3D human motion recovery from a single video using dense spatio-temporal features with exemplar-based approach

Leong, Mei Chee; Lin, Feng; Lee, Yong Tsui

2019

Leong, M. C., Lin, F., & Lee, Y. T. (2019). 3D human motion recovery from a single video using dense spatio-temporal features with exemplar-based approach. 2019 4th International Conference on Image, Vision and Computing (ICIVC 2019).

https://hdl.handle.net/10356/90230

3D Human Motion Recovery From A Single Video Using Dense Spatio-Temporal Features With Exemplar-based Approach

M. C. Leong, F. Lin, and Y. T. Lee

***Abstract*—This study focuses on 3D human motion recovery from a sequence of video frames by using the exemplar-based approach. Conventionally, human pose tracking requires two stages: 1) estimating the 3D pose for a single frame, and 2) using the current estimated pose to predict the pose in the next frame. This usually involves generating a set of possible poses in the prediction state, then optimizing the mapping between the projection of the predicted poses and the 2D image in the subsequent frame. The computational complexity of this approach becomes significant when the search space dimensionality increases. In contrast, we propose a robust and efficient approach for direct motion estimation in video frames by extracting dense appearance and motion features in spatio-temporal space. We exploit three robust descriptors - Histograms of Oriented Gradients, Histograms of Optical Flow and Motion Boundary Histograms in the context of human pose tracking for 3D motion recovery. We conducted comparative analyses using individual descriptors as well as a weighted combination of them. We evaluated our approach using the HumanEva-I dataset and presented both quantitative comparisons and visual results to demonstrate the advantages of our approach. The output is a smooth motion that can be applied in motion retargeting.**

***Index Terms*—3D pose estimation, feature descriptors, human motion recovery, motion tracking.**

I. Introduction

Single camera motion recovery is an active research topic due to its wide range of applications in surveillance, entertainment, sports science and healthcare. Especially when it is combined with motion modeling [1, 2], the recognized action can be effectively reconstructed. There exist major challenges in recovering monocular human motion due to the high degrees of freedom in an articulated structure, occlusions, 3D projection ambiguity, arbitrary camera viewpoints, as well as variations in human shape and appearance [3]. The traditional process of human motion recovery can be divided into two stages: 1) feature extraction and pose estimation, followed by 2) pose tracking/ prediction in the next frame. This usually involves generating a set of possible poses in the prediction state, then optimizing the mapping between the projection of the predicted poses and the 2D image in the subsequent frame. The computational complexity of this approach becomes significant when the search space dimensionality increases, as in most machine learning systems [4, 5]. Besides, wrong prediction can easily occur when there is occlusion or projection ambiguity in the frame. The pose prediction error from a single frame will be propagated to subsequent frames, causing accumulated errors for the remaining sequence.

In contrast to the conventional approach, we propose an exemplar-based framework for direct motion estimation in video frames, to address three main challenges in pose estimation and motion tracking: 1) iterative prediction-updating process in a single frame pose prediction, 2) occlusion and appearance ambiguity in video frames, and 3) error propagation in motion tracking.

Our system exploits coherence between a stack of consecutive frames by extracting dense appearance and motion features in the spatio-temporal space. Instead of predicting a single 3D pose at a time, our system estimates a sequence of 3D poses for up to 12 frames, by finding the best matching frames in the training dataset utilizing dynamic time warping. This approach is exemplar-based as it learns the human motion representation from the training images and does not require supervised training nor an underlying human model. The system ensures smooth motion reconstruction and is useful for motion retargeting to animate 3D models in AR/ VR environment [6], action recognition and motion prediction.

II. Literature Review

A number of surveys exist [3, 7-13] that review the work of 2D and 3D pose estimation from a single image, as well as the extension to motion recovery in video. In general, the approaches for 3D pose estimation can be divided into two categories: model-based and model-free [7, 9]. Model-based approach is defined as having an a priori 3D human model, mostly represented by a skeleton structure with kinematic chain. Model-free approach may be example-based where the 3D pose is estimated by finding its nearest neighbors from a set of reference poses, or learning-based, which uses machine learning methods to find mapping between input images and the output pose.

Shakhnarovich et al. [14] proposed an example-based method for fast pose estimation. Poses in an image are represented by multi-scale edge direction histograms, and a

Manuscript received April 30, 2019. This work is supported by the Interdisciplinary Graduate School, Nanyang Technological University, Singapore. It is also partially supported by the grant 2017-T1-001-053, Ministry of Education, Singapore.

M. C. Leong is with Institute for Media Innovation, Interdisciplinary Graduate School, Nanyang Technological University, Singapore (e-mail: mleong006@e.ntu.edu.sg).

F. Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: asflin@ntu.edu.sg)

Y. T. Lee is with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore (e-mail: mytlee@ntu.edu.sg).

parameter-sensitive hashing algorithm is applied to search for the nearest neighbors in the dataset. The final pose is estimated using a robust locally-weighted regression method. Agarwal and Triggs [15] described a learning-based approach to estimate 3D poses from training examples. Image observations are represented by a histogram of shape context descriptors, while body poses are parameterized by vectors of joint angles. Mori and Malik [16] recovered 3D body configurations using shape context extracted from sampled edge points. They implemented a hybrid of the example-based and model-based methods for 2D joint localization, followed by 3D pose estimation using the algorithm from [17]. Poppe [18] evaluated example-based pose estimation approaches where histograms of oriented gradients are used as the image descriptor. The approach is tested on the HumanEva dataset [19] and the results show that the estimation process is action and person-specific, where the accuracy decreases when variations in human appearance and unseen poses occur.

Jain et al. [20] are the first who used deep learning for feature learning in human pose estimation. They utilized a multi-layer convolutional network framework to learn image features and spatial priors between parts. Tompson et al. [21] combined a convolutional network (CNN) and graphical model in a unified model. The CNN is trained as a part detector, and its output is a heat map that shows possibilities of joint locations. To improve on the joint prediction, spatial relationships between joints are trained using a graphical model. Toshev and Szegedy [22] presented a cascaded Deep Neural Network to regress body joint location from images. The cascaded regressors managed to refine the final joint location based on previous estimation to achieve higher accuracy. Pfister et al. [23] considered a deep network using heat map regressors for joint location, spatial relation between joints, and temporal alignment with neighboring frames. The combined heat maps are pooled to predict the final joint locations. Newell et al. [24] proposed a novel stacked hourglass network for joint top-down and bottom-up processing. A single hourglass network is designed to process spatial information at different resolutions and output joint prediction in heat maps. Mehta et al. [25] demonstrated a real-time 3D human pose reconstruction using CNN regressor for joint estimation, combined with temporal tracking and 3D skeleton fitting. The results is claimed to be comparable to reconstructed pose from RGBD images. Human pose estimation using deep learning require large-scale training datasets to achieve reliable performance. As the HumanEva dataset has limited samples, experiments were not conducted for results comparison.

One of the most widely used tracking approach in 3D human pose estimation is particle filter [26]. However, it suffers from low prediction accuracy when sudden change of motion occurs [27]. To improve prediction accuracy, Liu et al. [27] worked on 3D human motion tracking by introducing an exemplar-based conditional particle filter method, which estimates a motion transition function from the previous pose to the current matched exemplar. Zhou and Li [28] investigated both the spatial and temporal information on human silhouette sequences to recover corrupted or occluded images. They proposed to learn the shape dictionary of a highly articulated object by representing each image as a linear combination of local shape features.

Commonly used spatio-temporal descriptors for video includes 3D volumes of Histograms of Oriented Gradients (HOG) [29], Histograms of Optical Flow (HOF) [30] and Motion Boundary Histograms (MBH) [30]. Uijlings et al. [31] evaluated the performances of these descriptors with various quantization methods in the context of video classification and proposed an efficient implementation of dense descriptors by dividing video volume into 3D blocks. Wang et al [32] proposed a dense representation based on trajectories and extracted features HOG, HOF and MBH that aligned with the trajectories as descriptors. Computed MBH along trajectories is used to capture and remove camera motion in dynamic video contents. Following the work of [32], Kantorov and Laptev [33] developed efficient video descriptors using sparse motion vectors to improve the speed of feature extraction. Local descriptors HOG, HOF and MBH are extracted for a sparse set of points positioned on the motion vectors. It can be seen that these descriptors are widely used in detection and video classification as they are robust and provide promising results. Our work exploits these descriptors and applies them to track and estimate human pose in a sequence of images.

Tekin et al [34] exploits motion information in a video sequence to reconstruct 3D human pose by training a regression model. Their work starts with aligning the subject’s motion in consecutive frames before extracting multi-scale 3D HOG features over the volume. For motion alignment, they trained CNNs to estimate the shift of subject’s position and iteratively refine it such that the subject is centered. The regression model is trained by finding the mapping function between the spatio-temporal features to the 3D poses. Given a video volume of 24 to 48 frames, the 3D pose in the central frame can be predicted using the trained model. Their work is implemented on several benchmark datasets and shows improvement over the state-of-the-art.

In contrast with previous work that utilized CNNs for iterative motion alignment, we implemented subject alignment by pre-processing background subtraction, followed by finding maximum overlapping area between two consecutive foreground subjects. Motion estimation with regression modeling [34] requires a stack of 24 to 48 frames as input to predict a single 3D pose. Conversely, our work focuses on direct motion estimation for up to 12 consecutive frames by utilizing exemplar-based approach. In addition, we ensure smooth reconstructed motion by interpolating the best matching exemplars in each frame. We evaluated our approach using the HumanEVA dataset and presented both quantitative comparisons and visual results to demonstrate the advantages of our approach over model-based and probabilistic methods.

III. Methodology

In view of the limitations of HOG in the spatio-temporal domain, we propose to exploit motion information by computing the optical flow of the moving subject. Unlike the work of [34] that computes only for the central pose given a sequence of images, we aim to recover the full motion in the sequence. To accomplish this task, we utilize the example-based method to store sub-video volumes (also called video patches), represented by a set of descriptors, with their corresponding 3D poses as exemplars. A video patch may contain a sequence of the original frames, or cropped frames where a full body pose is centered and aligned. The motion in a video patch can be estimated by finding the best matching exemplars in the training dataset. To ensure smooth motion between video patches, interpolation is performed to generate in-between poses. The overall work flow is depicted in Fig. 1. Detailed descriptions are presented in the following sections.

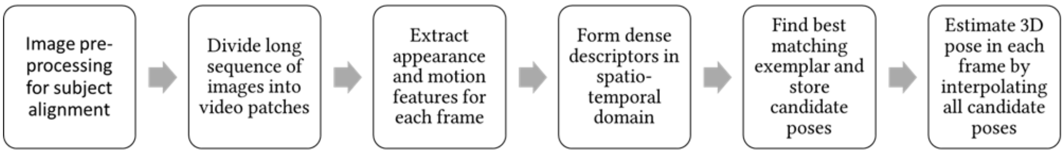
Image pre-processing for subject alignment
Divide long sequence of images into video patches
Extract appearance and motion features for each frame
Form dense descriptors in spatio-temporal domain
Find best matching exemplar and store candidate poses
Estimate 3D pose in each frame by interpolating all candidate poses

Fig. 1. Overview of the proposed 3D motion recovery in video volume using densely extracted spatio-temporal features.

A. *Subject Alignment*

We address the performance of feature descriptors on both the original frames (where the subject is moving around) and pre-processed frames where the subject is aligned and centered. Before aligning the subject in each frame, we first remove the background and then find the bounding box of the subject's silhouette. A Gaussian Mixture Model is employed to learn the probability distribution of the background pixels [19, 35]. Each Gaussian model is represented by the mean and variance of the color pixels (*r*, *g*, *b* values). A pixel that does not match with the Gaussian distribution is then labelled as foreground. However, this method cannot reliably remove shadow pixels from the foreground mask due to the changes in color intensities when shadows are cast. From the observation of [36], a shadow pixel can be taken to have significant intensity changes but without much changes in its chromaticity. The chromaticity can be computed as proportions of *r*, *g*, *b* colors in the image:

$$rc = r / (r + g + b);$$
$$gc = g / (r + g + b);$$
$$bc = b / (r + g + b) \qquad (1)$$

In computing the Gaussian Mixture Model, we now learn the mean and variance of background distribution using the chromaticity values *rc*, *gc* and *bc*. An example of a background image with mean chromaticity and the resulting foreground mask is shown in Fig. 2. One limitation of using chromaticity is that it could not differentiate the lightness of pixels – white and gray both have the same chromaticity with no color information [37]. This explains the mislabelling of the white and gray background as foreground pixels in Fig. 2(e). Pixels that are labelled as foreground in the RGB model but appear as background in the chromaticity model are treated as shadows and are removed from the foreground mask (Fig. 2(f)).

After obtaining the subject's silhouette from background subtraction, we can find the bounding box of the subject. We compare the bounding box with its previous frame to find a shifted position (in the *x* and *y* directions) that gives the maximum overlapping area between two silhouettes. The subject is aligned and centered across frames before cropping the image from the original dimension of 640×480 to 240×480. As the subject's height changes across the frames (subject walking in front and away from the camera position), we also scale the images to the same height before aligning and centering them in the frames with dimension 400×400.

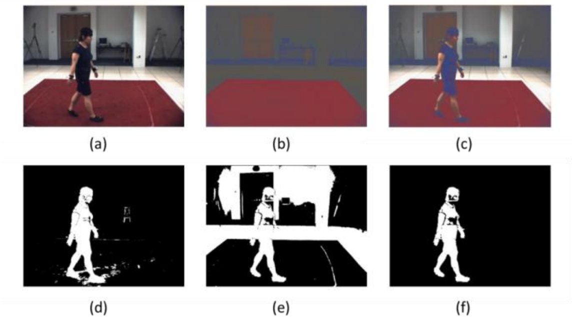

Fig. 2. (a) Original RGB image. (b) Background image with mean chromaticity. (c) Original image represented by its chromaticity. (d) Foreground mask obtained from RGB background model with cast shadows. (e) Foreground mask obtained from chromaticity background model (f) Final silhouette after background subtraction, shadow removal and noise removal

B. *Forming Video Patches and Spatio-Temporal Blocks*

We extract descriptors HOG, HOF and MBH from local video blocks and concatenate them to form the descriptor for a full video volume. At first, the video volume (Fig. 3(a)) is split into sub-volumes, which we call video patches. A video patch (Fig. 3(b)) contains frames from time *t* to *t* + *n*, where *n* denotes the number of frames to encode the motion features (we use *n* = 6 or 12). In our computation, we set the default interval between two patches as one frame. If a video volume has *N* frames and the interval between patches is *I*, the number of patches is (*N* – (*n*+1))/*I* + 1. Each frame is divided into 2D areas of pixels. By concatenating the areas in consecutive frames, we build spatio-temporal blocks for features encoding. Fig. 3 illustrates the terms for video volume, video patch and image areas.

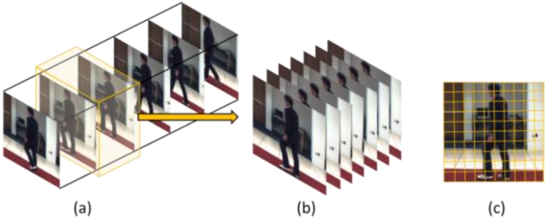

Fig. 3. A video volume that contains a sequence of image frames can be split into sub-volumes. (b) A video patch from time *t* to *t* + *n* frames. (c) Each frame is divided into areas for feature encoding.

C. *Feature Descriptors*

HOG is an edge and gradient based descriptor. For each pixel, the magnitude and orientation of the image gradient are computed and later quantized into a local histogram with eight orientation bins. HOG is used to capture image appearances, while HOF and MBH compute the optical flow between frames to encode motion information. The optical flow of a pixel is its displacement between two frames. The displacement magnitude and orientation are computed and binned into a histogram for each spatial block, to form the motion descriptor HOF. Another motion feature, MBH, treats optical flow in the vertical and horizontal direction separately, forming two separate histograms from the independent local gradients. MBH is developed to detect relative movement of the subject, suppressing constant motion from the camera and background [30].

After extracting the features for each frame, the responses are then accumulated and aggregated in the spatial and temporal domain based on a block's size. For a descriptor with 3×3×2 blocks, the responses in each block is concatenated to form the descriptor of dimension 3×3×2×8. Fig. 4 depicts the flow of features extraction (HOG, HOF and MBH), encoded features in local histogram for a spatio-temporal block, and finally forming a descriptor with size 3×3×2 or 3×3×1. The number of frames in a single video patch may be 7 (for descriptor of size 3×3×1) or 13 (for descriptor of size 3×3×2).

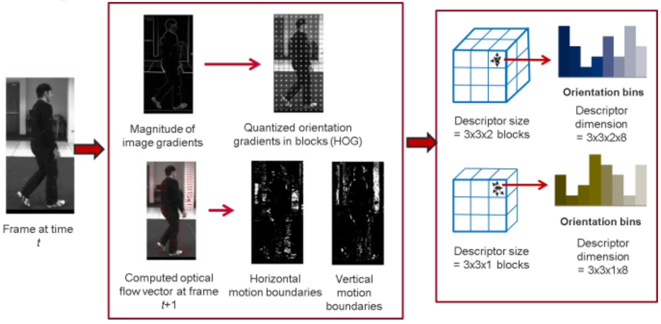
Frame at time
t
Magnitude of image gradients
Quantized orientation gradients in blocks (HOG)
Computed optical flow vector at frame *t*+1
Horizontal motion boundaries
Vertical motion boundaries
Orientation bins
Descriptor size = 3x3x2 blocks
Descriptor dimension = 3x3x2x8
Orientation bins
Descriptor size = 3x3x1 blocks
Descriptor dimension = 3x3x1x8

Fig. 4. Illustration of the process flow to compute HOG, optical flow and MBH descriptors.

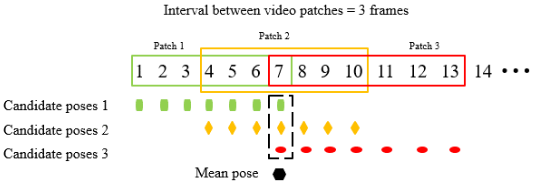

Fig. 5. An illustration of estimating the 3D pose for a single frame. From the test video patches, 3D poses from the best matching exemplar are stored as candidate poses for each frame. The final estimated pose is obtained by computing the mean pose from all candidate poses

D. *Feature Descriptors*

After computing the descriptor for spatio-temporal blocks, a video patch is represented by the concatenation and normalization of all these descriptors. Each video patch in the training dataset is stored together with their corresponding 3D poses. To estimate the motion of a test video patch from time *t* to *t* + *n* frames, we find the best matching exemplar in the dataset by comparing the absolute difference between descriptors of video patches. The test video patch and the exemplar may contain the same motion but vary in speed. To resolve this problem, we implemented dynamic time warping (DTW) [38], which uses similarity measures to find alignment of frames between two temporal sequences. Distances between frames are computed by re-using the HOG features. Frames with the shortest distance are matched, and the exemplar 3D poses are retrieved and stored as candidate poses in the test video patch. After obtaining all the candidate poses in the test sequence, the final pose is estimated by averaging all the candidate poses. Fig. 5 presents an example in which a video patch has *t* + 6 frames and the interval between patches is three frames. Therefore, a single frame in the video volume will be overlapped by three video patches, resulting in three candidate poses for interpolation.

IV. Experimental Setup and Results

A. *Implementation*

Our framework is implemented using MATLAB, built on the code released by Uijilings, et. al. [31] that extracts dense HOG, HOF and MBH in video blocks. We evaluated the performance of individual descriptors as well as a weighted combination of different features on HumanEva dataset [19]. The HumanEva dataset is commonly used for evaluating human pose estimation as they provide training and validation datasets of video sequences with synchronized 3D body performing different actions. We utilize the HumanEva-I dataset of walking motion captured from 3 different subjects (S1, S2 and S3) at camera C1.

B. *Experiment on Subject's Alignment and Weighted Descriptors*

In this experiment, we compare the performance of our method on 3 different input images: 1) original image, 2) subject is cropped and centered, and 3) subject is scaled to the same height and center aligned. The block size used is 12×12 pixels with 6 frames, and the descriptor contains 3×3×2 of these blocks. The video is divided into patches of 13 frames and their interval is set at one frame. For the first set of data, the original image size is 640×480, which gives a concatenated descriptor size of 1938×144 in a single video patch. For the second set of data, the size of the cropped image is 240×480, with a descriptor of dimension 684×144 in a video patch. As the third set of data involves scaling of the subject in the scene, this causes significant changes in the background. In order to focus on the human motion, we remove the dynamically changing background before feature extraction. To obtain a smooth foreground mask for background subtraction, the final silhouette (Fig.2 (f)) with internal holes are filled before applying a 2D Gaussian filter. The size of the scaled image is 400×400, and the size of a video patch descriptor is 961×144.

We evaluated the performance of pose estimation using individual descriptors – HOG, HOF, MBH in the vertical direction (MBH*y*) and MBH in the horizontal direction (MBH*x*), as well as their weighted combination.

Fig. 6 displays the results for the validation dataset of Subject 2 with walking action. The error metric used is the mean angular error between the joint angles in the reconstructed pose and the ground truth pose. The results show that images with scaled and aligned subject and background removed give the best outcome, followed by images with aligned subject (without scaling) and then the original image. Generally, HOG gives very high error rate as compared to the other descriptors, except when the background is removed. HOF gives high accuracy, followed by the combination of HOF and MBH. From this experiment, we observed that aligning the subject across frames is important such that the motion of body parts between frames can be captured accurately. Moving background in the video causes noise motion features that affect the performance of feature matching to find the best examplars. Fig. 7 illustrates the tracking results for Subject S2's torso and left hip angle in the test sequence.

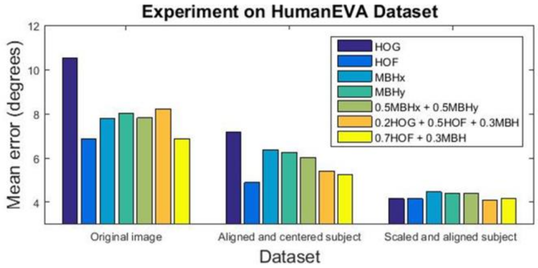

Fig. 6. Comparison of results on HumanEva dataset for Subject 2 walking action.

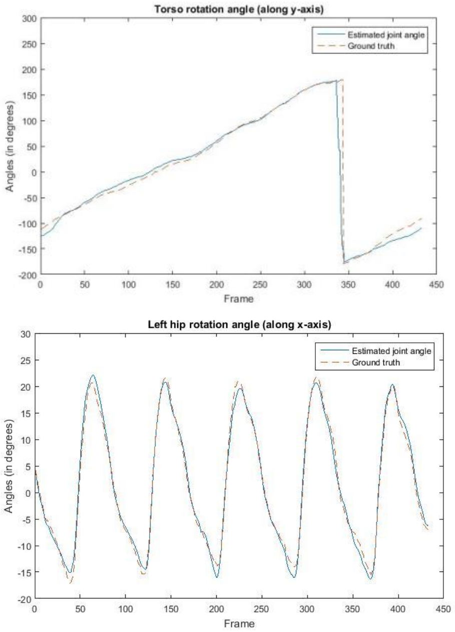

Fig. 7. Tracking results for Subject S2 in walking action in scaled and aligned dataset, using a weighted descriptors of 0.2*HOG + 0.5*HOF + 0.3*MBH. (Top) Torso rotation angle (Bottom) Left hip rotation angle.

C. Comparison with State-of-the-art

Comparison of 3D quantitative results is presented in Table 1. The error metric used is the mean absolute difference between 3D joints position. We utilize the limb's parameters and torso's 3D position provided in the dataset to generate 14-joints position – left and right shoulders, elbows, wrists, hips, knees, ankles, neck and pelvis. Result shown in Table 1 is not a direct comparison as different methods utilize different validation sequence and 3D model representations – 12-joints, 14-joints, 15-joints or 20-joints. Our work is compared with other pose estimation and tracking approaches that employed optimization on a set of sparse 3D poses [39], inference in a graphical model [40], a Bayesian framework for joint 2D and 3D models [41], probabilistic latent variable model [42], example-based method [18], CNN for 2D joint estimation and 3D pose update with geometric constraints [43], regression from spatio-temporal volume [34] and structured prediction with twin Gaussian processes [44]. Experimental results demonstrate that our method outperforms existing motion tracking methods, while supervised learning with CNN or regression model perform better than our exemplar-based approach. Examples of reconstructed 3D poses are shown in Fig. 8, where scaled and aligned subjects are used as input.

TABLE I: Quantitative results comparison by computing the mean 3D absolute error between estimated and ground truth joints position (in mm). Our method shows the average error obtained from the individual and combined descriptors of HOG, HOF and MBH

Authors	**Walking action**			
	S1	***S2***	***S3***	***Avg***
Wang et al. [39]	71.9	75.7	85.3	77.6
Sigal et al. [40]	66.0	69.0	-	67.5
Simo-Serra et al. [41]	65.1	48.6	73.5	62.4
Taylor et al. [42]	54.3	69.3	43.4	55.7
Poppe [18]	41.3	39.6	55.3	45.4
Our method	**44.5**	**27.0**	**54.9**	**42.1**
Zhou et al. [43]	34.3	31.6	49.3	38.4
Tekin et al. [34]	37.5	25.1	49.2	37.3
Bo et al. [44]	38.2	32.8	40.2	37.1

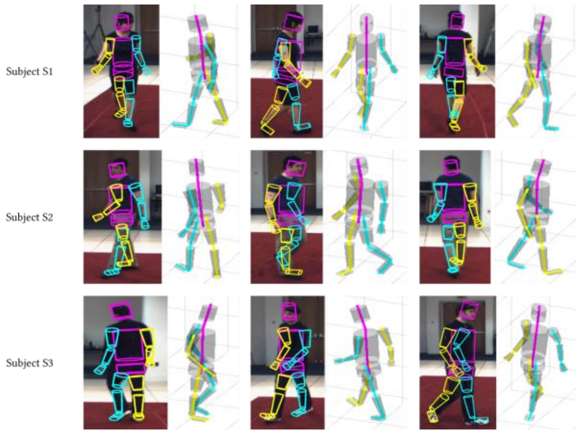

Fig. 8. Examples of reconstructed 3D pose for subject S1, S2 and S3 in walking motion, viewing from a different view point.

V. Conclusion

This paper presents a robust and efficient method in reconstructing human pose and motion in video frames by utilizing both appearance and motion features. Our result outperforms existing pose tracking and estimation methods and is comparable to the state-of-the-art approaches. The implementation is straightforward, and it ensures smooth motion across frames. Our method can be generalized to other actions besides walking motion and can be further extended to generate sequences of combined actions by merging and smoothing the motion from different video patches. This work is especially useful in animating 3D characters in short sequences where users may look for the desired motion in a 2D video and reconstruct its 3D motion to apply to the characters. For future work, we could detect and track body parts' trajectory for motion compensation instead of performing background subtraction. We could also reduce descriptor size from dense block size to sparse part-based descriptor.

References

[1] Cai, J., F. Lin, and H.S. Seah, *Graphical Simulation of Deformable Models*. Springer International Publishing Switzerland, 2016. ISBN 978-3-319-51030-9.

[2] Cai, J., et al., Modeling and dynamics simulation for deformable objects of orthotropic materials. The Visual Computer, 2017. 33(10): p. 1307-1318.

[3] Perez-Sala, X., et al., A survey on model based approaches for 2D and 3D visual human pose recovery. Sensors (Basel), 2014. 14(3): p. 4189-210.

[4] Stepanova, M., F. Lin, and V.C.-L. Lin, *A hopfield neural classifier and its FPGA implementation for identification of symmetrically structured DNA motifs.* The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 2007. 48(3): p. 239-254.

[5] Yu, J., et al., *Image classification by multimodal subspace learning.* Pattern Recognition Letters, 2012. 33(9): p. 1196-1204.

[6] Lin, F., et al., *Voxelization and fabrication of freeform models.* Virtual and Physical Prototyping, 2007. 2(2): p. 65-73.

[7] Moeslund, T.B., A. Hilton, and V. Krüger, *A survey of advances in vision-based human motion capture and analysis.* Computer vision and image understanding, 2006. 104(2): p. 90-126.

[8] Hen, Y.W. and R. Paramesran. Single camera 3d human pose estimation: A review of current techniques. in Technical Postgraduates (TECHPOS), 2009 International Conference for. 2009. IEEE.

[9] Ji, X. and H. Liu, *Advances in view-invariant human motion analysis: A review.* Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2010. 40(1): p. 13-24.

[10] Aggarwal, J.K. and Q. Cai, *Human motion analysis: A review.* Computer vision and image understanding, 1999. 73(3): p. 428-440.

[11] Wang, L., W. Hu, and T. Tan, *Recent developments in human motion analysis.* Pattern recognition, 2003. 36(3): p. 585-601.

[12] Weinland, D., R. Ronfard, and E. Boyer, *A survey of vision-based methods for action representation, segmentation and recognition.* Computer Vision and Image Understanding, 2011. 115(2): p. 224-241.

[13] Zhou, F. and F. De la Torre, Spatio-temporal Matching for Human Pose Estimation in Video. 2016.

[14] Shakhnarovich, G., P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. 2003. IEEE.

[15] Agarwal, A. and B. Triggs, *Recovering 3D human pose from monocular images.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006. 28(1): p. 44-58.

[16] Mori, G. and J. Malik, *Recovering 3d human body configurations using shape contexts.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006. 28(7): p. 1052-1062.

[17] Taylor, C.J. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. in Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. 2000. IEEE.

[18] Poppe, R., Evaluating example-based pose estimation: Experiments on the humaneva sets. 2007.

[19] Sigal, L. and M.J. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Brown Univertsity TR, 2006. 120.

[20] Jain, A., et al., Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302, 2013.

[21] Tompson, J.J., et al. Joint training of a convolutional network and a graphical model for human pose estimation. in Advances in neural information processing systems. 2014.

[22] Toshev, A. and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[23] Pfister, T., J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. in Proceedings of the IEEE International Conference on Computer Vision. 2015.

[24] Newell, A., K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. in European Conference on Computer Vision. 2016. Springer.

[25] Mehta, D., et al., VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. arXiv preprint arXiv:1705.01583, 2017.

[26] Isard, M. and A. Blake, *Condensation—conditional density propagation for visual tracking.* International journal of computer vision, 1998. 29(1): p. 5-28.

[27] Liu, J., et al., 3D Human motion tracking by exemplar-based conditional particle filter. Signal Processing, 2015. 110: p. 164-177.

[28] Zhou, X. and X. Li, Dynamic spatio-temporal modeling for example-based human silhouette recovery. Signal Processing, 2015. 110: p. 27-36.

[29] Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005. IEEE.

[30] Dalal, N., B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. in European conference on computer vision. 2006. Springer.

[31] Uijlings, J., et al., Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. International Journal of Multimedia Information Retrieval, 2015. 4(1): p. 33-44.

[32] Wang, H., et al., *Dense trajectories and motion boundary descriptors for action recognition.* International journal of computer vision, 2013. 103(1): p. 60-79.

[33] Kantorov, V. and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[34] Tekin, B., et al. Direct prediction of 3D body poses from motion compensated sequences. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[35] Akilan, T., Q.J. Wu, and Y. Yang, Fusion-based foreground enhancement for background subtraction using multivariate multi-model Gaussian distribution. Information Sciences, 2018. 430: p. 414-431.

[36] McKenna, S.J., et al. Tracking interacting people. in Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. 2000. IEEE.

[37] Bouwmans, T., et al., Background Modeling and Foreground Detection for Video Surveillance. 2014: Chapman and Hall/CRC.

[38] Müller, M., Information retrieval for music and motion. Vol. 2. 2007: Springer.

[39] Wang, C., et al. Robust estimation of 3d human poses from a single image. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[40] Sigal, L., et al., Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. International journal of computer vision, 2012. 98(1): p. 15-48.

[41] Simo-Serra, E., et al. A joint model for 2D and 3D pose estimation from a single image. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[42] Taylor, G.W., et al. Dynamical binary latent variable models for 3d human pose tracking. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010. IEEE.

[43] Zhou, X., et al., MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. arXiv preprint arXiv:1701.02354, 2017.

[44] Bo, L. and C. Sminchisescu, *Twin gaussian processes for structured prediction.* International Journal of Computer Vision, 2010. 87(1): p. 28-52.

Mei Chee LEONG is a PhD student in the Institute for Media Innovation under Interdisciplinary Graduate School, Nanyang Technological University. She received her BE degree from the National University of Malaysia, and her MSc degree from Nanyang Technological University. Her research interest include computer vision, 3D reconstruction and machine learning.

Feng LIN obtained his PhD degree in Computer Science and Engineering from Nanyang Technological University, 1996.

He has been working in the area of biomedical informatics, imaging and visualization, computer graphics and virtual reality. He has published more than 250 peer-reviewed technical papers and books including those in IEEE TPAMI, and holds a few Technology Disclosures.

Dr Lin is now an Associate Professor with Nanyang Technology University and an IEEE Senior Member. He once won the National Science and Technology Advancement award and other significant awards.

Yong Tsui LEE is an associate professor at the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. He obtained his bachelor and PhD degrees from the University of Leeds, and a master degree from the University of Rochester, New York. His research interests include computer graphics, geometric modelling, CAE, 3D recovery from 2D, and motion capture from monocular inputs.