3D Human Pose Estimation using 2D Body Part Detectors

Adela Bărbulescu [1,2] Wenjuan Gong [1] Jordi Gonzàlez [1] Thomas B. Moeslund [2] F. Xavier Roca [1]

[1] *Centre de Visió per Computador*
Universitat Autònoma de Barcelona

[2] *Aalborg University, Denmark*

Abstract

Automatic 3D reconstruction of human poses from monocular images is a challenging and popular topic in the computer vision community, which provides a wide range of applications in multiple areas. Solutions for 3D pose estimation involve various learning approaches, such as support vector machines and Gaussian processes, but many encounter difficulties in cluttered scenarios and require additional input data, such as silhouettes, or controlled camera settings.

We present a framework that is capable of estimating the 3D pose of a person from single images or monocular image sequences without requiring background information and which is robust to camera variations. The framework models the non-linearity present in human pose estimation as it benefits from flexible learning approaches, including a highly customizable 2D detector. Results on the HumanEva benchmark show how they perform and influence the quality of the 3D pose estimates.

1. Introduction

3D human pose estimation from monocular images represents an important and top researched subject in the computer vision community due to its challenging nature and widespread applications, ranging from advanced human computer interaction, smart video surveillance to arts and entertainment industry. The difficulty of the topic resides in loss of depth information that occurs during 3D to 2D space projection. Thus, a wide set of approaches have been proposed to tackle the problem of 3D configuration recovery from monocular images.

Due to the 2D-3D ambiguity, many approaches rely on well-defined laboratory conditions and are based on additional information such as silhouettes or edge-maps obtained for example from background subtraction methods [1, 2, 3, 4]. However, realistic scenarios present highly articulated human poses affected by self-occlusion, background clutter and camera motion, requiring more complex learning approaches.

A particular class of learning approaches use direct mapping methods from image features such as grids of local gradient orientation histograms, interest points, image segmentations to 3D poses [5, 6, 7, 8]. Another class of approaches maps the image features to 2D parts and then uses modeling or learning approaches to map these to 3D poses [9, 10]. Among these learning approaches, the most used ones are support vector machines, relevance vector machines and Gaussian processes. In [10] a comparison is presented between modeling and learning approaches in estimating 3D poses from available 2D data, using geometrical reconstruction and Gaussian processes.

This paper describes a two-stage framework which recovers 3D poses without requiring background information or static cameras. Image features are mapped to 2D poses using a flexible mixture model which captures co-occurrence relations between body parts, while 3D poses are estimated using a Gaussian process regressor. Experiments are conducted systematically on the HumanEva benchmark, comparing the 3D estimates based on different methods of mapping the image features to Gaussian process inputs.

2. Detector of 2D Poses

The dominant approach towards 2D human pose estimation implies articulated models in which parts are described by pixel location and orientation. The approach used by Ramanan [11] introduces a model based on a mixture of non-oriented pictorial structures.

The main advantages of using the articulated mixture model consist in the fact that it is highly customizable, using a variable number of body parts, and that it reflects a large variability of poses and appearances without requiring background or temporal information. Also, it outperforms state-of-the-art 2D detectors while requiring less processing time. The next sections describe the model proposed in [11]:

2.1. Part-based Model for Human Detection

The mixture model implies mixtures of parts or part types for each body part, in our case spanning different orientations and modeling the implied correlations. The body model can be associated with a graph $G = (V, E)$ in which nodes are represented by body parts and edges connect parts with strong relations.

Similar to the star-structured part-based model in [3], this mixture model involves a set of filters that are applied to a HOG feature map [12] extracted from the analyzed image. A configuration of parts for an n-part model specifies which part type is used from each mixture and its relative location. The score of a configuration of parts is computed according to three model components: co-occurrence, appearance and deformation [11]:

$$S(I,\ p,\ t) = \sum_{i \in V} b_i^{t_i} + \sum_{i \in V} w_i^{t_i} \cdot \Phi(I, p_i) + \sum_{i,j \in E} w_{ij}^{t_i} \cdot \Psi(p_i - p_j) \quad (1)$$

where the first term favors certain part type associations, the second term expresses the local appearance score by assigning weight templates associated to part i and part-type t_i to certain locations p_i, described by the extracted HOG descriptor, and the third term expresses the deformation score by assessing the part-type pair assignment parameters and the relative location between connected parts i and j.

As the model described is highly customizable, experiments have been deployed as to find a more efficient model structure by varying the number of part-types and mixtures. A full-body 26-part model (Figure 1) is chosen, as it shows increased performance due to the capture of additional orientation.

2.2. Inference and Learning

Inference using the mixture model described is obtained by retrieving the highest-scoring configuration, precisely by maximizing $S(I, p, t)$ (1) over all parts and part-types. Building the associated relational graph G as a tree allows for efficient inference with dynamic programming.

The solution used for training a model which generates high scores and outputs a set of parameters containing limb locations is a structural SVM, leading to a problem of quadratic programming (QP), which in this case is solved using dual coordinate-descent.

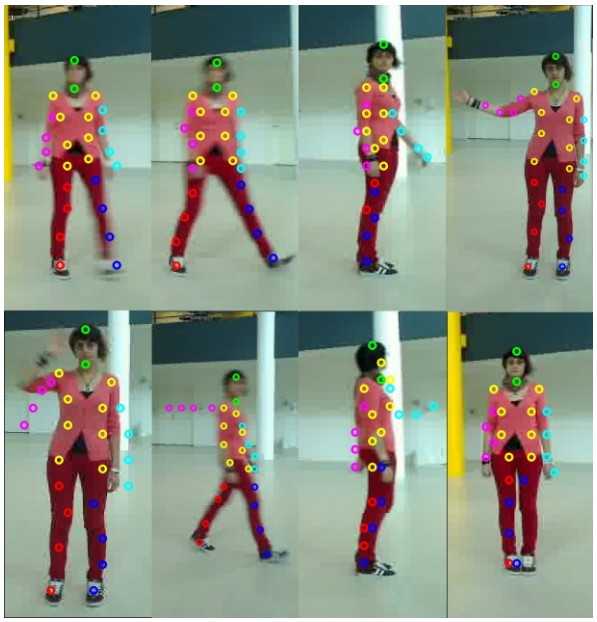


Figure 1. Person detected using a 26-part model, highlighting body part locations with circles. The upper row presents successful detections and the lower row presents limb misdetections.

Although the detector covers a wide variability of articulated poses, there are situations of limb misdetection, generated by self-occlusion, double-counting phenomena or background clutter.

3. Estimation of 3D Poses

Currently, Gaussian processes regression represents the most widespread learning method used in pose estimation, proving to be an effective approach for tackling the 2D to 3D mapping problem [5]. Given a prediction problem, Gaussian processes can be considered as a fine tool that extends a multivariate Gaussian distribution of the training data and which, using a correlation between observations and test data, maps the test data to new estimates. In our case, the input data is represented by the normalized and re-projected 2D body-part coordinates provided by the previously described detector and the output is represented by 3D pose estimates as direction cosines of limb orientations.

3.1. 3D pose representation

Considering the fact that the regressor outputs 3D poses, a robust representation is needed for the human pose. As training time is also an important factor, a smaller dimension representation is desirable. The human body is represented by a stick figure model composed of 13 body parts. As described in [13], a robust and efficient manner of representing 3D body limbs is the use of direction cosines. The angles of the limbs are considered with respect to a local coordinate system, fixed in the hip, with the *y* axis given by the torso, the *z* axis given by the hip line pointing from the left to right hip and the *x* axis given by the direction of their cross product.

The output is represented as a 36-dimensional vector:

$$v = \left[\cos\theta_1^x, \cos\theta_1^y, \cos\theta_1^z, \ldots, \cos\theta_{12}^x, \cos\theta_{12}^y, \cos\theta_{12}^z\right] \quad (2)$$

where $\theta_l^x, \theta_l^y, \theta_l^z$ represent the angles formed by a limb with the respective axes of the coordinate system. The use of direction cosines is robust and easily treatable as it prevents singular positions and discontinuities of angle values.

3.2. Gaussian process regression

Using Gaussian processes for prediction problems can be regarded as defining a probability distribution over functions, such that inference is defined in the function space-view. The training data observations $y = \{y_1, \ldots, y_n\}$ are considered samples from the *n*-variate Gaussian distribution that is associated to a Gaussian process and which is specified by a mean and a covariance function. Usually, it is assumed that the mean of the associated Gaussian process is zero and that observations are related using the covariance function $k(x, x')$. The covariance function describes how function values $f(x_1)$ and $f(x_2)$ are correlated, given x_1 and x_2. As the Gaussian process regression requires continuous interpolation between known input data, a continuous covariance is also needed. A typical choice for the covariance function is the squared exponential function:

$$k(x, x') = \sigma_f^2 \exp \frac{-(x - x')^2}{2l^2} \quad (3)$$

where σ_f represents the amplitude or the maximum allowable covariance, reached when $x \approx x'$ and $f(x)$ is very close to $f(x')$, and l represents the length parameter which influences the separation effect between input values. If a new input data x is distant from x' then $k(x, x') \approx 0$ and the observation x' will have a negligible effect upon the interpolation.

Therefore, Gaussian processes represent a flexible learning approach, capable of modeling the inherent non-linearity found in human pose estimation.

3.3. Testing and results

All experiments are carried on the HumanEva I dataset as it provides ground truth 2D and 3D information on subjects performing different actions. For every action, the image frames are equally divided in training and testing data, the input received being vectors of 2D coordinates. The measure of 3D estimation performance is computed using the average angle error and the average body part position error:

$$Err_{ang} = \frac{\sum_{i=1}^{J} \left|\theta_i - \hat{\theta}_i\right| \bmod 180^\circ}{J} \quad (4)$$

$$Err_{pos} = \frac{\sum_{i=1}^{M} \left|P_i - \hat{P}_i\right|}{M} \quad (5)$$

where J=3·14, for 3 Euler angles and 14 limbs, θ_i, $\hat{\theta}_i$, represent ground truth and predicted limb angles, M=3·15, for 3 coordinates per marker and 15 markers and P_i, $\hat{P}_i$, represent ground truth and predicted marker positions.

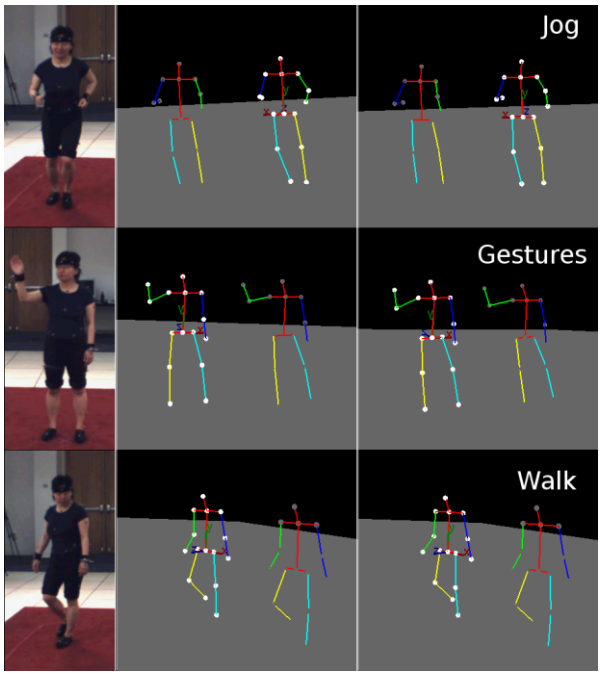
Experiments are conducted by varying the dimension of the input vectors containing the normalized 2D coordinates from the 2D detector. The final results are compared with an approach that uses a similar Gaussian process regressor and, as input, histograms of shape contexts obtained from extracted silhouettes [7]. As the silhouette-based experiments are carried in controlled conditions, requiring fixed cameras and background information, we will consider the method as ground truth experiment (GT).

The dimensions of the input are varied by manually choosing significant body parts and obtaining the associated coordinates by re-projecting the 2D coordinates. Ground truth data is obtained in a similar manner according to the HumanEva marker positions. The results show that using a simpler body representation for regression input performs better while training and prediction are less time consuming. Therefore, a 16-dimensional input is chosen containing normalized 2D coordinates corresponding to body parts: head, neck, upper and lower torso, two shoulders, two elbows, two wrists, two hips, two knees and two ankles. Results obtained for the two approaches are shown in Table 1:

Table 1. Results obtained on the HumanEva dataset

Input	Motion (CAM1, S1)	Err_{ang} [°]	Err_{pos} [mm]
Our system	Walking	1.85	41.50
	Box	2.68	45.45
	ThrowCatch	2.50	45.98
	Jog	2.64	49.93
	Gestures	0.89	12.07
GT	Walking	0.96	21.75
	Box	1.04	16.97
	ThrowCatch	1.08	19.19
	Jog	1.42	26.96
	Gestures	0.55	7.61

The shape context-based solution [7] outperforms the two-stage framework because of the increased reliability of the features extracted from silhouettes. The biggest error rate is obtained for the "Jog" database, where a bigger number of frames present self-occlusions and generate double-counting and limb misdetections. In the "Gestures" database the camera viewpoint is constant leading to a smaller error rate. Figure 2 presents visualizations of results for the HumanEva database:



(a) (b) (c)

Figure 2. (a) RGB human pose, (b) results using shape contexts (c) results of our approach. Estimated body parts are highlighted, while the simple body model represents 3D GT data.

4. Conclusion and future work

The paper presents learning approaches for the problem of 3D pose estimation from monocular images. The framework is composed of an articulated 2D detector with a varying number of body parts based on a structural SVM and a 2D to 3D Gaussian process regressor. Experiments carried on the HumanEva benchmark show that a simpler 2D body part model performs better, while the 3D estimates depend on the reliability of the 2D inputs. For future work, the 2D detector will be improved within the temporal context, using a "tracklets" approach [9] for different frame window sizes [10], followed by motion smoothing.

Acknowledgements

The authors acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010:MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133) and DiCoMa (TSI-020400-2011-55); along with the Spanish project TIN2009-14501-C02-02.

References

[1] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker. Detailed human shape and pose from images, *CVPR*, 2007
[2] J. Deutscher, I. Reid. Articulated body motion capture by stochastic search, *IJCV,* 2005
[3] L. Sigal, M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation, *CVPR*, 2007
[4] Th.B. Moeslund, A. Hilton, V. Kruger, L. Sigal. Visual Analysis of Humans: Looking at people, *Springer*, 2011
[5] A. Agarwal, B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 2006
[6] C. Ionescu, L. Bo, C. Sminchisescu. Structural SVM for visual localization and continuous state estimation. *ICCV,* 2009
[7] L. Bo, C. Sminchisescu. Structured output – associative regression. *CVPR,* 2009
[8] C. Ionescu, F. Li, C. Sminchisescu. Latent Structured Models for Human Pose Estimation. *ICCV*, 2011
[9] M. Andriluka, S. Roth, B. Schiele. Monocular 3d pose estimation and tracking by detection. *CVPR,* 2010.
[10] W. Gong, J. Brauer, M. Arens, J. Gonzàlez. On the Effect of Temporal Information on Monocular 3D Human Pose Estimation. *ICCV,* 2011
[11] D. Ramanan, Y. Yang. Articulated pose estimation using flexible mixtures of parts. *CVPR*, 2011
[12] Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection.1:886 –893 vol. 1, 2005
[13] I. Rius, J. Gonzàlez, J. Varona, and F. X. Roca. Actionspecific motion prior for efficient bayesian 3d human body tracking. *Pattern Recognition*, 2009