The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Cross-View Person Identification by Matching Human Poses Estimated with Confidence on Each Body Joint

Guoqiang Liang,[1, 2] Xuguang Lan,[1] Kang Zheng,[2] Song Wang,[2, 3] Nanning Zheng[1]

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, 710049

[2]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, 29208

[3]School of Computer Science and Technology, Tianjin University, Tianjin, China, 30072

gqliang@stu.xjtu.edu.cn, {xglan, nnzheng}@mail.xjtu.edu.cn, zheng37@email.sc.edu, songwang@cec.sc.edu

Abstract

Cross-view person identification (CVPI) from **multiple temporally synchronized** videos taken by multiple wearable cameras from different, varying views is a very challenging but important problem, which has attracted more interests recently. Current state-of-the-art performance of CVPI is achieved by matching appearance and motion features across videos, while the matching of pose features does not work effectively given the high inaccuracy of the 3D human pose estimation on videos/images collected in the wild. In this paper, we introduce a new metric of confidence to the 3D human pose estimation and show that the combination of the inaccurately estimated human pose and the inferred confidence metric can be used to boost the CVPI performance –the estimated pose information can be integrated to the appearance and motion features to achieve the new state-of-the-art CVPI performance. More specifically, the estimated confidence metric is measured at each human-body joint and the joints with higher confidence are weighted more in the pose matching for CVPI. In the experiments, we validate the proposed method on three wearable-camera video datasets and compare the performance against several other existing CVPI methods.

Introduction

Video-based surveillance has been widely used in many security, civil, and military applications. Traditional surveillance videos are captured by multi-camera network, where all the cameras are installed at fixed locations. Since they cannot move freely, they can only cover limited areas from pre-fixed view angles. In recent years, wearable cameras, like Google Glass and GoPro, have been introduced to many applications to expand the video coverage. Compared with fixed cameras, wearable cameras are mounted over the head of the wearers and can move with the wearers to better capture the scene of interest. For example, in a sport game or a protest event, multiple policemen can wear cameras to record videos at different locations and from different view angles, which can facilitate the detection of abnormal persons and activities.

One fundamental problem in analyzing these multiple videos taken by wearable cameras is **cross-view person identification (CVPI)** – identifying the same person from these multiple videos (Zheng et al. 2017). As in (Zheng et al. 2017), we assume all the videos are **temporally synchronized**, which can be achieved by sharing clock across all the cameras. Given the temporal synchronization, if the corresponding frames across multiple videos cover the same person, this person must bear a unique pose and motion in 3D space. As a result, we can estimate 3D pose and motion on each video and match them across these videos to achieve CVPI. As in many person re-identification methods, appearance feature matching can also be used for CVPI (Zheng et al. 2017), although the extracted 2D appearance features may vary under different views.

Previous work (Zheng et al. 2017) has shown the effectiveness of using appearance and motion features for CVPI, especially when using the view-invariant motion features extracted by supervised deep learning. It also showed that the appearance features and motion features can complement each other to improve the CVPI performance. However, the use of pose features for CVPI (Zheng et al. 2016) is not very successful due to the high inaccuracy of the 3D human pose estimation (HPE) on videos/images collected in the wild. For example, for the body joints in 3D HPE, the mean Euclidean distance between the ground-truth 3D locations and the estimations of Pavlakos et al. (2017) is 71.90 mm, which is about 1/7 of the length of human torso in Human 3.6M dataset (Ionescu et al. 2014), which is collected in a highly-controlled lab environment with very simple background. In outdoor environments, with more

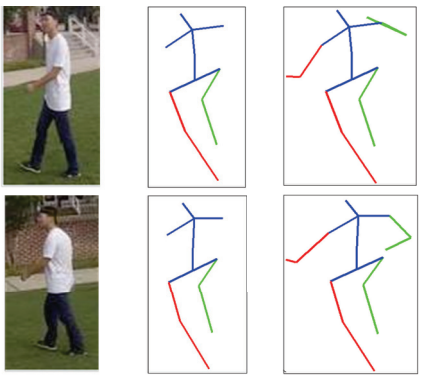


Figure 1: An illustration of 3D HPE with confidence. Left: Different views of the same person taken at the same time. Middle: (HPE-localized) joints with confidence larger than 0.7. Right: (HPE-localized) Joints with confidence larger than 0.1.

camera motion and view-angle change, the accuracy of 3D HPE can be much worse (Zhou et al. 2017).

In 3D HPE, the localization accuracy of different joints is highly inconsistent. As shown in Pavlakos at al. (2017), the 3D localization error at wrist is much larger than that at hip -- 101.48mm versus 28.81mm. This inconsistency comes from possible occlusions and different degrees of freedom. While using poorly estimated joint locations may significantly influence the accuracy of CVPI, the basic idea of this work is to identify a subset of accurately localized joints and then use them to improve CVPI.

To achieve this goal, we extend the 3D HPE to not only estimate the location of each joint, but also provide a confidence for each localized joint – the joints with higher confidence will be weighted more when matching 3D poses for CVPI. An example is shown in Fig. 1, where two images of the same person are taken from different views at the same time. Due to self-occlusions, the estimated locations of the right arm from a 3D HPE algorithm are largely incorrect. Without considering the right arm, we can match the estimated 3D poses from these two images much better. In this paper, we introduce a confidence metric for the estimated location of each joint in 3D HPE and combine HPE and the confidence metric to boost CVPI. Note that the goal of this paper is not to develop a new HPE algorithm with higher accuracy. Instead, we simply select one existing HPE algorithm, derive confidence at each joint and then apply them for CVPI.

In this paper, we derive the confidence at each joint from three aspects: (1) 2D confidence, which is derived from the process of 2D HPE; (2) 3D confidence, which refers to the certainty of 3D HPE from 2D joint heat-maps; (3) temporal confidence, which reflects the stability of joints' locations over time. These three aspects cover all the major steps of current 3D HPE methods. Confidence-weighted pose matching is performed between each pair of corresponded frames and then summarized over all the frames to obtain a pose-matching score between two videos. Then, for a given video, the matched video is the one with the smallest matching distance in the gallery dataset. Finally, we integrate the pose matching with appearance and motion feature matching for CVPI. Experimental results show that the use of the poses estimated with confidence can complement the appearance and motion features in CVPI.

Our main contributions include: (1) we introduce a new metric of confidence to 3D HPE; (2) we show the combination of the inaccurately estimated human pose and inferred confidence metric can improve the CVPI performance; (3) we achieve a new state-of-the-art performance of CVPI.

Related work

In this section, we briefly review the related works on person identification, human pose estimation, and confidence analysis.

Person identification

CVPI aims at associating person from temporally synchronized video taken by wearable cameras, which is proposed by Zheng et al. (2016, 2017). Compared with traditional person re-identification, the temporal synchronization brings new features for person identification. First, the 3D human poses of the same person are identical in the same frame of a pair of synchronized videos. Zheng et al. (2016) adapt the method in (Gupta et al. 2014) to estimate 3D human poses and use the distance of poses as the matching metric, but resulting in unsatisfactory CVPI performance. Besides pose, human motion is also consistent in 3D space if the same person present in synchronized videos. In order to compare the optical flow from different views, a triplet network is trained to learn view-invariant features (Zheng et al. 2017). Then, Euclidean distance between these features is used as a metric for CVPI. The combination of the appearance features and these learned motion features does lead to much better CVPI accuracy, but the cross-dataset accuracy is still lower than unsupervised methods.

Other related work is person re-identification, which aims to match persons captured in different time. The works on person re-identification can be roughly divided into two parts: effective feature extraction (Yang et al. 2017) and discriminative distance metrics learning (Li and Wang 2013; Yang et al. 2016). Recently, end-to-end CNN-based methods (Chen et al. 2017; Wang et al. 2016) have been developed to learn features and metrics at the same time.

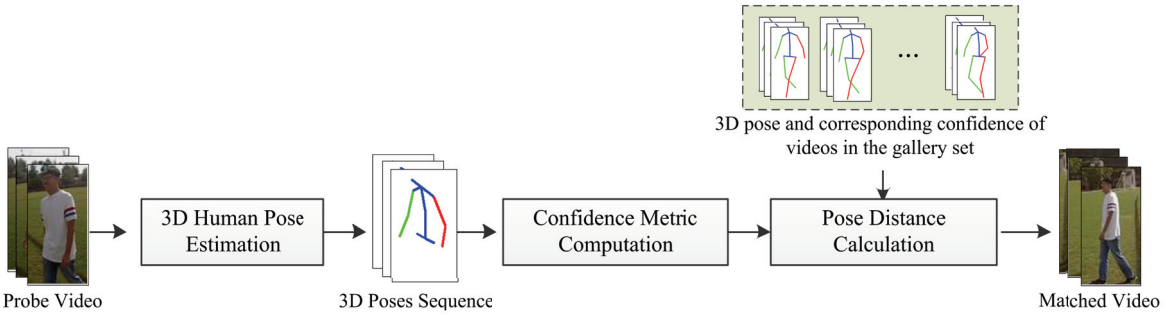


Figure 2: An illustration of the proposed framework for CVPI. First, we estimate 3D pose from the probe video. Then we compute its confidence metric. We then apply the same algorithms to estimate the 3D pose and the confidence metric for each video in the gallery set. Finally, we calculate the confidence-weighted pose distance between the probe video and each video in the gallery set. The video in the gallery set with the minimal distance is taken as the matched one.

Human pose estimation

Due to the powerful learning capacity, convolution neural network (CNN) has become a main building block for 3D HPE. As a fundamental step, 2D HPE from a single image has achieved high accuracy in various scenes (Newell, Yang, and Deng 2016; Insafutdinov et al. 2017; Andriluka et al. 2014). In contrast, the accuracy of 3D HPE is still far from satisfactory. A popular idea is to train a CNN to directly regress joint locations (Li and Chan 2014). After this, many improvements have been proposed, such as adding viewpoint prediction (Ghezelghieh et al. 2016), enforcing structural constraints (Tekin et al. 2016) and learning a way to fuse 2D and 3D information (Tekin et al. 2017). These algorithms are mainly developed for controlled lab environments. Their performances cannot be preserved when the images/videos are taken in outdoor environments or in the wild. To alleviate this problem, two-step approach is usually employed (Chen and Ramanan 2016; Wu et al. 2016; Zhou et al. 2016 and Zhou et al. 2017). The first step is to estimate 2D joint heat-maps, which can benefit from existing methods. Then, the 3D locations are regressed from the 2D joint heat-maps.

Our goal is to improve CVPI performance by introducing a confidence metric to each joint in the estimated 3D human pose instead of developing a new 3D HPE method. In this paper, we select one recent 3D HPE algorithm proposed by Zhou et al. (2016) to derive the confidence on each joint for CVPI. However, the proposed approach can be easily applied to other HPE algorithms.

Confidence Metric

Confidence reflects the certainty of an algorithm on its output. In this paper, we focus on confidence analysis over a single output instead of its statistical meaning over an output set. There have been existing works on confidence analysis for human pose estimation, face recognition, and other related topics (Jammalamadaka et al. 2012; Dutta, Veldhuis, and Spreeuwers 2015; Drevelle and Bonnifait 2013; Amin et al. 2014 and Zhang et al. 2014). Jammalamadaka et al. (2012) train a human pose evaluator to predict whether an algorithm returns a correct result given new test data. Amin et al. (2014) design a component to identify true positive pose estimation hypotheses with high confidence. These methods derive a confidence value for the entire pose estimation. In this paper, we focus on deriving a confidence value for each joint localized by HPE.

Cross-View Person Identification

In this paper, we propose to utilize 3D human pose for CVPI. Namely, CVPI is completed by matching the 3D human poses extracted from different videos. Considering the high inaccuracy of 3D human pose estimation, we introduce a metric of confidence to the 3D pose estimation. Specifically, the metric of confidence measures the certainty of each joint's location predicted by 3D human pose estimation method. Then, this confidence metric is used as weights in computing the Euclidean distance between a pair of 3D poses. The distance between two videos is the sum of pose distance over all the frames. For a probe video, the matched video is the one with the smallest distance in the gallery dataset.

As shown in Fig. 2, the proposed method can be divided into three parts: 3D human pose estimation from videos, confidence estimation on each joint, and person identification by matching 3D human pose weighted by confidence. We introduce the person identification in this section and the other two parts will be described in the following two sections. In order to compute the distance between poses, they should be normalized into the same view and scale. Our normalization includes three steps: rescale the limb's length to a constant value; translate the pelvis to the origin of axis; rotate the zenith and azimuthal of torso (the seg-

ment between pelvis and spine) to a constant angle because this body part can be assumed to be rigid. Assuming that we have obtained the normalized 3D poses and their confidence at each joint, the distance D_P between a pair of poses is defined as

$$D_P = \sum_{j=1}^{J} \min\left(W^1(j), W^2(j)\right) \left\| S^1(j) - S^2(j) \right\|_2 \quad (1)$$

where $S^1(j)$ ($S^2(j)$) represents the 3D location of j-th joint of the first (second) 3D pose, whose corresponding confidence value is $W^1(j)$ ($W^2(j)$), and J is the total number of joints. For each joint, we compare its confidence in the two considered poses and select the smaller one as the weight.

In matching two videos, we only use a subset of joints that are captured in both two corresponding frames to compute this distance. If a joint is detected in one video but not the other, this joint will not be considered in computing the distance. The final distance D_P between a pair of videos is the sum of distance over all the frames.

Further, we integrate the pose-based CVPI method with appearance- and motion-based CVPI methods by

$$D = D_M + \alpha D_A + \beta D_P \quad (2)$$

where D is the fused distance, D_M, D_A and D_P are the matching distance computed by motion, appearance and pose respectively, α, β are coefficients to balance the different value ranges of the three distances. Here, we use (Zheng et al. 2017) and (Yan et al. 2016) to compute the motion and appearance distances, respectively. Specifically, for each video, two feature vectors are derived to represent the motion or appearance features respectively. The former uses the optical flows while the latter uses the color and LBP features. Then, the Euclidean distance between motion (appearance) feature vectors derived from two videos are taken as the distance $D_M(D_A)$. The values of α and β are selected on validation dataset of SEQ1 (one of the datasets, see detail in the experiment section) and then we use these values for all three datasets. The final values for α, β are 0.5 and 10 respectively, since the pose-based distance D_P is typically much smaller than the appearance-based distance D_A. The influence of these values on CVPI performance will be analyzed in the experiment section.

3D Human Pose Estimation

As mentioned above, we select the 3D HPE method proposed by Zhou et al. (2016) to develop the proposed method of confidence estimation and pose-based CVPI. The 3D HPE consists of two steps: 2D HPE and 3D HPE from 2D heat-maps. Since the original method assumes all joints are captured by cameras, we adapt it to handle the case of varying number of captured joints in real videos. In the following, we first review this HPE algorithm and then introduce our adaptation.

2D Human Pose Estimation

Due to the remarkable performance, the stacked hourglass network architecture (Newell, Yang, and Deng 2016) is used for 2D HPE, which is performed on each frame independently. On each frame, the output Y is J heat-maps, each of which represents a 2D probability distribution for one joint. This network is trained by using the following Euclidean loss

$$L_{2D} = \frac{1}{J} \sum_{j=1}^{J} \left\| Y_j - Y_j' \right\| \quad (3)$$

where Y_j and Y_j' are the predicted heat-map and the ground-truth heat-map for j-th joint. The final 2D locations are the coordinates of peak value in heat-maps. Since this module is trained on the wild images, it shows great performance in the videos taken by wearable cameras.

3D HPE Using Heat-maps

After getting the 2D heat-maps on all the frames of a video, 3D HPE for a sequence is formulated as an energy minimization problem. Given a sequence of 2D poses over all the frames, the 3D pose is estimated by minimizing the following loss function with respect to C, R, T

$$L(\theta; P) = \frac{\nu}{2} \sum_{t=1}^{n} \left\| P_t - R_t \sum_{i=1}^{k} c_{it} B_i - T_t \right\|_F^2 + \mathcal{R}(\theta) \quad (4)$$

where $P_t \in R^{2\times J}$ is the 2D locations at time t obtained from the corresponding 2D heat-maps, c_{it} is the corresponding coefficient in the frame t for the i-th basis pose $B_i \in B$, B is the dictionary of 3D poses, R_t and T_t denote the camera rotation and translation. For notational convenience, we use $C = \{c_{it}\}$, $R = \{R_t\}$ and $T = \{T_t\}$ to represent the set of parameters in all frames. Finally, all these parameters are denoted as $\theta = \{C, R, T\}$, $\mathcal{R}(\theta)$ is the prior on parameters θ, $\|\cdot\|_F$ represents the Frobenius norm. Refer to (Zhou et al. 2016) for more details. The final 3D pose at time t is defined as

$$S_t = \sum_{i=1}^{k} c_{it} B_i \quad (5)$$

Since the 3D pose is rebuilt with a 3D pose dictionary, it always satisfies the structure constraints of human body.

Adaptation to Handle Missing Body Parts

Zhou et al. (2016) assume that all 2D joints are captured by the camera. In practice, some body parts may not be viewable in some frames of a video due to view-angle changes and occlusions. In this case, the above 2D HPE may return false locations for the missing joints, as shown in the top row of Fig. 3. Such incorrectly localized joints may violate the structure constraint of human body and prevent from rebuilding the correct 3D pose using 3D pose dictionary. Even if the algorithm produces an eclectic 3D pose, it will

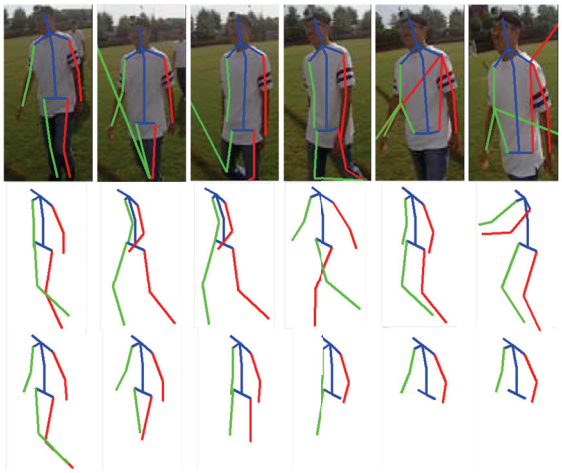


Figure 3: An illustration of 3D HPE results. Top: original image sequence with estimated 2D pose; Middle and Bottom: estimated 3D human pose without and with the adaptation to handle missing body parts. Green lines represent left limbs while the red lines represent the right limbs.

not conform to the true 3D pose in the image, as shown in the second row of Fig. 3. Using these false 3D poses may seriously hurt the performance of CVPI.

The peak value of heat-maps for joints with severely incorrect location is always smaller than that of normal joints. Based on this fact, we introduce the concept of visibility to the original algorithm. We assign a binary visibility label to every joint by comparing the maximal value of their heat-maps with a fixed threshold. Specially, the label is 1 if the peak value is larger than 0.1, and 0 otherwise. This threshold is selected through experiments on the validation dataset of SEQ1 and it is used to other two datasets. Then, we add the visibility label into the loss function (4), leading to

$$L(\theta;P)=\frac{\nu}{2}\sum_{t=1}^{n}\left\|\phi_t\left(P_t-R_t\sum_{i=1}^{k}c_{it}B_i-T_t\right)\right\|_F^2+\mathcal{R}(\theta) \quad (6)$$

where $\phi_t\in\{0,1\}^{1\times J}$ represents the visibility label of each joint at time t. Eq. (6) can be rewritten as

$$L(\theta;P)=\frac{\nu}{2}\sum_{t=1}^{n}\left\|\phi_t P_t-R_t\sum_{i=1}^{k}c_{it}\left(\phi_t B_i\right)-\phi_t T_t\right\|_F^2+\mathcal{R}(\theta) \quad (7)$$

As in the original algorithm, we minimize (7) through updating C, R or T alternately while fixing the others. The first term $\phi_t P_t$ in Eq. (7), which represents the 2D locations of visible joints, can be calculated in advance. Therefore, the update manner of minimizing Eq. (7) is the same as that of minimizing Eq. (4) in the original method except for changing the dictionary using the visibility matrix in advance. The parameters are kept the same as the original algorithm. Sample results of our adapted HPE algorithm are shown in the bottom row of Fig. 3.

Confidence Metric to 3D HPE

In this section, we define the confidence of each joint in a 3D pose estimation, which includes three aspects: confidence of 2D HPE, confidence of 3D HPE from 2D heat-maps and the temporal confidence. The first two focus on spatial space, while the last one is on the temporal space.

Confidence of 2D HPE

The confidence of 2D HPE is the certainty of the estimated 2D locations of joints. In fact, a heat-map is defined as a per-pixel likelihood for joint's locations. So we regard the value of heat-maps as the confidence of 2D joint locations. However the resolution of heat-maps is a quarter of that of the original images, indicating that the value of heat-maps is not very smooth. To improve robustness, the confidence of location p is defined as a weighted average of its four neighbors and itself. Supposing the j-th joint is detected at location p at time t, its 2D confidence is defined as

$$W_t^{2D}(j)=0.5\times Y(p)+0.5\sum_{p'\in N(p)}0.25\times Y(p') \quad (8)$$

where $W_t^{2D}(j)\in R$ is the 2D confidence of j-th joint at time t, $Y(p)$ denotes the value of heat-map Y at location p at time t, $N(p)$ represents the four neighbors of p. Such 2D confidence reflects the certainty of the map projecting visual appearance to 2D locations.

Confidence of 3D HPE

Since there may exist multiple possible 3D poses for a fixed 2D pose, 3D HPE is a selection process. As a result, confidence of 3D HPE denotes the certainty of 3D pose selection process given fixed 2D joint locations or heat-maps. The definition of this confidence is related to the 3D HPE algorithm.

Our current definition is based on the loss (7), which minimizes the distance between the estimated 2D pose from images and the projected 2D pose from 3D pose, i.e.,

$$W_t^{3D}(j)=-\left\|\phi_t(j)\left[P_t(j)-R_tS_t(j)-T_t(j)\right]\right\|_F^2 \quad (9)$$

where $W_t^{3D}(j)\in R$ is the 3D confidence of j-th joint at time t. Clearly, the 3D confidence of missing joints is zero, which means no impact of missing joints on the later pose matching.

Temporal Confidence

The above two kinds of confidence mainly focus on spatial space. To model the consistency of joint locations over time, we employ a temporal confidence, which describes

Table 1. CMC performance of the pose-based CVPI when using different confidence metrics.

dataset	SEQ1				SEQ2				SYN			
CMC Rank	1	5	10	20	1	5	10	20	1	5	10	20
Pose	12.81	38.60	54.56	65.61	18.64	42.27	55.00	66.14	52.50	73.17	81.15	89.90
Pose+2D	25.26	**68.95**	78.77	**90.53**	**30.68**	62.27	73.86	87.27	63.94	**87.98**	**92.79**	95.58
Pose+3D	22.98	66.49	77.19	86.32	24.77	54.09	69.09	82.05	59.42	82.31	91.83	94.90
Pose+T	22.46	59.12	75.79	87.02	28.86	55.91	70.68	82.95	57.02	81.93	89.62	94.13
CPose	**30.00**	67.02	**79.30**	88.77	29.55	**65.68**	**76.36**	**87.95**	**63.65**	85.96	92.31	**95.58**

the smoothness of a joint location over time. As a result, the temporal confidence is defined as the distance between the joints' locations in adjacent frames

$$W_t^T(j) = -\left\| S_t(j) - S_{t-1}(j) \right\|_2 \tag{10}$$

where $W_t^T(j) \in R$ is the temporal confidence of j-th joint at time t, whose 3D location is $S_t(j)$. As in Eq. (10), a sudden change of a joint's location means low confidence, which conforms to the motion pattern of human.

To fuse the confidence in different aspects, the final confidence is defined as

$$W(j) = W^{2D}(j) \times e^{\alpha^{3D} W^{3D}(j)} \times e^{\alpha^T W^T(j)} \tag{11}$$

where $W(j) \in R$ is the final confidence of j-th joint, α^{3D}, $\alpha^T \in R$ are the coefficients for different confidence, which fit the final confidence value to the interval of [0, 1]. For convenience, we omit the subscript t of confidence here. As shown in Eq. (11), we use the same coefficients for different joints. In experiments, these two coefficients are selected by using validation dataset of SEQ1 – α^{3D} and α^T are set to 0.2 and 0.5 respectively for all the datasets.

Experiments

In this section, we will describe the datasets, evaluation metrics and the experimental results including quantitative comparison and qualitative examples.

Dataset and Evaluation Metrics

We evaluate the proposed method on three datasets: SEQ1, SEQ2 (Zheng et al. 2016) and SYN (Zheng et al. 2017), all of which are human-walking videos. These datasets are taken by two temporally synchronized GoPro cameras with different views. As a result, these dataset consist of video pairs and each pair of videos actually capture the same walking subject from different views. The length of each video is 120 frames. All subjects wear similar clothes. SEQ1 and SEQ2 contain 114 and 88 video pairs performed by 6 subjects. In some videos, portions of human body parts are invisible due to occlusions and camera angles. SYN contains 208 video pairs performed by 14 subjects. Compared to the first two datasets, SYN has less camera motion. Besides, all the human body parts in SYN are visible. For fair comparison, the frame resolution of all the videos is normalized to 64×128.

While the proposed method is unsupervised, several of the selected comparison methods are supervised. Therefore, we randomly split the dataset into two equal-size datasets for training and testing respectively, as in Zheng et al. (2016 and 2017). Like previous methods, we employ the Cumulative Matching Characteristics (CMC) ranks as our metric for CVPI evaluation. One camera's videos are probe set and videos from the other are used as gallery set. We calculate the distance for each pair and rank them.

Effects of Confidence

In this section, we investigate the influence of different confidence metrics on CVPI performance. The results are shown in Table 1. In this table, 'Pose' denotes the CVPI performance just by using 3D human pose without any confidence, 'Pose+2D', 'Pose+3D' and 'Pose+T' are the performance using 3D pose weighted by the confidence of 2D HPE, the confidence of 3D HPE, and temporal confidence respectively. 'CPose' is the performance of the proposed method by using fused confidence. Since some body parts are occluded in the video of SEQ1 and SEQ2, the accuracy is much lower than that of SYN. The low accuracy shows the difficulty of accurate 3D HPE for videos captured by wearable cameras. For all three datasets, adding any one of the three confidences can improve the pose-based CVPI performance substantially. This verifies the combination of the inaccurately estimated pose with a joint-based confidence can boost the CVPI performance. From this table, we also find 2D confidence can help improve the CVPI performance more than the other two kinds of confidence. In most cases, the proposed method by using fused confidence leads to the best performance. In the following, we will employ 'CPose' for comparsion.

Quantitative Comparison

In this section, we compare our method (denoted as CPose) with other state-of-the-art methods, including DVR (Wang et al. 2014), 3DHPE (Zheng et al. 2016), RFA (Yan et al. 2016 and Zheng et al. 2017) and Flow (Zheng et al. 2017). In these methods, 3DHPE and our proposed method are unsupervised since they do not use the identity information

Table 2: Comparison of the proposed method with state-of-the-art methods on SEQ 1, SEQ 2 and SYN dataset in terms of CMC rank.

dataset	SEQ1				SEQ2				SYN			
CMC Rank	1	5	10	20	1	5	10	20	1	5	10	20
DVR	16.14	50.53	66.84	82.83	11.14	34.09	53.64	77.05	12.69	41.83	59.04	75.87
3DHPE	16.14	50.70	67.02	81.93	17.95	51.82	71.14	89.55	8.65	35.67	50.48	64.52
RFA	68.42	96.84	98.25	99.30	69.77	96.36	98.41	99.32	56.83	92.40	97.02	98.85
Flow	79.82	92.28	95.26	97.54	76.36	87.05	92.73	96.82	72.21	90.00	94.90	98.08
Flow+RFA	87.02	97.37	97.89	98.95	82.05	94.39	96.59	99.32	82.12	98.37	99.33	100
CPose	30.00	67.02	79.30	88.77	29.55	65.68	76.36	87.95	63.96	85.96	92.31	95.58
CPose+RFA	84.56	97.19	98.24	**99.65**	77.05	**98.19**	**99.32**	**100**	84.51	98.72	99.60	100
CPose+Flow	85.97	93.86	97.37	98.07	80.45	91.36	95.23	98.64	84.14	95.69	98.44	99.71
CPose+RFA+Flow	**91.75**	**97.37**	**98.42**	99.12	**86.36**	96.13	98.63	99.77	**91.54**	**99.81**	**100**	**100**

Note: 3DHPE and the proposed CPose are unsupervised methods.

Table 3: CMC performance of SEQ1 using different α and β.

α / β	0.2	0.3	0.4	0.5	0.6	0.7	0.8
7	96.50	96.45	96.40	96.14	95.57	95.31	95.04
8	96.65	96.66	96.64	96.58	96.45	96.18	95.83
9	96.45	96.49	96.58	96.67	96.67	96.49	96.40
10	96.58	96.49	96.49	96.67	96.67	96.54	96.40
11	96.40	96.54	96.58	96.58	96.67	96.58	96.62
12	96.45	96.45	96.45	96.54	96.62	96.67	96.62
13	96.40	96.54	96.49	96.45	96.40	96.49	96.58

of person. Other methods need to learn the parameters using training dataset. The results are shown in Table 2. The proposed method achieves much higher performance than DVR and 3DHPE, which validates the effectiveness of the combination of human pose with confidence metric. Since our method is unsupervised, the performance is inferior to that of Flow and RFA. Even so, we still outperform RFA in SYN in term of CMC rank 1. The reason may be that SYN contains less occlusions.

In the bottom four rows of Table 2, we give the results for combined methods. Adding pose to Flow (or RFA) leads to better performance than the original Flow (or RFA). Combining pose, Flow and RFA achieves the highest performance in most cases, which verifies that the human poses, although inaccurately estimated, can still complement motion and appearance features for improving CVPI.

To show the influence of the value of α and β on the CVPI performance, we conduct experiments on SEQ1 using different α and β. The results are shown in Table 3. For saving space, we only show the average scores over CMC rank 1, 5, 10 and 20. From this table, we can see the CVPI performance is not very sensitive when the values of α and β vary in the range of [0.2, 0.8] and [7, 13] respectively.

Table 4. Cross-data performance in terms of CMC rank.

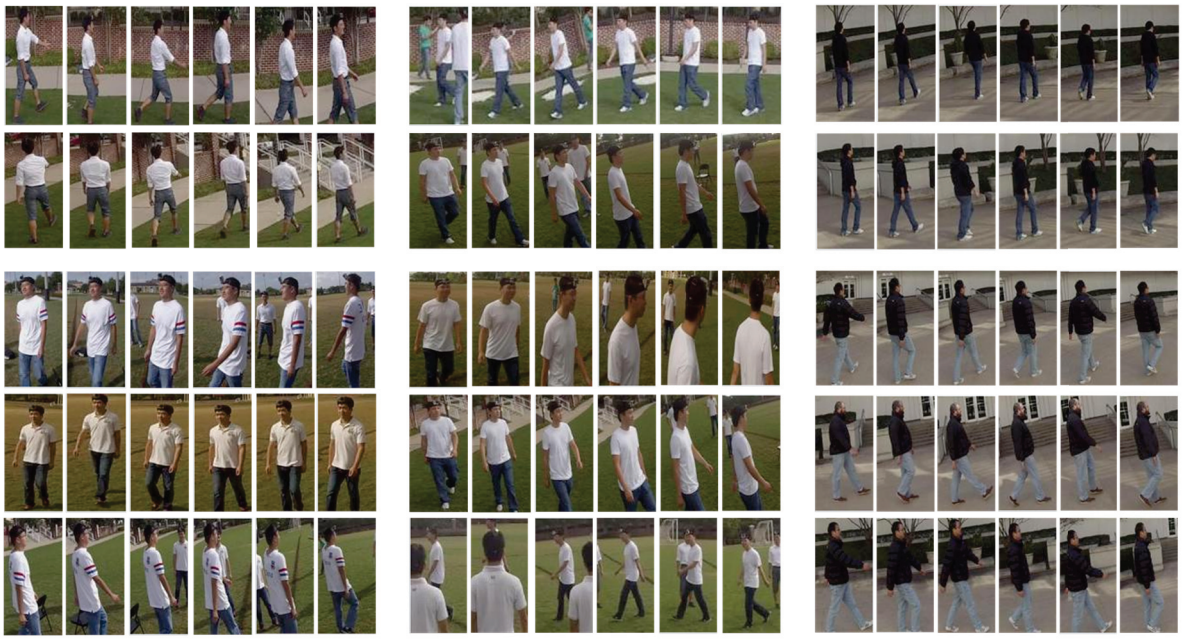
Rank	1	5	10	20
3DHPE	17.95	51.82	71.14	**89.55**
RFA	5.00	14.77	32.50	63.63
Flow	11.36	25.00	38.64	63.64
CPose	**29.55**	**65.68**	**76.36**	87.95

Cross-dataset Testing

Like Zheng et al. (2017), we also compare cross-dataset performance. For fair comparison, we perform this testing on SEQ2 using the parameters trained on SEQ1. The results are shown in Table 4. Note that for the two unsupervised methods, the CVPI performances keep unchanged. Our method achieves the best cross-dataset testing performance and the performance gain is over 12% in term of CMC rank 1. Besides, we can see that the two pose-based methods, 3DHPE and our proposed method, show much better cross-data testing performance than the appearance or optical-flow based methods. This shows that pose is a more robust cue with high generalization ability than optical or appearance in CVPI. Although SEQ1 and SEQ2 share similar background and subjects, the supervised methods still do not perform well on cross-dataset testing. In our opinion, this may be caused by over-fitting in training due to the large number of trained parameters and the small number of training samples.

Qualitative Results

In Fig. 3, we show the 3D human pose estimation results for a video sequence and we can see that some body parts of this sequence are not captured by cameras. Furthermore, we can see that the number of missing body parts is changing over time. The predicted 2D pose results are superposed on the images. As shown in the top row, the 2D locations for missing body joints are severely incorrect. The 3D pose obtained by the original algorithm is given in the second row. Due to the influence of false 2D locations of



(a) SEQ1 (b) SEQ2 (c) SYN

Figure 4: Sample matching results. (a), (b) and (c) are from SEQ1, SEQ2 and SYN dataset respectively. Top two rows are correct matching video pairs. The bottom three rows show failure cases. The third row is the probe input video. The fourth row is the returned matching video, which is incorrect, and the last row is the true matching video for the probe in the third row.

invisible joints, the 3D locations of visible joints are also wrong. In the last row, we give the estimated 3D pose with our adaptation. From Fig. 3, we can find our algorithm can return reasonable pose for visible human body parts at any time, which we use for pose-based CVPI.

Sample matching results are shown in Fig. 4. The three columns are samples from SEQ1, SEQ2 and SYN respectively. The top two rows are correct matching video pairs. Incorrect matching video pairs are shown in the bottom three rows, where the third row are probe inputs, the fourth row gives the returned false sequence and the true matching sequence is shown in the fifth row. The 3D human movements in the matched video pairs are completely consistent. In the correctly matched videos from SEQ1 and SEQ2, some body parts are missing. Nevertheless, our algorithm still returns the correct matching results. This shows that using some of body parts is sufficient for CVPI. For failure cases, the main reasons include the missing of too many key body parts and the overly large difference of the camera views. For example, the failed matching in SEQ1 in Fig. 4 may be caused by the totally opposite view angles of the truth matched video pairs, as indicated in rows 3 and 5 in Fig. 4(c). As a result, the visible parts in the probe are invisible in the true matching video, which leads to a false matching. Similarly, due to the camera-view difference, many key body parts are occluded in the probe video of SEQ2, which results in a false matching. The false matching video in SYN has very similar movement as the probe video.

Conclusion

In this paper, we developed a new metric of confidence to 3D human pose estimation (HPE), which measures the localization confidence of each joint, and used the estimated poses for cross-view person identification (CVPI), i.e., identifying the same person from temporally synchronized videos. Based on an existing 3D human pose estimation method, the confidence metric is defined in three aspects: 2D HPE confidence, 3D HPE confidence and temporal confidence. Then, we combined the inaccurately estimated human pose with the confidence metric for CVPI, by using confidence as weight in matching poses estimated from two videos. We found that the derived confidence can promote the pose-based CVPI. Finally, we integrated the estimated pose information into motion and appearance features and found that pose information well complements the motion and appearance features in CVPI and the integration of the pose, motion, and appearance features leads to new state-of-the-art CVPI performance.

There are two directions for the future work. First, compared to the current hand-crafted fusion of different confidences, supervised learning may be applied to produce a better fusion with further improved CVPI performance. Second, we can use CNN to learn the confidence metric, which can be combined with pose estimation through confidence weighted loss.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under grant 2016YFB1000903, NSFC No. 91748208, No. 61573268, No. 61672376, and NSF-1658987.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686-3693.

Amin, S.; Müller, P.; Bulling, A.; and Andriluka, M. 2014. Test-time adaptation for 3D human pose estimation. In *German Conference on Pattern Recognition*, 253-264. Springer International Publishing.

Chen, C. H. and Ramanan, D. 2016. 3D Human Pose Estimation = 2D Pose Estimation + Matching. In *arXiv preprint arxiv*:1612.06524.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. A Multi-Task Deep Network for Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 3988-3994.

Dutta, A.; Veldhuis, R.; and Spreeuwers, L. 2015. Predicting Face Recognition Performance Using Image Quality. In *arXiv preprint arxiv:*1510.07119.

Drevelle, V.; and Bonnifait, P. 2013. Localization confidence domains via set inversion on short-term trajectory. *IEEE Transactions on Robotics*, 29(5), 1244-1256.

Ghezelghieh, M. F.; Kasturi, R.; and Sarkar, S. 2016. Learning camera viewpoint using CNN to improve 3D body pose estimation. In *Fourth International Conference on 3D Vision*, 685-693.

Gupta, A.; Martinez, J.; Little, J. J.; and Woodham, R. J. 2014. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2601-2608.

Insafutdinov, E.; Andriluka, M.; Pishchulin, L.; Tang, S.; Levinkov, E.; Andres, B et al. 2017. ArtTrack: Articulated Multi-person Tracking in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6457-6465.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1325-1339.

Jammalamadaka, N.; Zisserman, A.; Eichner, M.; Ferrari, V.; and Jawahar, C. V. 2012. Has my algorithm succeeded? an evaluator for human pose estimators. In *European Conference on Computer Vision*. 114-128. Springer, Berlin, Heidelberg.

Li, S. and Chan, A. B. 2014. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, 332-347. Springer, Cham.

Li, W. and Wang, X. 2013. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3594-3601.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 483-499. Springer International Publishing.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 7025-7034.

Tekin, B.; Katircioglu, I.; Salzmann, M.; Lepetit, V.; and Fua, P. 2016. Structured prediction of 3D human pose with deep neural networks. In *arXiv preprint arxiv*:1605.05180.

Tekin, B.; Márquez-Neila, P.; Salzmann, M.; and Fua, P. 2017. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. In *arXiv preprint arxiv:*1611.05708.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *European Conference on Computer Vision* 688-703. Springer, Cham.

Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; and Zhang, L. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1288-1296

Wu, J.; Xue, T.; Lim, J. J.; Tian, Y.; Tenenbaum, J. B.; Torralba, A.; and Freeman, W. T. 2016. Single image 3d interpreter network. In *European Conference on Computer Vision*, 365-382. Springer International Publishing.

Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; and Yang, X. 2016. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, 701-716. Springer International Publishing.

Yang, Y.; Lei, Z.; Zhang, S.; Shi, H.; and Li, S. Z. 2016. Metric Embedded Discriminative Vocabulary Learning for High-Level Person Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3648-3654.

Yang, Y.; Wen, L.; Lyu, S.; and Li, S. Z. 2017. Unsupervised Learning of Multi-Level Descriptors for Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4306-4312.

Zhang, P.; Wang, J.; Farhadi, A.; Hebert, M.; and Parikh, D. 2014. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3566-3573.

Zheng, K.; Guo, H.; Fan, X.; Yu, H.; and Wang, S. 2016. Identifying Same Persons from Temporally Synchronized Videos Taken by Multiple Wearable Cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 105-113.

Zheng, K.; Fan, X.; Lin, Y.; Guo, H.; Yu, H.; Guo, D.; and Wang, S. 2017. Learning View-Invariant Features for Person Identification in Temporally Synchronized Videos Taken by Wearable Cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, 2858-2866.

Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K. G.; and Daniilidis, K. 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4966-4975.

Zhou, X.; Huang, Q.; Sun, X.; Xue, X., and Wei, Y. 2017. Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach. In *arXiv preprint arxiv*:1704.02447.