PROBABILISTIC HUMAN POSE RECOVERY FROM 2D IMAGES

F. Flitti, M. Bennamoun, D. Q. Huynh and R. A. Owens

The University of Western Australia
School of Computer Science and Software Engineering
E-mail: name@csse.uwa.edu.au

ABSTRACT
Image based human pose recovery has many applications in different industries such as games, entertainment, physiological rehabilitation and biometrics. This paper presents a new pose estimation algorithm from monocular images based on a nonlinear mapping of human silhouettes, coded using a collection of local image moments, to the pose space using a mixture of Neural Networks (NN) regressors. All parameters are estimated automatically. Experiments and comparative results show a superior performance of the proposed method.

Index Terms— Human pose, silhouette, image moments, Gaussian Mixture Model, Neural Networks.

1. INTRODUCTION
Interest in image-based human pose estimation is motivated by the need for automatic non-invasive systems in human movement analysis for applications such as games and entertainment industries, physiological rehabilitation, sport training, automatic intelligent surveillance and biometrics. The estimated pose permits the augmentation of real scenes by animating virtual avatars with the extracted human motion, the analysis of the movement of patients and athletes, the recognition of suspicious actions or the analysis of gait. Good recent literature overviews of this research field are provided in [1] and [2].

Work on pose estimation is mainly grouped into two categories: generative and discriminative approaches [2]. Generative methods use explicitly the geometry of the human body. They can be further divided into two groups: top-down and bottom-up methods. Top-down techniques project the human model into the 2D image space and measure a distance between the projection and the observed human image. The estimation process then minimizes the distance to find the best candidate in the pose space observing the physical constraints of the human body. This approach is however expensive in terms of computation resources, needs a good initialization of the solution and the distance function usually has many local minima [1, 2, 3]. Bottom-up techniques detect the parts of the human body and assemble them to reconstruct the human pose in accordance with the human body’s physical constraints. These methods require good body part detectors to maintain high estimation performance [2]. Discriminative methods define a direct mapping from the image descriptor space to the human pose space. Two main subclasses fall under this category, namely learning-based (LB) [3, 4] and example-based (EB) [5, 6] methods. LB techniques learn the desired mapping using a selected training data set. These approaches rely strongly on how representative is the training data set as well as the generalization power of the mapping, and can be coupled with generative techniques to provide an initial estimate. EB techniques use a discrete set of specific poses with their corresponding representations. The pose estimation is performed by interpolation using the subset of poses with the highest similarities. These techniques are generally greedy in storage resources.

In this paper, we present a new LB technique based on a nonlinear regression model that efficiently maps the human subject’s silhouette, coded using a set of image moments, to the human 3D pose represented by the joint angles between the human body parts. The new proposed mapping is a mixture of local regressors technique based on a multimodal modelling of the input and output data. A set of Neural Networks (NN) local regressors, associated with all input/output cluster pairs, are combined in a Bayesian framework according to the probabilities of the input/output cluster pairs learnt during the training process. The proposed method is evaluated on a large data set of 500 images generated using motion files, from the Carnegie Mellon University (CMU) motion capture (Mocap) data base [7], using the graphics package POSER from eFrontiers in a similar manner to the procedures in [3, 8, 9]. In [3], a pose estimation technique from silhouette images based on RVM and Shape Contexts (SC) [10] coding of the silhouette was used. In this paper we use much lower dimensional-shape descriptors which are the local image moments. In [11], a mixture of regressors was proposed as an extension of [3]. However, the authors used the mixture model jointly in the input and output spaces without modelling explicitly the probability of having an input/output cluster pair, as used in our model, which led them to a different regression formula. Besides, the number of clusters in [11] was not estimated automatically as proposed in this work.

This research is supported by the Australian Research Council DP00771294.
This paper is organized as follows. Section 2 details the proposed method with parameter estimation procedures. Section 3 reports experimental and comparative results. Section 4 concludes the paper.

2. THE PROPOSED METHOD

The pose estimation technique proposed in this work assumes, similarly to [3], that it is possible to extract the human’s silhouette from the background scene. This is a fair assumption since there exist many recent background subtraction algorithms, such as [12], that are very efficient especially with images where the variations of the background are negligible. The method is summarised in Fig. 1. The silhouette shape is coded using local image moments to generate description vectors of length 72 (Sect. 2.1). The Mixture of local input/output Neural Network (NN) regressors is then applied to the moment vectors to estimate the pose. Details of these two steps are given in the following subsections.

2.1. The local image moments for shape coding

The silhouette is the projection on the 2D camera plane of the 3D human convex hull associated with a given pose. It is much more relevant to pose recovery than the observed person’s texture or color [3, 4]. In addition, it can usually be extracted with good quality especially from controlled and low varying environment videos. It can however be locally corrupted by possible shadows in the scene and can suffer from some ambiguities caused by self occlusions. Due to the high dimensionality of the binary silhouette images, using an appropriate shape descriptor is crucial to reduce the dimensionality while keeping pertinent information. Many shape descriptors have been proposed in the literature [13, 14]. Fourier Descriptors (FD) [15], Shape Contexts (SC) [10] and Hu Moments (HM) [16] are among the descriptors used in pose estimation applications. In this paper, we propose to use local image moments (IMs) to code the silhouette’s shape. IMs are simple to compute and their use in different local regions of the silhouette should improve their performance and overcome their limitation in robustness to partial occlusion.

The normalised cropped silhouette image of size 128 × 64 is divided into five overlapping regions as shown in Fig. 2. For each region, in addition to the whole cropped image, 12 IMs are computed using the following formula:

\[
x^{pq} = \sum_{m} \sum_{n} m^n e^{j(m,n)}
\]

where \(m\) and \(n\) are coordinates in the considered image, \(e^{j(m,n)}\) is the intensity of the pixel at the location \((m,n)\) and \(p\) and \(q\) are integers in the interval \([0, 3]\). This gives 72 feature vectors \(x\) that need to be reduced further before being used for pose estimation. The dimensionality reduction is performed implicitly by the NN regressors in the proposed method.

2.2. The mixture of local input/output local mappings regression model

Let \(D = \{(x_i, y_i) : i = 1, \ldots, N, x_i \in \mathcal{R}^K, y_i \in \mathcal{R}^M\}\) be the set of pairs \((x_i, y_i)\), where \(x_i\) is an input vector, \(y_i\) is an output vector, \(i\) a pose vector. We seek a mapping \(f : \mathcal{R}^K \rightarrow \mathcal{R}^M\) that minimises the least squared error, \(J = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2\), between the estimates \(y_i\) and the true outputs \(y_i\).

We assume that both input and output have multimodal distributions which we model using Mixtures of Gaussian density models, \(p(x_i) \sim \sum_{j=1}^{L_x} p(\alpha_i = j)N(x_i; \mu^j_x, \Sigma^j_x)\) and \(p(y_i) \sim \sum_{j=1}^{L_y} p(\beta_i = j)N(y_i; \mu^j_y, \Sigma^j_y)\). The symbols \(\alpha_i\) and \(\beta_i\) refer to hidden variables associated with \(x_i\) and \(y_i\), respectively, indicating to which component of the mixtures they belong, \(L_x\) and \(L_y\) are the numbers of components in the mixtures, and \(N(z; \mu, \Sigma)\) is the multivariate Gaussian distribution with mean \(\mu\) and covariance matrix \(\Sigma\), evaluated at \(z\). The probabilities \(p(\alpha_i = j)\) and \(p(\beta_i = j)\) are simple to compute and their use in different local regions of the silhouette should improve their performance and overcome their limitation in robustness to partial occlusion.

The normalised cropped silhouette image of size 128 × 64 is divided into five overlapping regions as shown in Fig. 2. For each region, in addition to the whole cropped image, 12 IMs are computed using the following formula:

\[
x^{pq} = \sum_{m} \sum_{n} m^n e^{j(m,n)}
\]
p(β_i = j) are the a priori probabilities of the jth clusters in the input and output space respectively. Instead of using one global mapping, our method learns a local mapping for each input/output cluster pair. Accordingly, if x_i belongs to the the kth input cluster and y_j belongs to the mth output cluster, the output estimate y_{km} is obtained using a local NN regressor f_{km} between clusters k and m as follows:

\[ y_i = y_{km} + \epsilon_{km} = f_{km}(x_i) + \epsilon_{km}, \]

where \( \epsilon_{km} \) is the estimation error, assumed to be Gaussian with zero mean and covariance matrix \( \Phi_{km} \). Therefore, the probability of \( y_j \) given \( x_i \), \( \alpha_i = k \) and \( \beta_i = m \) is given by:

\[ p(y_i|x_i, \alpha_i = k, \beta_i = m) = N(y_i; f_{km}(x_i), \Phi_{km}) \]

(2)

From the above assumptions and definitions, the conditional probability of \( y_j \) given \( x_i \) is:

\[ p(y_j|x_i) \propto p(x_i,y_j) = \sum_k \sum_m p(x_i,y_i, \alpha_i = k, \beta_i = m) \]

\[ = \sum_k \sum_m N(y_i; f_{km}(x_i), \Phi_{km})p(x_i|\alpha_i = k)\pi_{km} \]

(3)

where the \( \pi_{km} = p(\alpha_i = k, \beta_i = m) \) is the a priori probability to have jointly the input and the output in the kth and mth clusters. The proportionality coefficient \( \frac{1}{\pi_{km}} \) is ignored as it is independent of \( y_j \). The inference of \( y_j \) is therefore possible using the Maximum A Posteriori (MAP) estimator, which gives:

\[ \hat{y}_i = \arg \max_y p(y|x_i) = \sum_k \sum_m f_{km}(x_i)p(x_i|\alpha_i = k)\pi_{km} \]

(4)

The parameters of this regression model are: the numbers of clusters \( L_x \) and \( L_y \), the parameters of the Gaussian mixture model of the input data, the a priori probability \( \pi_{km} \) and the parameters of the local NN regressor \( f_{km} \). The estimation of all these quantities will be addressed in the next section.

### 2.3. Model Parameter Estimation

We use the Akaikie Information Criterion (AIC) as a model selection criteria to determine the number of components in the Gaussian Mixture density Model (GMM) [17]. The estimation of the other parameters of the input and output data GMMs is carried out using the Expectation Maximisation (EM) algorithm [18, 17]. We use an empirical estimation of the prior based on the results of the EM algorithm. We cluster the input and output data sets by maximizing the a posteriori probabilities of the hidden variables associated with the GMMs modeling of the input and output data. Using this clustering, the a priori probabilities can be estimated by counting the number of occurrences of all possible cluster pairs’ combinations. This procedure is inspired by the segmental K-means algorithm for parameter estimation proposed by Rabiner et al. [19].

The local Neural Networks (NNs) used in this work are Multilayer Perceptrons [20]. To estimate the parameters of the local NNs, we use the result of the clustering obtained by the EM algorithm to select, for each regressor \( f_{km} \), the subset of data pairs, such that the input is in cluster \( k \) and the output is in cluster \( m \) as a training set. We use a two-layer feed-forward network with sigmoid hidden neurons and linear output neurons as a model for the local NN regressors. The number of neurons of the output layer is equal to the dimension of the output vectors while the number of neurons of the hidden layer is chosen through simulations. In our experiments, we found that 10 hidden neurons provides a good compromise between accuracy and complexity. The Levenberg-Marquardt (ML) backpropagation algorithm [21] is used to train the NN regressors.

### 3. EXPERIMENTS

To test our method, we used motion files from the CMU MoCap Database [7]. We used 11 major joint angles to represent the human body. The motion of each joint is characterized by three angles which gives a representation with dimension 33 for the output vectors of the poses. As all the joint angles are defined in the range \([-180°, 180°]\), abrupt changes of a joint angle can occur even though the movements are smooth. To overcome this problem, we replace the value of that angle by its sine and cosine values which usually have much smoother evolutions. We followed [3, 8, 9] and used the graphic package POSER from eFrontiers to generate a suitable video that corresponds to the motion files. This gives good quality silhouettes which allow us to assess reasonably the performance of the regression method independently from the silhouette extraction process. We randomly selected 5000 frames from different videos of walking sequences for training and 500 frames for testing. Our choice was motivated by the targeted application of physiological rehabilitation and gait analysis where walking movements are considered most of the time. To evaluate pose estimation algorithms, we followed [3] and used the average (over all M angles) Root Mean Squared (RMS) errors (in degrees) between joint angle vectors of the true and estimated poses as follows:

\[ \text{Err}(y_i, \hat{y}_i) = \frac{1}{M} \sum_{i=1}^{M} \| (y_i - \hat{y}_i) \| \mod 180° \]

where \( y \mod 180° \equiv (y + 180°) \mod 360° - 180° \) reduces angles to the interval \([-180°, +180°]\).

The average error obtained using the proposed method on the 500 test frames is 2.22°. Fig. 3 (a)-(b) provides two original and estimated poses using the proposed method for a visual appreciation of the performances. It can be easily seen that the pose is accurately recovered. Moreover, we tested a regression using a single global NN regressor and found an average error is 2.76° which demonstrates the effectiveness of the mixture modelling. The method presented in this work show better performances compared to those reported in [3, 11] using a similar testing approach and performance measures but with motion files from a different data set. One can note here that the local Image Moments used in this paper...
are very representative of the shape of the human silhouette, and their feature vectors exhibit a much lower dimensionality compared to the shape context distribution used in [3]. Fig. 3 (c)-(d) shows examples of recovered poses from real 2D images collected from the internet. It is worth noting here that even though these images are different from those used during training, the proposed method is able to fairly recover the poses.

![Fig. 3. Examples of recovered poses from the test sets. The CMU database: (a) the original pose and (b) the estimated pose. Real images from the internet: (c) the original image and (d) the estimated pose.](image)

4. CONCLUSION

A nonlinear regression technique based on a new formulation of a mixture of local NN regressors, each of which is associated with an input/output cluster pair, has been presented. This method makes explicit use of the multimodal nature of the input and output data that are common in complex regression problems. All model parameter estimations are carried out automatically. The application of the proposed method to human pose estimation shows an improvement in performance compared to the use of a global NN regressor. The obtained results also highlight the effectiveness of the local Image Moments in capturing the shape of the silhouette of the human subject. Current work includes the extension to pose estimation from multiple views and its tracking in videos.

5. REFERENCES


