Integration of Bottom-Up/Top-Down Approaches for 2D Pose Estimation Using Probabilistic Gaussian Modelling

Paul Kuo, Dimitrios Makris, and Jean-Christophe Nebel

Digital Imaging Research Centre, Kingston University, UK

***Abstract*— In this paper, we address the recovery of human 2D postures from monocular image sequences. We propose a novel pose estimation framework which is based on the integration of probabilistic bottom-up and top-down processes which iteratively refine each other: foreground pixels are segmented using image cues whereas a hierarchical 2D body model fitting constraints body partitions. Its main advantages are two-fold. First, the presented framework is activity-independent since it does not rely on learning any motion model. Secondly, we propose a confidence score indicating the quality of each estimated pose. Our study also reveals significant discrepancy between ground truth joint positions according to whether they are defined by humans or a motion capture system. Quantitative and qualitative results are presented on a variety of video sequences to validate our approach.**

***Index Terms*— Human body pose estimation, stochastic clustering, Gaussian mixture modelling, pattern classification, object recognition, confidence measure, ground truth.**

1. INTRODUCTION

Human pose recovery from a monocular camera is an important and challenging task in computer vision. Such research aims at allowing analysis of body postures for a range of applications from the study of athletes' performances during competitions to the detection of antisocial behaviours on images captured

from CCTV cameras. A robust system should be able to deal with the complexity of human poses, including self-occlusions, and observation variations due to different clothing, lighting and camera viewpoints. So far, techniques have only been proposed for constrained scenarios focusing on specific activities within a controlled environment. Therefore, a general solution remains an active research topic in computer vision.

The goal of pose recovery can be defined as the localisation of a person's joints and limbs in either an image plane (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 as it is the corner-stone of human motion analysis. For example, a sequence of 2D postures can be used for the study of linear gait [1]. It is also an essential step towards 3D pose recovery [2], which could be achieved, for example, by integrating camera self-calibration techniques [3][4]. The success of pose recovery is usually measured according to the accuracy of estimates of joint positions. However, we must accept some poses cannot be recovered using a single camera because of self-occlusions or certain view-points make this task impossible. Therefore, a robust pose recovery system should be able to evaluate the accuracy of joint estimates to recognise those difficult postures.

In this paper, we propose a novel probabilistic bottom-up/top-down approach for 2D pose recovery. The bottom-up module uses clustering technique to segment body parts from foreground pixels by integrating multiple image cues which consist of low-level image features (i.e. location, colour, edge orientation and optical flow) and advanced descriptors (haar-based adaboost responses, SIFT correspondences and omega model values). Clustering is driven by the top-down module of fitting a 2D human body model to obtain optimal segmentation.

The strength of our method is, unlike many state-of-the-art approaches, that it does not require any training stage, as body part characteristics can be extracted using selected image cues. This makes our pose recovery authentically ***activity-independent*** and, therefore, able to recover unusual poses. Since a key application of our technique is the initialisation of human body trackers, a ***probabilistic confidence***

measure is produced for each estimated pose so that initialisations could be performed only when postures are recovered with high confidence.

Another contribution of this paper is a statistical comparison of posture ground truths produced by either humans or a motion capture system. We identify a significant difference that needs to be taken into account when evaluating a training free method like ours.

The structure of this paper is as follows. After presenting related previous work, an overview of our framework is given. Detail discussion of the top-down module, the bottom-up module, and integration of these two modules are presented respectively in Section 2, 3 and 4. In Section 5, the probabilistic confidence measure for recovered poses is introduced. Quantitative and qualitative pose recovery results are discussed in Section 6. Finally conclusions and future work are addressed in Section 7.

1.1. Related Work

Human pose estimation has become one of the most active research topics in computer vision over the past decade. [5] [6] and [7] provides extensive surveys of human motion tracking and analysis methods, and a more recent review is available in [8]. In the broadest taxonomy, pose recovery algorithms can be divided into learning based or activity independent approaches depending whether training of either poses or observations is required. Pose recovery algorithms are also traditionally classified into top-down (model-based) and bottom-up (model-free) strategies.

In learning-based approaches, poses are estimated using either generative [9][10][11][12] or discriminative [13][14][15] approaches. Generative approaches use Bayesian rule to infer the pose configuration from learned state space to produce optimal image alignment to the observation, whereas discriminative approaches learn direct mapping functions from the visual observation to the pose configuration.

Whatever the approach, the main difficulty lies in the high dimensionality of the pose space. To tackle this, dimensionality reduction methods, such as Principal Component Analysis [16][17], Isomaps [18][10][9], Laplacian Eigenspace [19][11], Gaussian Process Latent Variable Model (GPLVM)

[20][21][9], Local Linear Embedding [22][10][23][9] and Diffusing map [24] have been investigated. In [23], view-dependent activity manifolds of human silhouettes and mapping function between the manifolds and 3D pose space were learned. Thus, a 3D pose can be recovered by projecting the visual input to the manifold and mapping to the learned 3D pose space. Instead of learning from image evidence, [21] learnt postures and motion dynamics, and embedded them in a low dimensional space using GPLVM. More recently in [9][10], both pose configuration and visual observation manifolds have been constructed so that generative pose inference can be performed more efficiently within the two low dimensional spaces.

A successful pose recovery algorithm also requires an efficient search strategy. The most popular ones include dynamic programming [25][26], Markov Chain [27][28], particle filtering [29][30][31], genetic algorithm [32] and Simulated Annealing [29][30][32]. In addition in [11], 3D voxel data is projected to a Laplacian Eigenspace so that nodes of body parts become discriminative. This allows body part search to be performed in a much smaller anthropometric-constrained space and removes part ambiguity. Howe's [27] silhouette-pose lookup approach is able to select a set of possible poses based on the given input silhouette. Then Markov chains modelling the temporal dependency of human motion are exploited to determine the most likely chain of pose sequences. [12] presents a top-down generative approach using a 3D physics-based motion prior so that searching of poses is constrained effectively among physically plausible hypotheses. [32] introduces an advanced search framework using stratified anneal genetic algorithm for poser recovery and tracking.

Most discriminative approaches [13][14][15] use nonlinear regression techniques to learn the mapping between pose configuration and visual observation. In [13] Bayesian mixture of experts with density propagation is used for learning and inferring poses. As a result of a comparison between a variety of regression methods including ridge regression, Relevance Vector Machine (RVM) regression, and Support Vector Machine, it was suggested RVM gives the best performance [14].

The main disadvantage of learning-based methods is their dependency on the training dataset, i.e. they can generally only recover poses belonging to a single activity and/or a specific actor. [15] proposed using

a mixture Gaussian Process kernel to handle efficiently very large training sets in order to be able to learn and infer several activities. However, under this scheme, results are only demonstrated for a unique actor.

Activity independent pose recovery algorithms exploit inherent human body characteristics which are present in the image, that are usually represented by low-level and mid-level image cues such as colour/skin [33][34], face [33], shape [35], optical flow [36][37], edge orientation [36][34] [38], and static foot point [39][2]. Since each individual cue is usually weak, cue combination is performed by either boosting [40] or clustering [34][36]. For example, [34] created a 30 dimensional feature vector from edge orientation and colour for body part clustering. In [36] the pose is estimated hierarchically; head and torso were located using separate cues and limbs are found by clustering feature vectors consisting of pixel locality, edge orientation, colour and optical flow.

These bottom-up methods usually apply geometric rules to filter out impossible pose candidates found using image cues. In [38], 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. [25][35] assemble shape pieces found from Normalised Cut to form a holistic human body using a predefined parsing rules. However, these bottom-up approaches are prone to noise, unless they receive feedback by top-down process as demonstrated in [11] and [34].

Our proposed method allows estimating poses independently from the type of human activity. The method consolidates both bottom-up and top-down pose recovery strategies so that strengths of top-down and bottom-up approaches can be combined, i.e. effectiveness of defining body structure and pose constraints, and accurate segmentation of individual body pieces respectively. Compared to [34] and [38], our bottom-up process uses a comprehensive set of image cues including pixel locality, edge orientation, optical flow and colour so that body parts can be segmented in many difficult scenarios where some cues are indiscriminative. Unlike [11] and [34] which combined the top-down/bottom-up sequentially, the main contribution of the proposed framework is to connect the top-down and bottom-up processes using a closed loop thus body model configuration and body part segmentation can be iteratively refined. The estimated

pose can then be derived from the body model when the configuration reaches a steady-state. In addition, each generated pose is assessed by a confidence score to quantify the accuracy of the pose recovery process.

1.2. Framework of Our Approach

The main characteristic of our algorithm lies in the integration of bottom-up and top-down approaches using probabilistic Gaussian modelling. The bottom-up module partitions the foreground area into body parts using a probabilistic clustering algorithm according to relevant image cues. The top-down module adapts dynamically a generic 2D body model to "fit" the segmented body parts by maximising a probabilistic objective function. The purpose of this is to impose anthropometric constraints to the segmented body parts. Our methodology does not require training and thus is activity-independent. Moreover, since poses are estimated probabilistically, a confidence score is generated to evaluate the success of pose estimation.

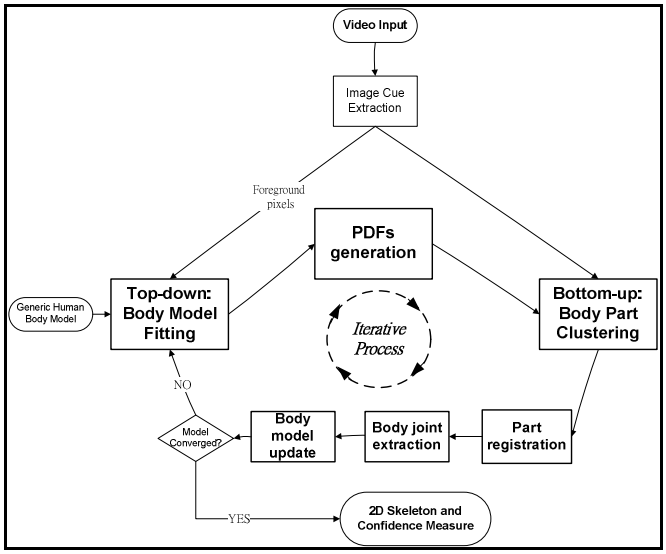


Figure 1 Flow diagram of the proposed pose recovery algorithm

As illustrated in Figure 1, our algorithm takes a video containing a moving human character as an input and generates for each frame a 2D skeleton with a confidence measure. The system iterates the processes of bottom-up and top-down pose recovery, i.e. ***body part clustering and body model fitting*** respectively. A

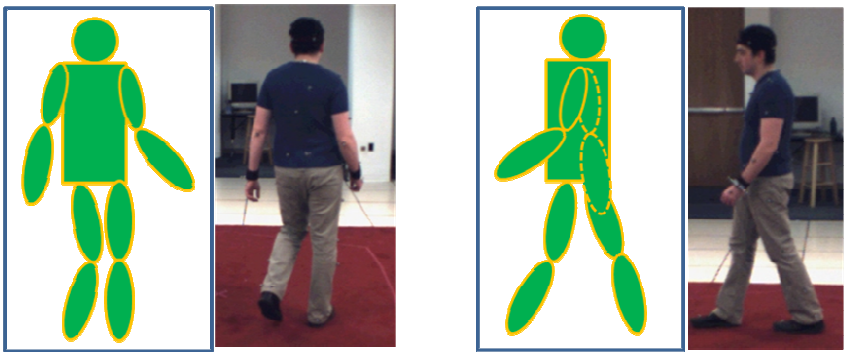
set of image cues describing body part characteristics are extracted from the input and used in the bottom-up module to probabilistically partition foreground pixels into a desired number of body pieces (section 3). The top-down module fits a 2D generic human body model onto the produced clusters by dynamically adjusting its scale, limb ratios and articulated configuration (section 2). In order to integrate the two main modules, a number of auxiliary modules are introduced. A module converting the fitted body model to a set of Gaussian distributions that embeds anthropometric constraints enforces body partitions (obtained from the clustering) to form a plausible human posture (section 4). After clustering, confidence scores is calculated by considering joint probabilities between the fitted body model and the body part partitions. This facilitates body part registration to the clusters, body joint extraction and thus updating the body model (section 5). The 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.

2. TOP-DOWN MODULE

The aim of this module is to initialise the bottom-up process by providing Gaussian mixtures where anthropometric constraints are implicitly embedded. It starts with fitting a body model to body part partitions and transfers the body model to Gaussian mixtures. Although the probabilistic modelling of body parts using Gaussians has been adopted by other groups [41][42], our integration of the top down/bottom up approaches allows optimising iteratively this process within a single frame. Two types of fully configurable body models are fitted separately and evaluated using confidence measures to choose the optimal model. In this section, the human body models are introduced. Then, this is followed by the detail description of the hierarchical fitting steps which include head, torso and limb fitting. Finally, conversion of the fitted body model to Gaussian mixtures is explained.

2.1. Human Body Model

In this work, a human body is defined as a kinematic chain of body parts. We represent this articulated structure using 2D generic configurable models, which can adopt any human posture. In order to be able to deal with any informative camera view[1], each element of the model (body part template) can be scaled independently so that alteration of body part ratios due to perspective can be represented on a 2D model. Moreover, since self-occlusion of an arm by the torso is very common and can last for many consecutive frames, this situation is specifically addressed in our framework. We employ two body models, ***full and profile models*** as shown in Figure 2, to process our data. As shown in our experiments (Section 6), most human poses can be represented by one of the two models: the selection of the more appropriate one is achieved by comparing their associated confidence scores. In the few cases, where none of the models is suitable, their poor confidence scores indicate the poses cannot be recovered. Apart from the torso which is modelled with a rectangle, all other body pieces of the body model are initially constructed using ellipses and standard human body ratios [43]. The difference between full and profile models is the number of pieces modelling human postures and some anthropometric constraints. As shown in Figure 2, the profile model has only one arm: this allows dealing with views, where the torso, which is the largest body piece, occludes one arm. Furthermore, models are associated with different anthropometric constraints, as listed in Table 1.



(a) (b)

Figure 2: (a) Full body model and corresponding pose. (b) Profile body model and corresponding pose. Note that self-occlusion of the far arm (marked as dashed ellipses) is addressed by using the profile template, which only models one arm

common constraints applied to both models	Head, upper arms and legs are connected to the torso Upper and lower pieces of each limb are connected tip-to-tip Head perimeter touches the midpoint of the torso's upper boundary	
specific constraints applied to either full or profile model	Full model	The connecting vertex of the upper left/right arm to the torso has 3 DOFs, i.e. planar rotation, V- and H- translation around the upper left/right corner of the torso The connecting vertex of the upper left/right leg to the torso has 2 DOFs, i.e. planar rotation and H- translation inwards from the lower left/right corners of the torso
	Profile model	The connecting vertex of the upper arm to the torso have 3 DOFs, i.e. planar rotation, V- and H- translation around the midpoint of the torso upper boundary The connecting vertex of the upper left/right leg to the torso have 3 DOFs, i.e. planar rotation, V- and H- translation around the midpoint of the torso lower boundary

DOF, H- and V- denote Degree of freedom, Horizontal and Vertical.

Table 1: Common and specific anthropometric constraints used in full and profile models.

2.2. Head Detection

The first step of body model fitting is the detection and localisation of the head. The distinct shape of the head and abundant features which can be extracted from the face, if visible, makes it the most reliable body part to identify. To ensure robust detection of this critical body part in our framework, we propose a fusion of "Ω" model head detection [44], AdaBoost face detection [45], and SIFT [46] head tracking (Figure 3). The "Ω" model detects the head by considering its unique shape, while Adaboost adds face information, if visible, to boost the head detection. Furthermore, SIFT exploits temporal consistency of detection of the

[1] We define an informative camera view as a view where, for a majority of poses, most body parts are not self occluded. For example, a top view is not informative, while a side view is.

head between successive frames to suppress false alarms.

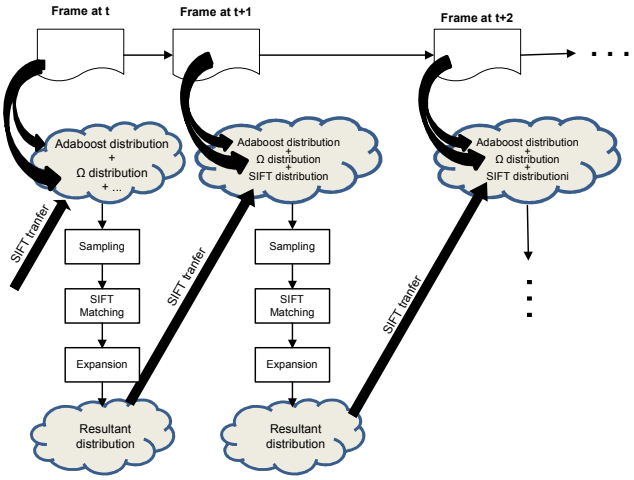


Figure 3: Pipeline for head detection and localization

For a given frame f_t, head and face, if relevant, are detected using estimates provided by "Ω" model [44] and AdaBoost [45] algorithm as implemented in OpenCV [47]. However, instead of using specific detection thresholds, we adapted these algorithms to obtain normalised distributions of head position and size. Let us denote the distribution of "Ω" model detection and AdaBoost detection as $\omega(p_i)$ and $AB(p_i)$ respectively, while $p_i = (x_i, y_i)$ denotes pixel coordinates. Since distributions are normalised and only pixels belonging to the foreground area are considered, $\omega(p_i)$, $AB(p_i)$ and p_i are defined as follows:

$$\begin{aligned} &\omega(p_i) \in [0,1] \\ &AB(p_i) \in [0,1] \\ &\forall p_i \in \Lambda \end{aligned} \qquad (1)$$

where Λ denotes the set of foreground pixels. A combined distribution of $\omega(p_i)$, $AB(p_i)$ and the SIFT transferred distribution, $\eta(p_i)$, from the previous frame (which is described below) is then generated by normalising the sum of the three distributions. Since points with high values indicate good confidence in the presence of the head, they are used to improve head detection in the subsequent frame. We introduce a SIFT-based framework to map these points between the current frame, f_t and the next frame f_{t+1}. Since SIFT

performs discrete mapping, confidence associated with the transferred points is interpolated to produce a continuous distribution (see next paragraph). This distribution will be integrated to the other head detection distributions, i.e. $\omega(p_i)$ and $AB(p_i)$, of the next frame.

In the current frame f_t, we extract N points (N=25 in the experiment), $P^s \in \{(x^s_n, y^s_n)\}$, $n=[1..N]$, corresponding to the N highest confidence values (s_n) in the current distribution and apply SIFT match [46] to relate the current samples to the next frame. SIFT is able to detect features points around the current samples and match them to the next frame so that the current samples, P^s, can be relocated to new positions $P^{s'} \in \{(x^{s'}_n, y^{s'}_n)\}$, $n=[1..N]$, in the next frame, as shown in Equation 2. Assuming the total number of SIFT feature points detected around each current sample is N^{all}_n, and the number of SIFT features matched in the next frame is N^{hit}_n, we design a transfer function, as defined in Equation 3 (a=0.25 in the experiment), to convey sample scores from the current frame to the next frame.

Then a new distribution is generated from the relocated samples. A Gaussian kernel (Equation 4) is used to expand the samples to N Gaussian distributions and then mix them using the weighting factors of s'_n to generate the resultant distribution, $\eta(p_i)$. In Equation 4, COV denotes the covariance matrix controlling expansion of the relocated samples and is proportional to the size of the foreground area $|\Lambda|$, as indicated in Equation 5. As an isometric Gaussian kernel is used to convert a scalar value to a probability density function, a diagonal matrix is used in Equation 5. C is a constant (0.01 in the experiment). $\eta(p_i)$ is then incorporated with the AdaBoost and "Ω" distributions of the frame, f_{t+1}, by normalising the sum of the three distributions to produce a combined distribution that will subsequently be sampled, matched, expanded and transferred to the following frame f_{t+2}. For each frame, the centre of the head is defined by the point having the highest score in the combined distribution and its size is calculated as the weighted average size of detected head from the "Ω" model, AdaBoost and from the previous samples transferred by SIFT. For the initial frame (f_0), only two distributions, $\omega(p_i)$ and $AB(p_i)$, are combined as $\eta(p_i)$ is not yet available.

$$(x^{s'}_n, y^{s'}_n) = SIFT\ ((x^s_n, y^s_n)) \qquad (2)$$

$$s'_n = \left(\frac{N_n^{hit}}{N_n^{all}} s_n \right)^a \quad 0 \le a \le 1 \qquad (3)$$

$$\eta(p_i) = \frac{1}{N} \sum_{n=0}^{N-1} s'_n \exp\{-\frac{1}{2}((x_n^{s'}, y_n^{s'}) - (x_i, y_i))^T COV^{-1}((x_n^{s'}, y_n^{s'}) - (x_i, y_i))\} \qquad (4)$$

$$COV = C \times |\Lambda| \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad (5)$$

This head detection algorithm has been tested in a variety of sequences, see section 6, showing it can be initialised for a wide range of head orientations as illustrated on Figure 4.

Figure 4: Results of head detection from various head orientations.

The circle and its centre (both in blue) indicate the size and centre of the head.

2.3. Detection of Torso Region

The torso region is detected by modelling a sample of torso colours using Gaussian Mixture Modelling (GMM) as previously used in [48][49][50]. First, "rectangular sampling regions" are created around the position of the head, which has been previously detected (see Figure 5). The area of these regions is set to 50% of the expected torso size, which is estimated using a standard body ratio [43], head size and foreground height. Since the torso must belong to the foreground, the number of relevant sampling regions is small (usually 2 or 3, see Figure 5). For each sampling region, a GMM composed of 3 mixtures is

estimated by using RGB colour of foreground pixels drawn from the sampling region. A statistical model is used to reject colour outliers [49]. Then for each sampling region, torso pixel candidates are detected from the foreground by using the trained GMM classifiers. Finally, sizes of detected candidate regions are compared to the expected torso size to select the most likely torso region [50]. Figure 6(a) shows a torso region extracted using the torso region detection procedure.

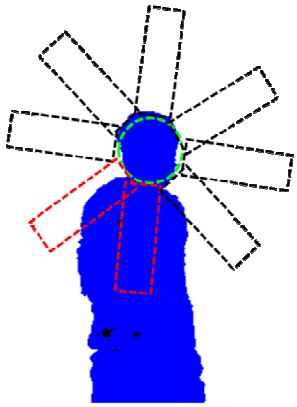
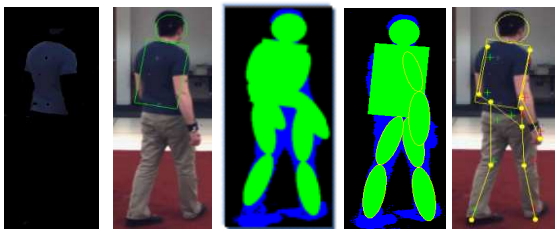


Figure 5: Torso sampling regions (dashed rectangles) defined around the head. Only two rectangular regions (marked in red) are considered because they contain a sufficient number of foreground pixels.

2.4. Body Model Fitting



(a) (b) (c) (d) (e)

Figure 6: Stages of body model fitting

The aim of body model fitting is to initialise the bottom-up pose recovery module (see Figure 1) by providing anthropometric constraints. Two 2D body models, full and profile models, introduced in Section

2.1, are fitted separately onto the probabilistically partitioned foreground pixels by maximising joint probabilities between the models and partitions. Since the two models bear different anthropometric constraints, model fitting results in two different confidence scores which allow for selection of the better model. The fitting is hierarchical, where the most reliable parts are fitted first, i.e. head and torso, and are used as references for fitting the limbs. This process is performed iteratively with an increasingly refined 2D model whose limb templates are resized independently to accommodate perspective effect according to the result of the bottom-up process (see Sections 4.3 and 4.4). During the first iteration, where the foreground partitions are not provided, generic models are fitted to the foreground by maximising the overlapping area.

The head is fitted with a circular template with the size and location found in Section 2.2. Since the detected torso region is only based on colour features (Section 2.3), this region may comprise some pixels which do not belong to the torso, e.g. sleeves on Figure 6(a). In addition, torso coverage may not be complete. Therefore, the definition of the torso is refined using a new constraint: a rectangular shape is fitted on the detected torso region adjusting position, orientation, scale and height/width ratio. This is achieved by maximising the overlap between pixels belonging to the rectangular shape and detected pixels of the torso region:

$$Overlap = \frac{\left| A_{m_{torso}} \cap A_{pixel\ torso} \right|}{\sqrt{\left| A_{m_{torso}} \right| \left| A_{pixel\ torso} \right|}} \quad (6)$$

where $\left| A_{m_{torso}} \right|$ and $\left| A_{pixel\ torso} \right|$ denote the pixel area of the rectangle and the number of detected torso pixels respectively. Figure 6(b) shows an example of fitted head and torso.

Once the head and torso have been fitted, they are used as constraints to perform limb fitting. Since this process requires optimising position, orientation and size of limb templates, it is computationally expensive. To mitigate this, first, we search for the optimal position and orientation and, then, we resize limb templates independently to accommodate perspective effects. Since the second step requires updating

the body model, it is performed after the bottom-up process in every iteration (see details in Section 4.3).

The 2D limb templates are translated and rotated within the search space defined in Table 1 to maximise the joint probabilities between limb templates and body partitions. Since in the first iteration the probabilities of pixels belonging to each body partition are unknown, we set them as uniform. Therefore, the maximisation of the joint probabilities is equivalent to finding the maximum overlapping area between the models and foreground. Although the pose estimated from initial fitting may not be satisfactory due to foreground imperfections, it is sufficient to initialise the clustering process which will allow iterations between the bottom-up and top-down stages. In subsequent fittings, where the probabilistic body partitions are produced by the bottom-up process, the joint probabilities, defined in Equation 15, between the limb models and partitions are maximised. The joint probabilities will be discussed in detail in Section 5.1.

Body model fitting is the most computationally expensive step in our framework since it is based on a hierarchical exhaustive search. To achieve real-time processing, the search strategy would need to be optimised. This could be achieved by using a pose tracking system: for example, importance sampling based trackers [51][52] can generate efficiently samples around the most likely body configurations. In such approach, an annealing framework [52][53] should be integrated to avoid local minima.

Figure 6(c) and (d) show the outcome of fitting both models on the foreground during the first iteration. The algorithm terminates when changes in body joint locations estimated from the fitted model between iterations become negligible. Then two confidence scores, one for each body model, are produced by concatenating the final joint probabilities of all body parts (Equation 14). Comparison of these confidence scores allow determining which pose estimate should be selected (Figure 6(e)).

2.5. Generation of Probability Density Function

The anthropometric constraints required for body model fitting are embedded in a set of Gaussian mixtures which are used to initialise the bottom-up pose recovery, i.e. probabilistic body part partition. This embedding is achieved by converting fitted templates (see Section 2.4) to a Probability Density Function (PDF) of Gaussian mixtures. Assuming the full body model fitting is considered, fitted body

templates are expressed by { $m_{head,}$ $m_{torso,}$ m_{lua}, m_{lla}, m_{rua}, m_{rla}, m_{rul}, m_{rll}, m_{lul}, m_{lll}}[2] and their centres are {$\mu_{head,}$ $\mu_{torso,}$ μ_{lua}, μ_{lla}, μ_{rua}, μ_{rla}, μ_{rul}, μ_{rll}, μ_{lul}, μ_{lll}}, $\mu_n \in \Re^2$. Equation 7, 8 and 9 express how the PDF of Gaussian mixtures are generated.

$$P(p_i \mid m_n) = \frac{w_n}{(2\pi)\,|COV_n|^{1/2}} \exp\{-\frac{1}{2}(\mu_n - p_i)^T COV_n^{-1}(\mu_n - p_i)\} \qquad (7)$$

$$w_n = \frac{|m_n|}{\sum_j |m_j|} \qquad (8)$$

$$COV_n = k \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) \\ \sin(\gamma) & \cos(\gamma) \end{bmatrix} \begin{bmatrix} L_n^2 & 0 \\ 0 & L'^2_n \end{bmatrix} \qquad (9)$$

where $n \in$ {*head, torso, lua… lll*}. $P(p_i|m_n)$ denotes the probability of a pixel, p_i, belonging to a template , m_n, i.e. it is the PDF of the Gaussian mixtures. W_n is the weight of the *n*-th mixture, and is proportional to the size of the template $|m_n|$. COV_n, in Equation 7, denotes the covariance matrix of the mixture which is proportional to the lengths of major and minor axes, i.e. L_n and L'_n, of the corresponding elliptical model. In the cases of the head and torso whose templates are not elliptical, $L_n = H/2$ and $L'_n = W/2$ for the torso, and $L_n = L'_n = r$ for the head, where *H* and *W* denote the height and width of the torso rectangular template and r is the radius of the head circular template. COV_n also depends on the angle, γ, between L_n and the horizontal axis *x* and a scaling constant, *k*, see Equation 9. $P(p_i|m_n)$ represents the statistical model of the anthropometric constraints obtained from the body model fitting. Figure 7(a) shows an example of initial Gaussian mixture density.

3. BOTTOM-UP MODULE

The aim of the bottom-up module is to partition the foreground probabilistically into a number of body parts. Here, we assume that the foreground has been reasonably well segmented, although it may contain

[2] *Apart from head and torso pieces, names are abbreviated by 3 letters denoting: "left" or "right", "upper" or "lower" and "arm" or "leg". For the profile model, only "ua" and" la" are used for denoting*

some imperfections. To achieve the partition of the foreground, first, a set of image cues describing the body parts are extracted from the image. Then probabilistic clustering of foreground pixels is performed according to the extracted cues. The following sections discuss image cue extraction and probabilistic clustering in detail.

3.1. Image Cue Extraction

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 a body part is defined by a continuous set of pixels, except in some cases of occlusion, pixel location provides a first low level cue. Moreover, human limbs are highly directional objects; hence they can be modelled by either sets of parallel lines or trapeziums [38] whose main orientations describe the underlying skeleton's main axes. Therefore, direction of edges is used as a cue to describe body parts. Because the human body is modelled as an articulated figure, distinctive changes of pixel motions occur at body part boundaries. Therefore optical flow representing pixel motion is used as a cue to partition body parts. The final cue for body part detection is pixel colour since each body part can usually be modelled by either homogenous colour or a low number of colour patterns [54].

Several image processing techniques have been employed to extract the image cues. Locations of the foreground pixels are obtained by conventional motion segmentation, along with shadow detection and foreground cleaning [47]. Orientation cues are calculated by populating orientations of main edges over all foreground pixels. First, foreground edges are detected by Canny Edge Detection. Then these edges are converted to line segments via Hough transform to obtain the main edge orientations and thus remove spurious edges and noises. Finally, orientations are populated to all foreground pixels according to the proximity between the pixel and line segment. Motion cues consist of two elements – speed and direction-,

---

single arm pieces: upper arm and lower arm.

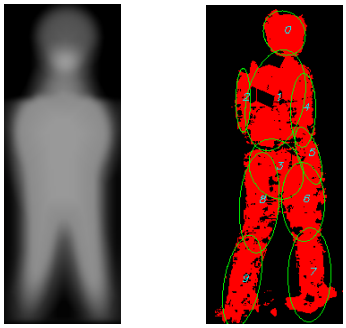
which are computed by Optical Flow. We adopted Lucas and Kanade's algorithm [55] to provide dense motion cues. Noise was suppressed by smoothing using a moving-average-of-5 temporal filter. Since preliminary experiments [36] showed that the colour space choice did not affect results in body part detection, colour cues are expressed by RGB values.

Since each individual cue provides only partial evidence of the presence of body parts, robust body part detection can only be achieved by cue combination. In the bottom-up pose recovery, body parts are partitioned using a clustering technique. Clustering is performed in a high dimensional space, called cue space, where each foreground pixel is represented by an 8-D feature vector, $p_i' = (x_i, y_i, \theta_i, v_i, \beta_i, r_i, g_i, b_i)$, whose elements are location (x, y), edge orientation (θ), speed (v), direction (β), and colour (r, g, b).

3.2. Probabilistic Partition of Foreground Pixels

Since our algorithm aims at producing a probabilistic confidence measure for each estimated pose, the bottom-up module has to comply with a probabilistic modality. For this reason, Gaussian Mixture Model (GMM) clustering is adopted. GMMs partition foreground pixels in the cue space into the desired number of body parts, i.e. 10 for the full model and 8 for the profile model, with soft boundaries. Taking the full model as an example, a set of 10 probabilities, $P(p_i|C_j), j \in [1..10]$ and $\sum_j P(p_i|C_j)=1$, is produced for each foreground pixel, p_i, indicating the likelihood of a pixel belonging to each of the 10 clusters, C_j. In our previous work [36], GMMs were initialised by K-means clustering where seeds were provided from the top-down module. In this work, we propose using a probability density function (PDF) (see Section 2.5) where anthropometric constraints are embedded to initialise GMM clustering. The PDF, $P(p_i|m_n)$, as shown in Figure 7(a), which is obtained from Equation 7, allows GMM clustering to be initialised at the Maximisation-step of its Expectation-Maximisation (E-M) computation. Then GMM partitions the foreground pixels according to the 8-D feature vectors in the cue space. Figure 7(b) illustrates the partition result where a 2-standard deviation boundary has been drawn to represent each cluster. Note that a spurious cluster located at the left hip should correspond to the left lower arm, which is actually hidden (see Figure

6(b). Fortunately, the poor confidence score associated to that cluster allows identifying this error.



(a) (b)

Figure 7: (a) PDF of Gaussian mixtures initially generated by converting model fitting result. (b) optimised GMM clustering resulting from EM optimisation

4. INTEGRATION OF TOP-DOWN/BOTTOM-UP MODULES

The integration of top-down/bottom-up modules aims at improving the accuracy of body model fitting by incorporating clustering results (see Figure 1). This is achieved by, first, associating body clusters to body parts. Then positions of body joints are extracted from the clusters and used to update the body model. Finally, top-down and bottom-up processes iterate until the body model converges towards a stable configuration. If convergence is not achieved within a certain number of iterations, this indicates that either the other body model should be used or the pose is too complex to be recovered from the provided camera view.

4.1. Part Registration

To make the body clusters meaningful, they need to be associated to the body parts. This insures the anthropometric constraints embedded in the body model are linked to correct clusters. This is important especially when two or more body parts are very close or even overlap on the image plane. The original cluster-body part associations established at the cluster initialisation stage becomes invalid as centres of clusters drift during the clustering process. Thus new registration is required.

The optimal one-to-one correspondence between the N clusters and the N body parts is achieved by the combinatorial maximisation of the sum of joint probabilities between the N cluster/body part pairs. If we define C as the ordered list of N clusters, $\{C_k\}$, and M_i as one of the $N!$ possible ordered lists of N templates , $\{m_k\}$, the optimal list of templates , M_o, which maximises the clusters-body parts association, is given by Equation 10:

$$M_o = \arg\max_{1 \le i \le N!} \sum_{k=1}^{N} P(m_{M_i[k]} \cap C_k) \qquad (10)$$

where $M_i[k]$ is the kth element of the list M_i and $P(m_i \cap C_j)$ is the joint cluster-body part probability, which is defined in Equation 15. To reduce the computational cost of estimating M_o, we restrict our search space so that only the 3 closest templates (smallest Euclidian distance between centroids) for each cluster are considered.

4.2. Extracting Body Joints from Body Clusters

After registration, a set of body joints can be extracted from the clusters. These joints are important as sizes of limbs are derived from them when updating the body model. We class body joints within two types: ***distal joints*** which are the endpoints of kinematic chains such as wrists and ankles; ***intermediate*** joints which are joints connecting adjacent body parts, i.e. shoulders, elbows, hips and knees. Locations of intermediate joints are defined as the clusters' probabilistic boundaries as expressed in Equation 11:

$$J_{j-k} = \arg\max_{p_i} \left\{ P(p_i \mid C_j) + P(p_i \mid C_k) - \left| P(p_i \mid C_j) - P(p_i \mid C_k) \right| \right\} \qquad (11)$$

where p_i, C_j and C_k denote the foreground pixel and the adjacent clusters. J_{j-k} is the intermediate joint between C_j and C_k. The conditional probabilities are given by GMM clustering.

The distal joints are located by using image cues, geometry constraints and, if available, appearance consistency between adjacent frames. As a first approximation, the expected distal joint should be positioned so that the distal and the adjacent intermediate joints are equidistant to the cluster centre. Moreover, since distal joints are defined as endpoints of limbs, they correspond to positions where pixel motion and colour changes abruptly. Finally, if the pose estimation in the previous frame is successful, as

indicated by the confidence score, locations of the distal joints are used to initialise SIFT features so that their recognition in the current frame can contribute to distal joint detection, see Section 2.2. Therefore, distal joint positions can be derived from the maximisation of Equations 12:

$$J_k = \arg\max_{p_i} \left\{ \Gamma_N(p_i) + \Delta_N(p_i) + \Omega(p_i) + \Psi_N{}^*(p_i) \right\} \quad (12)$$

"Normalised Colour Edge Map" ($\Gamma_N(p_i)$) measures the strength of colour change within an image frame. It is calculated by the following process. First, the input image is decomposed into three single band images (Red, Green and Blue) and Sobel edge detection is applied to each of them to generate three edge responses,($\varphi_r(p_i)$, $\varphi_g(p_i)$ and $\varphi_b(p_i)$). Then, the "*Colour Edge Map*" ($\Gamma(p_i)$) of the image is computed by taking the maximum edge response, i.e. $\Gamma(p_i) = \max(\varphi_r(p_i), \varphi_g(p_i), \varphi_b(p_i))$. Finally, $\Gamma_N(p_i)$ is produced by normalising $\Gamma(p_i)$, i.e, $\Gamma_N(p_i) = \Gamma(p_i) / \max(\Gamma(p_i))$.

$\Delta_N(p_i)$denotes *"Normalised Change-in-Motion Map"*. Lucas-Kanade optical flow algorithm is applied to the previous and next frames to generate motion vectors. They then undergo Sobel edge detection to generate the *"Change-in-Motion Map"* ($\Delta(p_i)$). Finally, standard normalisation is applied to produce $\Delta_N(p_i)$.

$\Omega(p_i) = -\left| \frac{\left|\overline{C_k} - p_i\right| - \left|\overline{C_k} - J_{j-k}\right|}{\left|\overline{C_k} - J_{j-k}\right|} \right|$ expresses the geometric constraints of the distal joints. $\overline{C_k}$ and J_{j-k} denote respectively the centre of cluster C_k and the intermediate joint between C_j and C_k. $\Omega(p_i)$ penalises p_i estimates which violate equidistance between $\overline{C_k}$ and J_{j-k}.

$\Psi_N{}^*(p_i)$ is the normalised score of SIFT matching. This measure is only used when the previous frame's pose has been recovered successfully (see Section 5).

The optimal p_i is searched from $\overline{C_k}$ to $\overline{C_k} + 2\left|\overline{C_k} - J_{j-k}\right|$ along the extended line of $\overline{J_{j-k}, \overline{C_k}}$ (towards the limb's distal end) to maximise Equation 12. Figure 8 shows examples of intermediate joints (here, shoulder and elbow) and the extraction of a distal joint, J_k (i.e. wrist).

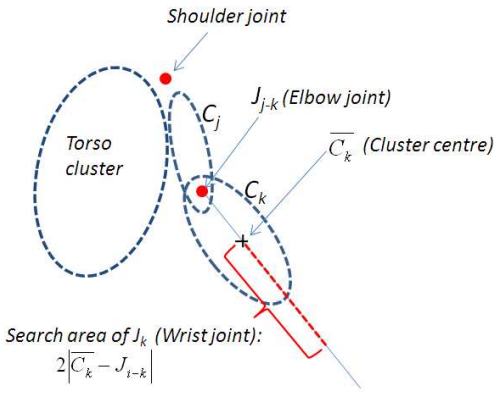


Figure 8: Body joint extraction

4.3. Body Model Update

The purpose of extracting body joints from the clusters is to update the body model so that, after each iteration of the top-down and bottom-up processes, better body part ratios are estimated to accommodate perspective effects. This is achieved by updating the body model using the putative body joint locations extracted from the clusters. Since head and torso templates were located using robust methods, only limb templates are updated. The length of the limb templates is resized according to the distance between the limb joints estimated from the clusters. If we define the length of a body part in the current iteration, *l*, by L_i^l, where $i \in \{lua, lla, rua, rla, rul, rll, lul, lll\}$ in the full model fitting mode, the updated length after an iteration is calculated using Equation 13.

$$L_i^{l+1} = (1 - \alpha_L) L_i^l + \alpha_L L_c^l \qquad (13)$$

where L_c^l denotes the length of the body part estimated from the joint positions extracted from the associated cluster and α_L indicates the model updating rate; ranging from zero (no updating) to one (full replacement by L_c^l). In our experiments, we set $\alpha_L = 0.3$.

4.4. Termination Criteria

The succession of clustering and model fitting processes iterates until the positions of joints extracted from the fitted body model converge. Convergence is defined as a joint variation between consecutive

iterations falling below a predefined consistency threshold. If this criterion is satisfied within a given number of iterations, a skeleton, as shown in Figure 6(e), is then extracted from the final body model and a probabilistic confidence measure is calculated. Otherwise, the pose cannot be recovered for the given body model. If none of body models is successful, the pose is considered as too complex for being estimated from a single camera view and is discarded from further analysis.

5. PROBABILISTIC CONFIDENCE FOR POSE ESTIMATION

5.1. Mathematical Formulation

An important feature of our method is that a confidence measure is provided for every pose estimate. This 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 quantify how much an observation can be trusted [56][59]. Our confidence measure is the probability that a pose is recovered successfully, *P(pose)*. If we assume this is determined by the success of recovering all body parts and their associated recovery probabilities are independent, it can be expressed by Equation 14.

$$P(\,pose\,) = \prod_{j} P(X_j) \qquad (14)$$

where $P(X_j)$ denotes the probability of body part, X_j, to be recovered successfully. We evaluate this by extending the definition of the overlap measure (Equation 6) to other body parts.

$$P(X_j) \sim Overlap\ (m_j, C_j) = \frac{\left|A_{m_j} \cap A_{C_j}\right|}{\sqrt{\left|A_{m_j}\right|\left|A_{C_j}\right|}} \qquad (15)$$

where A_{m_j} and A_{C_j} denote the sets of pixels belonging to the model part m_j and cluster C_j respectively. Norm $|\cdot|$ denotes the number of pixels in a set. Therefore, for each model part m_j, $|A_{m_j}| = \sum_{p_i \in m_j} 1$. Since the cluster C_j is defined by GMM clustering over the entire foreground pixels, F, its pixel area is conceptually equivalent to the sum of the probabilities of foreground pixels $p_i \in F$ belonging to that cluster:

$$\left|A_{C_j}\right| \sim \sum_{p_i \in F} P(p_i \mid C_j) \qquad (16)$$

Similarly, $\left|A_{m_j} \cap A_{C_j}\right|$ corresponds to the sum of the probabilities of model pixels $p_i \in m_j$ belonging to that cluster:

$$\left|A_{m_j} \cap A_{C_j}\right| \sim \sum_{p_i \in m_j} P(p_i \mid C_j) \qquad (17)$$

Therefore, $P(X_j)$ is expressed by

$$P(X_j) \sim \frac{\sum_{p_i \in m_j} P(p_i \mid C_j)}{\sqrt{\sum_{p_i \in m_j} 1 \times \sum_{p_i \in F} P(p_i \mid C_j)}} \qquad (18)$$

5.2. Normalisation of Confidence Scores

Confidence scores for the full model ($P_{full}(pose)$) and profile model ($P_{profile}(pose)$) are computed to determine the most suitable body model for pose recovery. For accurate comparison, a normalisation procedure is required as the two scores are formulated with different number of body pieces. As shown in Equation 19, the normalised confidence scores, $P_{full}^{N}(pose)$ and $P_{profile}^{N}(pose)$ are generated by considering the geometric mean of underlying fitting probabilities which consist of different number of body pieces according to the model:

$$P^{N}(pose) = [P(pose)]^{\frac{1}{k}} \qquad (19)$$

where $k=10$ or 8 for full or profile model fitting respectively.

6. EXPERIMENTAL RESULTS

6.1. Datasets

Three datasets were used to evaluate our pose recovery algorithm: (1) HumanEva I dataset [57] (2) outdoor walking sequences produced by Hedvig Sidenbladh [58] and (3) MuHAVi dataset [59]. All are benchmark datasets used by the computer vision community and are publicly accessible.

HumanEva I (HE I) dataset [57] consists of 4 human subjects performing 6 types of motions. The dataset

allows quantitative evaluations as it provides both motion capture and video data (progressive scan images; 640 x 480 pixels) 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 [60] is defined to evaluate pose estimations and an optimised motion segmentation algorithm is provided.

[58] contains a female subject performing circular walking and straight line walking (progressive scan images; 320 x 240 pixels). The outdoor setting of this dataset provides additional values and challenges. The human silhouettes were extracted and provided by [17].

The MuHAVi dataset [59] is one of the latest publicly accessible dataset for human motion modelling (interlace scan images; 720 x 576 pixels). It contains 14 actors performing a number of primitive actions, including walking, running, kicking and punching. This dataset also provides extracted foreground silhouettes. Images were deinterlaced before processing.

6.2. Interpretation of HumanEVA Ground Truth

Since our algorithm, like many other pose recovery algorithms in the literature, is evaluated against MOCAP ground truth, an important question is whether the ground truth provided by MOCAP data is consistent to what humans perceive. Since such investigation would allow refining the evaluation of pose recovery algorithms, an experiment of comparing HumanEva ground truth (HE GT) with human annotated GT (HA GT) is conducted.

6.2.1.Experiment Setting

The aim of this experiment is to evaluate how definitions of body joints varies between human perception and MOCAP data provided in HumanEVA I. Ten graduate subjects (4 females and 6 males), participated in the experiment. The sequence, ***Walk1_C1_S2*** (see Figure 10), is used as the test sequence. It contains an actor walking in a circular manner and thus it includes a variety of body postures seen from different viewpoints. We studied the sequence between frame 340 and 760 during which the actor

completes a full circle. Frames were down-sampled by 5, thus a total of 85 frames were manually annotated by the 10 subjects, who were asked to mark positions of 13 joints on the image using their own understanding of anthropometry. The joints, i.e. L/R shoulder, L/R elbow, L/R wrist, L/R hip, L/R knee, L/R ankle, and the head centre, correspond to HE GT.

6.2.2.Results of Evaluating HumanEVA Ground Truth

We use the error metrics defined in HumanEVA to measure the difference between HA GT and HE GT. Figure 9 shows the average error (in pixels) and associated standard deviation (s.d.) for each joint of human annotations against HE GT. The figure shows noticeable deviations for the head centre, left hip and right hip. Divergence regarding the position of the head centre was expected since it is not explicitly defined in the HE GT: it was estimated as the average position between the top and bottom of head as defined in HE GT. However, the large error at L/R hip indicates a disagreement regarding hip definition between human annotators and HE GT. Figure 10 shows the projection of the shoulder points and hip points defined in HE GT onto the image plane. Clearly human annotators think the hip points are too close to each other. Table 2 shows the average and standard deviation (s.d.) of error for each annotation and overall annotation. The overall average error is 11.0 pixels with a s.d. of 4.2 pixels. This provides a baseline for comparison of pose recovery algorithms evaluated against the HE GT.

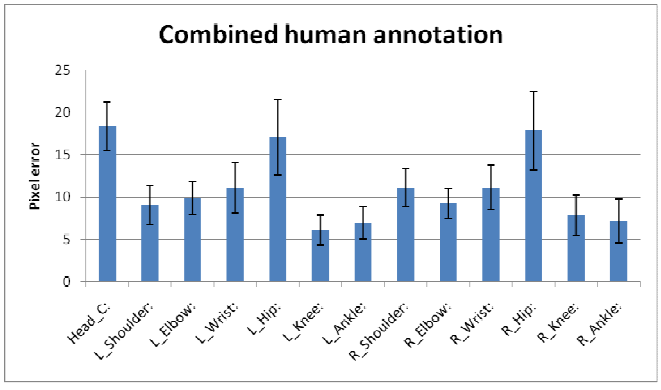


Figure 9: Average error and standard deviation for each joint from a total of 10 human annotations against HE GT.

Figure 10: Projection of the HE GT onto the image plane: lower green/red crosses, upper green/red crosses and white cross show locations of the hips GT, shoulders GT, and shoulder centre respectively.

	Average error (pixel)	Standard deviation
HA 1	11.3	3.9
HA 2	10.9	4.9
HA 3	10.1	2.8
HA 4	12.2	4.8
HA 5	10.2	3.9
HA 6	12.4	6.7
HA 7	11.8	4.8
HA 8	10.1	3.6
HA 9	9.7	4.4
HA 10	11.1	4.5
All	11.0	4.2

Table 2: Average error and standard deviation for each human annotation and overall annotation.

6.3. Quantitative Evaluation of the Pose Estimation Algorithm

6.3.1.Evaluation of HumanEVA Walking Sequence

Our pose recovery algorithm was first tested on the sequence used in Section 6.2. This sequence contains, in total, 420 frames during which the actor walked a complete circular path (see Figure 11). The resolution of the image is 640 x 480 pixels and the average human height is about 350 pixels. Figure 11 presents for each frame the confidence measures of pose estimates obtained using either the full (blue) or profile (red) body models. Since frames 340, 440 and 540 display typical profile, back and profile views, the better confidence scores correspond to the expected model. On the other hand, frames 640 and 740 are transition views where model score curves intersect. This confirms our suggestion that the confidence measure is a useful indicator for selecting the correct model for a given frame.

After selection of the best pose estimates using confidence scores, Figure 12 (a) shows accumulated number of image frames under defined error margins according to HE GT. The red, green purple and blue curves indicate the results obtained from (1) our previous work [36], which used full model only, (2) improved fitting of the full and (3) profile models using techniques presented in this paper, and (4) combines (2) and (3) according to the confidence scores. As can be seen in the figure, by refining the estimation framework, performances of pose estimation improve as the curve shifts leftwards. We can also notice that in this experiment, the profile model generally produces more accurate results than the full body model. We believe this can be explained by the fact that a walking motion, where arms swing back and forth, displays more views where an arm is occluded (even partially) than views where both arms are visible.

Another capability of confidence scores is to select good pose estimates. Figure 12 (b) shows positive correlation between confidence scores (blue) and pixel errors (red). The image frames are ranked by both confidence scores and errors produced from their poses estimation, and grouped by defined number of accumulated frames. Our proposed pose recovery method achieves an average error of 19.8 pixels when all pose estimates are considered, i.e. “All” bins. However, when one considers the top 50 frames, i.e. 12.5%

of all frames, as selected by our confidence score, an average error of 16.8 pixels is achieved. We have also calculated the correlation between the two curves. We obtained 0.71 for the Pearson Product-Moment Correlation Coefficient (PMCC) [61], which ranges from +1 (max. positive correlation) to -1 (max. negative correlation), between confidence scores and pixel errors. The green horizontal line shows the 'average error' in human annotation indicating what an 'optimal' algorithm should aspire to.

Figure 13 illustrates examples of estimated poses selected all from the "top 50" bins according to confidence scores. Since two body models are used, poses can be recovered in a large variation of viewpoints. This achieves significant improvement from our previous work [36] where poses shown in profile views were unlikely to be recovered successfully as only the full body model was used.

6.3.2.Evaluation of HumanEVA "Combo" sequence

One of the main characteristics of our algorithm is that poses can be recovered independently to the type of activities. To demonstrate this, ***HumanEVA Comb2_C1_S2*** is used. This sequence contains an actor performing several activities, altering continuously from one action to another. Here, we are particularly interested in the transitions between actions, since learning-based pose recovery algorithms would not be able to cope with them. To our best knowledge, this is the first quantitative experiment conducted for evaluation of pose estimation during activity change. We consider seven action transitions: Walk-to-run, run-and-turn, run-to-stop stop-to-short balance (left leg up), stop-to-short balance (right leg up), stop-to-long balance (left leg up) and stop-to-long balance (right leg up). For each transition, 20 to 50 frames were extracted and this resulted in processing 250 frames in total. After processing these sequences, we perform the same quantitative analysis as in Section 6.3.1. The overall average error is 23.7 pixels while the top 12.5% of total frames selected according to conference scores produces an average error of 21.1 pixels according to HE GT. There is still high correlation between group errors of selected frames by confidence scores and actual errors (PMCC=0.69).

Figure 14 illustrates the quality of estimated poses selected according to confidence scores.

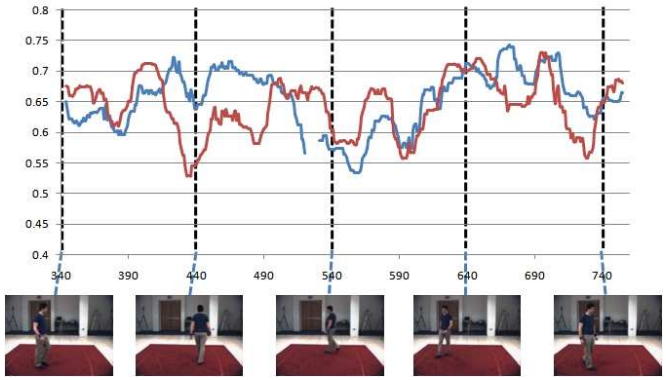


Figure 11: Confidence measure of the full (blue) and profile (red) models over frames

6.4. Qualitative Evaluation of the Pose Estimation Algorithm

Qualitative results of pose recovery on an outdoor sequence [58] and sequences showing various actions [59] are also provided. The outdoor sequence, as shown in the first row of Figure 15, is particularly challenging due to uncontrolled lighting condition and the subject's clothing whose colour is similar to human skin. The success of pose recovery on this sequence emphases the importance of cue combination; while colour is not discriminative, other cues, such as optical flow and edge orientation, provide useful indication to distinguish body parts. Second and third rows of Figure 15 illustrate pose recovery on primitive actions provided by MuHAVi dataset [59], i.e. walking, running, punching and kicking. The algorithm was unable to estimate the static arm correctly in kicking sequence as colour, motion and pixel location fail to provide discriminative cues. The confidence score of pose recovery on this sequence is relatively low compared to that of other sequences as none of the clusters can be matched accurately to the arm model. Therefore, a low fitting score of the arm is produced and results in an overall low confidence score for the estimated pose according to Equation 14. Despite this, the estimated poses are still informative if the "confidence scores" of individual limbs are considered (see Section 7). In contrast, the pose recovery on walking, running, and punching are successful, because pixel motion provides a useful cue for body part segmentation even though colour is ambiguous.

6.5. Summary of Evaluation

Our pose estimation method does not require any training and produce quantitatively and qualitatively convincing results. Although recent learning based approaches [62][63][64][31] achieve 5-15 pixels error for HumanEVA datasets, they are constrained by only being able to estimate poses with activities with which they have been trained. Therefore, these approaches cannot be applied in realistic day-to-day scenarios with natural human motions, such as the HumanEVA "Combo sequence" (see Section 6.3.2). While several other activity independent approaches have been suggested [34][35][38], they only report visual qualitative results, which do not allow objective comparisons. We should mention that activity independent pose tracking has also been proposed and tested on HE I dataset [56]. They achieved an average error of 13.2 pixels for this easier task which relies on manual initialisation and where they assume the character is performing bipedal motion.

Since learning-based approaches usually learn poses based on MOCAP data (for example, the Ground Truth provided by HumanEVA), they usually have a positive bias towards the references they have learned from. On the other hand, our method uses visual cues similar to the ones perceived by humans and used to recognise body postures. Therefore, comparison of our results with MOCAP based Ground Truth may result in larger errors than with body joint positions estimated by human beings as discussed in Section 6.2.

7. CONCLUSIONS AND FUTURE WORK

In this paper, a novel probabilistic bottom-up/top-down 2D pose recovery framework is proposed. It is an iterative process between a bottom-up stage, which partitions the foreground probabilistically using relevant image cues, and a top-down stage, which performs a hierarchical body model fitting to constrain segmented body partitions. Since a suitable set of local image cues is exploited to extract characteristics of body parts, there is no need for training poses. Consequently, our approach is totally activity-independent. Since both bottom-up and top-down processes are modelled probabilistically using Gaussian mixtures, a confidence score is generated for evaluation of the success of pose estimation. Our method has been

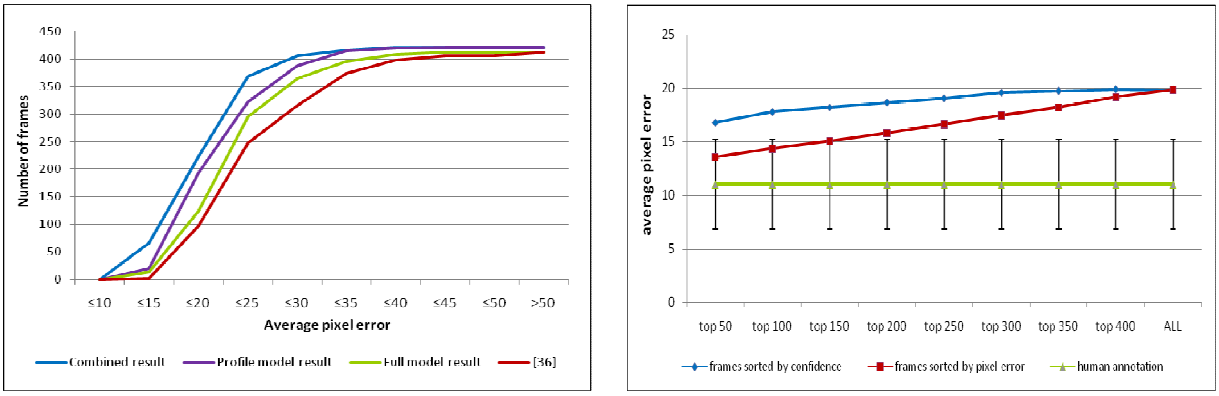
validated by various human motion sequences consisting of a wide range of activities, camera viewpoints within both indoor and outdoor scenarios. In addition, we identified some discrepancy between Ground Truth joint positions according to whether they are defined by humans or a motion capture system.

As results demonstrate, first, our confidence measure predicts 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 [56][62][63]. Even in cases that full pose recovery is unavailable, such as in the recovery of the kicking sequence shown in the last row of Figure 15, partial initialisation can still be achieved: since the confidence score is the product of each limb's fitting score, a tracker can be partially initialised across different frames where individual limbs are estimated accurately.

In future work, we intend to integrate our framework within a pose tracking system to increase its usability. Not only could poses, which currently cannot be recovered, be estimated by the tracker, but tracker predictions could contribute to pose evaluation. This would lead to smoother and continuous recovery of poses for a video sequence. Moreover, since we want to target real time applications, optimisation of the search strategies used in the framework is required.

ACKNOWLEDGEMENTS

This work was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) sponsored MEDUSA, and PROCESS projects (Grant No. EP/E001025/1 and EP/E033288 respectively)



(a) **(b)**

Figure 12: (a) Accumulative curve showing the average pixel error between the estimated joint locations and HE GT. Results of our previous work [36], fitting of the full model with the advanced technique, fitting of the profile model with the advanced technique and the combination of full and profile model fitting are presented with red, green purple and blue curves respectively. (b) Average pixel error of cumulated pose estimates. Blue and red curves show image frames sorted by confidence scores and actual pixel errors respectively. Green line represents errors in human annotation with s.d. shown as error bars.

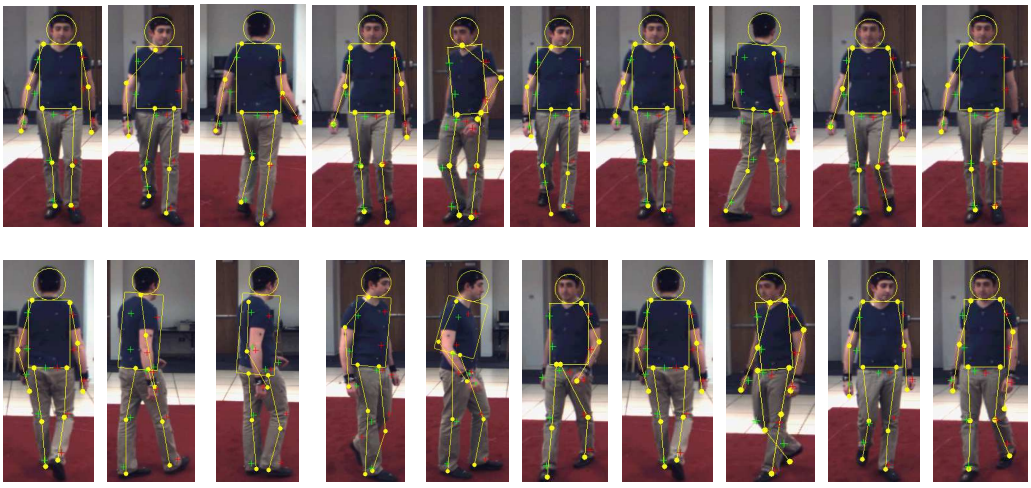


Figure 13: Results of pose recovery for HE I walking sequence [57]. Head and torso are shown by a yellow circle and rectangle. Yellow line segments and solid circles indicate the limbs and body joints; Crosses represent joints defined by HE GT.

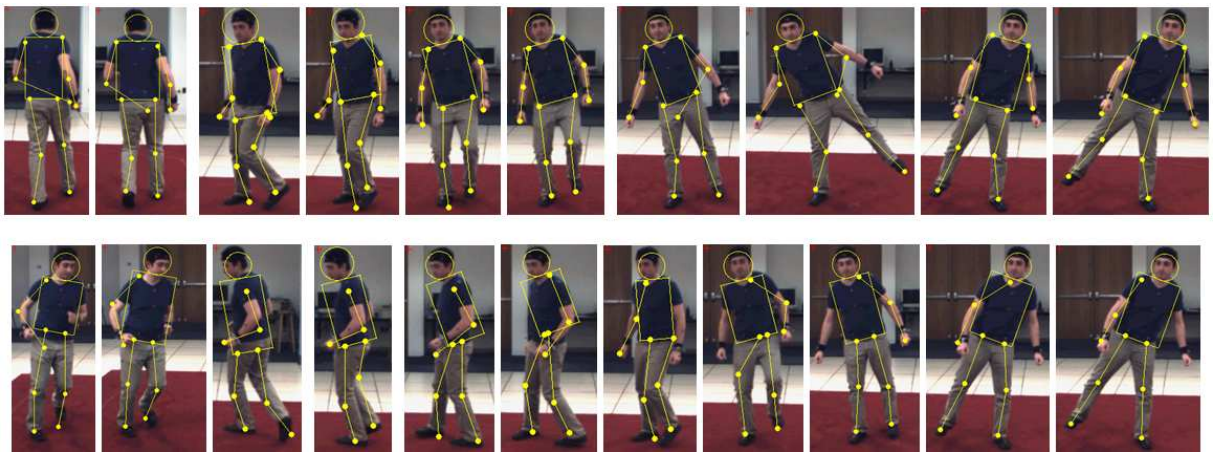


Figure 14: Results of pose recovery for HE I “Combo” sequence [57]. Head and torso are shown by a circle and rectangle. Yellow line segments and yellow solid circles indicate the limbs and body joints.

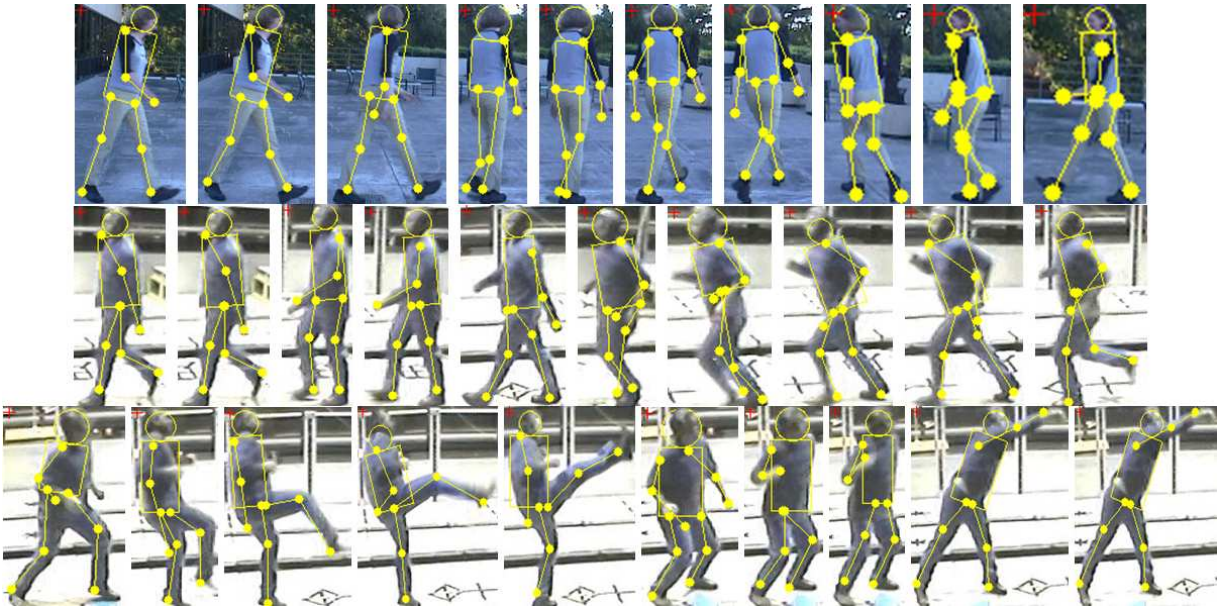


Figure 15: First row: results of pose recovery on the outdoor circular walking sequence [58]. Second row: results of pose recovery on straight-line walking (1st half) and running (2nd half). Third row: results of pose recovery on kicking (1st half) and punching (2nd half).

REFERENCES

[1] N. Spencer and J. Carter. “Towards pose invariant gait reconstruction”, ICIP’05 vol. 2, pp. 261-264, 2005.

[2] P. Kuo, A. Thibault, M Lewandowski, D. Makris, J.-C. Nebel, “Exploiting Human Bipedal Motion Constraints for 3D Pose Recovery from a Single Uncalibrated Camera”, Proc. VISAPP’09, 2009.

[3] P. Kuo, J.-C. Nebel and D.Makris. “Camera Auto-Calibration from Articulated Motion”, AVSS’07, pp.135-140. 2007.

[4] M. Armstrong, A. Zisserman and R. Hartley, “Self-Calibration from image triplets” ECCV’96 pp. 3-16, 1996.

[5] J.K. Aggarwal and Q. Cai, “Human Motion Analysis: A Review,” Computer Vision and Image Understanding, vol. 73, no. 3, pp. 428-440, 1999.

[6] D.M. Gavrila, “The Visual Analysis of Human Movement: A Survey,” Computer Vision and Image Understanding, vol. 73, no. 1,pp. 82-98, 1999.

[7] T.B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture, ” Computer Vision and Image Understanding, vol . 81, no 3, pp231-268, 2001

[8] T.B. Moeslund, A. Hilton, and V. Kru¨ ger, “A Survey of Advances in Vision-Based Human Motion Capture and Analysis,” Computer Vision and Image Understanding, vol. 104, no. 2, pp. 90-126, 2006.

[9] C.-S. Lee and A. Elgammal “Modelling View and Posture Manifolds for Tracking”, Proc. ICCV’07, 2007.

[10] A. Elgammal and C.-S. Lee “Tracking People on a Torus,” IEEE Trans. Pattern Analysis and Machine Intelligence vol 31, no3, pp. 520-538, 2009

[11] A. Sundaresan and R. Chellappa “Model-Driven Segmentation of Articulating human in Laplacian Eigenspace,” IEEE Trans. Pattern Analysis and Machine Intelligence vol 30, no 10, pp.1771-1785, 2008

[12] M. Vondrak, L. Sigal and O.C. Jenkins, “Physical Simulation for Probabilistic Motion Tracking”, Proc. CVPR’08, 2008

[13] C. Sminchisescu, A. Kanaujia, Z. Li and D. Metaxas, “Discriminative Density Propagation for 3D Human Motion Estimation,” Proc. CVPR’05, pp. 390 – 397, 2005.

[14] A. Agarwal and B.Triggs, “Recovering 3D human Pose from Monocular Images,” IEEE Trans. Pattern Analysis and Machine Intelligence vol 28, no1, pp. 44-58, 2006

[15] R. Urtasun and T. Darrell, “Sparse Probabilistic Regression for Activity-independent Human Pose Inference, ” Proc. CVPR’08, 2008

[16] T. Jolliffe, Principal Component Analysis. Springer Series in Statistics, 2 ed., 2002.

[17] G. Rogez, C. Orrite-Urunuela and J. Martinez-del-Rincon “A spatio-temporal 2D-models framework for human pose recovery in monocular sequences,” Pattern Recognition vol 41, pp. 2926-2944, 2008

[18] J.B. Tenenbaum, V. de Silva, and J.C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” Science, vol. 290, no. 5500, pp. 2319-2323, 2000.

[19] M.Belkin and P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation,” Neural Computation vol. 15, no 6, pp. 1373-1396, 2003.

[20] R. Li, T.-P. tian and S. Sclaroff, “Simultaneous Learning of Nonlinear Manifold and Dynamical Models for high-dimension al Time Series”, Proc. ICCV’07, 2007.

[21] S. Hou, A. Galata, F. Caillette N. Thacker and P. Bromiley, “Real-time Body Tracking Using a Gaussian Process Latent Variable Model” Proc, ICCV’07, 2007

[22] S.T. Roweis and L.K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[23] A. Elgammal and C.S. Lee, “Inferring 3D body pose from silhouettes using activity manifold learning”, CVPR’04 pp. 681-688. 2004.

[24] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, “Diffusion Maps, Spectral Clustering and the Eigenfunctions of Fokker-Planck Operators,” Proc. Conf. Neural Information Processing Systems (NIPS ’05), Dec. 2005

[25] G. Mori, X. Ren, A. A. Efros and J. Malik “Recovering human body configurations: Combing segmentation and recognition,” CVPR 2004 pp. 326-333, 2004

[26] X. Lan and D. P. Huttenlocher, “Beyond trees: common-factor for 2D human pose recovery”, Proc. ICCV’05, pp 470-477, 2005.

[27] N. Howe, “Silhouette Lookup for Automatic Pose Tracking” Proc. CVPR’04, pp15-22, 2004

[28] D.Weinland, E. Boyer, R. Ronfard, “Action Recognition from Arbitrary Views using 3D Exemplars”, Proc. ICC’07, 2007.

[29] J. Deutscher and I. Reid, “Articulated Body Motion Capture by Stochastic Search” Internal Journal of Computer Vision vol61, no2, pp.185-205, 2005

[30] A. O. Balan, L. Sigal and M. J. Black, “A Quantitative Evaluation of Video-based 3D Person Tracking”, Proc. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.349-356, 2005.

[31] Z. Husz, A. Wallace and P. Green, “Evaluation of a Hierarchical Partitioned Particle Filter with Action Primitives”, CVPR 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation ($EHuM_2$), 2007.

[32] X. Zhao and Y. Liu “Generative tracking of 3D human motion by hierarchical annealed genetic algorithm,” Pattern Recognition vol 41, pp. 2470-2483, 2008

[33] Z. Hu, G. Wang, X. Lin and H. Yan, “Recovery of upper body poses in static images based on joints detection,” Pattern Recognition Letters, Elsevier, vol 30, no 5, pp. 503-512, 2009

[34] D. Ramanan and D.A. Forsyth “Finding and tracking people from the bottom up” CVPR’03 pp. 467-474. 2003.

[35] P. Srinivasan and J. Shi. “Bottom-up recognition and parsing of the human body,”Prco, CVPR’07 pp.1-8, 2007.

[36] P. Kuo, D. Makris, N. Megherbi, J.-C. Nebel, "Integration of Local Image Cues for Probabilistic 2D Pose Recovery", ISVC’08, LNCS 5359, Springer-Verlag , 2008

[37] N. Howe, “Flow Lookup and Biological Motion Perception” Proc. ICIP’05, pp1168-1171, 2005

[38] X. Ren, A.C. Berg, and J. Malik, “Recovering Human Body Configurations using Pairwise Constraints,” Proc. ICCV’05, pp. 824–831, 2005

[39] I. Bouchrika and M.S. Nixon, “Gait-based Pedestrian Detection for Automated Surveillance,” Proc. ICVS’07, 2007.

[40] A. Bissacco, M.-H. TYang, and S. Soatto, “Fast Human Pose Estimation using Appearance and Motion via Multi-Dimensional Boosting Regression” Proc. CVPR ’07, pp.1-8, 2007

[41] C. Wren, A.Azarbayejani, T. Darrell and A. Pentland “Pfinder: Real-Time Tracking of the Human Body”, IEEE Trans. Pattern Analysis and Machine Intelligence vol 19, no7, pp. 780-785, 1997

[42] S. Park and J. K. Aggarwal, “Simultaneous Tracking of Multiple Body Parts of Interacting Persons”, Computer Vision and Image Understanding, vol. 102, no. 1, pp.1-21, 2006.

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

[44] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. CVPR’03, vol 2, pp. 459-466, 2003.

[45] P. Viola and M. Jones “Rapid Object Detection using a Boosted Cascade of Simple features”, CVPR’01 vol. 1, pp. 511-518, 2001

[46] D. Lowe, "Object recognition from local scale-invariant features", ICCV’99, vol 2, pp. 1150–1157. 1999

[47] http://opencv.willowgarage.com/wiki/MachineLearning

[48] C. Hu, X. Ma and X Dai. A “Robust person tracking and following approach for mobile robot”. Pro. Int. Conf. on Mechatronics and Automation pp. 3571-3576, 2007.

[49] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink and G. Sagerer “Audiovisual person tracking with a mobile robot”. IAS’04, pp. 898-906, 2004.

[50] S. J. Mckenna, Y. Raja and S. Gong “Tracking colour objects using adaptive mixture models” Image and Vision Computing, Elsevier, vol. 17 pp. 255-231. 1999

[51] C. Sminchisescu and B. Triggs, “Kinematic Jump Processes for Monocular 3D Human Tracking”. CVPR’01, vol. 1, pp. 69-76, 2003

[52] J.Deutscher, A. Blake and I. Reid, “Articulated Body Motion Captured by Annealed Particle Filtering”, CVPR ‘00,vol. 2, pp.126-133, 2000

[53] J. Deutscher, A. Davidson, and I. Reid, “Articulated Partitioning of High Dimensional Search Spaces associated with Articulated Body Motion Capture”, CVPR’01, vol. 2, pp. 669-676

[54] D. Ramanan.“Learning to parse images of articulated bodies”. NIPS. 2007.

[55] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision”, Proc. Imaging understanding workshop. pp. 121-130, 1981.

[56] J. Martinez-del-Rincon, J.-C. Nebel, D. Makris, C. Orrite, “Tracking Human Body Parts Using Particle Filters Constrained by Human Biomechanics”, BMVC’08, 2008.

[57] L. Sigal, A. O. Balan and M. J. Black, “HumanEVA: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion”, International Journal of Computer Vision pp. 4-27, 2009.

[58] Image sequences provided by Hedvig Sidenbladh, http://www.csc.kth.se/~hedvig/data.html, Last accessed 28 May 2009.

[59] MuHAVi: Multicamera Human Action Video Data, http://dipersec.king.ac.uk/MuHAVi-MAS, Last accessed 28 May 2009.

[60] 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.

[61] J. L. Rodgers and W. A. Nicewander, “Thirteen ways to look at the correlation coefficient”. The American Statistician, vol 42, no 1, pp 59–66, Feb 1988.

[62] N. R. Howe, “Recognition-Based Motion Capture and the HumanEva II Test Data”, in Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2), 2007.

[63] R. Poppe, “Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets”, in Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2), (2007).

[64] C.S. Lee, and A. Elgammal, “Body pose tracking from uncalibrated camera using supervised manifold learning”, in Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM), Whistler, Canada, 2006.

[65] A. Sundaresan and R. Chellappa, “Multi-camera Tracking of Articulated Human Motion Using Shape and Motion Cues”, IEEE Transactions on Image Processing, vol 18, no 9, pp 2114-2126, 2009.