Identifying Same Persons from Temporally Synchronized Videos Taken by Multiple Wearable Cameras

Kang Zheng, Hao Guo, Xiaochuan Fan, Hongkai Yu, Song Wang
University of South Carolina, Columbia, SC 29208
{zheng37,hguo,fan23,yu55}@email.sc.edu, songwang@cec.sc.edu

Abstract

Video-based human action recognition benefits from multiple cameras which can provide temporally synchronized, multi-view videos. ***Cross-video person identification****, i.e., determining whether at a given time, persons tracked in different videos are the same person or not, is a key step to integrate such multi-view information for collaborative action recognition. For fixed cameras, this step is relatively easy since different cameras can be pre-calibrated. In this paper, we study cross-video person identification for wearable cameras, which are constantly moving with the wearers. Specifically, we take the tracked persons from different videos to be the same person if their 3D poses are the same, given that these videos are synchronized. We adapt an existing algorithm to estimate the tracked person's 3D poses in each 2D video using motion-based features. Experiments show that, although 3D pose estimation is not perfect, the proposed method can still lead to better cross-video person identification than using appearance information.*

1. Introduction

Video-based human action recognition has many important civil, military and security applications. Traditional fixed-camera videos can only cover pre-specified small areas from fixed view angles. Wearable cameras, such as Google Glass and GoPro, provide a new perspective to capture human actions in a much larger area since they are worn by and moving with the camera wearers. With multiple camera wearers, e.g., several camera-wearing police officers working together to process an incident, we may collect *multiple, temporally synchronized* videos from different views. In addition, the wearers' perception and experience may get them to move to right positions and use the best views to capture the human actions of interest. These videos may provide mutually complementary information and lead to collaborative human action recognition.

Before we make use of such multi-video information for collaborative action recognition, we need to first perform *cross-video person identification*, i.e., to determine whether persons detected and tracked in different videos are the same person or not at a given time. While this task can be accomplished by prior camera calibration for multiple fixed cameras, it is a much more challenging problem for multiple wearable cameras because they are constantly moving and their external parameters cannot be accurately estimated over time. Appearance matching is a natural approach that can be used for cross-video person identification. However, in many scenarios, especially in the scenario of multiple or crowded people, different persons may show similar appearances [13, 21]. In addition, the same person may show different appearances when viewed from different angles. In this paper, we propose a new approach for cross-video person identification by matching the persons' poses in 3D space.

The proposed new approach is based on a special characteristic of the problem: the input multiple videos are temporally synchronized. As a result, at any given time, the same person's poses in different views must be identical and their motion must be synchronized in the 3D space. In addition, it is very rare for two different people to show identical 3D poses and synchronized 3D motion over time. Based on this observation, we propose to estimate the 3D poses for each person detected and tracked in each video and then match the estimated poses across videos for person identification. The proposed method is an off-line method, whose input are synchronized videos. Specifically, in this paper we adapt an existing method [9] for 3D pose estimation, followed by a normalization into the same canonical coordinate for cross-video pose matching and person identification. To handle the view change of wearable cameras, we temporally divide each video of a tracked person into a set of video segments and perform 3D human pose estimation on each video segment separately. Over a short video segment, we assume that the camera motion is small and the view change is negligible. On each segment, we use the data-driven method in [9] to estimate the possible 3D poses of the tracked per-

(a) (b)

Figure 1. An illustration of the difference between the person re-identification problem and the proposed cross-video person identification problem. (a) Example pair of videos of the same person in the person re-identification dataset, where the videos are taken by two fixed cameras and not temporally synchronized. (b) Example pair of videos of the same person in the cross-video person identification dataset, where the videos are taken by two wearable cameras and temporally synchronized.

son and then link the pose estimation results over the entire video by using Viterbi algorithm to enforce temporal consistency of poses between neighboring segments.

The proposed cross-video person identification shows certain similarities to the problem of person re-identification [18, 19, 22], which aims at determining whether persons shown in different images or videos are the same person or not. However, different from the proposed cross-video person identification, person re-identification usually handles fixed-camera videos without temporally synchronization, as shown in Fig. 1. In practice, most of the existing person re-identification methods use the traditional appearance matching for identifying the same person from different images or videos.

To demonstrate the effectiveness of the proposed method, we collect synchronized videos using two GoPro cameras. Experiment results show that, although 3D human pose estimation used in the proposed algorithm is not highly accurate, it can still improve the cross-video person identification and outperforms existing person re-identification methods, which mainly use the appearance matching. The main contributions of this paper are: 1) We introduce the problem of cross-video person identification for wearable cameras as well as a new dataset, 2) we propose to use 3D human pose estimation to address cross-video person identification, and 3) we propose to use Viterbi algorithm to improve the 3D human pose estimation from wearable-camera videos.

The remainder of the paper is organized as follows. Section 2 introduces prior related works. Section 3 elaborates on the proposed method. Section 4 reports the experimental results, followed by a brief conclusion in Section 5.

2. Related Work

In this section, we briefly overview the prior works on person re-identification and human pose estimation, that are related to the proposed research on cross-video person identification.

2.1. Person Re-identification

Person re-identification aims to match persons in different images or videos from different cameras. Without the temporal synchronization, it usually uses appearance features for cross-image or cross-video person matching. Motivated by human vision system, such a person matching can be achieved by identifying and matching a small portion of salient regions. Zhao *et al.* [19] propose to learn human salience from dense SIFT and dense color histogram in an unsupervised way by extracting feature patches that are distinctive and reliable across different views. To account for the variations across subjects, attribute-centric and part-based feature representations are proposed to learn adaptive weighted features for each individual [14, 11, 12]. Spatial information of body parts are also considered in person re-identification. Farenzena *et al.* [8] model human appearance with Symmetry-Driven Accumulation of Local Features (SDALF) and exploit all the information from different body parts for person re-identification. Bak *et al.* [6] propose an appearance model based on spatial covariance regions extracted from human body parts. The consideration of spatial information, especially human body parts, has been proved to be useful. However, many appearance/color-based approaches suffer from the lack of temporal information in multi-shot person re-identification scenarios. Recently, spatial-temporal information has been used in person re-identification. Wang *et al.* [18] extract HOG3D to represent video fragments for person re-identification. However, the extracted features are still in the 2D space and sensitive to the change of view angle. As in many other computer vision applications, deep convolutional neural networks have also been used to refine the features for person re-identification [4, 10, 20].

2.2. 3D Human Pose Estimation

In this paper, we will estimate 3D human pose for matching persons across different videos. Pose estimation in monocular image/video is an ill-posed problem without considering human body models, because human pose has many degrees of freedom and camera has intrinsic and extrinsic parameters. Agarwal *et al.* [2, 3] describe a learning-based method for 3D pose estimation from a single image or a monocular video. It recovers pose by direct conducting nonlinear regression against shape-descriptor vectors extracted automatically from the image silhouettes. Andriluka *et al.* [5] propose to track human body parts by exploiting 2D human pose estimation, then recover 3D pose with the

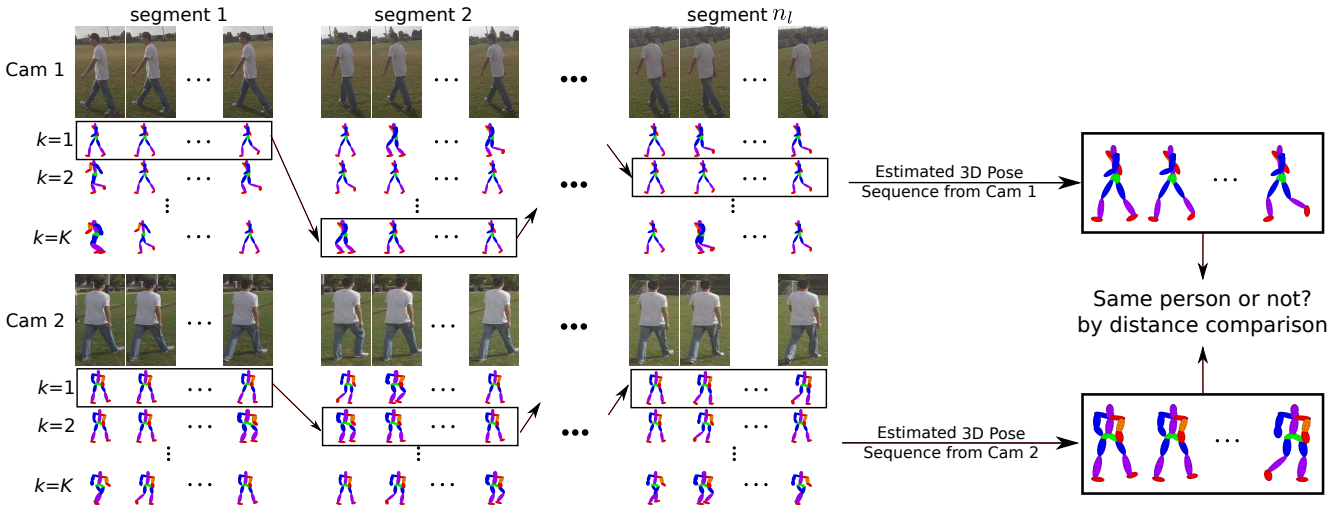

Figure 2. An illustration of the proposed framework. The input pair of videos are cropped such that only one person is contained in each of them. We estimate 3D poses for each video by adapting nCTE algorithm in [9]. For each video, we temporally divide it into segments. We perform nCTE for each segment to obtain top-K candidates of 3D pose sequences. Then Viterbi algorithm is used to enforce the temporal consistency between estimated 3D poses of neighboring segments. An optimal path is selected to obtain the 3D pose estimation for the whole video. Finally, we calculate the per-frame Euclidean distance between normalized 3D poses of the two videos. If the average distance is less than a pre-set threshold, then the videos are predicted to contain the same person.

help of 3D pose exemplars. Wang *et al.* [15] represent a 3D pose as a linear combination of learned 3D pose bases and use existing 2D pose estimation as the input. They recover 3D poses from a single image by minimizing the projection error between the 3D pose and the corresponding 2D pose detection. These methods are primarily used to estimate 3D human pose for a single image, in which temporal dynamics are not considered when they are applied to pose estimation through videos. Gupta *et al.* [9] propose a non-parametric model to recover 3D human pose by matching extracted dense trajectories with a database of synthetic trajectories [9]. Using the synthetic data, they successfully achieve the view invariance to the model. In this paper, we will adapt this method [9] and further apply the Viterbi algorithm to estimate the 3D human pose for each person tracked in a video.

3. Proposed Method

Cross-video person identification aims at identifying the same person from *temporally synchronized* videos taken by multiple wearable cameras from different, varying view angles. For convenience, we study a simplified case of two synchronized videos taken by two wearable cameras. The proposed method can be easily extended to the case of more than two videos by reducing it to person identification between each pair of videos and then combining the pairwise identification results. Furthermore, we assume that there is only one person and its bounding box well fills each frame of the video, as shown in Fig. 2. In the cases where the person's bounding box does not fill the frame well or each video contains multiple people, we can first apply person detection and tracking. We then crop each person out of each video in the spatial-temporal domain and use the proposed method to match each possible pair of persons cropped from the two videos, respectively.

The proposed method mainly consists of two parts: 3D human pose estimation on video and person identification via pose matching. The key step of the proposed method is to estimate the 3D human pose for *each frame* in each of the two input videos as described above. Given that two videos are temporally synchronized, we can compare the estimated 3D poses of a pair of videos at each corresponding frame to determine whether the persons in these two videos are the same person or not, as illustrated in Fig. 2. In this paper, we adapt an existing data-driven algorithm of non-linear Circulant Temporal Encoding (nCTE) [9] to estimate the 3D human poses from a video.

In [9], the input videos are assumed to be taken from fixed but unknown view angles. Motion features are extracted from each video. At the same time, a large number of 2D mocap sequences are synthesized from 3D mocap sequences provided in CMU mocap database [1], by projecting along a set of possible view angles. In mocap database, each 3D mocap sequence is represented by a sequence of 3D human joint locations over time. Similarly, a 2D mocap sequence is the corresponding projection of a 3D mocap sequence with a specific camera view angle. Motion features of 2D mocap sequences are also extracted. By aligning motion features of the input video and all 2D mocap sequences, we find the best matched 2D mocap sequence and then take

its underlying 3D human pose as the estimated 3D human pose for the input video.

However, one prerequisite of nCTE for 3D human pose estimation is that the person in the input video is viewed from a fixed angle when taking the video. If the camera is moving and the view angle changes over time, we could not simulate such view angle changes to synthesize the corresponding 2D mocap sequence and motion features for matching. In this paper, the input videos are taken by wearable cameras which are moving over time. Thus, the camera view angle is constantly changing. To address this problem, we propose a new strategy to adapt the nCTE, as illustrated in Fig. 2. Specifically, we first temporally divide each video into shorter video segments. By assuming that view angle is relatively stable in each shorter video segment, we can apply nCTE to estimate the 3D human pose in each segment. However, considering that a short video segment may not provide sufficient information for accurate 3D pose estimation, we take top-K ($K > 1$) best matched 2D mocap sequences and their underlying 3D poses as the estimation for each segment. This way, for each frame, we have K estimated poses. We then link the estimated 3D human poses across video segments by seeking an optimal solution over the whole video, using Viterbi algorithm. Viterbi algorithm is primarily used to enforce the temporal consistency of 3D poses over time. In the following sections, we elaborate on each component of the proposed 3D human pose estimation, including motion feature extraction, pose estimation for each video segment with nCTE, and optimal pose linking across video segments. Finally we will describe how we conduct cross-video person identification with estimated 3D poses.

3.1. Motion Features based on Dense Trajectories

Dense trajectories are very effective 2D motion features and they have been widely used for action recognition. Motion features based on dense trajectories also show certain level of robustness to camera motions. In this paper, we extract dense trajectories from an input video using a publicly available code in [16, 17]. Specifically, dense trajectories are extracted by tracking densely sampled points using optical flow fields. Given a trajectory of length L frames (here we use $L = 15$), it can be described by a sequence of displacement vectors $S = (\Delta P_t, \cdots, \Delta P_{t+L-1})$, where $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ and $P_t = (x_t, y_t)$ is the coordinate of a tracked feature point at frame t. Then the sequence is normalized by its magnitude as

$$S' = \frac{(\Delta P_t, \cdots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad . \tag{1}$$

Given a video described with dense trajectories, we use bag-of-words technique to compute per-frame motion descriptor for this video. For each frame of the video, we aggregate all the dense trajectories ending at this frame and quantize them into a frame descriptor with a learned dictionary. Then the per-frame motion descriptors are obtained for this video. The dictionary is learned from the dense trajectories of 2D mocap sequences, which will be discussed in the next section.

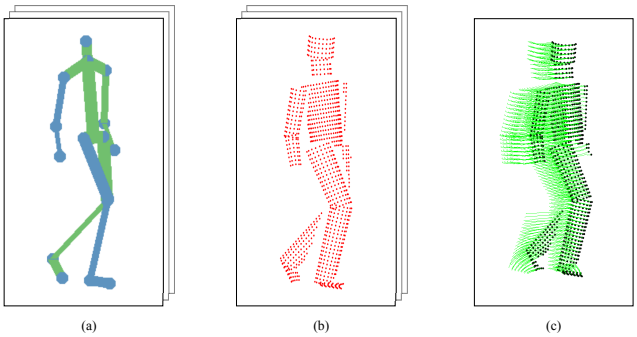

Figure 3. An illustration of projecting 3D mocap sequences to 2D sequence and computing dense trajectories from 2D mocap sequences. For each 3D mocap sequence (a), we approximate the 3D human body parts with cylinders. Points on the cylinder surface are sampled. We then project these points to 2D spaces using different camera view angles. (b) shows an example of the synthesized body surface points of the projected 2D mocap sequence. (c) shows the dense trajectories extracted from this 2D mocap sequence. They are extracted by tracking the body surface points and computing their inter-frame displacements over L consecutive frames. Normalization is applied to the displacement vectors afterwards as in Eq. (1).

3.2. Pose Estimation on Video Segments

As described earlier, to handle the view angle change in a wearable camera video, we divide each video into shorter video segments in the temporal domain. In this section, we follow the nCTE algorithm in [9] to estimate the 3D human pose for each video segment, by assuming that the camera view angle is relatively stable in each video segment. Specifically, we take a set of 3D pose sequences from mocap dataset, project each 3D pose sequence to 2D sequences and then compute trajectory-based motion features for each 2D sequence, as shown in Fig. 3. 3D human pose on a video segment can then be estimated by matching its motion features, as described in Section 3.1, against the constructed 2D mocap sequences.

In this paper, we project each 3D mocap sequence to 2D along 12 different directions: the azimuthal angle is selected from $\phi \in \{0, \pi/3, 2\pi/3, \pi, 4\pi/3, 5\pi/3\}$ and the polar/zenith angle is selected from $\theta \in \{\pi/12, \pi/6\}$. These 12 projection directions well cover the possible view angles to a person in practice. As described in [9], we use a cylinder to approximate each body part, such as a limb, the head, and the torso. Sampled points on the surface of such cylin-

ders are projected from 3D space to 2D space. To account for self occlusions, invisible surface points are removed by a hidden point removal operation. Finally, the dense trajectories of each 2D mocap sequence are produced by connecting the corresponding surface points over L consecutive frames and computing their inter-frame displacements.

To match the motion features between a video segment and a 2D mocap sequence, we also use the same bag-of-words technique to compute the per-frame motion descriptors for each 2D mocap sequence. The bag-of-words technique requires the construction of a dictionary to quantize motion features. In this paper, we construct such a dictionary with $d = 2,000$ words that are learned using K-means clustering over the dense trajectories extracted from all the projected 2D mocap sequences. This learned dictionary is used for both the considered video segment and the projected mocap sequences. This way, a 2D mocap sequence can be represented by a vector $\mathbf{z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n] \in \mathbb{R}^{d \times n}$, where n is the number of frames in the sequence and $\mathbf{z}_i$ is the descriptor for the 2D mocap sequence at frame i. Similarly, the considered video segment of m-frames can be represented by a vector $\mathbf{v} = [\mathbf{v}_1, \cdots, \mathbf{v}_m] \in \mathbb{R}^{d \times m}$, where $\mathbf{v}_i$ is the descriptor for the video at frame i. We then use nCTE to retrieve top-K best matched 2D mocap sequences for each video segment. The similarity between a 2D mocap sequence and a video segment is defined as:

$$s_\delta(\mathbf{z}, \mathbf{v}) = \sum_{t=-\infty}^{\infty} \langle \mathbf{z}_t, \mathbf{v}_{t-\delta} \rangle \quad , \tag{2}$$

where the vectors $\mathbf{z}_t$ (respectively, $\mathbf{v}_t$) are zero when $t < 1$ and $t > n$ (respectively, $t > m$). $s_\delta(\mathbf{z}, \mathbf{v})$ varies as δ varies and it reaches its maximum value when $\mathbf{z}$ and $\mathbf{v}$ are aligned. However, Eq. (2) assumes the dot product is a good similarity measure between $\mathbf{v}$ and $\mathbf{z}$, which is not the case for bag-of-words representations. Therefore, Gupta *et al* [9] extend this similarity by adding a kernel that transforms the data into the reproducing kernel Hilbert space of k:

$$\begin{aligned} s_\delta(\mathbf{z}, \mathbf{v}) &= \sum_{t=-\infty}^{\infty} k(\mathbf{z}_t, \mathbf{v}_{t-\delta}) \\ &= \sum_{t=-\infty}^{\infty} \langle \Phi(\mathbf{z}_t), \Phi(\mathbf{v}_{t-\delta}) \rangle \quad . \end{aligned} \tag{3}$$

In [9], a regularization term is further introduced. For efficient large-scale retrieval, the computation is accomplished using Fourier transform [9]. For each video segment, nCTE is performed over all the constructed 2D mocap sequences to obtain the top-K best matched 2D mocap sequences according to

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z} \in \mathbf{Z}} s_\delta(\mathbf{z}, \mathbf{v}) \quad , \tag{4}$$

where $\mathbf{Z}$ is the set of all the constructed 2D mocap sequences.

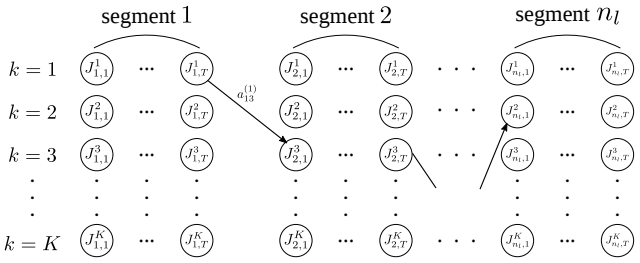

Figure 4. An illustration of using Viterbi algorithm for linking the estimated 3D poses across video segments for the final 3D pose estimation over the whole video.

More specifically, Eq. (4) is used to find the best matched 2D mocap sequence. We then remove this best matched 2D mocap sequence from Z and apply Eq. (4) again to find the second best matched 2D mocap sequence. To obtain top-K best matched ones, we repeat this process K times. Each 2D mocap sequence has an associated known 3D pose sequence. Therefore, we can get top-K candidates of 3D pose sequences for each video segment.

3.3. Pose Linking: From Video Segments to Whole Video

In this section, we will explain in detail how we link the estimated 3D poses across video segments to obtain the 3D poses through the whole video. Let's denote the considered whole video as V and we temporally divide it into n_l non-overlapping video segments $V = \{V_1, V_2, \cdots, V_{n_l}\}$. Each segment's length is T frames. As shown in Fig. 4, the detected top K candidates of 3D pose sequences for each V_i, are denoted by J_i^k where $1 \leq i \leq n_l$ and $1 \leq k \leq K$. J_i^k where $1 \leq i \leq n_l$ represents 3D poses of the k-th best matched 2D mocap sequence and J_i^k consists of a sequence of 3D poses $\{J_{i,1}^k, J_{i,2}^k, \cdots, J_{i,T}^k\}$. In our experiments, we choose $K = 10$ candidate 3D pose sequences for each segment. Then we apply Viterbi algorithm to find the optimal path $P = \{p_1, p_2, \cdots, p_{n_l}\}$ by selecting and linking the estimated 3D poses along all the video segments, where $p_i \in [1, K]$ is the selected candidate 3D pose sequence for i-th segment, as shown in Fig. 4.

The initial probabilities are assumed to be uniform for all candidate pose sequences. The emission probability of the person in V_i taking pose J_i^k is defined as the similarity between the motion features of the video segment and the k-th best matched 2D mocap sequence, as shown in Eq. (3). The transition probability is defined as:

$$a_{qr}^{(i)} = \frac{1}{d(J_i^q, J_{i+1}^r) + \epsilon} \quad , \tag{5}$$

where ϵ is a term to avoid dividing by zero, J_i^q and J_{i+1}^r are the q-th and r-th best matched 3D pose sequences for i-th and $\{i+1\}$-th segment respectively. We use Euclidean

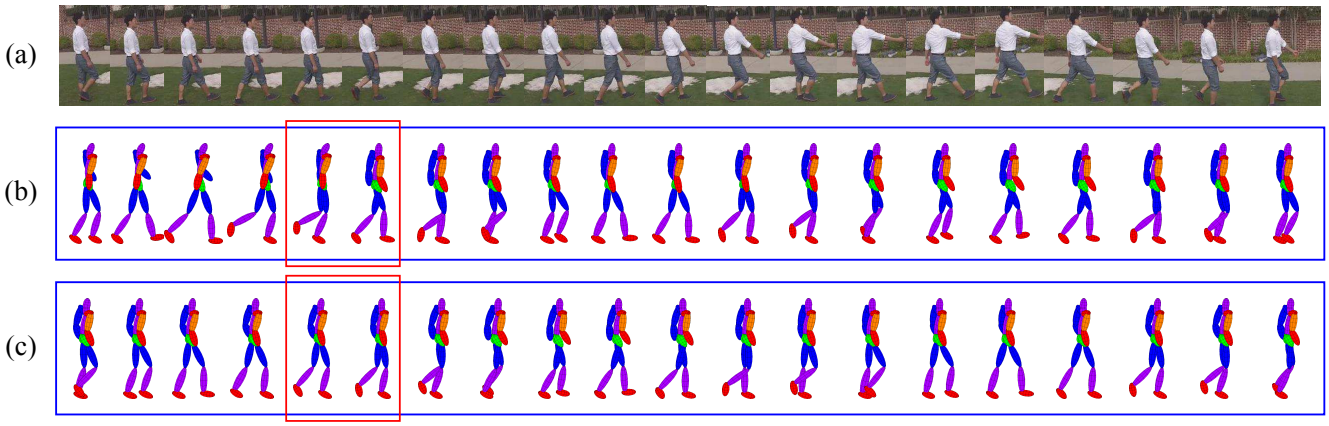

Figure 5. Human pose estimation comparison: (a) original image sequence. (b) estimated pose sequence with top-1 matched mocap sequence. (c) estimated pose with Viterbi algorithm applied to top-K matched mocap sequences.

distance to define $d(J_i^q, J_{i+1}^r) = \left\| J_{i,T}^q - J_{i+1,1}^r \right\|_2$, where $J_{i,T}^q$ and $J_{i+1,1}^r$ are the estimated 3D poses of last frame in i-th segment and first frame in $(i+1)$-th segment respectively.

Figure 5 shows a sample result of our pose estimation with and without the proposed Viterbi-algorithm based pose linking. In this figure, we only show frames 1, 4, 7, 10,$\cdots$ and the 3D poses estimated on these frames. We can clearly see that the poses of top-1 matched 3D pose sequence for each video segment (highlighted by red boxes) are not always correct and do not show good consistency between adjacent video segments. By including top K matched poses, as well as the proposed Viterbi-algorithm based pose linking, we can get more consistent 3D poses between neighboring video segments and the resulting 3D poses are more accurate estimations in terms of the underlying 3D poses. In the later experiments, we will report the quantitative results when using the proposed video division and pose linking for 3D pose estimation and cross-video person identification.

3.4. Cross-Video Person Matching with Estimated 3D Poses

With the estimated 3D poses for each video, we can achieve cross-video person identification by comparing 3D human poses between the two synchronized videos. To compare the coordinates of the joints in two 3D human poses, we normalize all the 3D human poses in the mocap sequences by following the steps suggested in [7]. This normalization step rescales the length of human limbs to the average length of all the subjects. We also normalize the zenith and azimuthal angles of the rigid part of human body, clavicles and hips, to be constants. After normalization, all 3D poses will be in the same canonical view. In addition, pose model used in mocap consists of 31 human joints which are overly detailed for the proposed task of person identification. In this paper, we only use 15 human joints, including head, neck, left/right shoulders, left/right elbows, left/right wrists, waist, left/right hips, left/right knees and left/right ankles, as shown in Fig. 6.

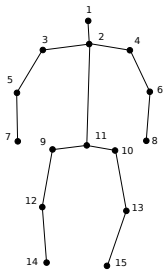

Figure 6. Human joints selected for cross-video person matching.

Given a synchronized video pair with estimated pose sequences, we can directly calculate the Euclidean distances between them after the above described normalization to the canonical view and a common scale. However, 3D human pose is a very challenging problem: even if we use Viterbi algorithm to search for an optimal pose linking across video segments, the estimated 3D pose sequence may still be inaccurate. To further improve the robustness, we repeat the Viterbi algorithm five times to get five optimal paths in Fig. 4, by removing the candidate poses along the obtained optimal path after each iteration. This way, we actually obtain five best pose estimations for each video. We then define the matching distance between a pair of synchronized video pair by searching for the best matched pose estimations between them, i.e.,

$$D = \min_{i,j} \left\| \mathcal{J}_1^i - \mathcal{J}_2^j \right\|_2 \quad , \tag{6}$$

where $\mathcal{J}_1^i$ and $\mathcal{J}_2^j$ are the i-th and j-th optimal pose sequences for the pair of videos respectively, with $i, j = 1, 2, \cdots, 5$. With the obtained distance between a synchronized video pair, we determine whether both videos depict the same person with a pre-set threshold: If their distance

D is smaller than the threshold, they are the same person; Otherwise, they are not the same person.

4. Experiments

In this paper, we collect our own video data for cross-video person identification since there is no publicly appropriate dataset that provides multiple, temporally synchronized videos by wearable cameras. In the data collection, each person is captured by two GoPro cameras worn by and moved with two viewers. The resulting videos are temporally synchronized using their pre-calibrated time-stamps. In total we use six different persons who always perform the action of "walking" for video dataset collection. Each collected synchronized video pair from the two GoPro cameras is of length 120 frames. In total, we collect 202 such temporally synchronized video pairs that capture a same person from different and varying video angles. In our experiments, we divide these 202 synchronized video pairs into two subsets, SEQ 1 and SEQ 2, which contains 114 videos pairs and 88 video pairs respectively.

Given N pairs of synchronized videos as described above, we take the N videos from one camera as the templates and the N videos from the other camera as the targets. For each target video, we match it to the N template videos one by one by calculating their matching distance as discussed in Section 3. We then select the top-R ranked template videos for this target video. If the synchronized video paired to this target video is among the top-R matched ones, we consider the matching for this target as correct. Repeating this for all the targets, we can calculate Cumulative Matching Characteristics (CMC), Precision, Recall and F-score to evaluate the performance of cross-video person identification. In our experiments, we compare the performance of the proposed method with Discriminative Video Ranking (DVR) [18], which uses the appearance matching and 2D motion features for person re-identification. We select DVR as the comparison method as it has shown to be superior over all other person re-identification approaches in multi-shot scenarios [18].

To estimate the 3D pose from a video, we use 60 mocap pose sequences from 14 different subjects in total. All of them are the sequences of the action "walking", but vary in speed, orientation, size and style. For our method, we divide each collected 120-frame video into 4 video segments for pose estimation. We set $K = 10$, i.e., take top-10 matched mocap sequences for each video segment as candidates. This process takes about 2 seconds when running on a 3.2 GHz computer using a single core. Then we use Viterbi algorithm to link the estimated 3D poses across the video segments. We select the top-5 optimally linked 3D pose sequences for each video and then use them for computing the matching distance D in Eq. (6) as discussed in Section 3.4.

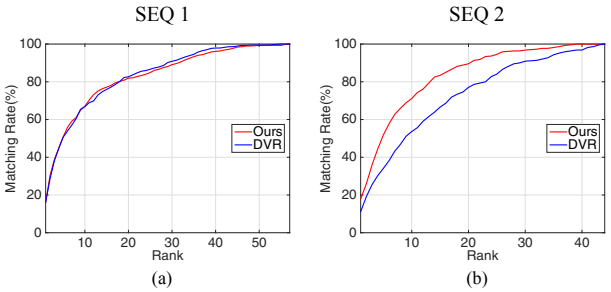

Figure 7. CMC performance of the proposed method and the DVR method.

Table 1. Rank (R) rates of the proposed method and the DVR method.

Dataset	Method	$R = 1$	$R = 5$	$R = 10$	$R = 20$	$R = 40$
SEQ 1	Ours	16.14	50.70	67.02	81.93	96.14
	DVR	16.14	50.53	66.84	82.63	97.89
SEQ 2	Ours	17.95	51.82	71.14	89.55	100.0
	DVR	11.14	34.09	53.64	77.05	96.82
Average	Ours	**17.05**	**51.26**	**69.08**	**85.74**	**98.07**
	DVR	13.64	42.31	60.24	79.84	97.36

The proposed method is unsupervised. But DVR, which we used for comparison, is a supervised method that requires training data to learn a model. Therefore, for DVR, we split the video dataset randomly into two subsets of equal size, one for training and the other one for testing. This process is conducted separately for SEQ 1 and SEQ 2. Pair videos are positive samples. If the template video and the target video are not a synchronized pair, then such a pair is a negative sample. To obtain more reliable results, we repeat the experiments 10 times, each of which uses different training and testing data, and report the average performance. Figure 7 and Table 1 show the CMC performance of the proposed method and the DVR method. We can see that the proposed method achieves comparable matching rates as the DVR method in SEQ 1 and much better matching rates in SEQ 2. This verifies that it is feasible to use 3D pose and pose change for person identification, even if the estimated 3D pose is not highly accurate. Figure 8 shows the relation between the rank-1 matching rate and the number of paired videos in the dataset for the proposed method and the DVR method. We can see that rank-1 matching rate decreases with the increase of the number of paired videos. This indicates that the problem of cross-video person identification becomes more challenging with the increase of involved subjects. Note that the rank-1 matching rate of the proposed method is always higher than the comparison method.

Precision, Recall and F-score of the proposed method are computed using varying thresholds to the matching dis-

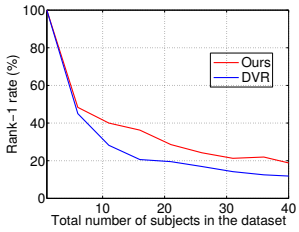

Figure 8. Rank-1 rate decreases as the total number of subjects increases in the dataset.

Table 2. Precision, Recall and F-score of the proposed method and the DVR method on the collected video dataset.

Dataset	Method	Precision	Recall	F-score
SEQ 1	Ours	29.45	23.51	24.98
	DVR	18.26	28.07	21.16
SEQ 2	Ours	13.33	35.23	17.93
	DVR	8.56	26.36	12.22
Average	Ours	**21.39**	**29.37**	**21.46**
	DVR	13.41	27.22	16.69

tance. For the DVR method, it uses a similarity metric for matching and therefore, a set of different thresholds are used for computing the Precision, Recall, and F-score. The average performance over 10 rounds of experiments are shown in Table 2, from which we can see that the proposed method performs better than the DVR method although we only use simple Euclidean distance to measure the similarity between the estimated 3D pose sequences of two videos.

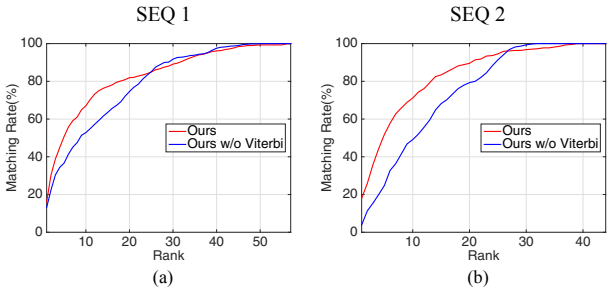

Figure 9. CMC curves of the proposed method with and without using Viterbi algorithm in 3D human pose estimation.

To show the effectiveness of the proposed strategy of using Viterbi algorithm for linking pose estimation over video segments, we perform a comparison study by removing the Viterbi-algorithm based pose linking step from the proposed method. More specifically, for pose estimation without using Viterbi-algorithm based pose linking, we simply choose the top-1 match in the mocap database as the estimated 3D poses for each video segment. Figure 5 shows samples of the 3D pose estimation results with and without the proposed Viterbi-algorithm based pose linking. We can see that the Viterbi-algorithm based pose linking can help the overall 3D pose estimation by enforcing the continuity of 3D poses across neighboring video segments. Quantitative results are shown in Fig. 9, we can see that the proposed Viterbi-algorithm based pose linking can substantially improve the performance of cross-video person identification. In our experiments, we also study the influence of the value of K, the number of selected candidate poses for Viterbi-algorithm based pose linking, to the final person identification performance. We tried $K = 5, 10, 15$. As shown in Fig. 10, the resulting CMC curves of different K are similar. Our method is not sensitive to the selection of the number of candidate poses.

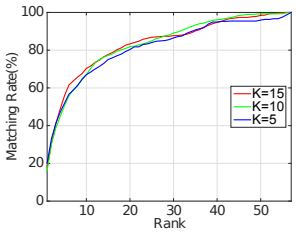

Figure 10. The influence of K over the CMC curves.

5. Conclusion

In this paper, we studied a new cross-video person identification problem for wearable cameras, i.e., determining whether tracked persons in temporally synchronized videos taken by multiple wearable cameras are the same person or not. Instead of using appearance matching, we proposed a new approach for person identification based on the 3D pose matching since it is rare for different people to show exactly identical and synchronized poses over a period of time. We adapted an existing data-driven algorithm to estimate 3D human poses from a video, by dividing the video into shorter video segments, estimating poses from each video segment, and finally using Viterbi algorithm to link the estimated poses across the video segments. We collected 202 synchronized video pairs using two GoPro wearable cameras for performance evaluation. Experiment results show that the proposed method achieves better performance, in terms of CMC rate, Precision, Recall, and F-measure, than a previous person re-identification method that is mainly based on appearance matching.

Acknowledgment This work was supported by NEH HK-50032-12 and AFOSR FA9550-11-1-0327.

References

[1] CMU Motion Capture Database. http://mocap.cs.cmu.edu. 3

[2] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *CVPR*, 2004. 2

[3] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *TPAMI*, 2006. 2

[4] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[5] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010. 2

[6] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010. 2

[7] X. Fan, K. Zheng, Y. Zhou, and S. Wang. Pose locality constrained representation for 3D human pose reconstruction. In *ECCV*, 2014. 6

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[9] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *CVPR*, 2014. 1, 3, 4, 5

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[11] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *ECCV Workshops and Demonstrations*, 2012. 2

[12] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012. 2

[13] Y. Lin, K. Abdelfatah, Y. Zhou, X. Fan, H. Yu, H. Qian, and S. Wang. Co-interest person detection from multiple wearable camera videos. In *ICCV*, 2015. 1

[14] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV Workshops and Demonstrations*, 2012. 2

[15] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3D human poses from a single image. In *CVPR*, 2014. 3

[16] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 4

[17] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 4

[18] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014. 2, 7

[19] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2

[20] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2

[21] K. Zheng, Y. Lin, Y. Zhou, D. Salvi, X. Fan, D. Guo, Z. Meng, and S. Wang. Video-based action detection using multiple wearable cameras. In *ECCV ChaLearn Workshop*, 2014. 1

[22] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013. 2