Integration of Local Image Cues for Probabilistic 2D Pose Recovery

Paul Kuo[1], Dimitrios Makris[1], Najla Megherbi[2], Jean-Christophe Nebel[1]

[1]Digital Imaging Research Centre, Kingston University, London, UK[1]
[2]Applied Mathematics and Computing Group, Cranfield University, UK

Abstract. A novel probabilistic formulation for 2-D human pose recovery from monocular images is proposed. It relies on a bottom-up approach based on an iterative process between clustering and body model fitting. Body parts are segmented from the foreground by clustering a set of images cues. Clustering is driven by 2D human body model fitting to obtain optimal segmentation while the model is resized and its articulated configuration is updated according to the clustering result. This method neither requires a training stage, nor any prior knowledge of poses and appearance as characteristics of body parts are already embedded in the integrated cues. Furthermore, a probabilistic confidence measure is proposed to evaluate the expected accuracy of recovered poses. Experimental results demonstrate the accuracy and robustness of this new algorithm by estimating 2-D human poses from walking sequences.

1 Introduction

Human pose recovery from a monocular camera is an important and challenging task in computer vision. Such technology would allow analysis of body postures for a range of applications from the study of athletes' performances during competitions to the detection of antisocial behaviours from images captured from CCTV cameras. A robust system should be able to deal with the complexity of human poses which includes a large posture space, self-occlusions and appearance which varies with individual, clothing and viewpoints. So far, techniques have only been proposed for constrained scenarios focusing on specific activities within a controlled environment. Therefore, a general solution remains a challenge for computer vision.

The goal of pose recovery is to localise a person's joints and limbs in either an image plan (2D recovery) or a world space (3D recovery), which usually results in the reconstruction of a human skeleton. In this work, we concentrate on 2D pose recovery. First, a sequence of 2D postures can be used for linear gait analysis [2]. Secondly, it is an essential step towards 3D pose recovery, which could be achieved by integrating camera self-calibration techniques [3][4]. The success of pose recovery is measured according to the accuracy of estimates of joint positions. However, we must accept some poses cannot be recovered because of self-occlusions or certain view-

[1] This work was partially supported by the EPSRC sponsored MEDUSA, PROCESS and REVEAL projects (Grant No. EP/E001025/1, EP/E033288 and GR/S98443/01 respectively)

points make this task impossible. Therefore, a robust pose recovery system should be able to evaluate the accuracy of joint estimates to detect those difficult postures.

In this paper, we propose a novel bottom-up method for 2D pose recovery. A clustering approach segments body parts from foreground pixels according to some relevant image cues. Clustering is driven by fitting a 2D human body model to obtain optimal segmentation. The model, as shown in Figure 2(a), consists of 10 body parts which is consistent with the segmentation resolution expected from CCTV images. Our method, unlike many state-of-the-art approaches, requires no training stage, as body part characteristics are already embedded in the selected image cues. Since a key application of our technique is the initialisation of human body trackers, a probabilistic confidence score is produced for each estimated pose. Therefore, initialisation is performed when postures are recovered with high confidence. Moreover, those scores could be used as prior observation probabilities to improve tracking.

The structure of this paper is as follows. After presenting related previous work, we detail in Section 2 our pose recovery algorithm. Then, in Section 3 a probabilistic confidence measure of pose recovery is introduced. Finally, quantitative results are given in Section 4 and conclusions and future work are addressed in Section 5.

1.2 Related Work

Estimation of 2D or 3D human body poses from either images or videos has been an active research topic in computer vision. There are many approaches to this task. They are usually divided into 2 main categories [5]: Bottom-up approaches attempt to piece together body parts to build a full body whereas Top-down ones start from a general structure which is broken down to gain insight into its compositional elements. Bottom-up methods include breaking an image into pieces according to salient edges by applying normalised graph cut (NCut) [6]. Then parsing rules are used to reconstruct human shapes. Other authors detect body parts by exploiting parallelism of part boundaries [7][8]. In [7], an edge map is computed by dividing edges in segments which are refined by constrained Delaunay triangulation. Then part candidates are identified by paring parallel lines according to anthropometric constraints. NCut and parallel lines have also been combined to recover body parts assembled using a Dynamic Programming approach [9]. Sigal and Black [10] estimate 3D poses using bottom-up 2D body part proposals. Primary part detectors (skin, head and limb detectors) are used to generate 2D proposals to infer 2D and then 3D poses. Similarly, a bottom-up parsing approach can construct multiple body model candidates [11]. All of them act as weak classifiers which are boosted to produce a final model. Finally, 2D poses have also been inferred by Data Driven Belief Propagation Monte Carlo algorithm using a variety of images cues, i.e. face, skin, shape and edges [12]. Among top-down approaches, a Pictorial Structures (PS) method has been proposed to model a holistic human body as a collection of parts arranged in a deformable configuration [13]. Poses are estimated by minimising a cost function consisting of individual body parts, and part paring. This was refined by adding constraints of symmetry and colour homogeneity in body parts [14] and tackling self-occlusion problem by using an extended body model containing occlusion likelihoods [15]. 3D recovery from stereo image sequences has also been suggested by linearly combing 3D pose proposals which are learned from 2D depth images and silhouettes from a large training set [16].

Recently, a top-down method has been used for body part tracking where trackers are initialised opportunistically by fitting a PS when a stylized pose is detected [17].

Top-down approaches are efficient in recovering general poses. However, they tend not to provide underlying body part segmentation and usually require complex search. On the other hand, bottom-up approaches are accurate in estimating individual limbs as local cues are used, but the full body structure may be estimated inaccurately due to wrong body part paring. In order to tackle this, we propose an algorithm which follows essentially a bottom-up approach to obtained accurate estimate of individual limbs, but also incorporates a body model fitting process to provide anthropometric constraints on detected body segments.

2 Methodology

2.1 Overview

Our pose recovery algorithm is based on clustering foreground pixels to segment body parts. Since initialisation is key to the performance of clustering algorithms, we propose an iterative scheme with refined initialisation to achieve optimal partition. This relies on a 2D body model which is fitted on the generated clusters. Not only does it introduce anthropometric constraints which only reveal possible human postures, but it also provides clustering initialisation points. The main advantage of our method is, in addition to recover poses, it is also provide a probabilistic measure of confidence of the pose recovery process.

The flow diagram of the pose recovery algorithm is shown in Figure 1. It takes a video containing a human figure as an input and generates for each frame a 2D skeleton with a confidence measure. The core of the algorithm is an iterative process involving body part **clustering** and human body **model fitting**. **Clustering** partitions foreground pixels belonging to a human character into a desired number of body parts according to a set of relevant image cues. Human body **model fitting** fits a 2D generic human body model on the produced clusters by adjusting its scale, limb ratios and articulated configuration. The centres of the fitted model pieces are used as estimates of centres of the body parts and therefore are used for clustering initialisation. These successions of clustering and model fitting processes iterate until the configuration of the fitted body model reaches a steady-state. Finally, a 2D skeleton representing the recovered 2D pose is generated with a confidence measure which expresses the expected quality of the pose recovery process.

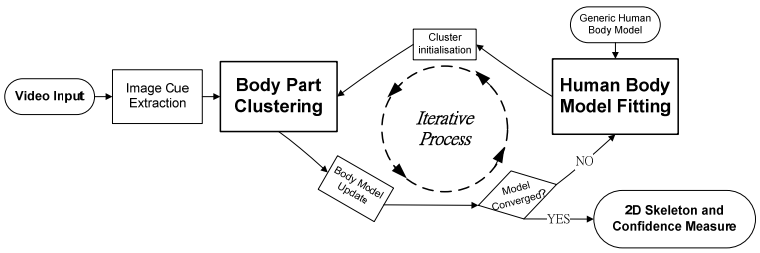


Figure 1: Flow diagram of the proposed pose recovery algorithm

2.2 Body Part Clustering

The aim of clustering is to partition foreground pixels into a predefined number of body parts according to a set of relevant image cues. In this section, image cues are introduced and then the clustering process is detailed.

2.2.1 Image Cues

In this work, we select **location**, **orientation**, **motion** and **colour** as the cues to partition the foreground pixels. These cues are collected and concatenated as feature vectors for each foreground pixel. This choice of cues aims at producing feature vectors which exhibit homogeneity within the body part and are distinctively different between adjacent body parts. Since body parts are displayed by a continuous set of pixels, except in some cases of occlusion, pixel **location** provides a first low level cue. Since human limbs are highly directional objects, they have been modelled as sets of parallel lines or trapeziums [7] whose main **orientations** describe the underlying skeleton. Therefore, directions of edges will be used as a cue to describe body parts. Because the human body is an articulated figure, pixel **motion** is usually discontinued from one body part to the other. The final cue for discriminating between body pieces is pixel **colour** since each body part can be modelled by homogenous colour or a low number of colour patterns [14].

Several imaging processing techniques have been employed to collect these image cues. Locations of the foreground pixels were obtained by conventional motion segmentation, along with shadow detection and foreground cleaning. Orientation cues were obtained by interpolating orientations of edges over foreground pixels. First, foreground edges are detected by Canny Edge Detection. Then, for the purpose of calculating their main orientation and removing spurious edges and noises, edges are converted to line segments via Hough transform. Finally, orientations are interpolated to all foreground pixels. Motion cues consist of two elements – speed and direction-, which are computed by Optical Flow. We adopted Lucas and Kanade's algorithm [18], as our algorithm requires dense motion information. Noise was suppressed by smoothing using a moving-average-of-5 temporal filter. Since preliminary experiments showed that the colour space choice did not affect results in body part detection, colour cues were expressed by RGB values.

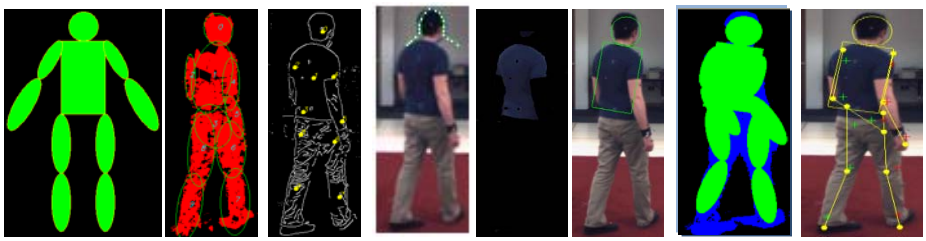
Since each cue can be considered as a weak classifier, robust body part detection can only be achieved by cue combination. Clustering is performed in a high dimensional space, called the cue space, where each foreground pixel can be projected to a location according to its feature vector. We define a projection function Φ, as shown in Equation (1), which transfers a 2D image pixel $p_i = (x_i, y_i)$ to an 8D feature vector $p_i' = (x_i, y_i, \theta_i, v_i, \beta_i, r_i, g_i, b_i)$ where $x_i, y_i, \theta_i, v_i, \beta_i, r_i, g_i, b_i$ are the associated feature vector components: location (x, y), orientation (θ), speed (v), direction (β), and colour (r, g, b).

$$\Phi : (x_i, y_i) \mapsto (x_i, y_i, \theta_i, \upsilon_i, \beta_i, r_i, g_i, b_i) \qquad (1)$$

2.2.2 Foreground Pixel Clustering

To produce confidence measures, we need to formulate our pose recovery problem in a probabilistic framework. Therefore, we adopt Gaussian Mixture Models (GMMs) to perform probabilistic clustering. The GMMs partition foreground pixels in the cue space into the desired number of body parts, i.e. 10. Therefore, a set of 10 probabilities, $P(p_i|C_j),\ j \in [0..9] so\ that\ \Sigma P(p_i|C_j)=1$, is produced for each foreground pixel p_i, indicating the likelihood pixels belonging to each of the 10 clusters, C_j. Figure 2(b) illustrates some partition results where a 2-standard deviation boundary has been drawn to represent each cluster.

GMMs are usually initialised by K-means clustering where the mean, weight and covariance of each of the Gaussian mixtures can be estimated. In standard K-means clustering, the process is initialised many times with random seeds so that optimal partition can be reached [19]. Since this is very time-consuming, we propose to use some prior knowledge of the body structure to initialise the clusters. A 2D articulated model, see Figure 2(a), is fitted onto either the foreground pixels for the first iteration or the produced clusters for subsequent iterations (see next section for details). This allows identifying the putative centres of the different body parts. These centre points are then projected to the cue space and used as the clustering seeds (Figure 2(c)).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 2: (a) 10-piece generic human body model. (b) Results of foreground (*red*) clustering (*green ellipses*) during first iteration. (c) Seeds (*yellow points*) generated by model fitting. (d) Head found by omega head detector (dotted *green*) and fitted by a circular head model (*blue*). (e) Pixels recovered by GMM torso detector. (f) Rectangular model (*green*) fitted on torso pixels. (g) Model fitted after first iteration. (h) Final posture estimate (*green and red crosses represent ground truth joint positions; yellow points and lines show recovered joints and skeleton*).

2.3 Human Body Model Fitting

The aim of human body model fitting process is to initialise the clustering process and to generate a 2D skeleton. A 2D body model is fitted onto either foreground or clustered pixels by maximising some overlapping costs. The process is hierarchical, where the most reliable parts are fitted first, i.e., head and torso, and are used as references for fitting the limbs. Fitting is performed iteratively with an increasingly accurate 2D model. During the first iteration, a generic model is fitted to the foreground to initialise the clustering. Then the model is resized according to cluster distributions.

In the subsequent iterations, the model is fitted on the clustered pixels which represent segmented body parts embedded in the foreground.

2.3.1 Human Body Model

The 2D generic human body model (M) we use is shown in Figure 2(a). It is an articulated figure consisting of 10 body pieces, $M=\{ m_{head}, m_{torso}, m_{lua}, m_{lla}, m_{rua}, m_{rla}, m_{rul}, m_{rll}, m_{lul}, m_{lll}\}$[2]. Similarly to other research groups', our model is composed of primary shapes [8][15][16]: the body model is made of a circle (head), a rectangle (torso) and 8 ellipses (limbs). Since this model provides anthropometric constraints, only plausible poses can be produced. The initial model is constructed using standard body part ratios [20] and its scale is estimated using the height of the segmented foreground. Then, during the model fitting process body part ratios and individual sizes are adjusted to be able fit a wide range of viewpoints and postures.

2.3.2 Model Fitting

The model fitting process starts by locating the head position. This is achieved using the omega head detection algorithm [21], where an Ω-shaped model is used to localise the head and shoulders in the image by minimising a Chamfer distance measure of edges. Location and size of the detected head are then represented by a circle in our body model, as shown in Figure 2(d). Then, the torso is detected by modelling its colours using GMM. The method proposed in [22][23][24] was implemented. First, a sampling region for the torso must be defined: a foreground region below the head is selected. The GMM is trained by using the colours drawn from this sampling region. Some colours are discarded as a statistical model is used to filter out noise [23]. Finally, torso pixels are detected from the segmented foreground by the trained GMM, as shown in Figure 2(e). The torso pixels are then fitted with a rectangular torso model, see Figure 2(f). Position, orientation, scale and height/width ratio are optimised, according to a "Mutually Overlapping Measure" (MOM):

$$MOM = \frac{A_{m_{torso}} \cap A_t}{A_{m_{torso}}} \times \frac{A_{m_{torso}} \cap A_t}{A_t} \quad (2)$$

Where $A_{m_{torso}}$ and A_t denote, respectively, the area of the torso model and the area of the detected torso pixels.

Finally, limbs need to be fitted. This process aims at translating and rotating the limb models to maximise some overlapping costs to the foreground or the body partitions, i.e. clustered pixels. In the initial fitting, the overlapping area between the models and foreground is maximised, while in the subsequent fitting where the clustering result is available, the joint probabilities, as defined in Equation (5), between the models and clusters are maximised. The joint probabilities will be discussed in detail in Section 3. Figure 2(g) illustrates the result of fitting the model on the foreground.

---

[2] *Apart from head and torso pieces, parts' names are abbreviated by 3 letters denoting: "left" or "right", "upper" or "lower" and "arm" or "leg".*

After clustering, the body model is updated by estimating the length of limbs from the produced clusters. This is achieved by locating the joints at the clusters' probabilistic boundaries using Equation (3): the length of a limb is defined as the Euclidean distance between the joints of adjacent clusters.

$$J_{j-k} = \arg\max_{i} \left\{ P(p_i \mid C_j) + P(p_i \mid C_k) - \left| P(p_i \mid C_j) - P(p_i \mid C_k) \right| \right\} \quad (3)$$

Where p_i, C_j and C_k denote the foreground pixel and the adjacent clusters. J_{j-k} is the joint between C_j and C_k. The conditional probabilities are given by GMM clustering. Since head and torso models were located using more robust methods, only limb models are updated according to the lengths of limbs estimated from the clusters.

The succession of clustering and model fitting processes iterates until the joint positions of the fitted body model reaches a steady-state. A skeleton, as shown in Figure 2(h), is then extracted from the final body model and a probabilistic confidence measure is calculated.

3. Probabilistic Confidence Measure for Pose Recovery

An important feature of our method is that a confidence measure is provided for every recovered pose. This probabilistic value is useful not only for pose evaluation but also for many applications built upon pose recovery. For example, body part tracking using either Kalman or Particle filter requires a prior probability to know how much an observation can be trusted [25]. Our confidence measure is constructed using Bays' theorem. If we assume the success of pose recovery is determined by the success of recovering all body parts and they are independent, the probability that a pose is recovered successfully *P(pose)* can be expressed by Equation (4).

$$P(pose) = \prod_{j} P(X_j) \quad (4)$$

Where $P(X_j)$ denotes the probability of each body part $X_j \in$ *{head, torso, rua, rla, lua, lla, rul, rll, lul ,lll}* to be recovered successfully and $P(X_j)$ expresses the likelihood that the model (m_j) and its corresponding clustering (C_j) overlap, i.e. the joint probability $P(m_j \cap C_j)$.

Assuming $p(m_j)$ is independent from $p(C_j)$, the joint probability can be obtained from the model fitting and clustering, according to Equation (5).

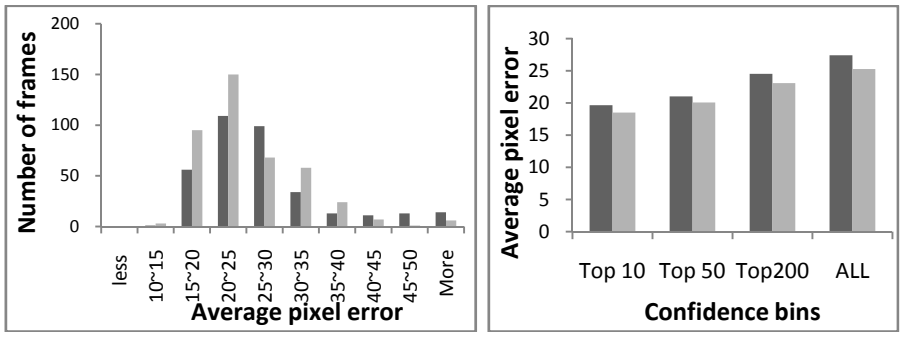
$$P(X_j) = P(m_j \cap C_j) = P(m_j \mid C_j) \times P(C_j \mid m_j) = \frac{\sum_i P(p_i \mid C_j \cap m_j)}{\sum m_j} \times \frac{\sum_i P(p_i \mid C_j \cap m_j)}{\sum_i P(p_i \mid C_j)} \quad (5)$$

Where $\sum m_j$ indicates the area of the model m_j, $\sum_i P(p_i \mid C_j)$ is the sum of the conditional probabilities of a pixel belonging to the cluster C_j (from the GMM clustering) over the entire foreground pixels and $\sum_i P(p_i \mid C_j \cap m_j)$ is the sum of the conditional probabilities of the foreground pixels fitted by the model m_j.

4 Experimental Results

The algorithm was tested over the HumanEva (HE) dataset [26], which is used as benchmark for pose recovery. It provides motion capture and video data which were collected synchronously. Therefore, motion capture data can be used as ground truth: since cameras are calibrated, 3D data points can be projected on the 2D sequences in order to evaluate quantitatively 2D pose estimates. Moreover, a standard set of error metrics [27] is defined to evaluate pose estimations. We conducted experiments with 2 walking sequences from these datasets: *S1 Walking (C1)* and *S2 Walking (C1)*. Since the original sequences are quite long, only one complete walking circle in both walking sequences are used, i.e. frame 280 to 700 for *S1 Walking (C1)* and frame 340 to 760 for *S2 Walking (C1)*. These sequences were chosen to include a variety of walking postures, i.e. a complete circle, seen from different viewpoints.

First, to evaluate the general performance of our algorithm, poses in all walking frames are estimated without considering their confidence measure. Figure 3(a) shows the histogram of the "average pixel error" between estimated joint locations and the ground truth for *S1 Walking* (dark gray*)* and *S2 Walking* (light gray) [27]. Most errors fall between a 15-30 pixels range which is comparable with results reported by other research groups [25]. Then to validate our confidence measure, estimated poses from both walking sequences were binned by descending order, into 5 classes: *Top 10*, *50*, *200* and *ALL* poses, depending upon their corresponding confidence measure. Figure 3(b) shows correlation between ground truth error and confidence measure supporting that the measure we propose is a useful indicator of the success of pose recovery. Figure 4 shows recovered poses from both sequences. The first 5 images from left are the ones with best confidence measures. As expected all of them are either frontal or back views as such views are the easiest to recover. The next 5 images show poses recovered from more difficult viewpoints, i.e. semi and full side views, which are also associated with good confidence measures, i.e. belonging to the Top 50 poses.



(a) (b)

Figure 3: (a) Histograms showing the average pixel error between the estimated joint locations and the ground truth. Results for *S1 Walking* and *S2 Walking* are shown in dark and light gray respectively. (b) Correlation between pixel error and confidence measure. Estimated poses are binned according to their confidence measures.

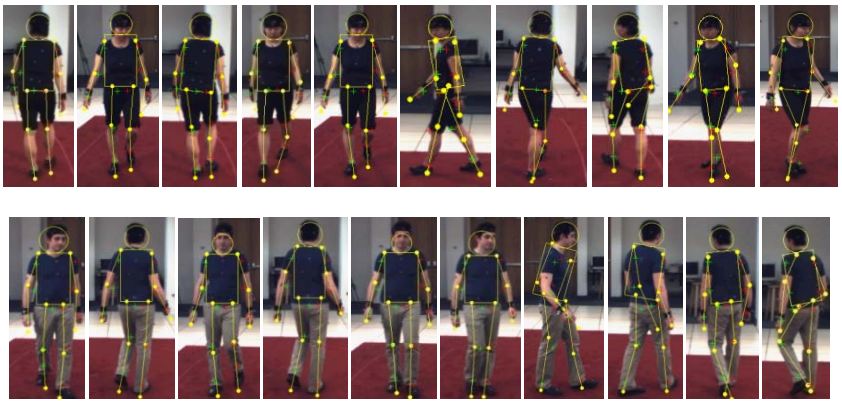


Figure 4: Results of pose recovery for both *S1 Walking* (*upper row*) and *S2 Walking* (*lower row*) sequence. The first 5 images from the left are the ones with best confidence measures and the next 5 images, which are also selected with good confidence measures (Top 50), show the recovery from more difficult viewpoints.

5. Conclusions and Future Work

In this paper, we proposed a probabilistic 2D pose recovery method using a combination of image cues. It is an iterative process of partitioning the foreground by clustering and fitting a human body model on the clusters. The clustering is initialised by the model fitting, while the body model is updated by the clustering. Since fitting and clustering are probabilistic, the estimated pose, obtained when the iterative process has converged, is associated with a confidence measure indicating the accuracy of the recovery. Our method was tested in two walking sequences containing a variety of postures seen from different viewpoints. Results demonstrate, first, our confidence measure can predict the accuracy of recovered postures and, secondly, our method is able to estimate reliably a substantial number of 2D poses. Therefore, the presented framework appears particularly suited to regular (re-)initialisations of body trackers.

Since the best recovered poses come usually from either frontal or back views, we plan to introduce a second body model, i.e. a side-view model, which should increase significantly the number of poses which are recovered with high accuracy. We also intend to integrate our method in a limb tracker framework [25]. Finally, we want to extend our work to 3D pose recovery by incorporating camera auto-calibration techniques [3].

6. References

[1] A. Elgammal and C.S. Lee. Inferring 3D body pose from silhouettes using activity manifold Learning. CVPR (2): 681-688. 2004.

[2] N. Spencer and J. Carter. Towards pose invariant gait reconstruction ICIP (2): 261-264. 2005.

[3] P. Kuo, J.-C. Nebel and D.Makris. Camera Auto-Calibration from Articulated Motion AVSS:135-140. 2007.

[4] M. Armstrong, A. Zisserman and R. Hartley. Self-Calibration from image triplets ECCV: 3-16. 1996.

[5] D. M. Gavrila. The visual analysis of human movement: A survey. Journal of computer Vision and Image Understanding. Vol.73. No.1. 82-98. 1999

[6] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body CVPR: 1-8. 2007 .

[7] X. Ren, A.C. Berg and J. Malik. Recovering human body configurations using pairwise constraints. ICCV: 824-831. 2005

[8] D. Ramanan and D.A. Forsyth. Finding and traking people from the bottom up CVPR (2): 467-474. 2003.

[9] G. Mori, X. Ren, A. A. Efros and J. Malik. Recovering human body configurations: Combing segmentation and recognition CVPR (2): 326-333. 2004

[10] L. Sigal and M. J. Black. Predicting 3D people from 2D pictures. AMDO. 2006.

[11] Y. Wang and G. Mori. Boosted multiple deformable trees for parsing human poses. HUMO: 16-27. 2007.

[12] G. Hua, M. H. Yang and Y. Wu. Learning to estimate human poses with data driven belief propagation. CVPR (2): 747-754. 2005

[13] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. IJCV: 55-79. 2005.

[14] D. Ramanan. Learning to parse images of articulated bodies. NIPS. 2007.

[15] L. Sigal and M.J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. CVPR (2): 2041-2048. 2006.

[16] H. D. Yang and S. W. Lee. Reconstructing 3D human body pose from stereo image sequences using hierarchical human body model learning. ICPR (3) 1004-1007. 2006.

[17] D. Ramanan, D. A. Forsyth and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. CVPR (1): 271-278. 2006

[18] B. D. Lucas and T. Kanade An iterative image registration technique with an application to stereo vision. Imaging understanding workshop: 121-130. 1981.

[19] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. Applied Statistics. 28 (1): 100-108. 1979.

[20] L. Da Vinci, Description of "Vitruvian Man", 1492.

[21] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. CVPR (2): 459-466, 2003.

[22] C. Hu, X. Ma and X Dai. A Robust person tracking and following approach for mobile robot. International Conf. on Mechatronics and Automation 3571-3576, 2007.

[23] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink and G. Sagerer. Audiovisual person tracking with a mobile robot. IAS, 898-906, 2004.

[24] S. J. Mckenna, Y. Raja and S. Gong. Tracking colour objects using adaptive mixture models. Image and Vision Computing (17) 255-231. 1999

[25] J. Martinez-del-Rincon, J.-C. Nebel, D. Makris, C. Orrite. Tracking Human Body Parts Using Particle Filters Constrained by Human Biomechanics, BMVC 2008.

[26] http://vision.cs.brown.edu/humaneva. HumanEVA dataset, Brown University.

[27] L. Sigal and M. J. Black, HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Tech. Report CS0608, Brown Univ. 2006.