Tracking Human Poses With Head Orientation Estimation

TIAN Jinglan[1,*], WANG Zhengyuan[2], LI Ling[3], LIU Wanquan[3]

(1.*School of Information Science and Engineering, University of Jinan, Jinan* 250022, *China*;
2.*Jinan Tengyue Electronic Co. Ltd.*, *Jinan* 250000, *China*;
3.*Department of Computing, Curtin University, Perth* 6102, *WA, Australia*)

Abstract: Lots of progress has been made recently on 2*D* human pose tracking with tracking-by-detection approaches. However, several challenges still remain in this area which is due to self-occlusions and the confusion between the left and right limbs during tracking. In this work, a head orientation detection step is introduced into the tracking framework to serve as a complementary tool to assist human pose estimation. With the face orientation determined, the system can decide whether the left or right side of the human body is exactly visible and infer the state of the symmetric counterpart. By granting a higher priority for the completely visible side, the system can avoid double counting to a great extent when inferring body poses. The proposed framework is evaluated on the HumanEva dataset. The results show that it largely reduces the occurrence of double counting and distinguishes the left and right sides consistently.

Key words: Human Pose Tracking; Head Orientation; Tracking by Detection

1 Introduction

Human pose tracking for 2D monocular sequences is still challenging in computer vision due to appearance variations, background clutter, illumination, motion blur and occlusion. Recent approaches on human pose tracking tend to estimate pose in each frame and integrate temporal coherence between pose in successive frames to remove the temporal noise. For achieving better tracking performance, many researchers put their efforts on improving the performance of pose estimation either by modelling a robust (accurate) appearance model[1, 2, 3, 4] or building more reliable body structure representations[5, 6]. Lu et al. propose an approach to learn an uncontaminated appearance model for each body part by removing background pixels through pixel-analysis strategy[4]. Some Researchers [7, 8] propose assembling more realistic body part dependencies by a hierarchical human body structure, e.g., poselets. Although these approaches are effective for pose estimation, one essential issue happens frequently during tracking is the confusion between the left and right limbs. For example, Fig.2 shows several tracking results from the framework proposed in[9] with a walking sequence from the well-known tracking dataset HumanEva[10]. Although the frontal (or back actually) pose can be accurately detected especially when all body parts are visible, the left and right limbs are often confused especially when the human body is lateral. Intuitively, in Fig.2(a), the pink colour bounds the left side of the body while the bounding boxes in light blue colour are for the right side. However, the left and right legs are incorrectly identified for the lateral pose shown in Fig.2(b) and (c) due to the overlapping of these body parts and the similarities of their appearance in terms of colour and shape.

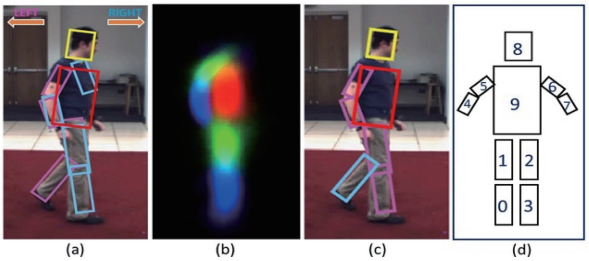

Fig. 1 With the proposed framework, the double counting and left/right confusion issues for pose estimation shown in (a) are corrected (see (c)). The body part order is illustrated in (d)

There is another big challenge remaining in 2D human pose tracking when dealing with occlusion in a video sequence, i. e., self-occlusion. Different parts occlude each other frequently because the human body is an articulated and symmetric structure. Body parts are often occluded by the torso or their symmetric counterparts, e.g., the left arm is occluded by the torso, or the left-upper leg is occluded by the right-upper leg, etc. in Fig.1(a). During occlusions, the appearance symmetry of the human body can cause the so-called 'double counting' problem in tracking: the same image evidence is used to explain the location of both symmetric parts. It is noted that temporal continuity priors can be used to deal with occlusions and the double counting problem[9]. Varun et al. [6] incorporate temporal reasoning about occlusion into a multi-target tracking framework to track human poses. These methods utilize motion consistency to force the tracker to find image evidence to support a smooth path when occlusion occurs.

The problems described above could be inherited from the base model used for 2D human pose tracking - the Pictorial Structures (PS) model [11, 12]. The left and right body parts in PS are defined only depending on their relative locations in the image coordinate system because the PS is originally designed to infer body poses for one frame. Moreover, the double counting problem often occurs for lateral pose in the PS framework because it always infers the body pose using the candidates from all the body part detectors. In this work, we propose using the head orientation (looking left or right) to provide instructive information to address the problems. A head-yaw-estimation step is introduced into the tracking framework to serve as a complementary tool to assist the human pose estimation. Yaw rotation of the head is one important type of head pose and it attracts much attention because its estimation has many potential applications[13]. Many image representations and models have been extracted to character the head pose, such as Bag-of-Words[14], Fisher vector[15] and VoD representation[13]. Moreover, some techniques of inference have been provided to detect the yaw rotation. In this work, accurate estimation of the head yaw angle is not necessary because we only need an indication on whether the human body is facing left or right, a simple skin colour detector and a set of threshold templates are hence used to roughly identify it.

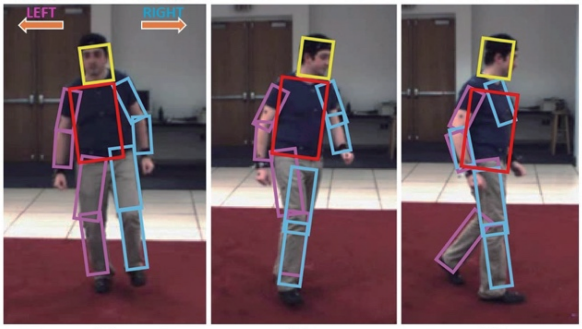

Fig. 2 Several tracking results from the framework[9]. In (b) and (c), the detections of left and right legs are not consistent with those in (a) because they are inversely identified

Generally, occlusions are more likely to happen when the tracked targets are in lateral poses. With the head orientation determined, if it is not front or back side, the system can first determine whether the left or right side of the body is fully visible and then deduce the situation of the other side based on image evidences. For example, for the side pose in Fig.1, the head pose detector indicates that the pose is facing right. The system can then identify that the left side of the body is definitely visible. Accordingly, the system will grant higher priority for the left body parts and assign the posterior with higher score to them. The confidence map of the posteriors is shown in Fig.1(b). The posteriors for the right side limbs will then be searched and allocated with reference to their left counterparts. If all posteriors for a certain right body part score very low, this part will be regarded as being occluded. Therefore, the double counting problem is effectively avoided and the confusion about left and right body parts is cleared up

(see Fig.1(c)). The next section provides description of the overall human pose tracking system, with an emphasis on the proposed novel approach utilizing head orientation detection.

2 Methodology

For an articulated body, the configuration is denoted with $L = [l_1, \ldots, l_N]$ and the image observation is $D = [d_1, \ldots, d_N]$, where N is the defined number of body parts. At time instance t, the configuration of an articulated body is denoted by L_t and the observation is denoted by D_t. Given an image sequence, human pose tracking is to infer the posterior $p(L_t \mid D_t)$ across all frames, i.e., to estimate the optimal tracks of each part, which corresponds to find the maximum a posteriori:

$$L^* = \underset{L}{\operatorname{argmax}}(p(L_{1:T}^{1:N} \mid D_{1:T})). \quad (1)$$

The general approach to optimize the above function is the tracking-by-detection manner, i.e., detecting body poses in each frame based on the Pictorial Structures (PS) model. Following a 1st-order Markov model and a pictorial structure model, the tracker is represented as:

$$p(L_{1:T}^{1:N} \mid D_{1:T}) \propto \prod_{t=1}^{T} \prod_{i=1}^{N} p(L_t^i \mid L_{t-1}^i) p(L_t^i \mid L_{t-1}^{\pi(i)}) p(D_t \mid L_t^i). \quad (2)$$

As a notation convention, the superscripts denote body parts (i ranges over the torso plus left/right upper/lower arms/legs) and subscripts denote frames $t \in 1, \ldots, T$. The term $\pi(i)$ denotes the parent of part i, following the tree structure. As mentioned before, there are two common problems in this kind of framework, i.e., double counting and left/right inconsistency. To avoid these issues during tracking process, we propose using a detector to roughly estimate the head orientation and then infer the human body poses by a stage-wise Pictorial Structures (s-PS) model depending on the estimated head orientation.

2.1 Stage-wise Pictorial Structures Model

The configuration of an articulated body is denoted with $L = [l_1, \ldots, l_N]$, where l_i for each individual body part is parameterized by the location (x, y) and the orientation θ, N is the defined number of body parts. In our case, $N = 10$, i.e., a 10-part model is used for the full body: head, torso and left/right lower/upper arms/legs. The image observation associated with L is denoted by D.

Human pose tracking aims at inferring $p(L_t \mid D_t)$ with PS model across all frames. Specifically, for each frame,

$$p(L_t \mid D_t) = \sum_{i=1}^{N} E^u(l_t^i; D) + \sum_{i \sim j} E^p(l_t^i, l_t^j; D),$$

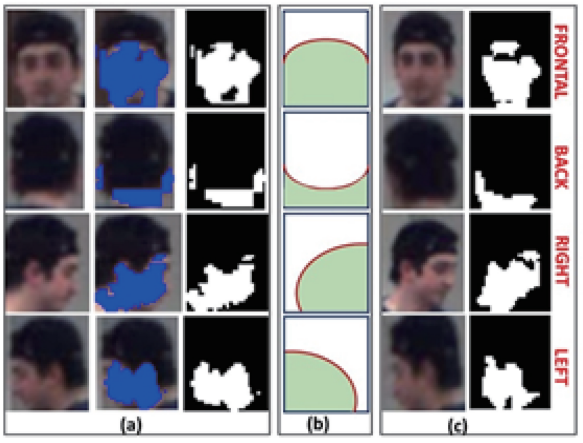

Fig. 3 The face detector is shown in (a). The skin-map is overlaid onto the image marked in blue colour and the binary skin-maps are shown in the third column (b) shows the set of face templates. The testing results are shown in (c)

where E^u is the unary term and E^p is the pairwise term of the pictorial structures model, and i ~ j denotes the neighborhoods relationship between the body parts. The unary term denotes the image likelihood based on a set of pre-trained shape-based appearance models for all body parts[16]. In the generic PS model, the pairwise term represents the information about relations between the connected body parts.

2.2 Estimating the Head Facing Orientation

The goal of this step in our work is to roughly decide whether the tracked target is facing left/right/front/back, rather than estimating the accurate rotational angle of the head. Therefore, before inferring

body poses in the PS model, a simple algorithm is utilized to roughly identify the head orientation of the tracked target, such as front, back, left or right.

Given the head bounding box for each frame shown in the first column of Fig.3, it is noted that the absolute location of the face area (C) or the relative location of the

Algorithm 1 Tracking human poses with head orientation estimation

Generate all part proposals with part detectors.
Generate head orientation from head-yaw estimation.
for *all frames* do
Sample posterior candidates for all body parts
if *is right facing pose* then
Prioritize the left-side parts and then infer the right side pose with reference to the left side parts.
else if *is left facing pose* then
Prioritize the right-side parts and then infer the left side pose with reference to the right side parts.
else
Directly infer the pose with 10-part PS model.
end if
end for

face and hair regions (ρ) are essential clues for the head orientation estimation. Considering the small size of the head area in this kind of applications and to ensure simplicity of the algorithm, we utilize a skin colour detector to select the face area which can produce a binary skin-map and highlight patches of skin-like pixels for a given image (see Fig.3 (a)). The hair region is not detected separately. The head image is firstly transformed from RGB colour space to YCbCr colour space and the resultant image is comprised of intensity component (Y) and chrominance components (Cb and Cr).

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4}$$

The YCbCr colour space is effective and efficient for the separation of image pixels in terms of colour and can be applied for complex colour images with uneven illumination.

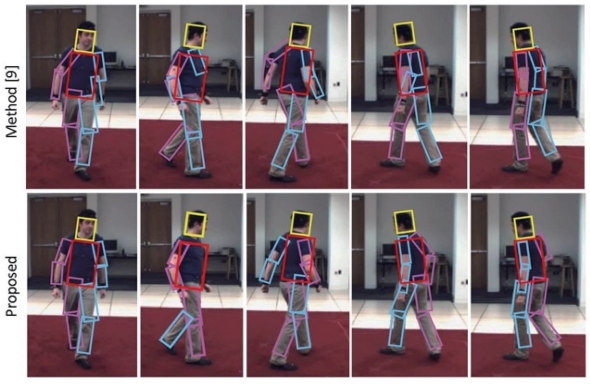

Fig. 4 Qualitative comparison. Several screenshots show the improvement on the sequence HE Combo_S2

To determine the head orientation, a set of templates for different head orientation are pre-set and illustrated in Fig.3(b). The boundary of the face area is assumed roughly elliptic and the region crossed with the head bounding box is considered as the face area, which is marked in green. The head resolution is set as $33 * 28$. Intuitively, if the skin region accounts for a great part of the bounding box and locates in the central, it is considered that the target is looking forwards. In contrast, if there is a small number of skin-like pixels and the region of them locates at the bottom of the head bounding box, it is considered as a backward pose. If the facing area is located only at the right-bottom or the left-bottom area of the bounding box, we consider that the head is looking right or left and then the sideway pose is likely to face right or left accordingly. It is noted that in this work, it is assumed that the pose orientations are always in accordance with the head orientation during tracking. Other cases, such as frontal pose with sideways head are not considered.

2.3 Tracking

In the proposed framework, we first perform inference using the tree-based PS model, and then sample from the posterior margins of each part according to the Gaussian Prior and the proposed sampling approach. The fact that the posterior margins of

each part in the tree-structure model approximately satisfy a Gaussian distribution has been proposed and proved in[9]. Based on this fact, the posterior marginal set (named M_t^i) can be sampled and more posteriors are sampled at the locations near the center of the tracked body part in the previous frame. The marginal set sampled this way is named $M_{t\text{-}g}^i$.

The proposed system: With the marginal set $M_{t\text{-}g}^i$ and the given head orientation, the system samples the posteriors and performs inference again to obtain the final pose estimation. If the body pose is determined as not being frontal or back side, the system can first determine whether the left or right side of the human body is fully visible according to the head orientation and grant higher priority to that side for the subsequent pose inference. For example, if the head pose detector indicates that the pose is facing right, the system can identify that the left side is fully visible. Accordingly, the system will grant higher priority for the left side and sample posteriors of body parts in the order $[l_0, l_1, l_4, l_5, l_8, l_9]$.

Table 1 The Performance Comparison in Percentage for HE-combo-S1 Sequence

Method	Tor	Head	LLL	RLL	LUL	RUL	LFA	RFA	LUA	RUA
[9]	100	100	82.6	86.1	89.7	90.5	67.3	69.1	72.2	74.8
Proposed	100	100	93.1	94.5	95.2	96.1	78.7	81.6	88.3	89.7

Table 2 The Performance Comparison in Percentage for HE-combo-S2 Sequence

Method	Tor	Head	LLL	RLL	LUL	RUL	LFA	RFA	LUA	RUA
[9]	100	100	80.7	84.1	83.5	86.8	65.4	66.3	69	70.7
Proposed	100	100	91.9	93.2	94.5	94.7	76.1	79.4	84.3	85.9

The next step is to select the sampled posteriors for the right side limbs $[l_2, l_3, l_6, l_7]$ and allocate them to different places from the left counterparts. It is noted that if all posteriors for a certain right body part are small, this part will be regarded as being occluded. In contrary, if the pose is facing left, the order will be $[l_3, l_2, l_7, l_6, l_8, l_9]$ and $[l_0, l_1, l_4, l_5]$. The arrangement of the part order is demonstrated in Fig.1(d). If the body pose is determined as being frontal or back side, it will be inferred directly with the complete 10-part PS model.

3 Experimental Results

In this section we evaluate the performance of the proposed framework and compare its performance with the framework proposed in [9].

Datasets: Two sequences, HE_combo_S1 and HE_combo_S2, selected from the popular HumanEva dataset [10] are used in the experiments. Both show a person moving in a circle and contain several non-lateral motions, such as jumping, kicking, leaning and stretching. The main difference between them is that the HE_combo_S2 sequence contains more side-facing poses than the HE_combo_S1 sequence.

Evaluation Metric: We use the well-known PCP metric introduced in [17] to numerically evaluate the performance of the proposed framework. This is commonly used as an evaluation metric in human pose estimation and tracking. In [9], the accuracy was calculated without considering the consistency during tracking. In fairness, we recalculated the performance of that framework with the left and right consistency during tracking.

Results: Several screenshots of the tracking results from the two frameworks are shown in Fig.4. Row 1 shows results from the framework [9] and the selected screenshots show the inconsistent detections of the left and right body parts during tracking. In addition, the results are also affected by the mis-

placed body parts when occlusion occurs and the double counting problem for some side-facing poses. The performance of the proposed approach is demonstrated in Row 2, where the accuracy is significantly increased since most of the left/right confusion and the double counting problems are avoided, it benefited from the stage-wise pose inference process with head orientation estimation.

The quantitative performance of both frameworks on the two Combo sequences is shown in Table 1 and Table 2. For both sequences, the proposed approach improves the tracking performance consistently in every single case by around 10%. The left/right consistency in the proposed framework is significantly improved, so that it decreases the detection errors in the previous framework [9]. The stage-wise pose inference approach further increases the accuracy for limb tracking by eliminating the double counting errors caused by self- occlusions.

4 Conclusion

A human pose tracking framework that complemented by a head orientation detector has been presented in this paper. Although it is a very simple and efficient method, it addresses one of the biggest problems in 2D human pose tracking, i.e., the confusion of the left and right parts due to the over-lapping and occlusion when the human is facing sideways. For lateral poses, a simple head orientation detector is proposed determining the head facing direction, and then determining the orientation of the whole body. The fully visible side of the body can be determined and given higher priority for the subsequent inference. The other side will be inferred with reference of the visible side, with some body parts confidently determined to be occluded. The stage-wise pose estimation system significantly decreases the left/right confusion and effectively eliminates the double counting errors. Experimental results show that the proposed framework is able to achieve high detection rate for very complicated video sequences involving large variations of motions and orientation. Future work will aim at modelling the probability of an occlusion state and dealing with inferences for the occluded parts.

References

[1] Ramanan D., Forsyth D., and Zisserman A. (2007). Tracking People by Learning Their Appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), pp. 65-81.

[2] Ferrari V., Jime 813 0857 M., and Zisserman A. (2009). 2d Human Pose Estimation in TV Shows. *Statistical and Geometrical Approaches to Visual Motion Analysis*, pp. 128-147.

[3] Johnson, S., and Everingham, M. (2011). Learning effective human pose estimation from inaccurate annotation. In: *Computer Vision and Pattern Recognition*, 42, pp.1465-1472.

[4] Lu Y., Li L., and Peursum P. (2012). Background Suppression for Building Accurate Appearance Models in Human Motion Tracking. In: *International Conference on Digital Image Computing Techniques and Applications*, pp.1-6.

[5] Sapp B., Weiss D., and Taskar B. (2011). Parsing Human Motion with Stretchable Models. In: *Computer Vision and Pattern Recognition*, pp. 1281-1288.

[6] Ramakrishna V., Kanade T., and Sheikh Y. (2013). Tracking Human Pose by Tracking Symmetric Parts. In: *Computer Vision and Pattern Recognition*, pp. 3728-3735.

[7] Tian J., Li L., and Liu W. (2014). A Robust Framework for 2d Human Pose Tracking with Spatial and Temporal Constraints. *IEEE Conference on Digital lmage Computing: Techniques and Applications*, 2014, pp. 1-8.

[8] Johnson, S., and Everingham, M. (2010). Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In: *British Machine Vision Conference*, Aberystwyth, pp.1-11.

[9] Tian J., Li L., and Liu W. (2015). Monocular Human Motion Tracking with Non-connected Body Part Dependency. In: *IEEE Conference on Digital lmage Computing: Techniques and Applications*, 2015, p. 17-23.

[10] Sigal L., Balan A., and Black M. (2010). Humaneva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated

Human Motion. *International journal of computer vision*, 87(1), pp. 4-27.

[11] Fischler M. and Elschlager R. (1973). The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, 100(1), pp. 67-92.

[12] Felzenszwalb P. and Huttenlocher D. (2005). Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1), pp. 55-79.

[13] Ma B., Huang R., and Qin L. (2015). Vod: A Novel Image Representation for Head Yaw Estimation. *Neuro computing*, 148, pp. 455-466.

[14] Sivic J. and Zisserman A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In: *IEEE International Conference on Computer Vision*, pp. 1470-1477.

[15] Perronnin F. and Dance C. (2007). Fisher Kernels on Visual Vocabularies for Image Categorization. In: *Computer Vision and Pattern Recognition*, pp. 1-8.

[16] Andriluka M., Roth S., and Schiele B. (2012). Discriminative Appearance Models for Pictorial Structures. *International Journal of Computer Vision*, pp. 1-22.

[17] Ferrari V., Jimenez M., and Zisserman A. (2008). Progressive Search Space Reduction for Human Pose Estimation. In: *Computer Vision and Pattern Recognition*, pp. 1-8.

Authors' Biographies

Tian Jinglan, received the BEng degree from Harbin Engineering University in 2006 and the PhD degree in computer science from the University of Western Australia in 2016. From 2006 to 2008, she worked in Inspur Group Co., Ltd. as an engineer. She is currently a lecture in University of Jinan. Her research interests include computer vision, machine learning and artificial intelligence.

Email: ise_tianjl@ujn.edu.cn

Wang Zhengyuan, received the Bachelor's degree in electronic and information engineering from Qufu Normal University in 2005. He is currently working in Tengyue Electronic Co. Ltd. as the director of Technical Centre.