistical shape representation of non-rigid and articulated body contours. To accommodate large viewpoint changes, a mixture of a finite number of view-dependent models is employed.

1.2. Overview of the work

This paper presents a novel probabilistic spatio-temporal 2D-models framework (STMF) for human motion analysis. In this approach, the 2D-shape of the entire body is associated to the corresponding stick figure allowing the joint segmentation and pose recovery of the subject observed in the scene. The first step of this work, described in Section 2, thus relies on the construction of the "shape–skeleton" training data set: contour parameters are associated to the corresponding 2D joints extracted from many different training views of the same walking sequences (varying azimuth angle of the camera).

The framework construction is then detailed in Section 3. First, a novel technique is presented for shape clustering that establishes dynamics correspondences between the different training views. Basically, a structure-based clustering of the training shapes is achieved by partitioning the 3D pose parameters subspace, thus dividing the gait cycle into a series of basic steps. The resulting labelling is then used to construct the non-linear models in each training view where a mixture of PCA models is learned using the expectation maximization (EM) algorithm [10,27,28], the clusters being used at initialization. The method is compared to other two approaches previously developed to deal with non-linearity: nearest neighbor (NN) classifier and independent component analysis (ICA).

Using the motion-based partitioning and the spatial clustering directly provided by the training views, a *spatio-temporal clustering* is obtained in the global shape–skeleton eigenspace: the different clusters correspond in terms of dynamic (temporal clusters) or viewpoint (spatial clusters). A local 2D-model is then built for each spatio-temporal cluster, generalizing well for a particular training viewpoint and state of the considered action. All those models are concatenated and sorted, what leads directly to the construction of the global STMF presented in Fig. 1.

Given this huge amount of data, an efficient search method is needed. In that way, *temporal* and *spatial constraints* are considered to build a probabilistic transition matrix (PTM). This matrix limits the search in the feature space by giving a frame to frame estimation of the most probable local models to be considered during the fitting procedure. This constraint-based search is described in Section 4.

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 background subtraction, the model is fitted to jointly segment the body silhouette and infer the posture. This model fitting is explained in Section 5.

Experiments are presented in Section 6 where both segmentation and 2D pose estimation are tested. The main goal of

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

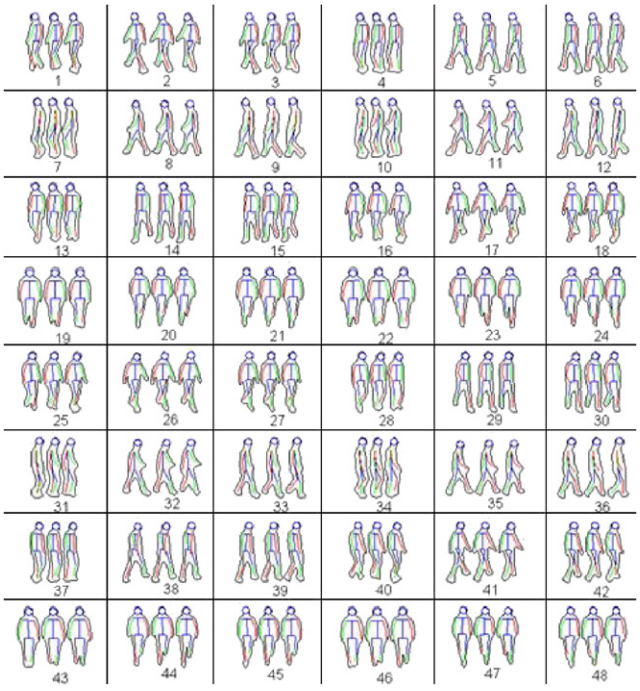


Fig. 1. Spatio-temporal shape–skeleton models framework: 1st variation modes of the 48 local models that compose the framework. The columns of this matrix correspond to the gait steps (temporal clusters) while the rows represent the eight camera views (spatial clusters).

this part is to test the robustness of the approach w.r.t. the viewpoint changes with realistic conditions: indoor, outdoor, cluttered background, shadows, etc. In that way, the following hypothesis will be considered to select the different testing sequences: only one walking pedestrian per sequence, with no occlusions but with important viewpoint changes. Note that both training and testing sets comprise of hand-labelled data. The CMU MoBo database [29] will be used for training and real video-surveillance sequences for testing [30]. The HumanEVA data set, recently introduced by Sigal and Black [31], will be considered for numerical evaluation of the pose estimation.

Section 7 finally concludes with some discussions and ideas for future work.

2. Shape–skeleton training database

The goal is to construct a statistical model which represents a human body and the possible ways in which it can deform. Point distribution models (PDMs) are used to associate silhouettes (shapes) and the corresponding skeletal structures.

2.1. Training database construction

The generation of the 2D deformable model follows a procedure similar to [32]. CMU MoBo database [29] is considered for the training stage: good training shapes are extracted manually trying to get accurate and detailed approximations of human contours. Simultaneously, 13 fundamental points corresponding to a stick model are extracted: head center, shoulders, elbows, wrists, hips, knees and ankles. The skeleton vectors are then defined as

$$\mathbf{k}_i = [x_{k1}, \ldots, x_{k13}, y_{k1}, \ldots, y_{k13}]^\top \in \Re^{26}, \quad (1)$$

with $i = 1, \ldots, N_v$, N_v being the number of training vectors. Two gait cycles (low and high speed) and four viewpoints (frontal, lateral, diagonal and back views) are considered for each one of the 15 selected subjects. This manual process leads to the generation of a very precise database but without shape-to-shape landmark correspondences.

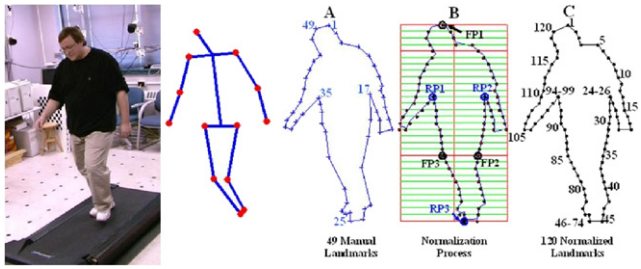


Fig. 2. From left to right: MoBo image, 2D skeleton and shape normalization: (A) hand-labelled landmarks, (B) rectangular grid and (C) 120 normalized landmarks, part of them grouped at "repository points": 24–26 at RP2, 46–74 at RP3 and 94–99 at RP1.

2.2. Shapes normalization

The good results obtained by a PDM depend critically on the way the data set has been normalized and on the correspondences that have been established between its members [33]. Human silhouette is a very difficult case since people can take a large number of different poses that affect the contour appearance. A big difficulty relies on establishing correspondences between landmarks and normalizing all the possible human shapes with the same number of points. In this work, it is proposed to use a large number of points for defining all the contours and "superpose" those points that are not useful (see Fig. 2).

A rectangular grid with horizontal lines equally spaced is applied to the contours database. This idea appears as a solution to the global verticality of the shapes and the global horizontality of the motion. The intersections between contours and grid are then considered. The shapes are then divided into three different zones delimited by three fixed points: the higher point of the head (FP1) and the intersections with a line located at $\frac{1}{3}$ of the height (FP2 and FP3). A number of landmarks is thus assigned to each segment and a repository point (RP) is selected to concentrate all the points that have not been used. In this paper, all the training shapes are made of 120 normalized landmarks:

$$\mathbf{s}_i = [x_{s1}, \ldots, x_{s120}, y_{s1}, \ldots, y_{s120}]^\top \in \Re^{240}, \quad (2)$$

with $i = 1, \ldots, N_v$.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

2.3. Shape–skeleton eigenspace—PCA model

Shapes and skeletons are now concatenated into shape–skeleton vectors:

$$\mathbf{v}_i = [\mathbf{s}_i^\top \ \mathbf{k}_i^\top]^\top \in \Re^{266}, \tag{3}$$

with $i = 1, \ldots, N_v$. This training set is aligned using Procrustes analysis (each view being aligned independently) and principal component analysis (PCA) is applied [10] for dimensionality reduction on the four view-based training sets. In that way, four view-dependent shape–skeleton models are constructed by extracting the mean vector and the variation modes:

$$\mathbf{v}_i \simeq \bar{\mathbf{v}}_\theta + \mathbf{\Phi}_\theta \mathbf{b}_i, \tag{4}$$

where $\bar{\mathbf{v}}_\theta$ and $\mathbf{\Phi}_\theta$ are, respectively, the mean shape–skeleton vector and the matrix of eigenvectors for the training viewpoint θ. $\mathbf{b}$ is the projection of $\mathbf{v}_i$ in the corresponding eigenspace i.e. a vector of weights $\mathbf{b}_i = [b_1, b_2, \ldots, b_n]^\top$. The main problem is that the PCA assumes a Gaussian distribution of the input data. This supposition fails because of the inherent non-linearity of the feature space and leads to a wrong description of the data: the resulting model can consider as valid some implausible shape–skeleton combinations. Other approaches have to be taken into account to generate the "shape–skeleton" model and adequately represent the training set.

3. Spatio-temporal 2D-models framework

Many researchers have proposed approaches to non-linear PDM [10,11]. The use of Gaussian mixture model (GMM) was first proposed by Cootes and Taylor [10]. They suggested modelling non-linear data sets using a GMM fitted to the data using the EM algorithm. This provides a more reliable model since the feature space is limited by the bounds of each Gaussian that appear to be more precise local constraints:

$$p_{\text{mix}}(\mathbf{b}) = \sum_{j=1}^{m} \omega_j \mathcal{N}(\mathbf{b} : \bar{\mathbf{b}}_j, \mathbf{S}_j), \tag{5}$$

where $\mathcal{N}(\mathbf{b} : \bar{\mathbf{b}}, \mathbf{S})$ is the p.d.f. of a Gaussian with mean $\bar{\mathbf{b}}$ and covariance $\mathbf{S}$.

Bowden [11] proposed first to compute linear PCA and to project all shapes on PCA basis. Then do cluster analysis on projections and select an optimum number of clusters. Each data point is assigned to cluster and separate local PCA are performed independently on each cluster. This results in the identification of local model's modes of variation inside each Gaussian distribution of the mixture: $\mathbf{b} \simeq \bar{\mathbf{b}}_j + \mathbf{\Phi}_j \mathbf{r}$ (see Eq. (4)). Thus a more complex model is built to represent the statistical variations. Given the promising results described in Ref. [11], a similar procedure is followed in this work, the main difference relying on the way the feature space is clustered: the proposed methodology consists in partitioning the complete shape–skeleton feature space using only the dynamical information provided by the pose parameters. The contour parameters are not taken into account for clustering since they do not provide any additional information on dynamics and can lead to ambiguities as stated in Ref. [20].

3.1. Structural clustering

While in Ref. [25], the clustering of the shape feature space was based on a similarity measure derived from the registration procedure, here it is proposed to use the structural information provided by the pose to cluster both shape and skeleton training sets, thus establishing dynamical correspondences between view-based data.

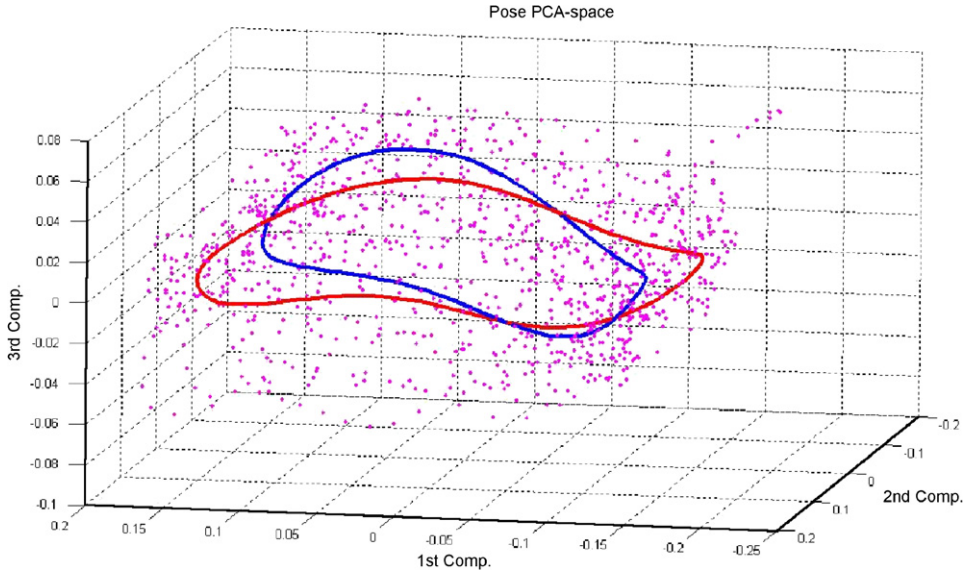


Fig. 3. Low and high speed gait cycles represented on the three first modes of the pose eigenspace.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

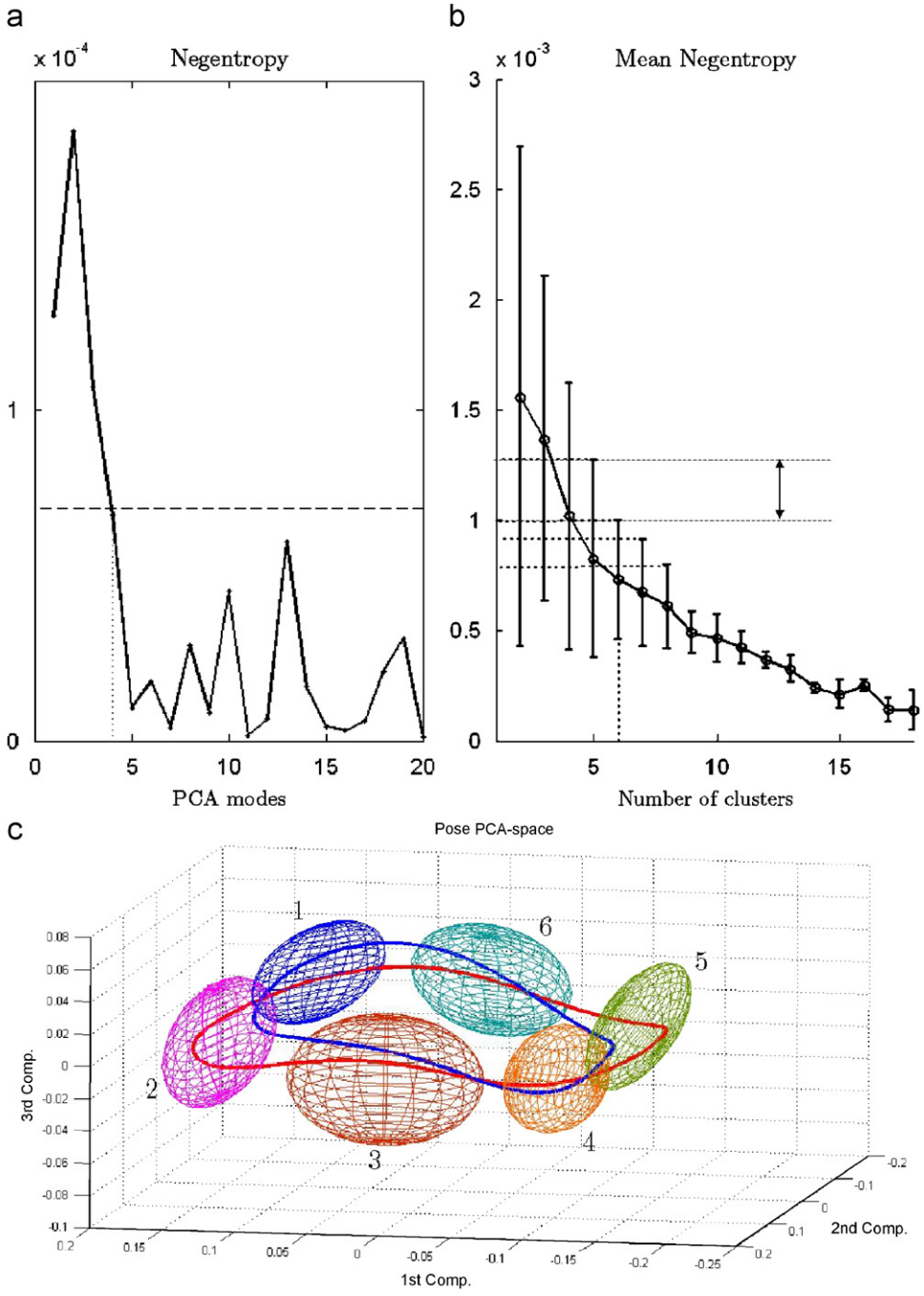


Fig. 4. (a) Negentropy of the 20 first modes of the pose eigenspace, (b) Negentropy of the GMM (mean and std. dev.) vs. number of clusters and (c) resulting GMM for $k = 6$, represented in the pose eigenspace together with the gait cycles.

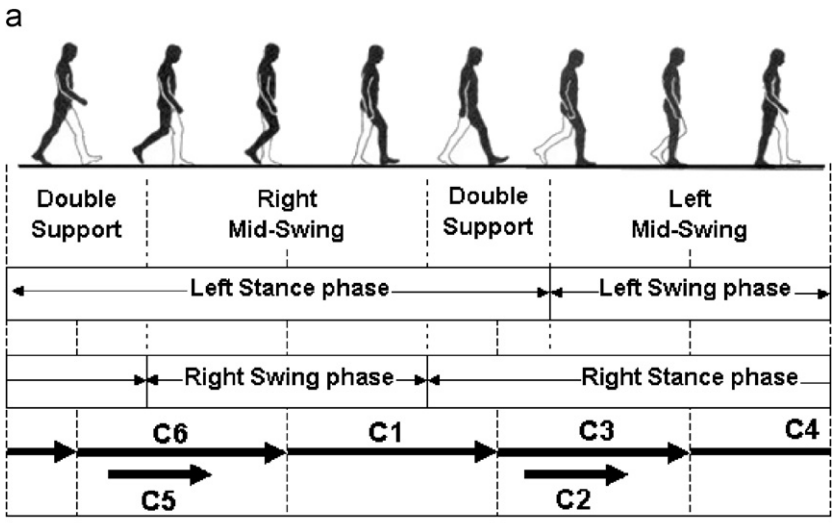
3.1.1. Pose eigenspace for clustering

The information provided by the 3D poses is used for clustering: for each snapshot of the training set, the 3D skeleton is built from the corresponding 2D poses $\mathbf{k}_i$ of the four views, by reconstructing the 3D position of the joints using the four 2D-projections and Tsai's algorithm [35]. The resulting set of 3D poses is then aligned using Procrustes and reduced by PCA obtaining the pose eigenspace (Fig. 3) where the dynamic-based clustering will be operated.

3.1.2. Principal components selection

The non-linearity of the training set is mainly localized in the first components of the PCA that capture the dynamics, as shown in Fig. 3. These components are really influential during the partitioning step while the last ones, more linear, only model local variations (details) and do not provide so much information for clustering. Only the first non-linear components are thus selected to perform the clustering of the data in a lower dimensional space. For components selection, the non-Gaussianity of the data is measured on each component. There

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012



b

	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

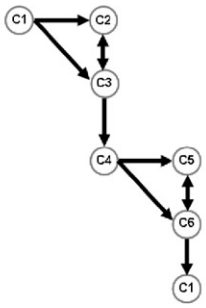


Fig. 5. (a) Correspondences between gait cycle and the six clusters obtained, (b) Markov state transition matrix and (c) state diagram.

are different methodologies to test whether the assumed normal probability distribution accurately characterizes the observed data or not. Skewness and kurtosis are two classical measures of non-Gaussianity.

A more robust measure is given by the Negentropy, the classic information theory's measure of non-Gaussianity, whose value is zero for Gaussian distribution [34]. Fig. 4a shows how the Negentropy converges to 0 and oscillates when considering lower modes. This oscillation between 0 and 0.75×10^{-4} starts from the 4th mode. It can be observed how the first three modes present a much higher Negentropy compared to the other modes. According to this analysis, we select the first three components for clustering.

3.1.3. Determining the number of clusters

K-means algorithm is used fairly frequently as a result of its ease of implementation. K-means clustering splits a set of objects into a selected number of groups by maximizing between variations relative to within variations. The main disadvantage of this algorithm is its extreme sensitivity to the initial seeds. A solution could be found by applying K-means several times, starting with different initial conditions and then choosing the best solution. But this supposes a supervision that makes the process more ad hoc. To make the clustering independent from the initial seeds, the K-means algorithm is ran many times and the total results are clustered as in Ref. [36]. For each case ($K = 2, \ldots, N$), a GMM is fitted to the pose eigenspace using the EM and a local PCA's is applied on each cluster. Since local modes of variation inside each Gaussian distribution of the mixture are expected, one of the aspects that should be evaluated when determining the optimal number of cluster is the global Gaussianity of the GMM. All the points of the training set are then projected onto the corresponding local PCA space and the Negentropy is computed for each cluster.

In Fig. 4b, the evolution of the mean Negentropy can be observed for K varying from 2 to 18. The curve decreases and converges logically to 0. It is desired to create as few clusters as possible and obtain some clusters as Gaussian as possible. A good compromise between number of clusters and Gaussianity is reached at $K=6$ where the std. of the Negentropy substantially decreases compared to the one at $K = 5$. Fig. 4c shows the GMM obtained with $K = 6$, represented in the pose eigenspace. This graphical representation shows the accuracy of GMM only by simple visual criteria: comparing with Fig. 3, it can be observed how well the GMM limits the feature space.

This leads to the recognition of basic gait cycle phases [37], as illustrated by Fig. 5, 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 left mid-swing. C4 follows until the second double support of the cycle which ends with C6. C2

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

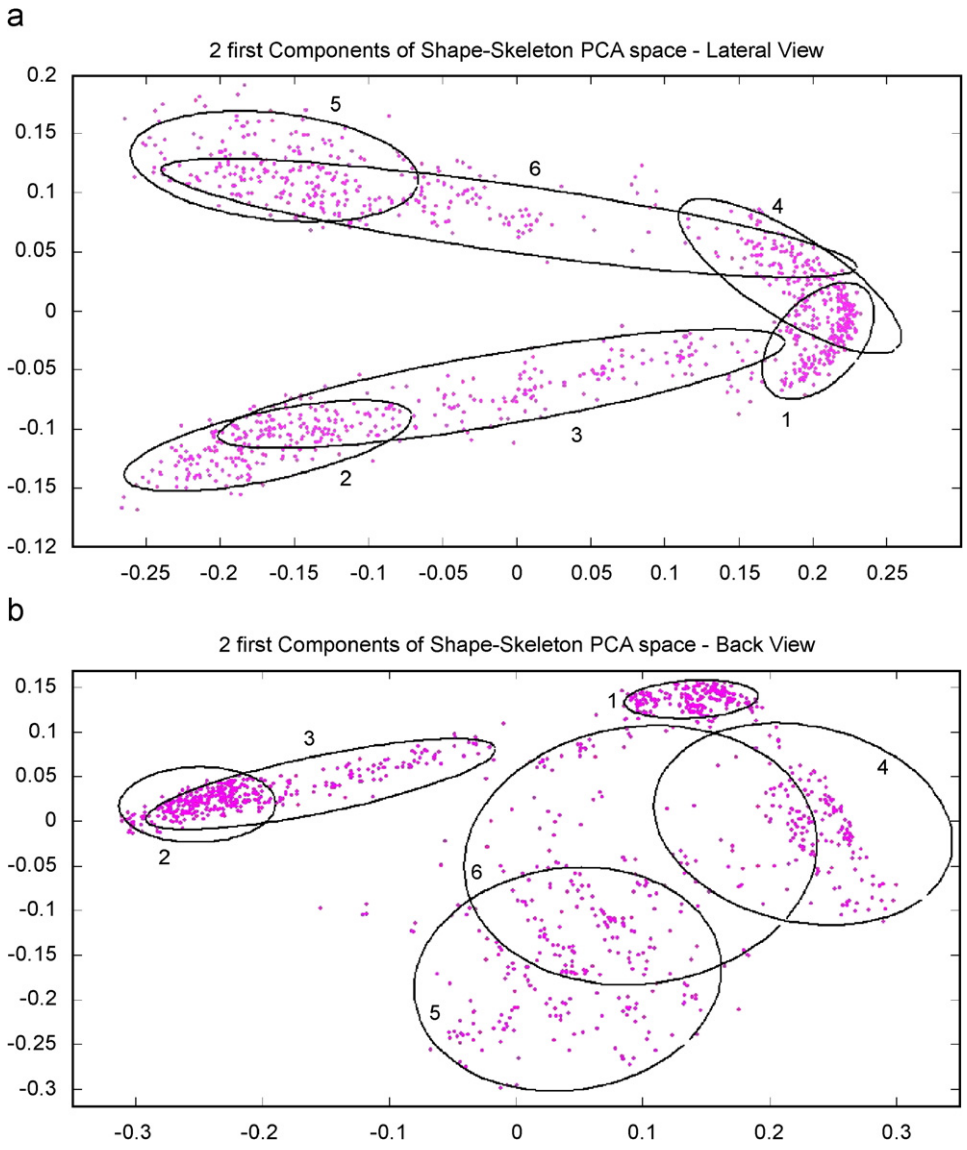


Fig. 6. GMM represented on the two first components of the shape–skeleton eigenspace for the (a) lateral and (b) back views.

and C5 complete C3 and C6 phases in case of a higher speed gait with larger step. A Markov state transition matrix (STM) [38] is then constructed (Fig. 5b), associating each sample to one of the six patches. Each temporal cluster corresponds to a state in the Markov chain. This gives the state transition probabilities, valid for the four sets (views) of SS-vectors.

3.2. View-based non-linear models

A view-based mixture of PCA is now fitted to the four shape–skeleton eigenspaces, using the structure-based clustering obtained before. Fig. 6 shows how the different mixtures limit the feature spaces: the clustering imposes a particular location of the Gaussian distribution (represented as ellipsoids) that consequently treat some unseen data as valid by interpolating. Fig. 7 shows how both shape and skeleton deform linearly in each one of the cluster of the view-based GMM. Dynamic correspondences are obtained between Gaussian models of the four mixtures, each cluster corresponding to one of the six basic gait phases.

3.2.1. Joint estimation of shape and skeleton

In Ref. [22], Grauman inferred 3D structure from multi-view contour. Following the same idea, when presented a new shape, the unknown 2D structure (structural parameters) is treated as missing variables in an SS-vector. The corresponding $\mathbf{b}^*$ is then computed from Eq. (4) and the nearest cluster, defined by eigenvectors $\mathbf{\Phi} = [\Phi_1, \ldots, \Phi_t, \ldots, \Phi_T]$ and eigenvalues λ_t, is selected. Thus the closest allowable SS-vector from the model is constructed by finding $\mathbf{r}$ so that

$$\mathbf{r} = \mathbf{\Phi}^{-1}(\mathbf{b}^* - \bar{\mathbf{b}}) \quad \text{and} \quad -\beta\sqrt{\lambda_t} \leqslant r_t \leqslant \beta\sqrt{\lambda_t}. \tag{6}$$

To ensure a valid SS-vector generation, the weight vector $\mathbf{r}$ is constrained to lie in the hyper-ellipsoid representing the linear

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

subspace model [32]. This leads to a model-based estimation of both shape and skeleton (cf. Fig. 9).

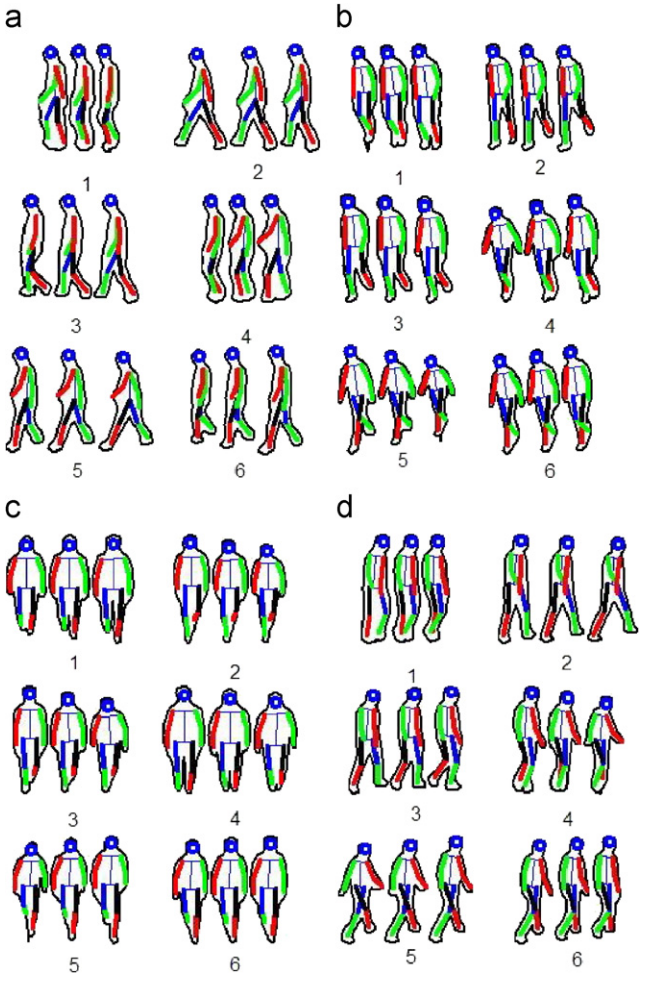


Fig. 7. Principal modes of variation of the six corresponding Gaussian models for the four view-based GMMs: (a) lateral, (b) diagonal, (c) frontal and (d) back views.

3.2.2. Non-linear models testing

The first approach we followed to cope with the non-linearity of the eigenspace was to select the closest allowable shape from the training set by means of an NN classifier [39].

This technique always warranties a valid contour but is imperfect because it cannot generate new shapes absent from the training data. Moreover, the computational cost makes this approach infeasible with a very large database. In Ref. [36] we proposed to use ICA for human shape modelling. The dynamic-based GMM developed in this paper will be compared to both methods.

For the evaluation of the view-based models, four gait sequences whose viewpoints correspond more or less to the four training views (cf. Fig. 9) are selected from the Caviar database [30]. On the one hand, groundtruth data are constructed by manually extracting the silhouettes of selected people appearing in the scene and on the other hand, human blobs are calculated by motion detection. Errors will be calculated as Euclidean distances between groundtruth and estimated shapes.

Two kinds of errors can be estimated: reconstruction and fitting errors. The first one is calculated by projecting and reconstructing a groundtruth shape with the model: this error

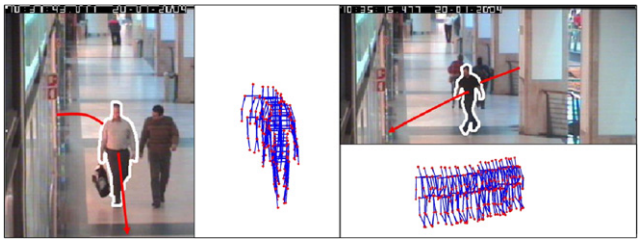


Fig. 9. Selected Caviar sequences for testing frontal (left) and diagonal (right) views-based GMM. For each of the two sequences, a frame with fitted shape is presented as well as the 2D poses automatically generated when applying the SS-model along the sequence.

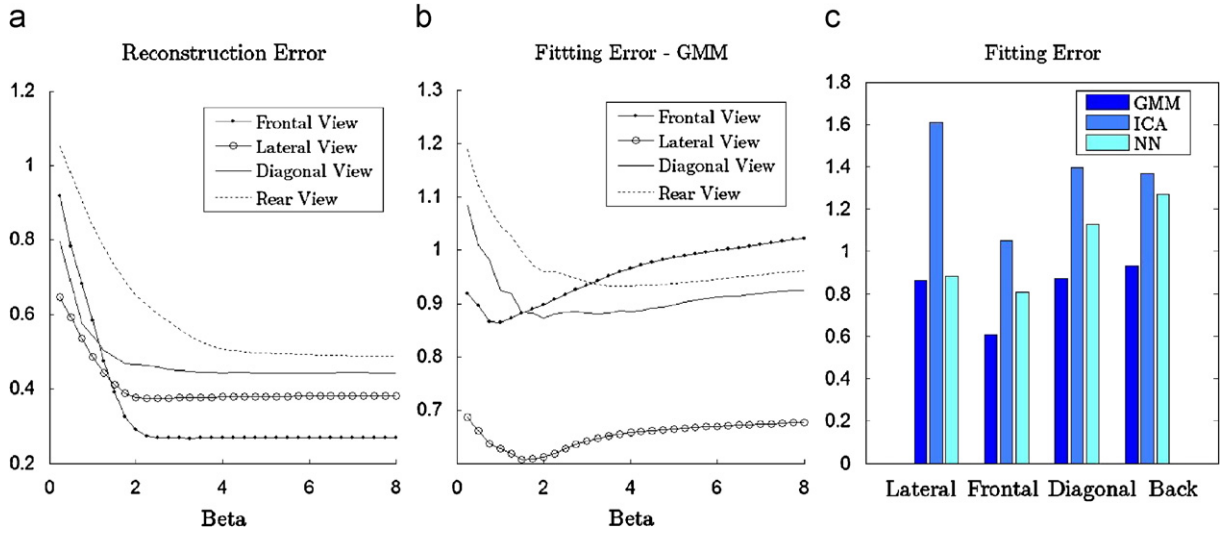


Fig. 8. (a) Reconstruction error, (b) fitting error obtained applying our GMM on the four Caviar sequences and (c) comparative results for the NN, ICA and GMM.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

characterizes the ability of the model to generate new silhouettes. The reconstruction error decreases and converges logically for the four models when augmenting parameter β from Eq. (6) (see Fig. 8a). The fitting error is calculated by correcting the shape extracted from the human blob with the model: this error characterizes the ability of the model to correct bad shapes. In Fig. 8b, it can be observed how the reconstruction error decreases until a minimum value and then starts increasing for the four models when augmenting β. This allows us to determine the optimal value of β for every view-based GMM. In Fig. 8c, fitting errors obtained when applying GMMs, NN and ICA are compared for the four views (four Caviar sequences).

GMM exhibits best results than both ICA and NN methods, and shows a better capability to reconstruct unseen shapes. Moreover computational cost of GMM mainly appears during the off-line stage (model construction) while the NN method requires an online comparison to the training exemplars. This makes this approach much more feasible for real-time applications with large databases of different poses and motions.

These four training views are obviously not sufficient to model all the possible orientations of the subject w.r.t. the camera and a more complete model must be built, considering other camera viewpoints. All the resulting models will be included in a global multi-view 2D-models framework.

3.3. Construction of the global 2D-models framework

Recently some authors have proposed a common approach consisting in discretizing the space considering a series of view-based 2D models [26,40]. In the same way, eight different viewpoints will be considered, uniformly distributed between 0 and 2π, thus discretizing the frontal view (vertical image plane) into eight sectors. For each sequence, the four training viewpoints used up to that point are now completed by a 5th supplementary back view that is also manually labelled. Finally, the last three missing views are interpolated using the periodicity and symmetry of human walking. By this process, a complete training database is generated encompassing more than 20 000 shape–skeleton vectors, SS-vector (more than 2500 vectors per viewpoint). The resulting eight view-based shape–skeleton associations for a particular snapshot of the CMU MoBo database are presented in Fig. 10.

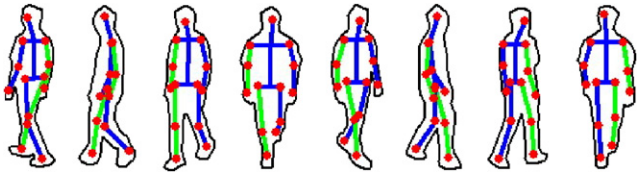


Fig. 10. Training views considered for framework construction.

The complete set of SS-vectors is concatenated in a common space (the eight views together) whose dimensionality is reduced using PCA, obtaining

$$\mathbf{v}_i \simeq \bar{\mathbf{v}} + \Phi_v \mathbf{a}_i, \tag{7}$$

where $\bar{\mathbf{v}}$ is the mean SS-vector, Φ is the matrix of eigenvectors and $\mathbf{a}_i$ is the new vector represented in the eigenspace. Let us call $\mathscr{A}$ the shape–skeleton eigenspace $\{\mathbf{a}_i\}$.

A series of local dynamic motion models has been learnt by clustering the structural parameters subspace. As mentioned in Section 3.1, the gait cycle is divided into six basic steps, providing the temporal clusters C_j, while the eight training views directly provide the spatial clustering (clusters R_r). The different clusters correspond in terms of dynamics or viewpoint. Using this structure-based partitioning and the correspondences between training viewpoints, 48 spatio-temporal clusters $\{\{T_{j,r} = C_j \cap R_r\}_{j=1}^{6}\}_{r=1}^{8}$ are obtained in the global shape–skeleton feature space where all the views considered are projected together.

Thus, following Ref. [11], a local linear model is learnt for each spatio-temporal cluster $T_{j,r}$ and a mixture of PCA is fitted to the clustered $\mathscr{A}$ space, obtaining a new STMF. For each cluster, the local PCA leads to the extraction of local modes of variation, in which both shape and skeleton simultaneously deform (see Fig. 12). Parameters for the 48 Gaussian mixture model components are determined using EM algorithm. The prior shape–skeleton model probability is then expressed as

$$p_{\text{mix}}(\mathbf{a}) = \sum_{j,r} \omega_{j,r} \mathcal{N}(\mathbf{a} : \bar{\mathbf{a}}_{j,r}, \sigma_{j,r}), \tag{8}$$

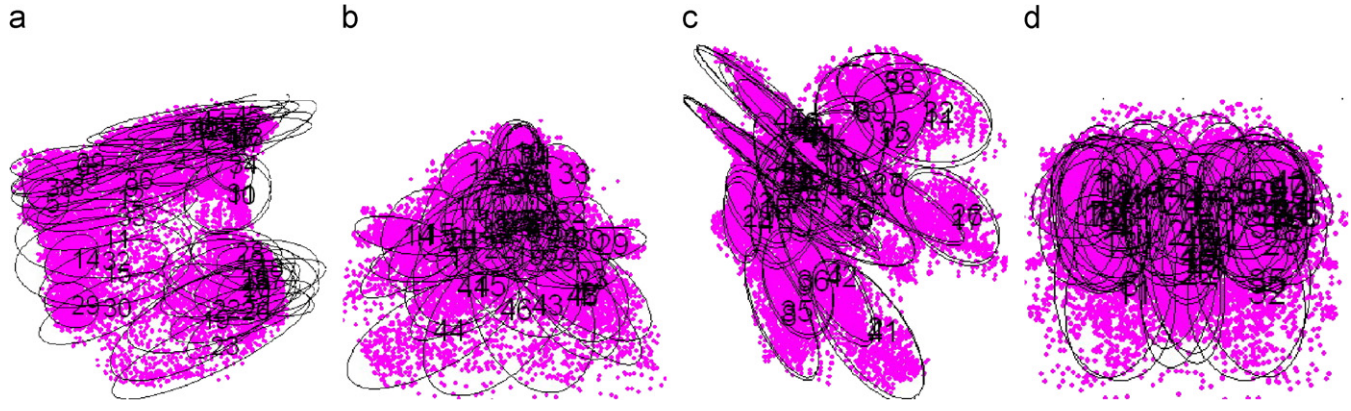


Fig. 11. 48-Clusters Gaussian mixture model plotted with training data projected onto the planes defined by (a) 1st and 2nd, (b) 3rd and 4th, (c) 5th and 6th and (d) 7th and 8th components of the shape–skeleton eigenspace.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

where $\mathbf{a}$ is the eigendecomposition of the shape–skeleton vector, $\mathcal{N}(\mathbf{a} : \bar{\mathbf{a}}, \sigma)$ is the p.d.f. of a Gaussian with mean $\bar{\mathbf{a}}$ and covariance σ and $\omega_{j,r}$ is the mixing parameter corresponding to $T_{j,r}$.

Fig. 11 shows the mixture projected onto various planes of the eigenspace space $\mathscr{A}$. The 48 hyper-ellipsoids corresponding to the 48 local spatio-temporal models are also plotted. It can be appreciated how well the GMM delimits the subspace of valid SS-vectors.

Given this huge amount of data, an efficient search method is required. In that way, temporal and spatial constraints will be considered to constrain the evolution through the STMF along a sequence and limit the feature space only to the most probable models of the framework.

4. Constraint-based search

The total space has been clustered following temporal approach (clusters C_j) as well as spatial approach (clusters R_j) as described in the previous section. The first one partitions the dynamics of the motion, and the second one, the viewpoint i.e. the direction of motion in the image. The purpose of the following probabilistic modelling is to obtain a transition matrix combining both spatial and temporal constraints.

4.1. 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 this state space is partitioned into N clusters $\mathscr{C} = \{C_1, \ldots, C_N\}$, 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 $t-1$ [11]. The $N \times N$ state transition matrix (STM) computed in the previous section points out the probabilities density function (pdfs).

4.2. Modelling spatial constraint

In this paper, a novel spatial prior $p(D_t|D_{t-1,\ldots,t-m})$ is introduced 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 its m previous directions of motion $(D_{t-1}, D_{t-2}, \ldots, D_{t-m})$. In this approach, the continuous values of all possible camera viewpoints are discretized. Consequently, the direction of motion in the image plane D_t takes a fixed set of values corresponding to the discrete set of M training viewpoints and M clusters $\mathscr{R} = \{R_1, \ldots, R_M\}$ in the feature space.

Let $\Delta_t = [R_{k_0}^t, R_{k_1}^{t-1}, \ldots, R_{k_m}^{t-m}]$ 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. It has to be noted that some of these k_i labels might be the same. Consequently, $p(R_j^t|\Delta_{t-1})$ is 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, a reasonable assumption is made that this direction of motion follows a normal distribution, with expected value equal to the local mean trajectory angle $\bar{\theta}_t$ and variance calculated as a function of the sampling rate:

$$p(R_j^t|\Delta_{t-1}) = p(R_j^t|R_{k_1}^{t-1}, R_{k_2}^{t-2}, \ldots, R_{k_m}^{t-m}) \sim \mathcal{N}(\bar{\theta}_t, \sigma), \quad (9)$$

where

$$\bar{\theta}_t = \frac{1}{m+1} \sum_{i=t}^{t-m} \theta_i,$$

m being a function of the sampling frequency.

4.3. Combining spatial and temporal constraints

Let T be the $N \times M$ matrix, whose columns represent the N temporal clusters and rows correspond to the M spatial clusters. Thus the probability $p(C_j^t \cap R_r^t) = p(T_{j,r}^t)$ 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 $p(T_{j,r}^t|C_k^{t-1}, \Delta_{t-1})$, 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 6 and 7. This leads to the following simplified equation:

$$p_{j,r} = p(T_{j,r}^t|C_k^{t-1}, \Delta_{t-1}) \propto p(C_j^t|C_k^{t-1}) p(R_r^t|\Delta_{t-1}). \quad (10)$$

The resulting $N \times M$ matrix is the PTM that gives, at each time step, the pdf that limits the region of interest in the STMF to the most probable models. Considering the cyclic nature of the walking action and the circular distribution of the training viewpoints, the resulting PTM is a toroidal matrix (Fig. 12) whose lines correspond to the M training view-based gait manifolds. Its 3D and 2D representations are illustrated in Fig. 12. All the different models can be ordered and classified according to their direction of motion and state, thus putting in evidence the correspondences with the PTM as shown in Fig. 12. Spatial and temporal relationships can be appreciated between local models from adjacent cells.

The content of the PTM can be visualized by converting it to gray scale image as will be shown in next sections. To compute this PTM and constrain the evolution through the STMF along a sequence, only previous viewpoint and previous state are required at each time step. Note that our approach shares some similarities with the one recently proposed by Lv and Nevatia [41]. In this paper, the authors model an action as a series of 2D poses rendered from a wide range of viewpoints and represent the constraints on transition by a graph model

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

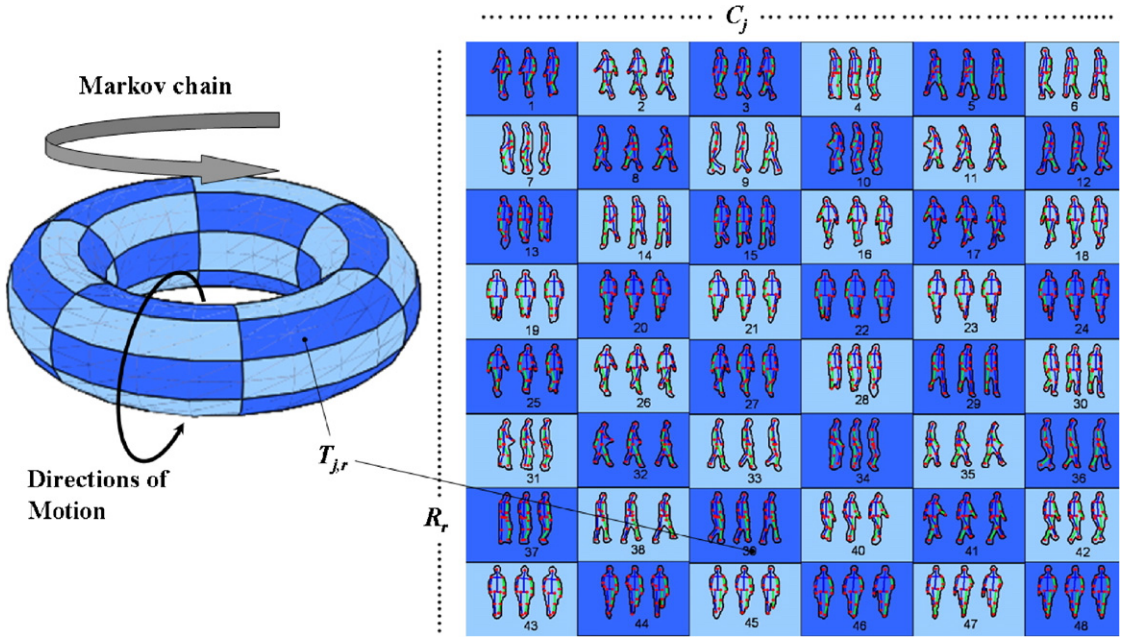


Fig. 12. 3D (left) and 2D (right) representations of the toroidal probabilistic transition matrix (PTM). The 1st variation modes of the 48 local models that compose the framework are superposed with the 2D representation of the PTM: the six columns correspond to the six temporal clusters C_i while the eight rows represent the eight spatial clusters R_i.

where they assume a uniform transitional probability for each link.

5. Joint segmentation and pose estimation

A discriminative detector as the ones proposed in Ref. [42] could be used to initialize the shape model-driven algorithm presented next. In this work, scene context information is considered to roughly limit the feature space only to the "logical" 2D-models from the framework. For example, if an object appears in the scene from the right side (right-to-left direction of motion), only the first three lines of the PTM will be considered.

Once the system has been initialized, each frame of the sequence is processed individually by applying *Segmentation-PoseEstimation* (Algorithm 1), taking advantage of previous information (trajectory angle θ, state *index*, background B) that is used to treat the current frame.

Algorithm 1. $(\mathbf{s}, \mathbf{k}, B, \theta, index) = SegmentationPose\ Estimation(B, I, \theta, index,\ nIter)$

$m[i] = ModelsSelection(\theta, index)$;
initialize shape $\mathbf{s} \leftarrow 0$;
initialize pose $\mathbf{k} \leftarrow 0$;
for $n \leftarrow 1$ to $nIter$ **do**
  blobsList = *AdaptiveBackgroundSubtraction* $(B, I, \mathbf{s}, n, nIter)$;
  Silhouette = *BlobsProcessing*(blobsList);
  $\mathbf{s} = ContourExtraction$(Silhouette);
  $[\mathbf{s}, \mathbf{k}, \text{index}] = ShapeSkeletonCorrection(\mathbf{s}, \mathbf{k}, m[i])$;
end for
$B = BackgroundUpdate(B, I, Silhouette)$;

Algorithm 2. blobsList = $AdaptiveBackgroundSubtraction(B, I, \mathbf{s}, n, nIter)$

if $(n == 1)$ **then**
  $D = BackgroundSubtraction(B, I, thr)$;
else
  Mask = $ShapeToMask(\mathbf{s})$;
  $t_{low} = DecreaseThreshold(thr, n, nIter)$;
  $D_{low} = BackgroundSubtraction(B, I, t_{low})$;
  $t_{high} = IncreaseThreshold(thr, n, nIter)$;
  $D_{high} = BackgroundSubtraction(B, I, t_{high})$;
  $D = D_{low} \times Mask + D_{high} \times \overline{Mask}$;
end if
blobsList = $BlobsLabelling(D)$;

The prediction of the most probable models from the GMM is estimated in *ModelsSelection* by means of the PTM. It allows a substantial reduction in computational cost and can solve some possible ambiguities by considering a limited number of models.

In *ShapeSkeletonCorrection*, the extracted shape $\mathbf{s}$ and an estimate for the skeleton are concatenated in $\mathbf{v} = [\mathbf{s}^\top\ \bar{\mathbf{k}}^\top]^\top$ and projected into the SS-eigenspace obtaining $\mathbf{a}$. Then, for each one of the most probable clusters given by the PTM $p_{j,r}$, we update the parameters to best fit the "local model" defined by its mean, eigenvectors and eigenvalues, as done in Section 3.2.1, obtaining $\mathbf{a}^*$. The distance between extracted and corrected shapes is then calculated for each one of the estimations in order to select the best estimation. We then project the vector $\mathbf{a}^*$ back to the feature space obtaining $\mathbf{v}^*$ which contains a new estimation of both shape $\mathbf{s}^*$ and skeleton $\mathbf{k}^*$: $\mathbf{v}^* = [\mathbf{s}^{*\top}\ \mathbf{k}^{*\top}]^\top$.

Aside from the models and the constraint-based search proposed in this work, some novelties appear in the segmentation

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

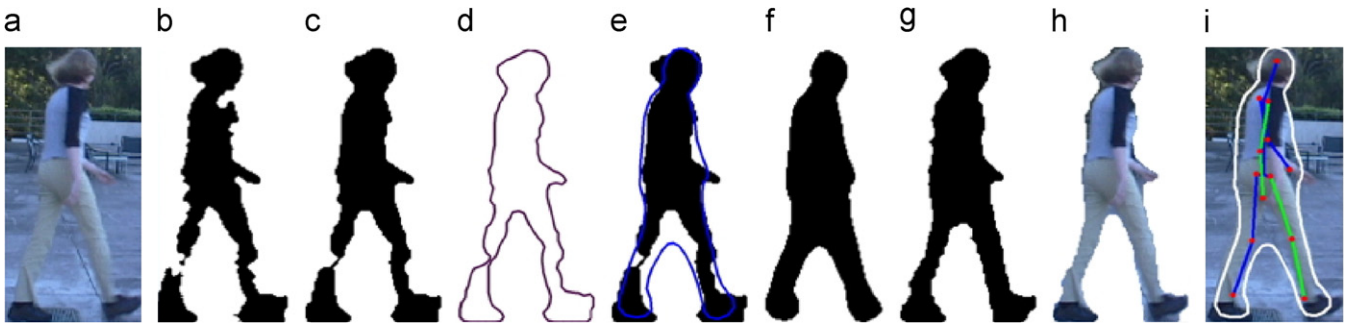


Fig. 13. Model fitting: (a) original input image, (b) silhouette resulting from background subtraction, (c) silhouette after being processed by *BlobsProcessing*, (d) contour extracted by silhouette erosion, (e) corrected shape represented on the silhouette and corresponding mask, (f) used for finer background subtraction, (g) resulting silhouette after six iterations, (h) corresponding segmentation and (i) resulting shape and pose plotted on the original input image.

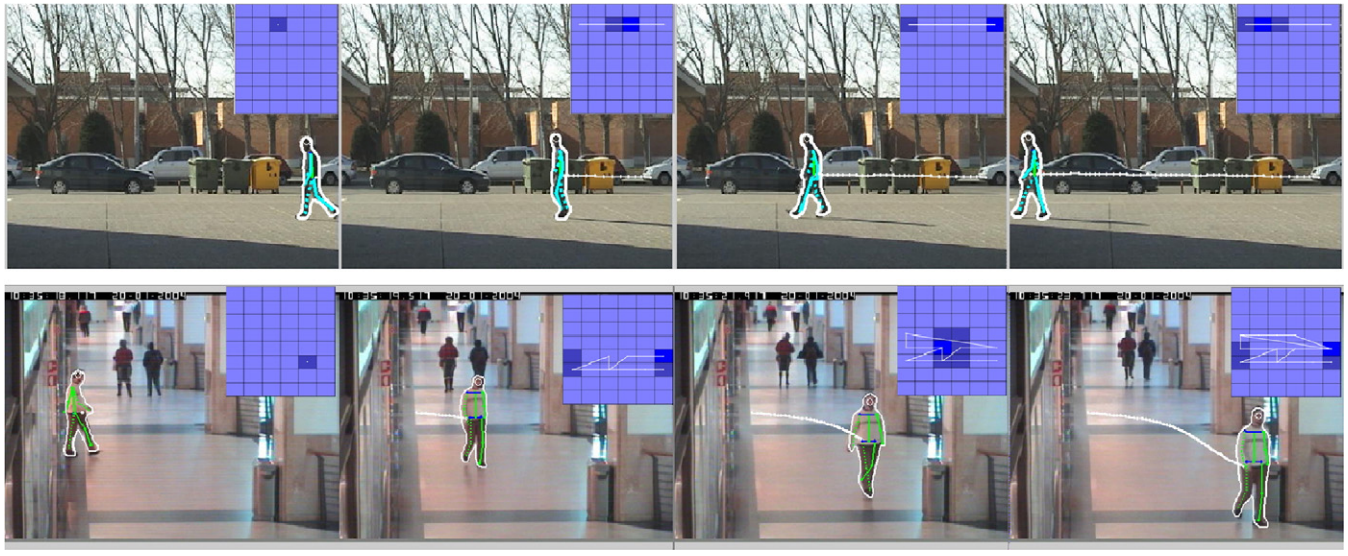


Fig. 14. (up) Outdoor straight-line walking sequence at constant speed and (down) Caviar sequence with slight bend trajectory.

algorithm. The first one referred to the shape extraction task (*ContourExtraction* in Algorithm 1): while it is usually extracted from the blob looking along straight lines through each model point as in Ref. [4], here the shape is directly obtained by eroding the human blob and normalize the resulting contour following the shape normalization proposed in Section 2.2. This allows a direct, precise and faster registration of the shape in the image. The only drawback of this shape registration is that it requires an entire and non-fragmented silhouette. The *BlobsProcessing* function thus previously applies some common morphological operations to the result of *AdaptiveBackgroundSubtraction* and connect the possible fragments.

Another novelty of the fitting process appears in *AdaptiveBackgroundSubtraction* that aims at reconstructing the binary silhouette resulting from the background subtraction using the "corrected" shape returned by *ShapeSkeletonCorrection*. It is achieved by decreasing/increasing the detection threshold inside/outside the shape. This leads to an accurate silhouette segmentation, improving considerably the results specially when there is no significant difference between background and foreground pixels.

The last novelty relies on the way the background is updated: the final segmented silhouette, the foreground, is used to actualize the background more finely, eliminating shadows from the foreground and improving the segmentation in next frames.

The different steps of the process are depicted in Fig. 13 for a particular frame.

6. Experiments

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 only model fitting and pose estimationwill be tested, and not the tracking in the image, the system is provided with the bounding-box taken from groundtruth avoiding the possible problems due to the tracking. In the PTMs from Fig. 14 (as well as from Figs. 18 and 19), the colored cells represent the probability $p_{j,r}$ from Eq. (10). 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 PTM indicates the orientation of the pedestrian with respect

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

to the camera. Additionally, both trajectory and previous states are, respectively, plotted in the image/matrix with a white line.

As illustrated in Fig. 14(up), the resultant vectors from a pedestrian crossing the scene straight ahead without stopping or turning towards anything all belong to models from the same row of the PTM. Any change of direction is observed as a progressive change of row (see Fig. 14(down)).

In Fig. 15, the results obtained with two challenging frames are presented: in (a) the pedestrian is carrying a bag and in (b) he is partially occluded by the wall. In both cases, a satisfactory estimation is made of both shape and pose.

6.1. Framework validation

To validate the framework, the 2D poses and 2D shapes of three different sequences with different characteristics of interest are hand-labelled: an outdoor straight-line walking sequences at constant speed (Fig. 14(up)), an outdoor "Walkcircle" sequences with constant speed and constant viewpoint and scale evolution (Fig. 18) and an indoor sequence with viewpoint and speed variations (Fig. 19). Note that the subjects turn and move "in depth" so that both apparent scale and viewpoint vary. A top-down estimation of depth is directly provided by the "winning" model that points out the motion direction in 3D space (see Fig. 14 and later Figs. 18 and 19).

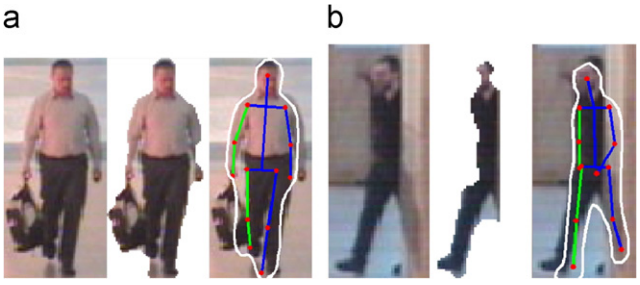


Fig. 15. Results obtained with two challenging frames. For each of the two examples, original image (left), segmentation (center) and resulting pose and shape are represented (right).

6.1.1. Segmentation

Quantitative validation is performed by comparing with manually segmented solutions, both the segmentation obtained by simple background subtraction and the one resulting from the proposed model-based approach. Denote the manual segmentation in the images as $S_{groundtruth}$, and the results as $S_{estimated}$. We define the false negative (FN) ratio to indicate the fraction of silhouette that is included in the groundtruth segmentation but missed by the automatic method:

$$FN = \frac{|S_{groundtruth} - S_{estimated}|}{S_{groundtruth}}. \quad (11)$$

The false positive (FP) ratio indicates the amount of foreground falsely identified by the algorithm as a fraction of the total silhouette in the groundtruth segmentation:

$$FP = \frac{|S_{estimated} - S_{groundtruth}|}{S_{groundtruth}}. \quad (12)$$

The true positive (TP) describes the fraction of the total silhouette in the true segmentation that is overlapped with the proposed method:

$$TP = \frac{|S_{estimated} \cap S_{groundtruth}|}{S_{groundtruth}}. \quad (13)$$

Example segmentation results are shown in Fig. 16 and average statistics compiled in Tables 1 and 2. On the outdoor sequence, the segmentation results produce the following: FN ratio is improved by 3.8%, FP by 14.48% and TP by 3.8%. On the indoor sequence, only FP ratio is improved by 7.03% while FN and TP stay unchanged. In both cases, we can observe how part of the shadow is eliminated with the proposed method what leads directly to the FP ratio improvement.

6.1.2. Pose estimation

Fig. 17 shows the pose estimation results for the three tested sequences. The mean position error (in pixels) is calculated as the feet-distance between the skeleton estimated by the algorithm and the hand-labelled one. Some peaks can

Fig. 16. Segmentation results for (left) "Walkcircle" outdoor sequence and (right) for "elevator" indoor sequence, (up) the original image, (center) the result obtained by simple background subtraction and (down) the result obtained by applying the proposed model-based algorithm are represented for each example.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

be noticed in this figure. For instance, in the indoor sequence (center) 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 "Walkcircle" sequences (right) the model fails because of the stationary behavior 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 behavior of the input motion. This is due to the very low shape variability in the back view where it is very complicated to distinguish a state from another. 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 of independency of spatial and temporal event, made in Section 4.3, is reasonable.

6.2. Numerical evaluation with HumanEva data set

For numerical evaluation of the framework, the four walking sequences of the HumanEva-II data set [31] are considered: subjects S2 and S4 observed from cameras C1 and C2.

Table 1
Segmentation results for outdoor "Walkcircle" sequence

	Background subt. (%)	Model-based segm. (%)
FN	9.31	5.51
FP	27.92	13.44
TP	90.69	94.49

Table 2
Segmentation results for indoor "Elevator" sequence

	Background subt. (%)	Model-based segm. (%)
FN	6.79	6.80
FP	20.83	13.80
TP	93.21	93.20

Note that the groundtruth is not available for these sequences and that for each frame, the bounding-box is estimated using a simple Kalman filter. The good results obtained with such setting demonstrate that the method behaves quite well even if it is not provided with the exact bounding-box taken from groundtruth.

Segmentation and estimated 2D poses resulting from the proposed model-based approach are presented together in Fig. 20 while numerical evaluation is given in Fig. 21. This evaluation has been obtained using the on-line evaluation system and the metrics provided for 2D pose estimation i.e. the average distance in pixels over all the 13 2D key-points of the stick model. For each sequence, this error is shown for all the processed frames in Fig. 21 and the average error per sequence (over all the frames) is given in Table 3.

In the four sequences, the human body was segmented and tracked successfully as can be seen in Fig. 20, maintaining the sequentiality of the motion even if some pics can be observed in the error curve. However, the average difference is quite high in all the frames even when the result is shown to be visually accurate in Fig. 20. This can be explained by the differences in defining the joint centers in the proposed skeleton model (constructed from hand-labelled data) and in the marker-based system, which causes an offset clearly observable in Fig. 21.

7. Conclusions and discussions

This paper has presented a novel probabilistic spatio-temporal 2D-models framework for human motion analysis. In this approach, the 2D-shape of the entire body has been associated to the corresponding stick figure allowing the joint segmentation and pose recovery of the subject observed in the scene along a sequence.

The problem of non-linearity has been solved by fitting a Gaussian mixture model (GMM) to several training views.

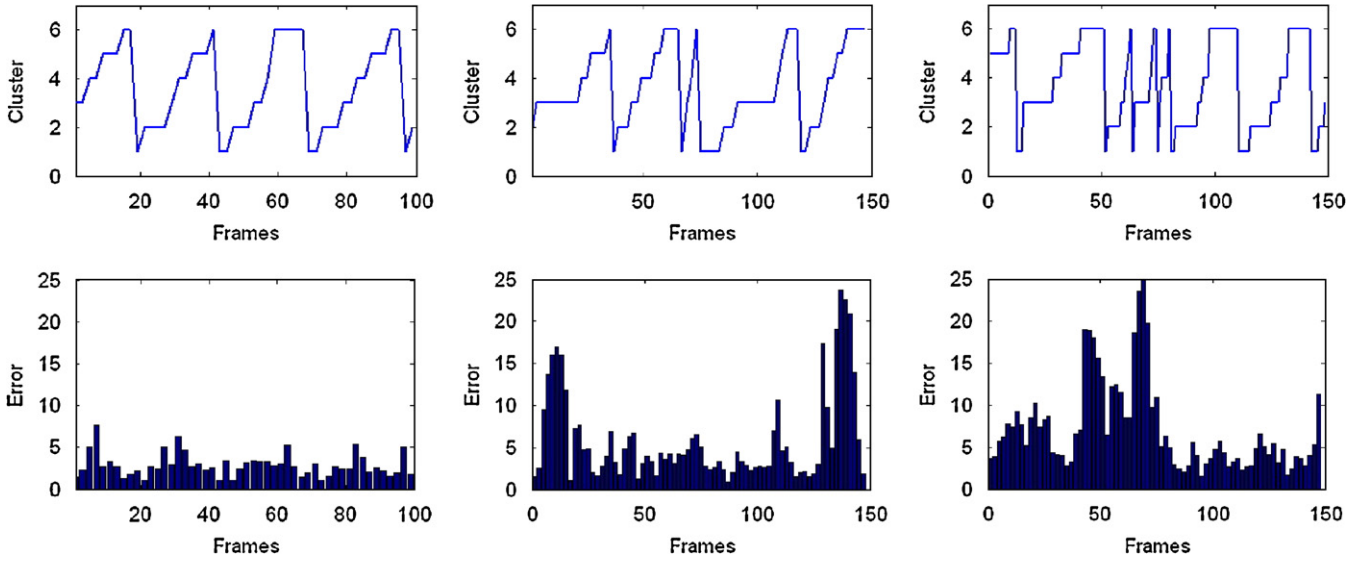


Fig. 17. Feet position error in pixels (bottom) and temporal clusters (top)—given by the column of the PTM corresponding to the "winning" model—of the straight-line walking (left), indoor (center) and "Walk-circle" (right) sequences.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

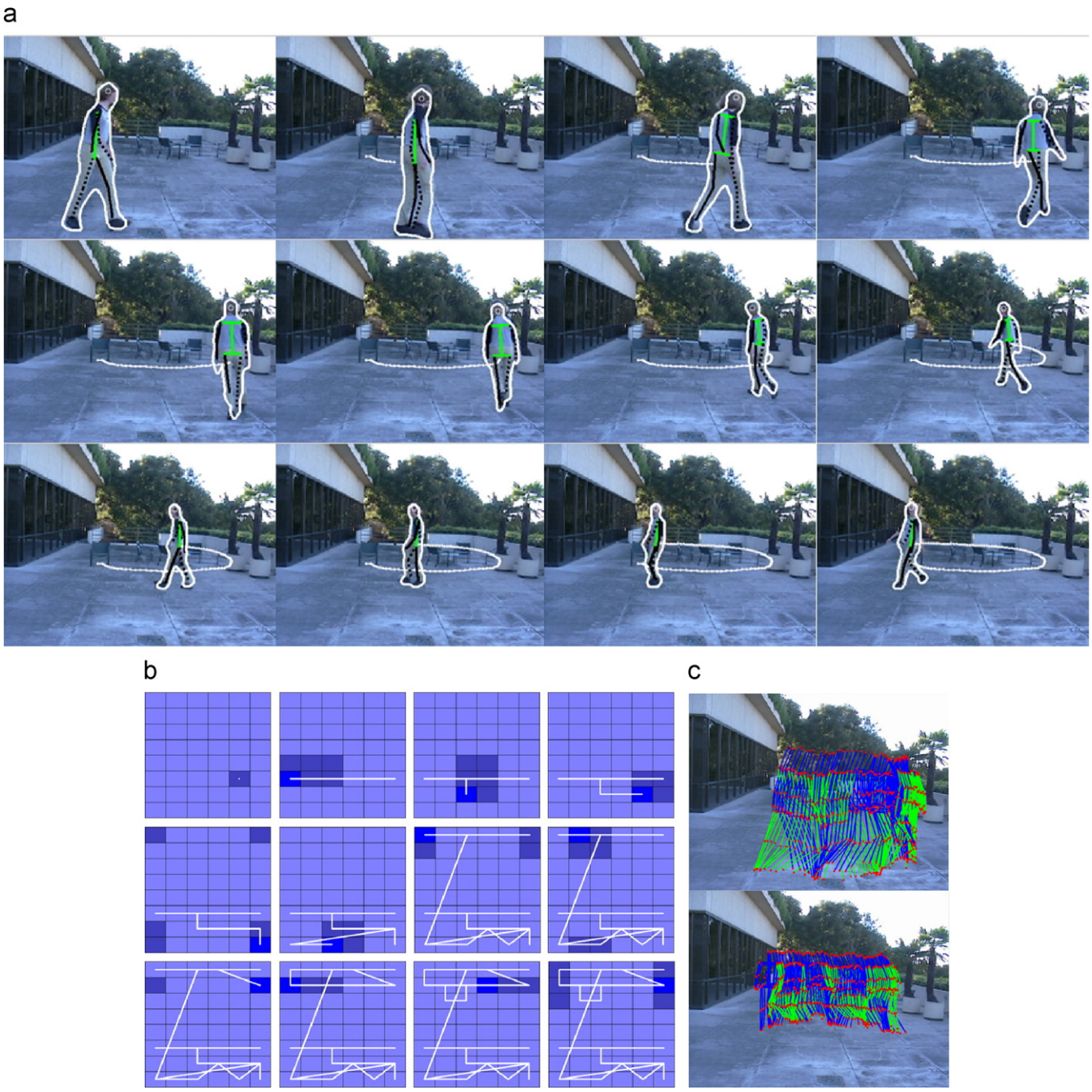


Fig. 18. Results obtained for the outdoor "Walkcircle" sequence with constant speed and constant viewpoint and scale changes from Ref. [3]: (a) estimated shapes and poses represented on the original image for frames 1, 15, 25, 40, 50, 60, 75 90, 100, 115, 130 and 140, (b) 12 corresponding PTM matrices and (c) 2D poses estimated along the complete sequence.

Since shape variations of articulated objects are closely linked to the pose evolution along time, the total training set has been clustered using only the structural information projected in the pose eigenspace. In order to simplify the problem, only the most non-linear components have been selected to perform the clustering of the data in a lower dimensional space. The optimal number of clusters has been determined by considering the mean Gaussianity of the GMM. This approach has been compared to other two methods developed to cope with shape models non-linearity: GMM exhibits best results than both ICA and NN methods, and shows a better capability to reconstruct unseen shapes.

To cope with the restriction to the viewpoint, local spatio-temporal 2D-models corresponding to many views of the same sequences were trained, concatenated and sorted in a global framework (a multi-view GMM). When processing a sequence, temporal and spatial constraints are considered to build the probabilistic transition matrix (PTM) that gives the frame to frame prediction of the most probable models from the framework. The proposed fitting algorithm, combined with the new

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

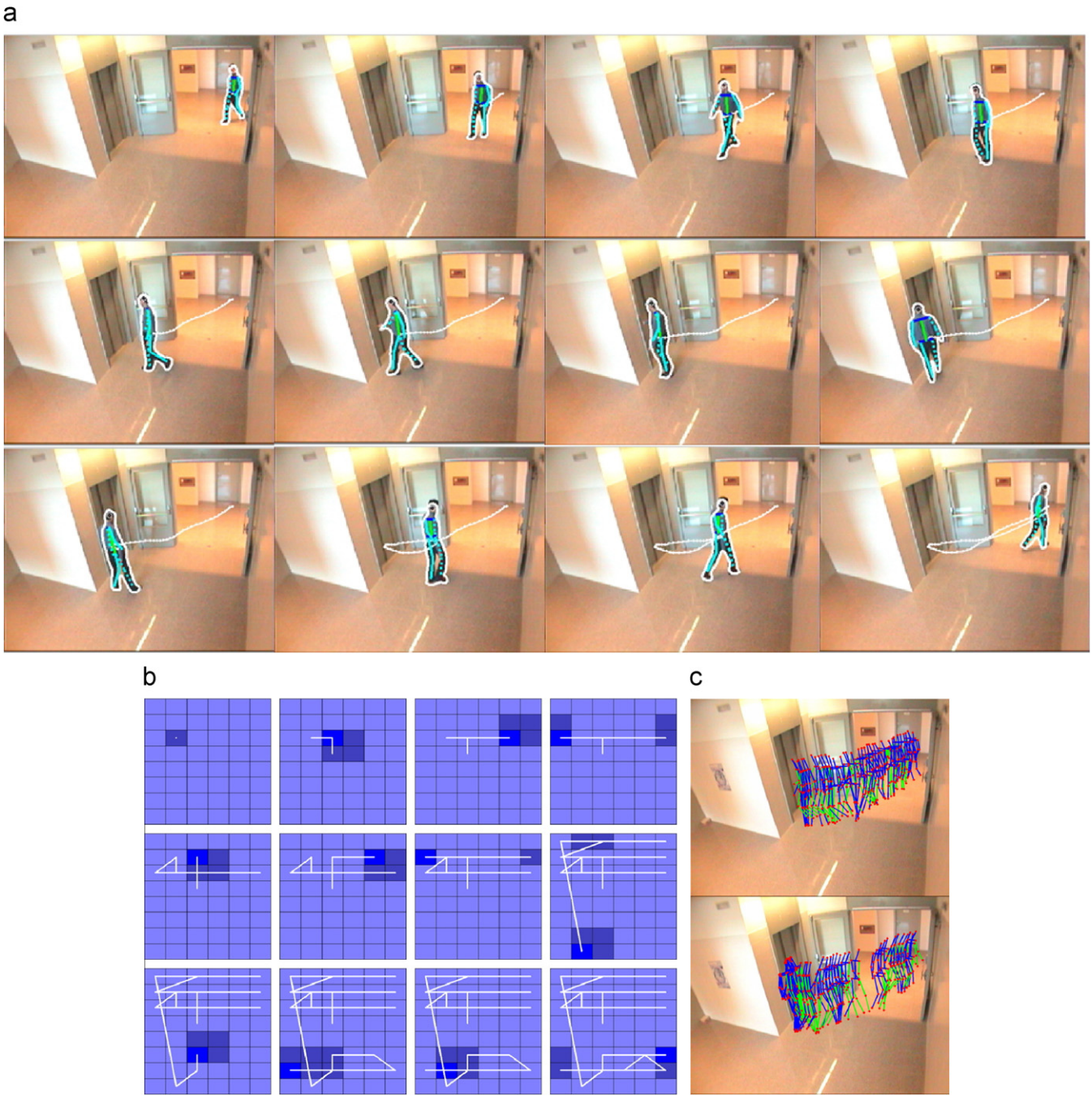


Fig. 19. Results obtained for the indoor "Elevator" sequence with viewpoint and speed changes: (a) estimated shapes and poses represented on the original image for frames 1, 18, 30, 38, 48, 58, 68, 88, 100, 122, 128 and 142, (b) 12 corresponding PTM matrices and (c) 2D poses estimated along the complete sequence.

probabilistic models, allows a faster and more reliable estimation of both pedestrian Silhouette and stick figure in real monocular sequences. The experiments carried out on both indoor and outdoor sequences have demonstrated the ability of this approach to adequately segment the pedestrians and estimate their postures independently of the direction of motion during the sequence. They have also demonstrated that the method responds quite robustly to any change of direction during the sequence. However, further work must be done.

The main ongoing work relies on addressing the tracking issue to estimate both location of the person in the image and model parameters during the sequence. The case of multiple people tracking with occlusions has to be considered. Moreover, the assumption has been made 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.

In the presented example, only one value has been considered for the elevation angle, due to practical reasons. To deal with different tilt angles, a preprocessing stage can be considered to remove the perspective effect as we proposed in Ref. [43].

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

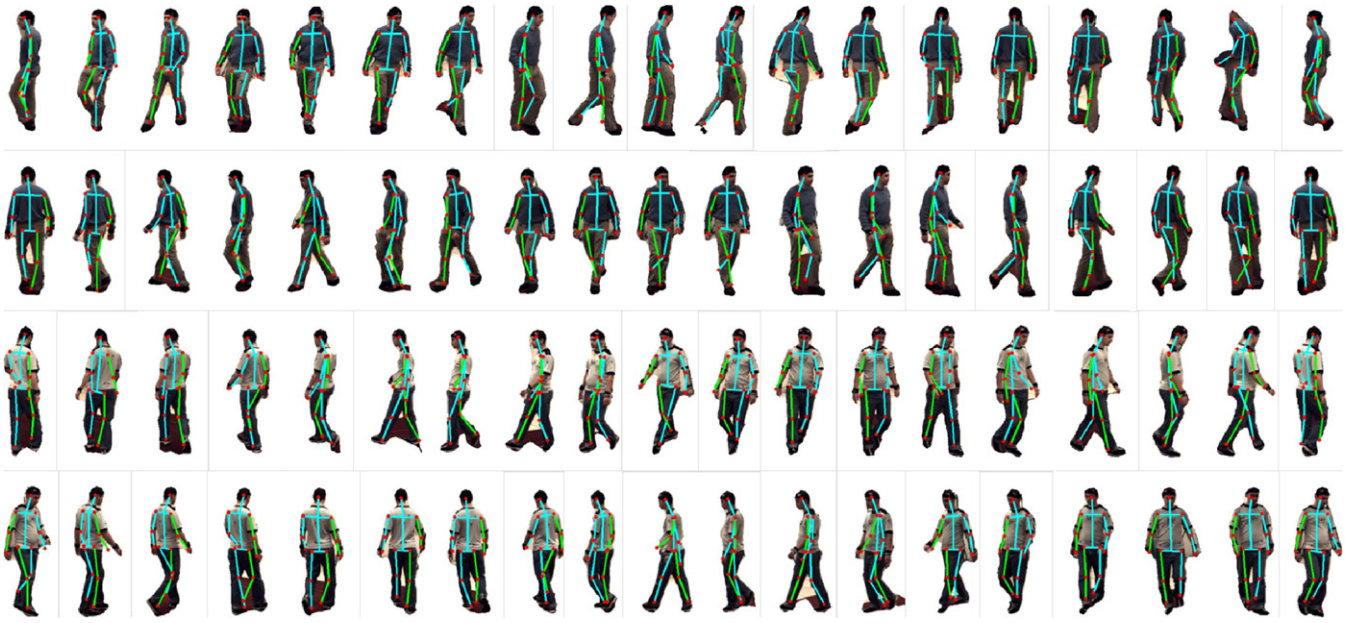


Fig. 20. Segmentation and 2D pose estimation obtained for the four HumanEva testing sequences. From up to down: Subject S2, camera views C1 (*1st line*) and C2 (*2nd line*), and Subject S4, camera views C1 (*3rd line*) and C2 (*4th line*). For each sequence, frames 1, 20, 40, 60, 80 …300, 320, 340 and 350 are represented.

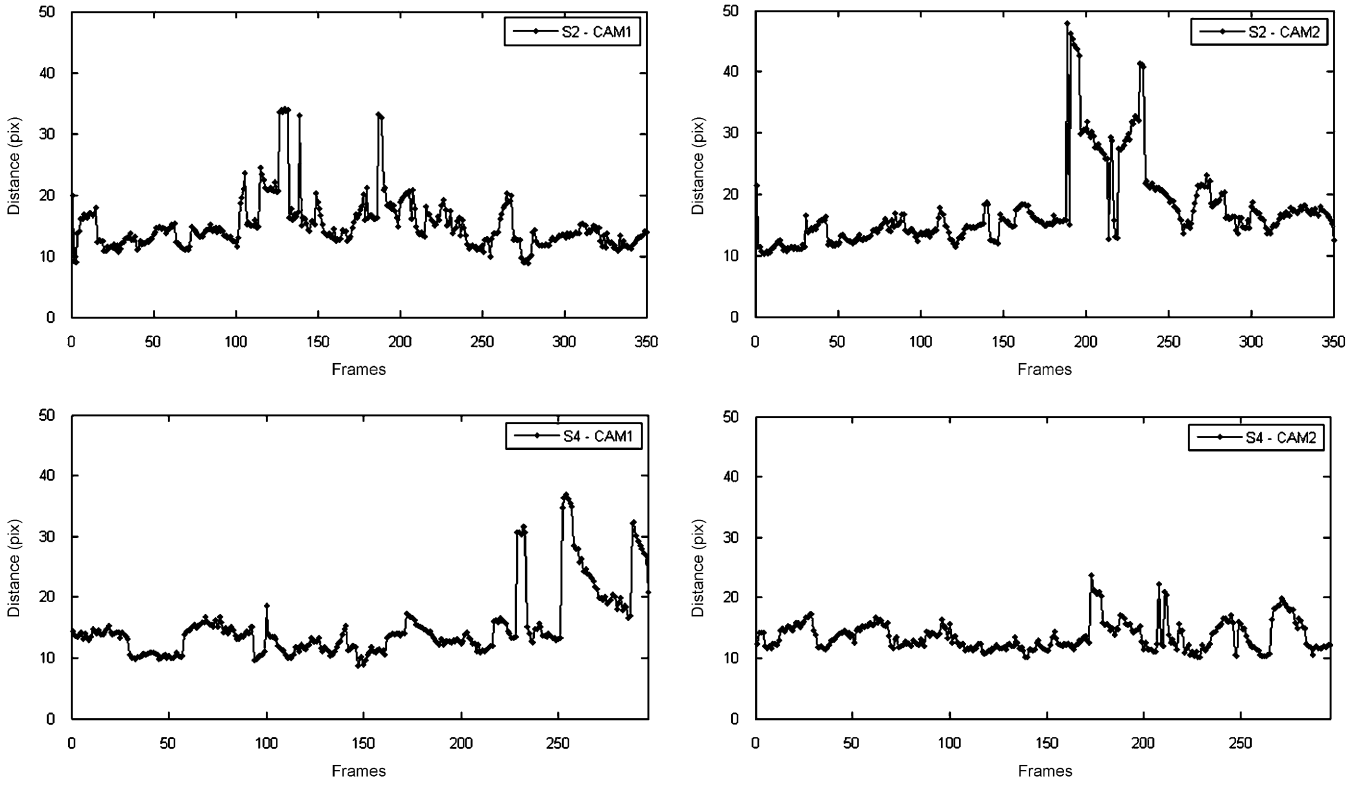


Fig. 21. Numerical results obtained for the four HumanEva testing sequences: for Subject S2 (up) and S4 (down). In both cases, the average error of 2D pose reconstruction is given for camera views C1 (left) and C2 (right).

Another possibility to handle large viewpoint changes (when using roof-top cameras for example) is to train the model with several values of this tilt angle as in Ref. [41]. The supplementary angle variation could then be represented by an additional third dimension in the toroidal transition matrix in order to keep the spatial continuity between viewpoints of connected cells.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

Table 3
2D pose average error on HumanEVA data set

Subject	Camera	Start	End	Mean error (pix)
S2	C1	1	350	16.96
S2	C2	1	350	18.53
S4	C1	1	290	16.36
S4	C2	1	290	14.88

Even though it has been tested with the specific gait motion, the presented approach is generic and could be applied to any other action. A large human motion capture database and a 3D computer graphics human model will be used for synthesizing automatically training pairs of 2D and 3D representations. In this paper, a way has been provided to transition between view-based manifolds of a same action. Transitions between different activities sub-manifolds embedded in a global one will have to be considered. Finally, even with large amounts of training data (with numerous viewpoints), classification cannot expect to be perfect. Other shape matching techniques could reduce misclassifications and lack of correspondences between input and training views.

Acknowledgments

This work is supported by Spanish Grants TIC2003-08382-C05-05 (MCyT) and TIN2006-11044 (MEyC) and FEDER. Grégory Rogez is funded by the Spanish Ministry of Education under FPU Grant AP2003-2257 and Jesús Martínez-del-Rincón under FPI Grant BES-2004-3741.

References

[1] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Comput. Vision Image Understanding 104 (2006) 90–126.
[2] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.
[3] H. Sidenbladh, M.J. Black, L. Sigal, Implicit probabilistic models of human motion for synthesis and tracking, in: Proceedings of the IEEE European Conference on Computer Vision, vol. 1, 2002, pp. 784–800.
[4] A.M. Baumberg, D. Hogg, Learning Flexible Models from Image Sequences, Lecture Notes in Computer Science, vol. 800, Springer, Berlin, 1994, pp. 299–308.
[5] J. MacCormick, A. Blake, A probabilistic exclusion principle for tracking multiple objects, Int. J. Comput. Vision 39 (1) (2000) 57–71.
[6] N.T. Siebel, S.J. Maybank, Fusion of multiple tracking algorithms for robust people tracking, in: Proceedings of the IEEE European Conference on Computer Vision, 2002, pp. 373–387.
[7] T.H.W. Lam, R.S.T. Lee, D. Zhang, Human gait recognition by the fusion of motion and static spatio-temporal templates, Pattern Recognition 40 (9) (2007) 2563–2573.
[8] I.A. Kakadiaris, D.N. Metaxas, Model-based estimation of 3D human motion, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1453–1459.
[9] C. Sminchisescu, B. Triggs, Kinematic jump process for monocular 3D human tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, June 2003.
[10] T.F. Cootes, C.J. Taylor, A mixture model for representing shape variation, in: A.F. Clark (Ed.), BMVC, Essex, UK, 1997, pp. 110–119.
[11] R. Bowden, T.A. Mitchell, M. Sarhadi, Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences, Image Vision Comput. 18 (2000) 729–737.
[12] H. Ning, T. Tan, L. Wang, W. Hu, People tracking based on motion model and motion constraints with automatic initialization, Pattern Recognition 37 (2004) 1423–1440.
[13] C.J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Comput. Vision Image Understanding 80 (2000) 349–363.
[14] L. Sigal, S. Bhatia, S. Roth, M.J. Black, M. Isard, Tracking loose-limbed people, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, July 2004, pp. 421–428.
[15] R. Urtasun, D.J. Fleet, A. Hertzmann, P. Fua, Priors for people tracking from small training sets, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 403–410.
[16] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 65–81.
[17] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff, Estimating 3D body pose using uncalibrated cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, 2001, pp. 821–827.
[18] G. Mori, J. Malik, Recovering 3D human body configurations using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1052–1062.
[19] E.J. Ong, A.S. Micilotta, R. Bowden, A. Hilton, Viewpoint invariant exemplar-based 3D human tracking, Comput. Vision Image Understanding 104 (2006) 178–189.
[20] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 44–58.
[21] A.M. Elgammal, C.S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. II, July 2004, pp. 681–688.
[22] K. Grauman, G. Shakhnarovich, T. Darrell, Inferring 3D structure with a statistical image-based shape model, in: Proceedings of the IEEE International Conference on Computer Vision, October 2003, pp. 641–648.
[23] L. Fan, K.K. Sung, T.K. Ng, Pedestrian registration in static images with unconstrained background, Pattern Recognition 36 (2003) 1019–1029.
[24] A. Veeraraghavan, A.K. Roy-Chowdhury, R. Chellappa, Matching shape sequences in video with applications in human movement analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1896–1909.
[25] J. Giebel, D. Gavrila, C. Schnörr, A Bayesian framework for multi-cue 3D object tracking, in: Proceedings of the IEEE European Conference on Computer Vision, May 2004, pp. 241–252.
[26] J. Zhang, R. Collins, Y. Liu, Bayesian body localization using mixture of nonlinear shape models, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 725–732.
[27] D. Ponsa, F.X. Roca, A novel approach to generate multiple shape models for tracking applications, in: Proceedings of the International Conference on Articulated Motion and Deformable Objects, July 2002, pp. 80–91.
[28] A.A. Al-Shaher, E.R. Hancock, Learning mixtures of point distribution models with the EM algorithm, Pattern Recognition 36 (12) (2003) 2805–2818.
[29] R. Gross, J. Shi, The CMU motion of body (MoBo) database, Technical Report CMU-RITR, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2001.
[30] EC Funded CAVIAR project IST 2001 37540, ⟨homepages.inf.ed.ac.uk/rbf/CAVIAR/⟩.
[31] L. Sigal, M.J. Black, HumanEva: synchronized video and motion capture dataset for evaluation of articulated human motion, Technical Report CS-06-08, 2006.
[32] A. Koschan, S. Kang, J. Paik, B. Abidi, M. Abidi, Colour active shape models for tracking non-rigid objects, Pattern Recognition Lett. (2003) 1751–1765.
[33] Rh.H. Davies, C.J. Twining, P. Daniel Allen, T.F. Cootes, C.J. Taylor, Building optimal 2D statistical shape models, Image Vision Comput. 21 (2003) 13–14.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012

[34] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, New York, 2001.
[35] R.Y. Tsai, An efficient and accurate camera calibration technique for 3D machine vision, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 1986, pp. 364–374.
[36] G. Rogez, C. Orrite-Uru nuela, J. Martínez del Rincón, Human figure segmentation using independent component analysis, in: IbPRIA, 2005, pp. 300–307.
[37] V.T. Inman, H.J. Ralston, F. Todd, Human Walking, Williams and Wilkins, Baltimore, USA, 1981.
[38] T. Heap, D. Hogg, Wormholes in shape space: tracking through discontinuous changes in shape, in: Proceedings of the IEEE International Conference on Computer Vision, 1998, pp. 344–349.
[39] C. Orrite Uruñuela, J. Martínez del Rincón, J.E. Herrero Jaraba, G. Rogez, 2D silhouette and 3D skeletal models for human detection and tracking, in: ICPR, 2004, pp. 244–247.
[40] X. Lan, D.P. Huttenlocher, A unified spatio-temporal articulated model for tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, July 2004, pp. 722–729.
[41] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and Viterbi path searching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2007.
[42] M. Dimitrijevic, V. Lepetit, P. Fua, Human body pose detection using Bayesian spatio-temporal templates, Int. J. Comput. Vision Image Understanding 104 (2006) 127–139.
[43] G. Rogez, J.J. Guerrero, J. Martínez, C. Orrite, Viewpoint independent human motion analysis in man-made environments, in: Proceedings of the 17th British Machine Vision Conference, vol. II, September 2006, pp. 659–668.

About the Author—GRÉGORY ROGEZ graduated in Physics from Ecole Nationale Superieure de Physique de Marseille in 2002. He received the M.Sc. degree in Electrical Engineering from the University of Zaragoza in 2005. He is currently completing his PhD studies in the Computer Vision Laboratory of the Aragon Institute of Engineering Research (I3A). His main research interests include computer vision, learning and human motion analysis.

About the Author—CARLOS ORRITE-URUÑUELA received the master degree in Industrial Engineering at the Zaragoza University in 1989. In 1994, he completed the master degree in biomedical engineering working in the field of medical instrumentation for several industrial partners. In 1997 he did his PhD on computer vision at the Zaragoza University. He is currently associate professor at the Department of Electronics and Communications Engineering, at the University of Zaragoza and carries out his research activities in the Aragon Institute of Engineering Research (I3A). His research interests are in the area of computer vision and human–machine interface. He has participated in several national and international projects. He supervises several MSc and PhD students in the area of computer vision, biometrics and human motion analysis.

About the Author—JESÚS MARTINEZ-DEL-RINCON received the MSc degree from the University of Zaragoza specializing in Biomedical Engineering in 2006. He previously graduated from the University of Zaragoza in Telecommunication in 2003. He is currently pursuing Doctoral studies specializing in computer vision, motion analysis and human tracking.

Please cite this article as: G. Rogez, et al., A spatio-temporal 2D-models framework for human pose recovery in monocular sequences, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.02.012