2.5D human pose estimation for shadow puppet animation

Shiguang Liu[1*], Guoguang Hua[2] and Yang Li[1]
[1] School of Computer Science and Technology, Division of Intelligence and Computing, Tianjin University
Tianjin, 300350 - China
[e-mail: lsg@tju.edu.cn]
[2] School of Information and Electrical Engineering, Hebei University of Engineering
Handan, 056038 - China
[e-mail: huaguoguang@foxmail.com]
*Corresponding author: Shiguang Liu

Received July 27, 2018; revised October 11, 2018; accepted October 28, 2018; published April 30, 2019

Abstract

Digital shadow puppet has traditionally relied on expensive motion capture equipments and complex design. In this paper, a low-cost driven technique is presented, that captures human pose estimation data with simple camera from real scenarios, and use them to drive virtual Chinese shadow play in a 2.5D scene. We propose a special method for extracting human pose data for driving virtual Chinese shadow play, which is called 2.5D human pose estimation. Firstly, we use the 3D human pose estimation method to obtain the initial data. In the process of the following transformation, we treat the depth feature as an implicit feature, and map body joints to the range of constraints. We call the obtain pose data as 2.5D pose data. However, the 2.5D pose data can not better control the shadow puppet directly, due to the difference in motion pattern and composition structure between real pose and shadow puppet. To this end, the 2.5D pose data transformation is carried out in the implicit pose mapping space based on self-network and the final 2.5D pose expression data is produced for animating shadow puppets. Experimental results have demonstrated the effectiveness of our new method.

Keywords: Human pose estimation, shadow puppet, mapping network, 2.5 pose data, CNN

http://doi.org/10.3837/tiis.2019.04.017 **ISSN : 1976-7277**

1. Introduction

Shadow play has a long history in China, Indonesia, India, Greece, etc. As a form of entertainment for children, adults and old people, shadow play is popular in many other countries around the world. We focus on Chinese shadow puppet in this paper. Chinese shadow play contains rich cultural elements, which is one of the most famous folk arts. Flat structure shadow puppet consists of several parts and its joints are connected by threads. Puppeteers manipulate the shadow puppets through sticks and the shadows are projected on a simple illuminated cloth screen to create moving pictures [1].

Because of the need for operational skills and experience, Chinese shadow play is becoming less known to the public. New techniques are required urgently to give new life to the Chinese shadow puppet. Fortunately, digital shadow puppet can help solve this problem. The most common approaches to driving digital shadow puppet include 1) controlling the puppet with a digital glove [2], 2) using computer vision for tracking marks in some objects that controls the shadow puppets [3], 3) using a multi-touch surface for direct manipulation of bi-dimensional shadow puppets [4], and 4) using body gestures to control the puppets with Kinect sensor [5] [6], etc. These researches contributed to greater knowledge on digital shadow puppet, however, some of them are complex to use or difficult to implement, and others need expensive equipment. In contrast, our motivation is to propose an easy method to generate interactive shadow puppet animation by real human pose data.

Human pose estimation in video is common in 2D plane or 3D scene. However, such type of human pose data extracted from these dimensions cannot be directly applied to drive shadow puppets. As shown in **Fig. 1**, the puppets body component adopts the frontal view of the human body (the 3/4 sides of the body) and is a rigid plane component in shadow play scene, which is a special scene between 2D and 3D scenes. The movement of shadow puppet is limited in 2D space, but it cannot be considered the traditional 2D space, it is the compression of the movement of human body in the side direction in the 3D space, so we call the scene as 2.5D. The 2.5D pose is a simplified 3D (x, y, z) surface representation that contains at most one depth (z) value for every point in the (x, y) plane.

The extracted human pose data by directly using 2D pose estimation method cannot better control shadow puppet, because the method might lose the number of detailed information, such as texture and depth cues. By contrast, the 3D pose estimation method contains a number of useful depth information. However, 3D human pose data can not be applied for shadow puppet control, because there is difference in movement and composition structure between human pose data of real scenario and shadow play. Some extraction methods can be generalized. The first one is to estimate human pose based on 2D scene, and then perform some depth recovery from depth dimension. The pose estimation is carried out from 2D pose methods and recovery from depth data. This is the process of increasing the dimension once again from the simplified data. This will lead to detail loss of the gesture form the process of simplification, which will have a serious impact on subsequent mapping operations. The second one is to estimate human pose in the 3D scene, and then restrain in depth dimension and map 3D pose into the 2.5D space built by difference information between real scenario and shadow puppet scene. However, training a network to obtain highly accurate 3D human pose estimation will cost a huge amount of computation. Some research work combines the two projects. For example, Tekin et al. [7] proposed a method of human pose estimation based on Convolutional Neural Network (CNN) for 2D and 3D human pose data fusion. In addition,

Tekin et al. [8] also used the structured relationship between body parts to improve the accuracy and speed of 3D human pose estimation on their own research.

To maintain the performance pattern of Chinese shadow play during the shadow play manipulation process, we need to simulate all the confining actions in the puppet style. There are several basic puppet motion patterns, such as walk, fight, nod, laugh, wave, etc. Beside, the real puppet is controlled with three sticks which are fixed on the puppet's neck and two hands separately, and the motion pattern of other puppet parts is affected by gravity. Recently, there are some research works on shadow puppets focusing on the user body interaction with the virtual shadow play, i.e., the puppet's motion imitates the user action [9] [10] [11]. But this method can not maintain some puppet's specific action style. For manipulation, we identify lots of motions for the animation and collect instances from a set of shadow puppetry videos. The conversion guideline is constructed by the collected instances. In the end, we train the self-organizing network by the conversion guideline, and then obtain the final 2.5D pose data.

Fig. 1. Example of the real performance scene of Chinese shadow play

In this paper, different from conventional 2D and 3D human pose estimation methods, we combine the advantages of these two methods for human pose estimation for driving characters in Chinese shadow play. Specifically, we consider human pose data from the 3D human pose estimation method as baseline features, and map the baseline features into the 2.5D space according to the conversion guideline built by difference between real human pose and shadow puppet. Finally, we obtain some special limited human pose data that can better drive shadow puppet. In addition, we propose some operations to optimize human pose estimation network to get accurate and robust human pose data, such as spatio-temporal consistency, self-organization and HOG3D feature, etc. In dealing with the appearance feature of video frames at the same time, we generate clue information in time domain, and complete the 3D pose estimation from single images. Then, the constraint mapping of the 3D pose data is performed according to difference guide information. Our contributions are two-fold:

(1) We propose a new method to obtain 3D human pose data as baseline data, which combines the advantages of 2D human pose estimation methods and 3D human pose estimation methods.

(2) We design a special translation scheme for mapping pose trajectory to 2.5D space. We first constrain the 3D pose data into the 2.5D scale space, and then train a transformation network according to conversion guideline to get the final 2.5D data. Beside, some optimization schemes are also designed to make the translation pose data more stable, quick and accurate in driving shadow puppet.

2. Related Work

2.1 Puppet Animation

Recently, some research works have been conducted on digital puppetry. As a visualization tool for traditional cinematic animation, digital shadow puppetry transforms the movements of a performer to the actions of an animated character to provide live performance. Producing an animated film with shadow puppets by puppeteers is laborious and time-consuming. Recently, the solution of animation performed by human pose data emerges. Fu et al. [12] designed a skeletal structure for driving digital shadow play. Leite et al. [13] proposed an anim-actor technique, which is a real-time interactive puppets control system using low-cost motion capture based on human body movements of non-expert artists. Lin et al. [14] proposed a method based on semantic tagging script to create the drive data of shadow puppets in the Kinect environment. Hsu et al. [15] introduced a motion planning technique which automatically generates the animation of 2D puppets. Tan et al. [16] presented a method for interactive animation of shadow play puppets by real-time visual simulating using texture mapping, blending techniques and blurring effects. Kim et al. [17] presented a 2D shape deformation of the triangulated cartoon which is driven by its skeleton and the animation can be obtained by re-targeting the skeleton joints to the shape. However, this method based on 2D pose might lose the micro-depth information and some special action pattern, such as the entry exit, the horizontal rotation of the body and arm waving.

Recently, the solution of puppet animation performed by 3D appears. Robert Held et al. [18] presented a method that allows users to quickly create a animations by performing the motions with their own familiar puppets. Shin et al. [19] proposed to transfer motion capture data to animated characters using the notion of dynamic importance of an end effecter. Theses methods are memory and time consuming. In this paper, we focus on constrained compression of 3D human pose data and mapping guideline from 3D to shadow puppet.

2.2 Human Pose Estimation

In the extraction of human pose data, the key step is human pose estimation. The 3D human pose estimation method is mainly based on image and single feature. Traditional methods [20] [21] [22] [23] rely on manual feature engineering to construct the posture of the human body. Manual features of human pose were aggregated into pose sets, and then the method of generating model search was used to obtain a reasonable inference corresponding to the parts of the body. The traditional methods of 3D pose estimation based on graph structure to get the estimation accuracy are less satisfactory. The deep learning framework is becoming the mainstream method. A powerful automatic functional network is built by the deep learning framework. Abundant low levels of expression characteristics is regressed to construct a mapping from image to 3D human pose, and using a variety of pose characteristics of joint prediction estimation of 3D human pose data in principle return or detection [24] [25] . Shotton et al. [26] proposed a method to predict 3D positions of body joints from a single depth image. They take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Finally, they generate confidence scored 3D proposals of several body joints by re-projecting the classification result and finding local modes. However, this method relied on single image feature, and the estimation error occurs in some fuzzy situations such as self-occlusion, mirror image and distortion caused by projection.

On the basis of the further integration of body parts information structured encoding, Li and Chan [27] and Ionescu et al. [28] took into account the global location information at the same time, and the structural features of parts of the human body, improving the 3D human pose estimation accuracy. Yasin et al. [29] proposed a method of RGB image based on single input dual pose data combined with 2D and 3D pose data, because only considering the global location information, the actual match is not accurate.

Later studies have found that the addition of global or structured logical features to the deep learning network makes the results of the 3D human pose estimation more accurate. It is means that researcher need utility 2D poses to predict 3D poses [30] [31] [32] [33] [34]. These approaches usually generalize better poses estimation since these methods can benefit from the state-of-the-art 2D pose estimators or methods. For example, Chen and Ramanan [31] proposed a method to predict 3D pose based on 2D pose. They handle the 2D pose estimation considering the camera coordinates and then the estimated poses are matched to 3D representations by means of a nonparametric shape model. Martinez et al. [34] proposed a simple fully connected residual network to directly regression 3D coordinates from 2D coordinate.

However, the above method has certain dependence on the 2D human pose data processing depth information, but the information may be missing some of the camera's perspective, resulting in the actual matching is not accurate, and only consider the location information of the feature in the whole process, so there are the problems of instability of the result of pose estimation inaccuracy and time on the domain. Therefore, the method of combining the global location information regression and the joint detection with other additional features, namely the method of 3D human pose estimation using multiple features is proposed. In the 3D pose estimation of video, we also use the method of maintaining the inherent consistency of space and time, and combine the 3D pose estimation method with multiple features. Tekin et al. [35] used structured characteristics of the relationship between body parts to improve the precision of 3D human pose estimation, but before the work is carried out in a single image above, also considering the 3D human pose estimation in consecutive video frames, and is added on the basis of the motion characteristics of human body image method 3D pose estimation, the image cube from short sequences extracted from the spatial and temporal characteristics, and the degradation of 3D pose, to capture the temporal information. Theobalt [36] proposed a new method combining 2D and 3D real-time attitude motion estimation, this method not only considers the global location information, but also considered to use motion features to make the spatio-temporal domain more stable overall detection.

In the paper, we focus on human pose data for driving virtual Chinese shadow play in 2.5D mapping scene obtained from real scenario. We consider the use of 3D human pose estimation method, while considering the actual difference between real scene and shadow play scene. The depth feature of 3D human pose estimation method is classified into an implicit feature, and the implicit feature is mapped in 2.5D space according some rule studied from difference between real scene and shadow play scene, and the stereo and effect drive data is obtained.

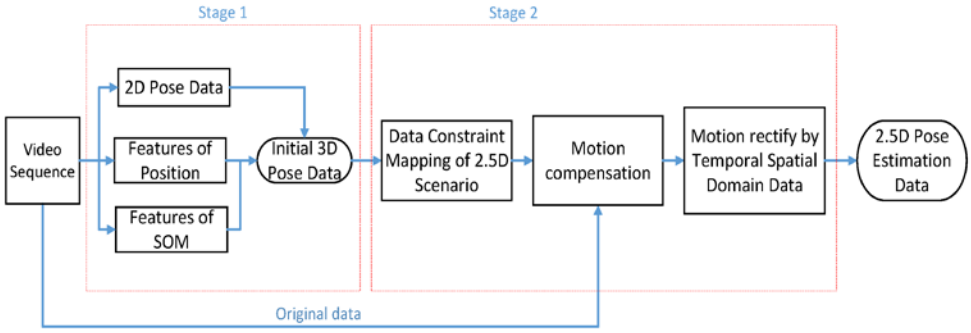

Fig. 2. The workflow of the main method with two stages for 2.5D pose data

3. Overview

As shown in **Fig. 2**, our method can be divided into two stages. The first stage is mainly used to obtain the initial 3D pose estimation data as the baseline, and the second stage is designed for 3D pose data constraint mapping to 2.5D pose from the 3D baseline data. In order to compensate for the loss of data features in the previous step for constraint, the skip scheme is proposed, and this branch contains pose trajectory of human in the time domain. The scheme compensates continuity of motion avoiding action fusion. The motion data is rectified by the HoG3D feature. Finally, relying on difference between real human motion pose and shadow puppets motion, the 2.5D pose driving data is generated. Technically, in the first stage, we obtain the 2D human pose data on single video frame by the method based on global appearance features and structured human body part features. And then the roughly 3D human pose data is obtained by the CNN framework from the 2D pose data and the original global appearance features. We build the network by using the multi-class network and the self-organizing feature map network. After obtaining the 3D pose data, it is transformed into the 2.5D scene by mapping guidelines. In order to improve the fluency of puppet's movements driven by 2.5D pose data, we will track the pose data by motion compensation and rectify the 2.5D pose data by a skip branch. Finally, the animation of the shadow motion effect of the shadow play model is obtained. In the next section, we will describe each step in detail.

4. Main Method

4.1 Initialization of 3D Pose Estimation Data

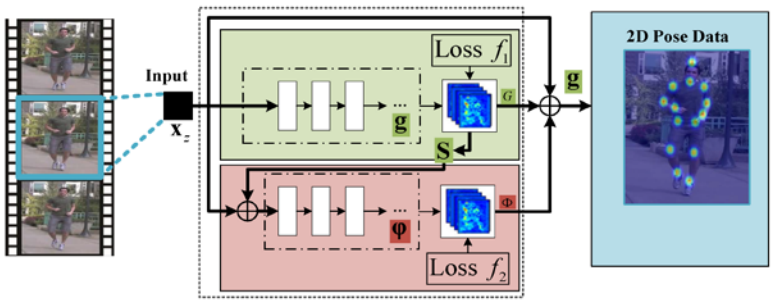

Fig. 3. Example of 2D pose data acquisition. The processing combines the global image appearance features and human structure features.

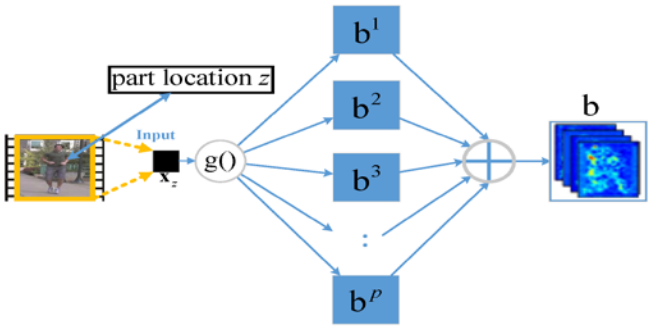

Fig. 4. Multiclass classifier module. Note that b is the confidence score feature map of each joint part.

The acquisition of the initial 3D pose data needs the help of the 2D pose data. With the assistance of 2D pose data, the joint data is extracted from the known 2D position of joint points by regression, which includes three ingredients, namely 2D pose estimation, self-organization, and regression for 3D pose data.

2D pose estimation: In order to obtain the 2D human pose data, we use the procedure as shown in **Fig. 3**. The overall network [37] can be divided into two parts, and the training process is also carried out separately. The first part is a multi-classifier module [38]. As shown in **Fig. 4**, the output definition of this multi-classifier module is $g(\bullet)$. This multi-classifier is designed for prediction of each body joint parts, and getting the confidence scores. Here we use the Gaussian peak function to represent the confidence map of the body joint part $p \in \{1,2,\cdots,12\}$. And the Gaussian peak function is defined as follows:

$$G_i^*(p) = \exp\left(-\frac{\left\|g(x_i) - y_i\right\|_2^2}{\sigma}\right) \tag{1}$$

where y_i denotes the ground-truth position of body joint part p, and σ controls the spread of the peak. We assume that 1-stage has J confidence maps ($J \in \{1,2,\cdots\}$) and use L_2 distance as the loss function f_1:

$$f_1(p) = \sum_J \sum_{i=1}^{p} W(p) \left\| G_i^j - G_i^* \right\|_2^2 \tag{2}$$

where $W(p)$ denotes the penalty item in the training process and it makes the loss function converge better. The structured features S are obtained after the structured spatial consistency learning process [39]. The second part of the training process also uses L_2 distance. The only difference is that there is an original input to reduce the vanishing of the feature gradient:

$$f_2(p) = \sum_{j=1}^{J} \sum_{i=1}^{p} W(p) \left\| \Phi_i^j - G_i^* \right\|_2^2 \tag{3}$$

$$\Phi = \varphi_p(x_z; \mathrm{S}) \tag{4}$$

Finally, the overall loss function is:

$$f = \sum_{t=1}^{T} (f_1 + f_2) \tag{5}$$

where t is the prediction over successive stages, $t \in \{1, \cdots, T\}$. The 2D pose data of each body joint part will have a confidence score feature map, which is derived from

$$b(p) = g^{p}(x_z; G; \Phi) \tag{6}$$

SOM extraction: Self-organized feature map (SOM) is a feature extracted by self-organizing network automatically [40]. And we use it to get abundant low-level pose data. SOM network is a fully connected self-organizing network. It can fulfill automatic unsupervised learning, which is similar to an auto-encoder network. In the whole process of the network, it can automatically extract the image features. It adopts the principle idea of sparse coding. The high order characteristics are sparsely coded and the features are reconstructed to achieve the goal of constructing rich low-level abstract image features. The operation principle of the self-coder is similar to the principle of principal component analysis (PCA), focusing on the principal component features of the image, which can realize the function of noise removal. And then we use the self-organization network in deep convolutional neural networks to obtain the initial 3D pose data by regression.

Regression Method for 3D Pose Data: In the previous steps, we have obtained 2D pose data and low-level features abstracted from SOM self-organizing features. So we can combine 2D human pose data to generate 3D human pose data based on the underlying features. In order to obtain 3D human pose data, we need to perform two steps, that is to obtain 2D human pose data, and then combine it with 3D human pose estimation network to obtain 3D pose estimation. The 3D pose estimation network consists of two parts, namely SOM network and location regression network.

Let Y be encoded as 3D pose position coordinate vector, X is the feature vector of the image. We use X to infer Y by regression methods [41] [42] [43], so the regression model function is:

$$Y_i = f(X_i) \tag{7}$$

where i represents the i_{th} body joint point. We optimize the L_2 distance of the 3D pose vector directly using the predicted results and the ground-truth results by the training procedure. The optimization formula is to minimize it by

$$\arg\min \sum_{i} \left\| Y_i - Y_i^{*} \right\|_2^2 \tag{8}$$

where Y_i^{*} indicates the i_{th} ground-truth body joint part position data. The model parameters of the regression network are trained. However, there are some difficulties in the actual 3D human pose estimation due to the problem of self-occlusion or mirror ambiguity in the image. So here we only get a rough initial 3D human pose estimation data.

4.2 Data Mapping and Rectification in 2.5D Space

2.5D Space Data Constraint Mapping: The initial 3D pose data we get is similar to coordinate (x, y, z) of each joint. Then, we are going to map it to the 2.5D space scene according to difference between real scene and shadow play scene. First, we would map it to a 2.5D space scene, so the data on the z dimension of the 3D point coordinate is limited to a certain range. Here, we restrict the range of depth dimension z between $[-10, 10]$ by standardization :

$$z_i' = -10 + \mathrm{k}(z_i - z_{\min}) \tag{9}$$

where k denotes the normalization coefficient calculated by

$$k = 20 / (z_{\max} - z_{\min}) \tag{10}$$

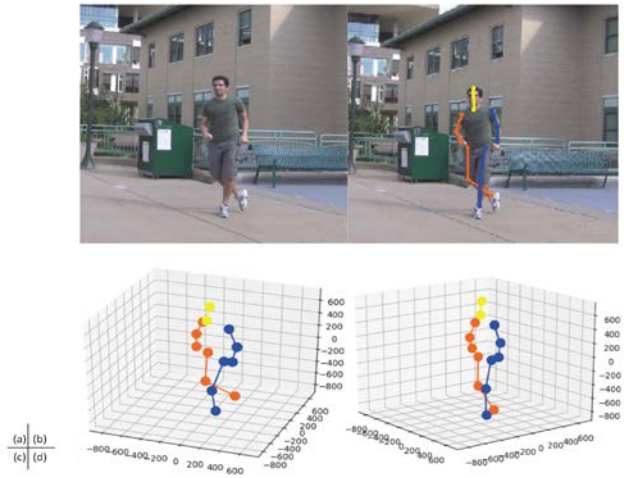

Fig. 5. A visual example of pose estimation data. (a), (b), (c), and (d) denote the original image, 2D pose estimation data ,3D pose estimation data, and constrained 2.5D pose estimation data, respectively.

As shown in **Fig. 5**, our 3D human pose estimation data is obtained based on 2D human pose estimation and some low-level features. Through the constraint normalization, the axis data can be constrained to $[-10, 10]$. In order to establish the coordinate mapping relationship between the joint points of the real human pose data and the shadow puppets, we need establish the 2.5D scene of shadow play according to the distribution of the joints of the real human pose. In addition, we need to consider the movement of shadow puppets, because the shadow puppet is driven by the joints in the puppet model skeleton. In short, the structure and movement of shadow puppet strictly follow programming norms, so we should design the standard according to the analysis of difference between real human pose and shadow puppet. The body structure of shadow play is divided into eleven components and twelve components based on literary and martial arts, respectively. In this paper, we take the twelve component puppet as an example. And then we learn the difference mapping between real human pose and shadow puppet by auto-encoded network with conversion guideline.

Action of conversion guideline: It is easy to control the puppet with 2.5D pose data. However, this is only true for simple human movements, such as moving legs, jumping, waving hands, etc. Those stereoscopic turn and complicated actions, e.g., back flips, splits and turn, could not be performed by the normal 2.5D pose data. We should define the puppet motion patterns. We show the part conversion guideline in **Table 1**.

Table 1. The conversion guideline list for puppet

Motion pattern of the puppet	Real human pose
To walk left	Moving the leg step to right (including the side and the front)
To walk right	Moving the leg step to left (including the side and the front)
To roll left	Sharply preparatory action of rolling left or forward
To roll right	Sharply preparatory action of rolling right and backward
Left somersault	Arms stretching up-left or up-forward
Right somersault	Arms stretching up-right or up-backward
Splits	Splitting two legs with an angle of $\geq$ 60 degrees

Rectification: The pose data obtained by the above method would suffer from unnatural phenomena, so it is necessary to further performing the process of rectification. So we will manipulate the overall data in the time domain of the video. Motion compensation is performed on the shadow play video frames to counteract the unnatural movement caused by data constraints and data mapping. we use the calculation features for analyzing 3DHoG feature [44] [45] [46], because it has the appearance of encoding information and motion information of human body in the image, which is composed of a group of equally spaced fine grained data cells, and calculate the gradient of each 3D spatiotemporal characteristics and the histogram. The characteristics of the overall space division are divided into a plurality of cube, and each cube is an independent unit. Then for the same partition, we will select a small cube which is divided into smaller units. The HoG features can be described as a polygon and can be used to calculate the final average gradient in this small unit. We will apply this feature to rectify the final driving data.

5. Experiment and Discussion

Based on the above algorithm, we conducted various experiments. All experiments were run on a computer configured as follows: Intel Core i5-4460 3.20 GHz CPU, NVIDIA GeForce GTX745 GPU, 8 GB RAM. In our experiment, we created Chinese version models with the software of 3DS Max.

5.1 Datasets

We ran our method on several datasets, including YouTube Pose, Human Eva-I [48], Outdoor Pose [49], MPII Human Pose Dataset [50], and Leeds Sports Pose Dataset [51] etc. Our experimental process is divided into three parts including the 2D human pose estimation, 3D human pose estimation, and the 2.5D scene after constraint and mapping. We focus on experimental analysis and comparison on Human Eva-I and Outdoor Pose datasets. Finally, we will show a series of visual data in 2.5D scene, and compare the results with 3D real human pose estimation results.

5.2 Experiments and Analysis

Experiment 1: 2D human pose estimation

In this section, we compared and analyzed some state-of-the-art methods of 2D human pose estimation. As shown in **Fig. 6** and **Fig. 7**, we have compared with methods of Park and Ramanan [52], Ramakrishna et al. [49], Chen and Yuille [53], Zhang and Shah [54] on two datasets of Human Eva-I and Outdoor Pose. From the visualization, we can find that previous methods have some pose estimation errors on two datasets. When the human limbs are occluded or articulated, the pose estimation error of the comparison methods would be easily generated. And the quantitative analysis on Outdoor Pose Dataset is shown in **Table 2**. It can be seen that our method outperforms existing methods and can achieve an average PCP score of 0.984. These results demonstrate that our methods can greatly improve the accuracy in some parts, and some of self-occlusion pose case can be better inferred.

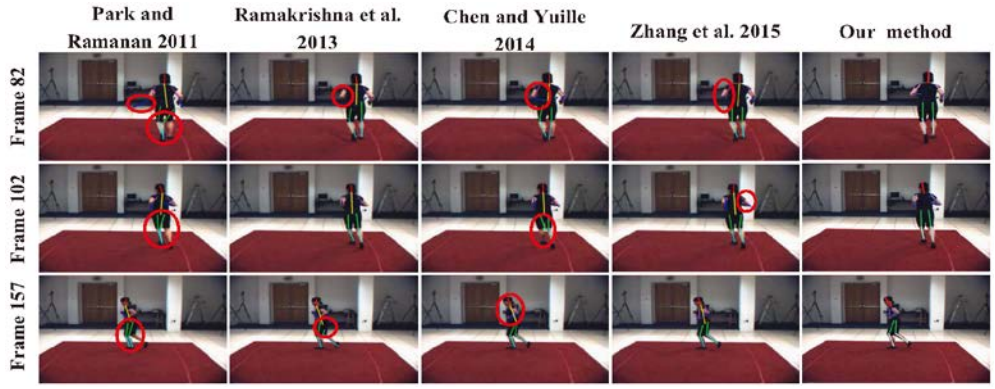

Fig. 6. 2D pose estimation comparisons with the state-of-the art methods for the Human Eva I. The red circle on the image is an apparently error estimation.

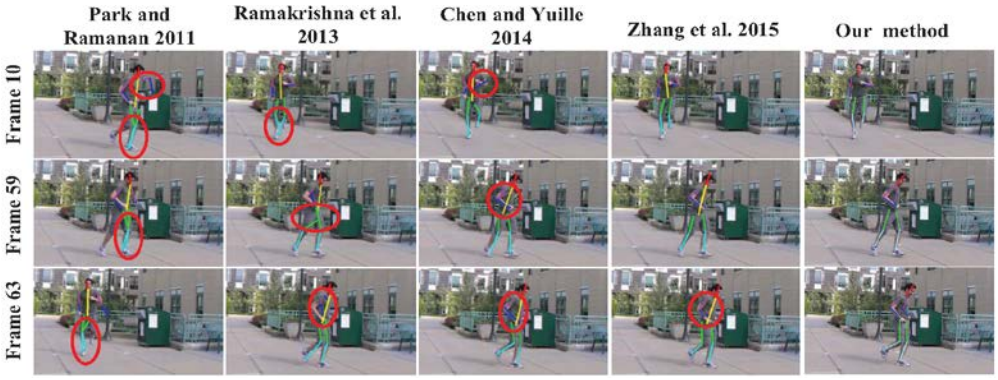

Fig. 7. 2D human pose estimation comparisons with the state-of-the-art methods for the Outdoor Pose Dataset. The red circle on the image is an apparently error estimation.

Table 2. Precision comparison on Outdoor Pose Dataset with methods the state-of-the-arts

Method	Head	Torso	U.L	L.L	U.A	L.A	Average
Park and Ramanan [52]	0.99	0.83	0.92	0.86	0.79	0.52	0.82
Ramakrishna et al. [38]	0.99	0.86	0.95	0.96	0.86	0.52	0.86
Chen and Yuille [53]	1.00	1.00	0.98	0.94	0.94	0.85	0.95
Zhang and Shah [54]	1.00	1.00	0.97	0.98	0.95	0.88	0.96
Li and Liu [55]	1.00	1.00	0.98	0.98	0.96	0.98	0.98
Our method	1.00	1.00	0.98	0.98	0.96	0.96	0.98

Experiment 2: 3D human pose estimation

In this experiment, we first present some qualitative results. Fig. 8 shows the results on challenging examples from the Outdoor Pose Dataset. We choose examples with self-occlusion. To demonstrate the accuracy of the 3D human pose estimation predictions, we visualize with novel viewpoints. It can be seen that our method can produce valid results for challenging images with self-occlusion and other challenging poses. This implies that our method can reliably estimate 3D human pose estimation.

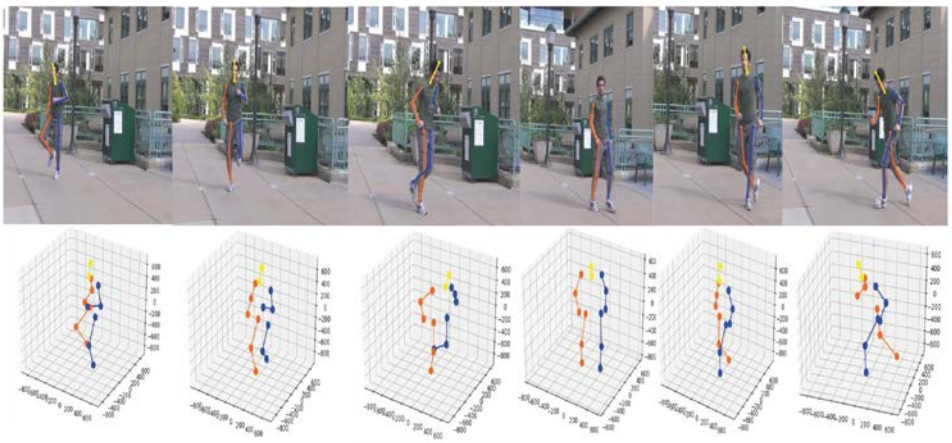

Fig. 8. 3D pose estimation results on the Outdoor Pose Dataset

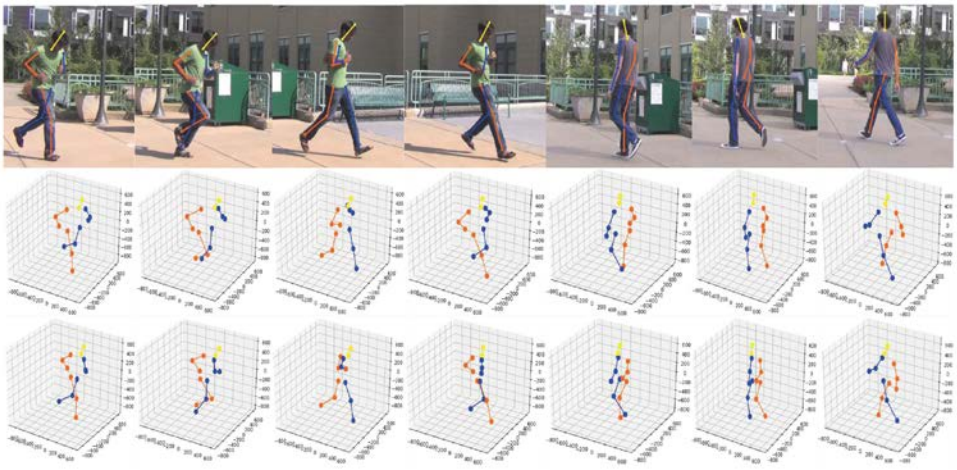

Fig. 9. Visual results of 3D pose estimation and 2.5D pose estimation on the Outdoor Pose Dataset. The bottom row is the 2.5D pose estimation results.

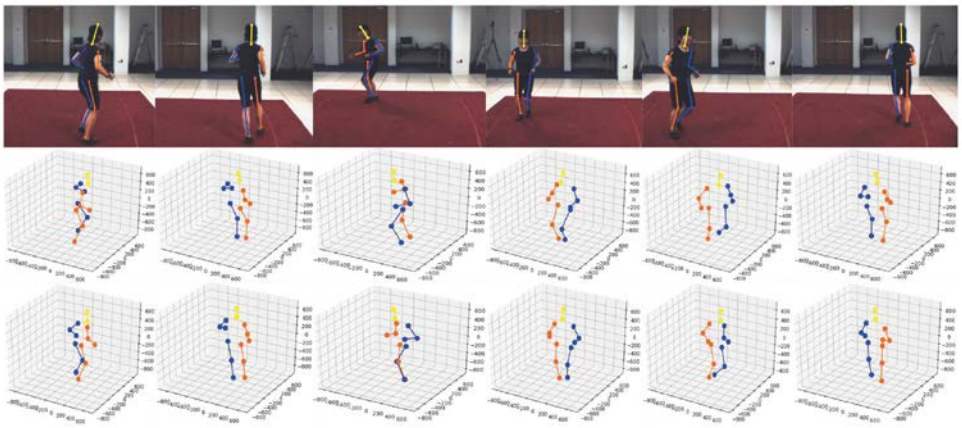

Fig. 10. Visual results of 3D pose estimation on the Human Eva I dataset. The bottom row is the 2.5D pose estimation results.

Comparison with state-of-the-art 3D pose estimation methods:

Different from the network structure in [27], our network replaces the front part of the network in [27] with two parallel networks, i.e., a new 2D pose feature network and a SOM network. The time complexity of our network is $O\left(M^2 \cdot K^2 \cdot C_{in} \cdot C_{out}\right)$, where M is the output feature map, K is the length of the side kernal, C_{in} is the channel of every convolution kernal and C_{out} is the output channel. If the time complexity is too high, the training and prediction of the model will consume a lot of time. In comparison with [27], we adopt abundant 1×1 convolution kernals to extract features. Therefore, our method can obtain real-time performance and better accuracy as shown in **Table 3**.

Compared with the methods in [29] and [31], our structure has no advantages in accuracy. However, our goal is to drive virtual Chinese shadow play in a 2.5D scene, we not only need to take accuracy into consideration, but also account for speed and the number of parameter. Both the methods in [29] and [31] use a large of 3D pose dataset to assist structure in matching 3D pose. And the method in [29] still does not achieve real-time performance when the number of main parameter K reaches the minimum. In contrast, we only use human Eva-I 3D motion capture data to train our extraction of data network and the parameter will obviously reduce by introducing smaller convolution kernal such as 3×3 and 1×1.

Table 3. Comparison among state-of-the-art 3D pose estimation methods and our method

Methods	Training time	Prediction time	Prediction accuracy	Training dataset
[27]	slow	Non-real time	low	Large(human 3.6M)
[29]	slow	Non-real time	**high**	Large(human 3.6M and human evaI)
ours	**fast**	**Real-time**	medium	**Little(human eva I)**

Experiment 3: 2.5D pose estimation by pose data constraints and mapping

This experiment aims to demonstrate the effectiveness of our final 2.5D driving pose data from 3D human pose estimation data. We show the results based on Human Eva-I, Outdoor Pose, and Human3.6M dataset. As shown in **Fig. 9** and **Fig. 10**, the 2.5D driving pose data constraints and mapping from 3D human pose estimation data is more suitable for the special 2.5D scene of shadow play. In this paper, we adopt the nine-part puppet design, i.e. the digital puppet consists of nine parts linked by eight joints and in total joints has 12 degrees of freedom. The eight joints are named according to their positions: neck, left and right shoulder, left and right elbow, waist and front and rear knee. These joints are rotation joints. Note that the waist has an additional prismatic joint along vertical axis of the upper body. The puppet parts are different from real human body. For driving shadow puppet, we should map the pose data according to mapping guideline. As shown in **Fig. 11** and **Fig. 12**, this is some image sequences that are captured from shadow play animation controlled by pose estimation data.

Fig. 11. Motion control results of shadow puppet example 1.

Fig. 12. Motion control results of shadow puppet example 2.

5. Conclusion and Future Work

This paper has proposed a new method to drive virtual Chinese shadow play in 2.5D mapping scene using human pose data obtained from real 3D scenarios. In order to obtain the 2.5D control data to drive the shadow puppets, we use the 3D human pose estimation method to obtain the initial data. In the following process of transformation, we constrained the initial data in a fixed range to shadow puppet side phenomenon. The feature transformation was carried out according to difference between real human pose data and shadow puppet driving data. Then, we can get the driving and mini stereoscopic expression data. In addition, we used the process of motion compensation and correction to maintain the optimization of overall pose trajectory flow and the optimization makes the pose trajectory data more smooth and stable to insure shadow play fluency. The 2.5D driving data were used in shadow puppets control and other special scenes. The automatic control animation of shadow play models were obtained by this method. The method of pose data extraction has reached higher estimation accuracy and shadow puppets movement is more fluent driven by 2.5D pose data. However, our method also has room for improvements. Our method consumes a lot of time in the extraction of pose data, because it can not automatically obtain the 2.5D pose data, which

needs two steps to process and optimize to get the final driving data. In the future, we would like to take more consideration to the effect of real-time estimation and control of shadow puppets.

References

[1] F. P. L. Chen, "AVisions for the masses: Chinese shadow plays from shaanxi and shanxi," *East Asia Program*, vol. 64, no. 2, pp. 324-327, 2004.

[2] A. Barlev, A. M. Bruckstein, G. Elber, "Virtual marionettes: a system and paradigm for real-time 3D animation," *Visual Computer*, vol. 21, no. 7, pp. 488-501, 2005. Article (CrossRef Link).

[3] A. Sirota, D. Sheinker, O. Yossef, "Controlling a virtual marionette using a web camera," *Mahmoudzeidan Com*, vol. 28, no. 5, pp. vii-ix, 2004.

[4] I. Takeo, I. Yuki, "Implementing as-rigid-as-possible shape manipulation and surface flattening," *Journal of Graphics Gpu and Game Tools*, vol. 14, no. 1, pp. 17-30, 2009. Article (CrossRef Link).

[5] H. Zhang, Y. Song, Z. Chen, "Chinese shadow puppetry with an interactive interface using the Kinect sensor," in *Proc. of International Conference on Computer Vision*, vol.7583, pp. 352-361, 2012. Article (CrossRef Link).

[6] KM. Lee, HM. Won, "Dynamic gesture recognition using a model-based temporal self-similarity and its application to taebo gesture recognition," *KSII Transactions on Internet and Information Systems*, vol. 7, no. 11, pp. 2824-2838, 2013. Article (CrossRef Link).

[7] B. Tekin, P. Marguez-Neila, M. Salzmann, "Learning to fuse 2D and 3D image cues for monocular body pose estimation," in *Proc. of the IEEE international Conference on Computer Vision (ICCV)*, pp. 3961-3970, 2007. Article (CrossRef Link).

[8] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3D human pose with deep neural networks," in *Proc. of International Conference British Machine Vision Conference (BMVC)*, pp.130.1-130.11, 2016. Article (CrossRef Link).

[9] L. Luis, V. Orvalho, " Shape your body: control a virtual silhouette using body motion," in *Proc. of ACM CHI Extended Abstracts on Human Factors in Computing Systems*, pp. 1913-1918, 2012. Article (CrossRef Link).

[10] Y. S. Iin, C. K. Shie, S. C. Chen, "Action recognition for human-marionette interaction," in *Proc. of the ACM international conference on Multimedia*, pp. 39-48, 2012. Article (CrossRef Link).

[11] H. Zhang, Y. Song, Z. Chen, "Chinese shadow puppetry with an interactive interface using the Kinect sensor," in *Proc. of the 12th international conference on Computer Vision (ECCV)*, pp. 352-361, 2012. Article (CrossRef Link).

[12] L. Fu, J. P. Cai, "Research and application of digital shadow-play bones animation," *Computer Engineering and Design*, vol. 34, no. 1, pp. 241-246, 2013. Article (CrossRef Link).

[13] L. Leite and V. Orvalho, "Anim-actor: understanding inter with digital puppetry using low-cost motion capture," in *Proc. of the 8th International Conference on Advances in Computer Entertainment Technology*, pp.65, 2011. Article (CrossRef Link).

[14] M. Lin, Z. Hu, S. Liu, "eHeritage of shadow puppetry: creation and manipulation," in *Proc. of the 21st ACM international conference on Multimedia*, pp. 183-192, 2013. Article (CrossRef Link).

[15] S. W. Hsu, T. Y. Li, "Planning character motions for shadow play animations," in *Proc. of the International Conference on Computer Animation and Social Agents (CASA)*, pp. 184-190, 2005.

[16] K. Tan, A. Talib, M. Osman, "Real-time visual simulation and interactive animation of shadow play puppets using openGL," *World Academy of Science Engineering and Technology*, vol. 47, no. 2008, pp. 212-218, 2008.

[17] D. H. Kim, M. Y. Sung, J. S. Park, "Realtime control for motion creation of 3d avatars," in *Proc. of the 6th Pacific-Rim Conference on Advances in Multimedia Information Processing*, pp. 25-36, 2005. Article (CrossRef Link).

[18] R. Held, A. Gupta, B. Curless, "3D puppetry:a kinect-based interface for 3D animation," in *Proc. of the 25th annual ACM symposium on User interface software and technology*, pp423-434, 2012. Article (CrossRef Link).

[19] H. J. Shin, J. Lee, S. Y. Shin, "Computer puppetry: An importance-based approach," *ACM Transactions on Graphics (TOG)*, vol. 20, no. 2, pp. 67-94, 2001. Article (CrossRef Link).
[20] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, H. W. Haussecker, "Detailed human shape and pose from images," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007. Article (CrossRef Link).
[21] L. Sigal, M. Isard, H. Haussecker, M. J. Black, "Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15-48, 2012. Article (CrossRef Link).
[22] S. Gammeter, A. Ess, T. Jäggli, K. Schindler, B. Leibe, L. V. Gool, "Articulated multi-body tracking under egomotion," in *Proc. of the 10th European Conference on Computer Vision*, pp. 816-830, 2008. Article (CrossRef Link).
[23] J. Gall, B. Rosenhahn, T. Brox, H. P. Seidel, "Optimization and filtering for human motion capture," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 75-92, 2010. Article (CrossRef Link).
[24] N Bruce Xiaohan, P. Wei, S. C. Zhu, "Monocular 3D Human Pose Estimation by Predicting Depth on Joints," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3467-3475, 2017. Article (CrossRef Link).
[25] Pavlakos, Georgios, et al., "Learning to Estimate 3D Human Pose and Shape from a Single Color Image," *arXiv preprint arXiv:1805.04092*, 2018.
[26] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finoc-chio, A. Blake, M. Cook, R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116-124, 2013. Article (CrossRef Link).
[27] S. Li, A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Proc. of Asian Conference on Computer Vision*, pp. 332-347, 2014. Article (CrossRef Link).
[28] C. Ionescu, J. Carreira, C. Sminchisescu, "Iterated second-order label sensitive pooling for 3d human pose estimation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1661-1668, 2014. Article (CrossRef Link).
[29] H. Yasin, U. Iqbal, B. Kruger, A. Weber, J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2010, no. 1, pp. 4948-4956, 2016. Article (CrossRef Link).
[30] X Zhou, M Zhu, S Leonardos, et al. "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4966-4975, 2016. Article(CrossRef Link).
[31] CH Chen, D Ramanan, "3D human pose estimation=2D pose estimation+matching," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5759-5767, 2017. Article (CrossRef Link).
[32] F Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1561-1570, 2017. Article (CrossRef Link).
[33] BX Nie, P Wei, SC Zhu. "Monocular 3D human pose estimation by predicting depth on joints," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3467-3475, 2017. Article (CrossRef Link).
[34] J Martinez, R Hossain, J Romero et al. "A simple yet effective baseline for 3d human pose estimation," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2659-2668, 2017.
[35] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3D body poses from motion compensated sequences," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991-1000, 2016. Article (CrossRef Link).
[36] C. Theobalt, "VNect: real-time 3D human pose estimation with a single RGB camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 44, 2017. Article (CrossRef Link).
[37] S Liu, Y Li, G Hua. "Human Pose Estimation in Video via Structured Space Learning and Halfway Temporal Evaluation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp.1,

2018. Article (CrossRef Link).
[38] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," In *Proc. of ECCV*, pp. 33-47, 2014. Article (CrossRef Link).
[39] X. Chu, W. Ouyang, H. Li, X. Wang, "Structured feature learning for pose estimation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4715-4723, 2016. Article (CrossRef Link).
[40] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, 1982. Article (CrossRef Link).
[41] A. Agarwal, B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proc. of the 2004 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR)*, pp.882-888, 2004. Article (CrossRef Link).
[42] C. Bregler, A. Hertzmann, H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. of the 2000 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR)*, pp.690-696, 2000. Article (CrossRef Link).
[43] A. Kanaujia, C. Sminchisescu, D. Metaxas, "Semi-supervised hierarchical models for 3d human pose reconstruction," in *Proc. of the 2007 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007. Article (CrossRef Link).
[44] X. Fan, K. Zheng, Y. Zhou, S. Wang, "Pose locality constrained representation for 3D human pose reconstruction," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 174-188, 2014.Article (CrossRef Link).
[45] R. Urtasun and T. Darrell, "Sparse probabilistic regression for activity-independent human pose inference," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2008. Article (CrossRef Link).
[46] A. Klaser, M. Marsza ek, C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. of British Machine Vision Conference (BMVC)*, pp.99.1-99.10, 2008. Article (CrossRef Link).
[47] D. Weinland, M.Ozuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. of the 11th European conference on computer vision conference on Computer Vision (ECCV)*, pp. 635-648, 2010. Article (CrossRef Link).
[48] L. Sigal, A. O. Balan, M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4-27, 2010. Article (CrossRef Link).
[49] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Tracking human pose by tracking symmetric parts," in *Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3728-3735, 2013. Article (CrossRef Link).
[50] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686-3693, 2014. Article (CrossRef Link).
[51] S. Johnson, M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. of British Machine Vision Conference (BMVC)*, pp. 1-11, 2010. Article (CrossRef Link).
[52] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2627-2634, 2011. Article (CrossRef Link).
[53] X. Chen, A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *Advances in Neural Information Processing Systems*, pp. 1736-1744, 2014.
[54] D. Zhang and M. Shah, "Human pose estimation in videos," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2012-2020, 2015. Article (CrossRef Link).
[55] Y. Li and S. G. Liu, "Temporal-coherency-aware human pose estimation in video via pre-trained res-net and flow-cnn," in *Proc. of International Conference on Computer Animation and Social Agents (CASA)*, pp. 150-159, 2017.

Shiguang Liu is a Professor at School of Computer Science and Technology, Tianjin University, P.R. China. He graduated from Zhejiang University and received a PhD from State Key Lab of CAD & CG. His research interests include modelling and simulation, realistic image synthesis, image/video editing, computer animation, and virtual reality, etc.

Guoguang Hua graduated from School of information and Electrical Engineering, Hebei University of Engineering, P.R. China. His research interest is computer vision.

Yang Li graduated from School of Computer Science and Technology, Tianjin University, P.R. China and got his Master degree. His research interest is computer graphics.