2D Silhouette and 3D Skeletal Models for Human Detection and Tracking

Carlos Orrite-Uruñuela, Jesús Martínez del Rincón, J. Elías Herrero-Jaraba, Grégory Rogez
Aragon Institute for Engineering Research
University of Zaragoza, María de Luna 1, 50018 Zaragoza, SPAIN
e-mail: {corrite, jesmar, jelias, grogez}@unizar.es

Abstract

In this paper we propose a statistical model for detection and tracking of human silhouette and the corresponding 3D skeletal structure in gait sequences. We follow a point distribution model (PDM) approach using a Principal Component Analysis (PCA). The problem of non-lineal PCA is partially resolved by applying a different PDM depending of pose estimation; frontal, lateral and diagonal, estimated by Fisher`s linear discriminant. Additionally, the fitting is carried out by selecting the closest allowable shape from the training set by means of a nearest neighbor classifier. To improve the performance of the model we develop a human gait analysis to take into account temporal dynamic to track the human body. The incorporation of temporal constraints on the model increase reliability and robustness.

1. Introduction

Visual analysis of human motion is currently one of the most active research topics in computer vision and has led to several surveys. The last relevant review was probably due to Wang et al. [1], however, some previous reviews should be mentioned: Cedars and Shah [2] and Aggarwal and Cai [3]. Following these reviews the entire field of computer vision based human analysis is traditionally divided into: face analysis, gesture analysis and body analysis. In this work we study body movements. According to [2], there exist several models to work with the human body. Some of them are based on stick figures and others on volumetric ones. In our case, we employ a skeleton-model to represent the human body: we use a 3D model of stick figure to matching in 2-D data. We have selected 13 fundamental points that define human motion: center of the head, shoulders, elbows, wrists, hips, knees and ankles are found to build the body skeletal structure.

The present approach can be divided into two phases: an off-line stage to generate the shape model from a training set using them to extract the mean shape and variation modes applying Principal Component Analysis (PCA); and an on-line stage consisting in locating and fitting the model in the image, and correcting it to the closest plausible model. The 2D silhouette of a moving human and the corresponding 3D skeletal structure are encapsulated within a point distribution model (PDM). Some previous works in learning deformable models for tracking human motion are applied in [4], [5], [6].

The fitting process begins with a roughly human figure detection based on image difference and background subtraction. Then, we determine the pose of the figure to apply a specific model depending of the view. Fitting is an iterative process, where the model uses suggested movement from control points to find natural resting place. Once we have the silhouette we obtain automatically from the training set the skeletal structure corresponding to the matched contour of the database.

The rest of this paper is organized as follows: in section 2, we briefly present the model learning algorithm followed. In section 3 we explain the model fitting procedure to obtain the silhouette and skeleton in an image. In section 4 we introduce a human gait analysis carried out to take into account temporal information in a video sequence. Some results are presented in section 5. Finally, in section 6 we present some conclusions.

2. Human shape model

Our main goal is to construct a mathematical model which represents a human body and the possible ways in which it can deform. Point distribution models (PDM) are used to obtain a complete silhouette of a non-rigid object, as well as the corresponding skeletal structure.

2.1. Training

The generation of a 2D deformable model of the human contour follows a similar procedure as described in [4] and [5]. For the training stage we use the CMU Motion of Body (MoBo) database [7]. Good training contours, which are taken manually, are located combining two different grids, one rectangular and other semicircular used for taking point between the legs, to extract 49 landmark points for the human silhouette. Simultaneously, we extract 13 points corresponding to a stick model. So, we obtain 3600 contours with their respective stick models, corresponding to 9 subjects in 4 different views (frontal, lateral, diagonal backwards and

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

diagonal forwards). We align all the training set using Procrustes analysis to free the model of position, size and rotation (each view must be independently aligned). Finally, we apply PCA analysis to reduce the number of parameters.

2.2. Pose estimation

The possible shapes generated by the combination of the primary modes of variation are not indicative of the training set due to its inherent non-linearity. In order to produce a model more accurate for practical applications, we divide the global training set in several cluster attending to the pose of the human figure. It is clear that the silhouette of a human figure taken from a frontal view is quite different from a lateral one. To reduce the non-linearities of a global PDM, we first estimate the most probable pose of the human view, and then we apply a specific PDM for this view. We have chosen four different pose: frontal, lateral, diagonal backwards and diagonal forwards, which encompass most of the silhouettes we can find in real situations. To carry out this pose classification, we first determine a global PDM taking into account all different views from the training data set. Thus, by means of a LDA, we make a supervised (cluster-based) classification of the four distinct poses. To classify a new silhouette shape we use the Mahalanobis distance. We have tested the pose classification method with other images taken from the MoBo database obtaining a mismatch error lower than 0.35%.

3. Model fitting

As model fitting we call the process by means we find the position of the silhouette and skeleton points in a new image by matching the model generated off-line in the blobs detected in the current image.

First of all some image processing is performed to obtain the blobs corresponding to humans. Then we determinate the pose of the human, so we can apply a specific PCA model to the blobs that have been found in the previous state. Finally, once we obtain an acceptable silhouette, the skeleton of the figure is located since we have correlated silhouette with skeleton in the training process.

3.1. Blob detection

A widely used technique for separating moving objects from theirs background is based on subtraction and thresholding. Assuming the camera is stationary with fixed lighting conditions and good contrast, the method can be used to segment moving objects in a scene. In this approach an image $B(x,y)$ of the background is stored before the introduction of a foreground object. Then, given an image $I(x,y)$ from a sequence, feature detection of moving objects are restricted to areas of $I(x,y)$ where $|I(x,y)- B(x,y)| > \sigma$, where σ is a suitable chosen noise threshold.

Blob detection is the most critical state of the model fitting. Experimentally we usually have found distorted and split blobs corresponding to a single person. To improve the blob segmentation we first apply a model fitting to obtain a roughly contour and then we reduce the threshold value inside to detect more foreground pixels. This process is repeated iteratively to convergence in a more reliable silhouette, as depicted in figure 1.

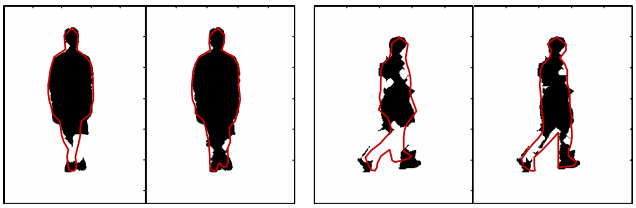

Figure 1. Blob detection and completion.

3.2. Silhouette matching

PCA model is used to constrain the shape of the PDM when applied to an image. The fitting process is an iterative one, where the model uses suggested movement from control points to find natural resting place. Movement of the model is allowed through the relocation of the model within the image plane using translation, rotation and scaling. Deformation of the model is also permitted by finding the closest allowable shape as determined by the bounds of the mathematical model of deformation. Given a new shape x, the closest allowable shape from the model is constructed by finding b such as: $b = \Phi^T(x - \bar{x})$. As pointed in [5], each component of vector b can be clipped in a range given by $-3\sqrt{\lambda_i} \le b_i \le +3\sqrt{\lambda_i}$, obtaining vector b^*.

Following this approach, the closest allowable shape can then be reconstructed as $x \approx \bar{x} + \Phi b^*$.

To this point, we have used a linear PCA; however, non-linearities are present in a model where parts of the contour display rotational characteristics in the image plane around some pivotal point. In this work we have used a simplification to avoid non-linearities. As mentioned previously, given a new shape x, the closest allowable shape from the model is constructed by finding b such as: $b = \Phi^T(x - \bar{x})$, but we select in each iteration the closest allowable shape from the training set by means of a nearest neighbor classifier. This technique always warranties acceptable contour determination, reducing noise effect errors in the blob detection.

3.3. Skeletal structure of human body

During the training state we have selected manually 49 points corresponding to the contour in four different views, and simultaneously, the 13 points of the skeleton in the four views. We have correlated all points: silhouette and skeleton, so, once we have located the contour of the human figure, the skeleton is recovery immediately from the corresponding silhouette of the training set (to one silhouette corresponds a skeleton).

Moreover, as we have correlated the skeletal points between different views, we are able to construct a 3D skeleton model, just inverting the camera calibration process carried out in the generation of the MoBo database. In figure 2 we represent a walking cycle corresponding to one subject in several views.

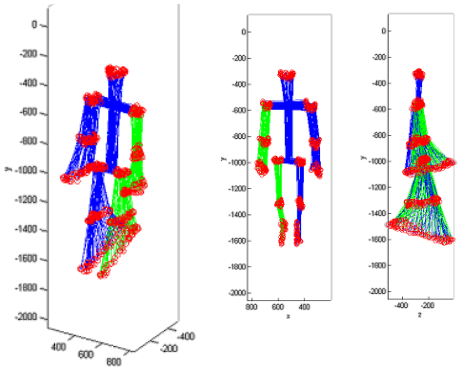

Figure 2. 3D skeleton model.

4. Human gait analysis

Up to this point we have presented a statistical model of the human figure based on a single image, without taking into account the dynamic of the gait encompassed in a temporal sequence. It is obvious that temporal information will improve the model fitting reducing the number of possible candidates from the database to match. Considering the human gait, we have divided a natural cycle into four states: left foot stop and right foot moving behind, left foot stop and right foot moving ahead; and vice versa, right foot stop and left moving behind or ahead. We have used a k-means clustering algorithm to the skeleton points to divide the whole space for every pose into four sub areas in relation to the previous steps defined in a gait cycle. In the clustering analysis we have considered only the six points of the skeleton corresponding to the hips, knees and ankles. Once we have all skeleton classified we apply a Fisher Linear Discriminant analysis to separate the characteristic space into disjoined regions, see figure 3. This state diagram, with four steps and four transactions, produces a more accurate and robust model and at the same time reduces the neighborhood space used to fit a new figure.

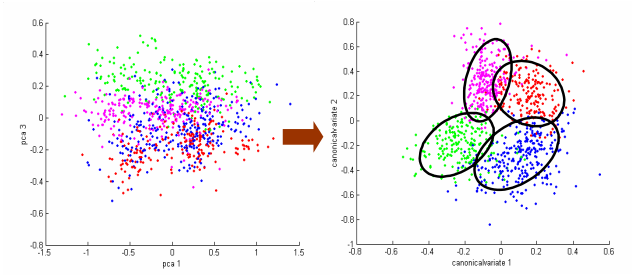

Figure 3. Human gait clustered into four steps.

5. Results

In our experiments we have verified that the human contour is a good indicator to determinate the skeleton of the person. Obviously, as more precise is the silhouette more exact will be the corresponding stick figure as depicted in figure 4.

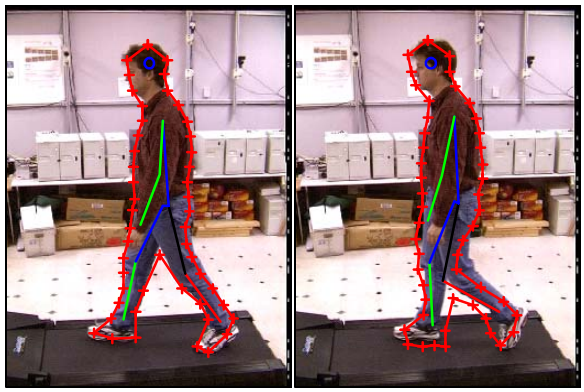

Figure 4. Silhouette and skeleton fitting.

As we have generated a 3D stick model of the human figure, we can employ this model to recovery the position of some hidden arms and legs in natural movements. In figure 5 we show some results obtained in different camera positions where an arm is completely occluded by the body.

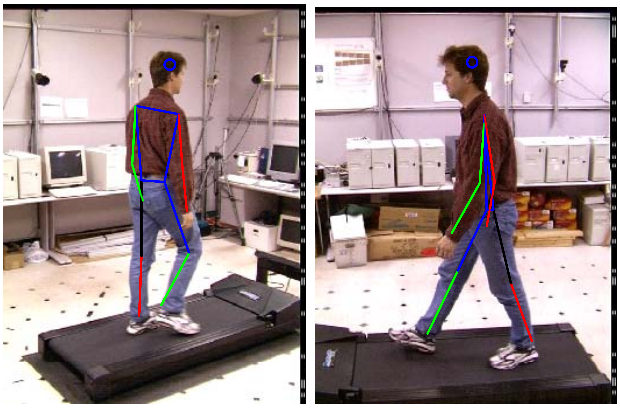

Figure 5. Hidden limb detection by 3D skeleton projection.

In figure 6 we present the silhouette and skeleton of a person in a completely different sequence in relation to the pictures from the MoBo database, with different camera point of view, frame rate and image size. To improve the segmentation process we have used a Kalman filter to track the bounding box corresponding to the human blob. We track the centroid of the blob and the width and height. This tracking process reduces significantly problems related with the blob location in very poor contrasted foreground in relation to background, as well as computational cost.

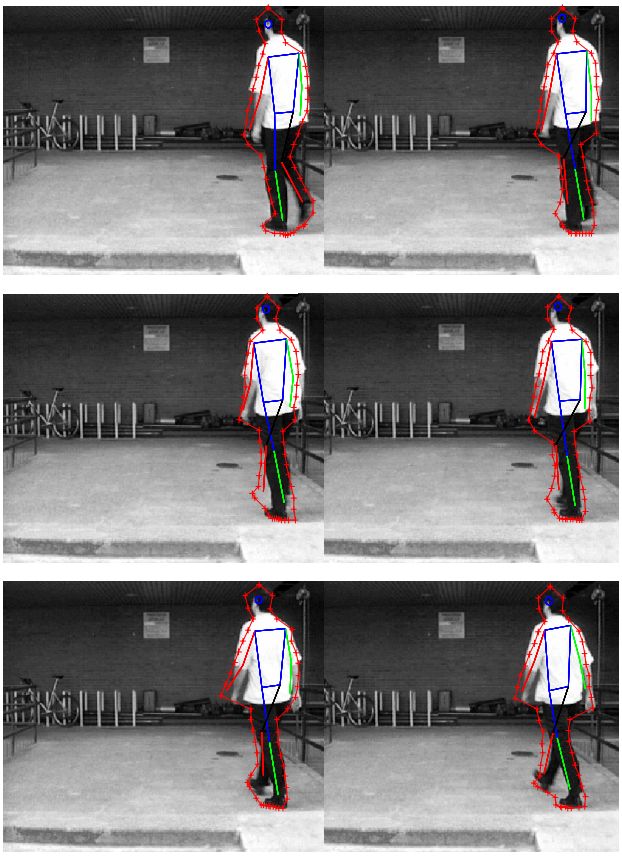

Figure 6. Skeleton detection and tracking.

6. Conclusions

In this paper we have developed a statistical model for human silhouette and the corresponding 3D skeletal structure. This model can be used to determinate the pose and structure of the human body from a monocular view.

The problem of non-linear principal component analysis is partially resolved by applying a different PDM depending of pose estimation; frontal, lateral, diagonal, estimated by a linear discriminant analysis. Additionally, the fitting is carried out by selecting the closest allowable shape from the training set by means of a nearest neighbor classifier. However, to cope with this problem we consider the necessity to use non-linear statistical models as proposed in [8].

To obtain binary regions we do not base the final result to the appropriate selection of a threshold, but we automatically update dynamically this parameter in relation to the human silhouette we are iteratively matching.

To improve the performance of the model we develop a human gait analysis to take into account temporal dynamic to track human body. The incorporation of temporal constraints on the model increase reliability and robustness.

A truly 3D skeletal structure model allows us to predict hidden human body parts in 2D images.

Experimental results show the goodness of the present method. However, to improve tracking of a person in complex images where small body movements can cause huge discontinuities in the feature points, we have to consider a more complex human motion analysis and also to use a particle filter for tracking.

Acknowledgments

This work is partially supported by grants TIC2003-08382-C05-05 from (MCyT) and FEDER.

References

[1] Wang, L., Hu, W., Tan, T.: Recent development in human motion analysis. Patter Recognition. 36 pp. 585-601, (2003).

[2] Cedras, C., Shah, M.: Motion-based recognition: a survey. Image Vision Comput. 13 (2) pp. 129-155, (1995).

[3] Aggarwal, J.K., Cai, Q.: Human Motion analysis: a review. Proceedings of the IEEE Workshop on Motion on Non-Rigid and Articulated Objects, pp. 90-102, (1997).

[4] A. Koschan, S. Kang, J. Paik, B. Abidi, M. Abidi, "Color active shape model for tracking non-rigid objects", Pattern Recognition Letters 24, pp. 1751-1765, (2003).

[5] L. Marcenaro and C. Regazzoni, "Dynamic Shape Detection for Multiple Camera Systems", Multimedia video-based Systems: Requirements, Issues and Solutions, Kluwer Academic Publishers, pp. 71-83, (2000).

[6] A.M. Baumberg, "Learning Deformable Models for Tracking Human Motion", University of Leeds, (1995).

[7] The CMU Motion of Body (MoBo) Database, http://www.hid.ri.cmu.edu, under "databases".

[8] 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 and Vision Computing 18, pp. 729-737, (2000).