DANCING LIKE A SUPERSTAR: ACTION GUIDANCE BASED ON POSE ESTIMATION AND CONDITIONAL POSE ALIGNMENT

Yuxin Hou, Hongxun Yao, Haoran Li, Xiaoshuai Sun

School of Computer Science and Technology, Harbin Institute of Technology, China.

ABSTRACT

Action Guidance (**AG**) aims at scoring how accurate the action is and giving guidance to the learners on how to correct their actions according to the standard instructive videos. **AG** has plenty of real-world applications such as sports training, rehabilitation treatment, and dance teaching. However, the problem of assessing the accuracy of action has almost no effective solution. In this paper, we describe a two-stage framework for action guidance. Firstly, we estimate the poses in the test video and standard video using person detection method and convolutional pose machines (CPMs). As for action guidance, we propose a network to compute the essential differences between two poses under different sizes, views, and locations. Extensive experiments with real-world and synthetic datasets demonstrate the effectiveness of our framework.

Index Terms— action guidance, conditional pose alignment, pose estimation

1. INTRODUCTION

Pose estimation and action recognition are widely explored in computer vision to understand the human motion. However, the high-level problems of human motion analysis such as scoring how accurate people perform the action and giving guidance for correction have almost no effective solution. However, solving these problems is critical to a lot of real-world applications such as sports training, rehabilitation treatment, and dance teaching. Action guidance is not an easy task because of the large variation in target size, view, and location in different pose sequences. How to map two poses to the same situation is very important. Moreover, action guidance relies heavily on accurate person detection and pose estimation to get the action information as precisely as possible.

Traditional methods[1, 2, 3, 4, 5, 6, 7, 8] for pose estimation model the pictorial structures of body parts which mainly based on handcrafted features. These methods have successful results on the person who are visible without any occlusion on body parts. Recent progress in pose estimation largely relies on the development of convolutional neural networks. The Convolutional Pose Machines [9] combine the pose machine architecture [10] with convolutional architecture. CPMs use CNN to learn the feature representations and spatial context directly from the data.

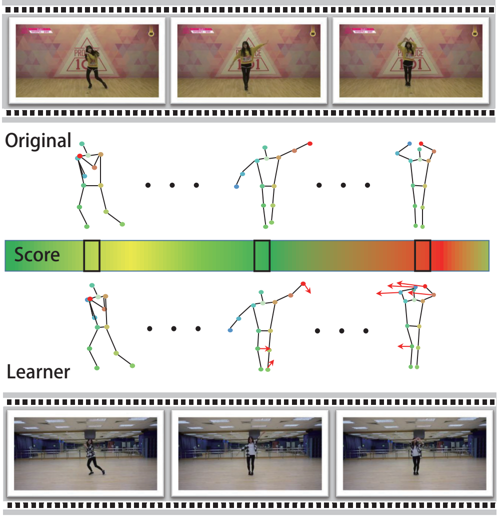

Fig. 1. The action guidance task analyzes a pair of videos from original performance and cover performance. Our method can assess the accuracy of the learner's poses in every frame comparing with the original poses and give the final score (green means the learner performs accurately, red means the similarity between original dance and cover dance is small) as well as the guidance on how to correct the action(red arrows on the pose articulation model).

With the popularity of human motion analysis in computer vision and the mature techniques, the demand for action guidance has been increasing rapidly. There are a few works [11, 12, 13] trying to assess the actions which are specific to some certain domains. Recently, [14] propose a learning-based framework to assess the quality of actions. However, the goal of the score and the feedback result in this work is to let the action looks good given a specific kind of sport, which does not care about the similarity between the input action and a certain standard action. Generally, most applications need a strategy to guide the imitation with standard actions. For example, if a dance learner or a gymnast want to learn how to perform a dance or a gymnastic movement, what they need is letting his action looks similar to a certain standard action but not just visually appealing.

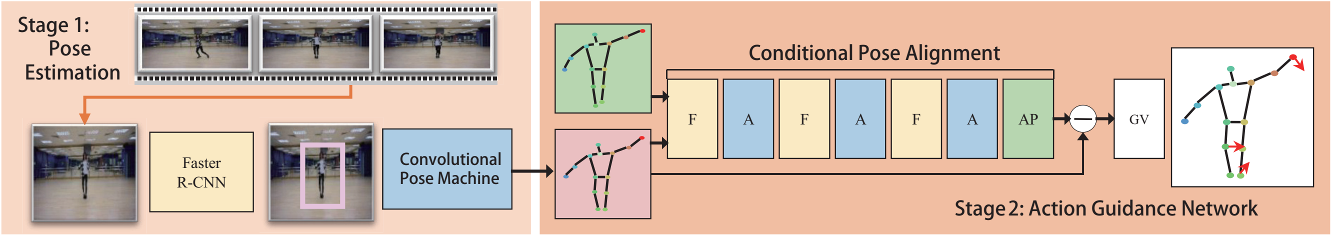

Fig. 2. We present a two-stage framework for action guidance task. In the first stage, Faster R-CNN is used to detect the person in the input frames of the video. Then the person candidate (pink bounding box) is analyzed by CMPs to get the pose articulation. In the second stage, we propose an action guidance network to compute guidance vectors (GV) for the comparison against the standard pose. (F: Fully-Connected, A: Activation, AP: Aligned Pose)

In order to address the problem of action guidance, we introduce a two-stage framework (see Fig.2). In the first stage, we use Faster R-CNN [15] to detect the person in every frame of two input videos and CMPS to estimate the poses. In the second stage, we propose an action guidance network to align and compare the standard video with the test video which has the ability to model poses under different sizes, views, and locations. The network we proposed can also provide guidance on how to correct the action in each frame. The primary contributions of this paper can be summarized as follows: 1) we solve a new task called Action Guidance to score and guide action given a standard reference, and 2) we propose a new two-stage action guidance network to assess how accurate the learner copy the reference action and, 3) we release a new dance dataset for action guidance which includes the original performances and the corresponding cover dances.

2. METHOD

In this section, we describe our two-stage framework for action guidance. First we introduce the pose estimation method. After getting the pose, we show our action guidance network.

2.1. Pose Estimation

It's crucial to get the accurate pose information because action guidance is based on the pose estimation. We use CPMs which is the state-of-the-art method for pose estimation. CPMs predict the heat map by using the stacked CNN and iterated predicting strategy. However, CPMs has its own applicable conditions, the result will be better if the target person is in the center of the image and of the suitable size.

Considering that the CPMs are sensitive for input images, we design a integrate automatic pose estimation pipeline by getting the person location prior from person detection. We use Faster R-CNN to detect a person and get a bounding box for pose estimation.

2.2. Action Guidance

We show how the accuracy score $\mathcal{S}$ and guidance vectors $\mathcal{G}$ of action can be provide by our network. We denote the body parts configuration $x_i = \{x_i^1, x_i^2, ..., x_i^{14}\}$. We compute the differences between cover dance sequence $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ and original dace sequence $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ to provide the score $\mathcal{S}$ which describe the accuracy of action and guidance $\mathcal{G} = \{g_1, g_2, ...g_n\}$ which is a set of vectors pointing how to correct to a more accurate location of each body part.

Action accuracy is actually a measurement for two pose structures. There are two parts in computing the accuracy score $\mathcal{S}$. The first part is the distance between the original pose and the learner's pose from the certain frame of the input videos. The distance between two structures can be divided into two parts: alignment distance $\mathcal{S}_a$ which is caused by the view, size, location and other environmental impacts and structure distance $\mathcal{S}_g$ which is the essential differences between two pose structures after mapping to the same condition. In our action guidance task, we want to catch the structure distance and ignore the alignment distance. The second part $\mathcal{S}_t$ is the temporal feature which computing the changes between the neighboring frames to measure the intensity and the velocity of the action. Therefore, the two parts make up the total score of the cover performance:

$$\mathcal{S} = \alpha\mathcal{S}_t + \beta\mathcal{S}_g \tag{1}$$

where α and β are the weights of the two parts.

The action guidance network (stage2 of Fig.2) which takes two poses from a certain frame of original performance video and cover performance video as input and outputs the guidance vectors overlying on the learner's pose. Therefore we map the two poses into the condition of the cover performance, which means our network align the original pose to the cover pose. We use $\mathcal{Y}|\mathcal{X} = \{y_1|x_1, y_2|x_2, ..., y_n|x_n\}$ to represent the original pose sequence after alignment. The guidance $\mathcal{G}$ can be represented as following:

$$\mathcal{G} = \mathcal{Y}|\mathcal{X} - \mathcal{X} \tag{2}$$

In order to get the alignment result $\mathcal{Y}|\mathcal{X}$, we design a 4-layers stacked fully-connected conditional pose alignment network showing in Fig.2. In the network, the original pose of the certain part will be aligned to the same condition of the cover performance pose including size, view, and location.

We have already analyzed the guidance vector $\mathcal{G}$ which related to the structure distance. In this period, we will introduce the temporal feature $\mathcal{T}$. An action is not a static structure, the accuracy of the action should not be described only by the pose of every frame but also the intensity, velocity, and continuity of the action. For example, the pose of the current frame can be same with the original pose, however, the process of the action is not suitable. The temporal feature of the body moving plays the same important role in action analysis. For each frame, we compute the change from the previous frame to the current frame, and compare the change between original action and cover action:

$$\begin{aligned} t_i &= (y_i|x_i - y_{i-1}|x_{i-1}) - (x_i - x_{i-1}) \\ &= (y_i|x_i - x_i) - (y_{i-1}|x_{i-1} - x_{i-1}) \\ &= g_i - g_{i-1} \end{aligned} \tag{3}$$

From this function, we can easily see the relationship between $\mathcal{T}$ and $\mathcal{G}$. We can regard $\mathcal{T}$ as a dynamic description of the guidance, and $\mathcal{G}$ is static description.

Since $||\mathcal{T}||^2$ and $||\mathcal{G}||^2$ is none negative. And the smaller the value is , the better the performance is. We design the score functions of $\mathcal{S}_g$ and $\mathcal{S}_t$ for two input videos with n frames as following:

$$\mathcal{S}_t = \frac{1}{n}\sum_{i=1}^{n} 100 * e^{-||t_i||^2/\eta} \tag{4}$$

$$\mathcal{S}_g = \frac{1}{n}\sum_{i=1}^{n} 100 * e^{-||g_i||^2/\eta} \tag{5}$$

The η is a regular factor to make the score interpretable.

The methods of how to get the accuracy score and the guidance vector are shown above. In this period, we introduce how to train the conditional pose alignment network. We map the two poses into the condition of the cover performance, which means our network align the original pose to the cover pose. In training, we create the training data we need. Firstly, we get a lot of pose structures from pose estimation. Then for each pose p_i we get, we randomly change the location of some parts to get a new pose q_i. Then we randomly change the location, size, and view of q_i but do not change the structure of the pose and get many different r_i. Therefore, q_i is the alignment result of r_i mapping to the condition of p_i. We define there are m pairs of poses in the training dataset, the loss of training is Mean Square error(MSE):

$$\begin{aligned} \mathcal{L} &= \sum_{i=1} ||f(r_i, p_i; W) - r_i|p_i||^2 \\ &= \sum_{i=1} ||f(r_i, p_i; W) - q_i||^2 \end{aligned} \tag{6}$$

The goal of the conditional pose alignment network is to get the alignment result $y|x$ when giving the certain pose pair x and y. We achieve this goal by using mini-batch SGD to minimize $\mathcal{L}$ and get the W for testing.

3. EXPERIMENTS

To test the effectiveness of our framework, we design our experiments on both real-world and synthetic data. Due to the limitation of CPMs mentioned in the previous section, we use Faster R-CNN to detect the person first to prove the accuracy of pose estimation. In this section, we mainly show the experimental results of the conditional pose alignment and action guidance.

3.1. conditional pose alignment

Firstly, we introduce our experiments on synthetic dataset designing for testing the effectiveness of the conditional pose alignment. We test our method on the MPII [16] dataset and a new dancing dataset named PickMe. For every pose $p \in R^{(14,2)}$ in the dataset, we produce a random (14,2) disturbance vector d. Each value in the disturbance vector is randomly sample from uniform distribution during [-20,20]. The new pose is denoted as $q = p + d$. For each pose q, we change the view, size, and location randomly and get a new pose r which is not aligned to p.

$$\begin{aligned} r_i^T &= \mathcal{A}_i(q^T) \\ &= s_i * \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} * q^T + l_i \end{aligned} \tag{7}$$

In this formulation, $\mathcal{A}$ is the change of pose q. s, θ and l are the changes on the scale, angel and translation of q. By using this, we can get 2888300 pairs pose $(p, r_i) \to q$.

While training the network on the MPII dataset, we use Xavier to initialize the weight of the network. Since MSE is hard to converge at the beginning, we train the network using Mean Absolute Error(MAE). When the network converges to the locally optimal solution, we begin to use MSE to train.

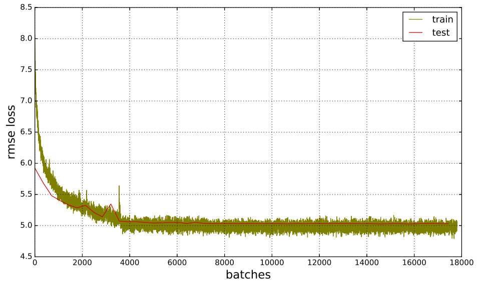

Fig. 3. We finetune the pretrained model in the PickMe real-world dataset, the model converged to a low loss which only is about 4-5 pixels.

dataset	head	neck	Rsho	Relb	Rwri	Lsho	Lelb	Lwri	Rhip	Rkne	Rank	Lhip	Lkne	Lank	MAP
MPII	58.90	61.84	61.29	59.25	55.47	61.72	59.37	55.42	59.53	61.08	58.76	59.22	61.80	59.22	59.49
PickMe	68.55	74.19	73.82	75.23	71.93	74.31	75.28	73.12	77.75	72.70	64.62	78.60	73.03	64.96	72.72

Table 1. The mean Average Precision @ 0.5 head segment length.

dataset	head	neck	Rsho	Relb	Rwri	Lsho	Lelb	Lwri	Rhip	Rkne	Rank	Lhip	Lkne	Lank	mAP
MPII	95.63	96.93	96.74	95.59	93.82	96.69	95.60	93.74	95.80	96.45	94.39	95.62	96.35	94.82	95.58
PickMe	99.90	99.97	99.96	99.94	99.86	99.98	99.92	99.84	99.93	99.85	99.81	99.96	99.95	99.76	99.90

Table 2. The PCKh @ 0.5 head segment length.

We use mini-batch momentum stochastic gradient descent (SGD) to optimize the model, The learning rate is 0.000001, weight decay is 0.00001, momentum is 0.99. The learning rate times 0.1 after each 10 epochs. After pre-training, we finetune the model in the PickMe dataset (Fig.3).

We test our method on the MPII and PickMe dataset, and we use mean average precision@0.5(mAP@0.5) and PCKh@0.5 to evaluate our model. As shown in Table.1, the mAP@0.5 before alignment is about 17% while the alignment result is 59.49%. The PCKh@0.5 and mAP@0.5 demonstrate that our model can align the pose accurately.

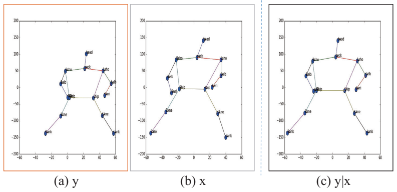
(a) y
(b) x
(c) y|x

Fig. 4. The alignment result: (a) is the original pose, (b) is the test pose, and (c) is that alignment result.

As shown in Fig.4, we see the model can align the original pose by scaling and translating the original pose conditionally based on the test pose.

Secondly, we introduce our experiments on a real-world dataset which aim at giving score and guidance to the input video pair to help the learner perform the action more accurately. Our PickMe dataset is a new dance dataset including 60 pairs of dance videos(original dance and cover dance). In this dataset, one pair of videos shows the same dance performed by different people. We first estimate the pose of every frame in the videos. Then we put the poses into the action guidance network to get the score and guidance vector.

As shown in Fig.5, the pose results are overlapping on the images. We take two frames from each video as an example. For the first frame, we can see from the figure that the learner performs similarly to the original dance, so the T score ($\mathcal{S}_t$) and the G score ($\mathcal{S}_g$) are high (green means high while red low). As for the second frame we choose, the learner forgets how to dance. Therefore, the G score of the neighboring frames are very low. Although the T score in the second frame is low, it's better than the G score. The reason is that the movement of the neighboring frames in the original dance is small. Although the locations of her body parts are wrong, the mistake in displacement is not so big. The guidance vectors can provide how to correct the action more accurately.

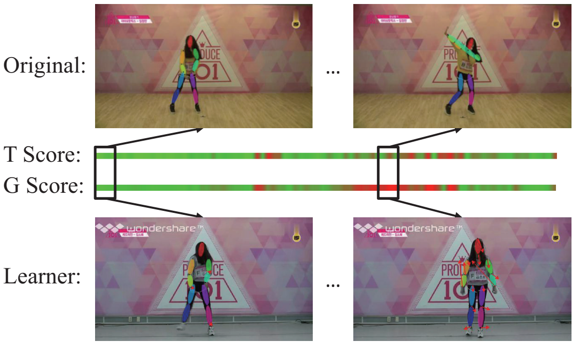

Fig. 5. Example guidance result in PickMe dataset. Red arrows(→) on the learner's pose are the guidance vectors.

4. CONCLUSION

Action guidance is a challenging task but of great real world interests which can enable various applications such as sports training, rehabilitation treatment, and dance teaching. In this paper, we propose a two-stage framework for action guidance which can score how accurate people perform the action and give proper guidances to the learners on how to correct their action according to the standard video. The experimental results indicate that our action guidance network can not only compute the essential differences between two poses under different scales, views, and locations but also have the ability to provide score and guidance in action imitation applications.

In our future work, further improvement will be focused on better pose estimating in videos and more advanced networks for pose alignment. We also hope this work can interest more people to explore the relative areas of action guidance.

5. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Project No. 61472103.

6. REFERENCES

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1014–1021.

[2] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[3] Sam Johnson and Mark Everingham, "Clustered pose and nonlinear appearance models for human pose estimation.," in *BMVC*, 2010, vol. 2, p. 5.

[4] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1385–1392.

[5] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3487–3494.

[6] Ben Sapp and Ben Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

[7] Fang Wang and Yi Li, "Beyond physical connections: Tree models in human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 596–603.

[8] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.

[9] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," *arXiv preprint arXiv:1602.00134*, 2016.

[10] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*. Springer, 2014, pp. 33–47.

[11] Andrew S Gordon, "Automated video assessment of human performance," in *Proceedings of AI-ED*, 1995, pp. 16–19.

[12] Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič, "Trajectory based assessment of coordinated human activity," in *International Conference on Computer Vision Systems*. Springer, 2003, pp. 534–543.

[13] Matej Perše, Matej Kristan, Janez Perš, and Stanislav Kovačič, *Automatic Evaluation of Organized Basketball Activity using Bayesian Networks*, Citeseer, 2007.

[14] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba, "Assessing the quality of actions," in *European Conference on Computer Vision*. Springer, 2014, pp. 556–571.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3686–3693.