BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

Animal Exercise: A New Evaluation Method

Yu Qi[1*] Chongyang Zhang[1] Hiroyuki Kameda[2]

1. The Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Japan
2. The School of Computer Science, Tokyo University of Technology, Japan

ARTICLE INFO

Article history
Received: 27 May 2022
Accepted: 10 June 2022
Published Online: 15 June 2022

Keywords:
Motion transfer
Animal exercise
Evaluation method
Monkeys
Target scale normalization

ABSTRACT

At present, Animal Exercise courses rely too much on teachers' subjective ideas in teaching methods and test scores, and there is no set of standards as a benchmark for reference. As a result, students guided by different teachers have an uneven understanding of the Animal Exercise and cannot achieve the expected effect of the course. In this regard, the authors propose a scoring system based on action similarity, which enables teachers to guide students more objectively. The authors created QMonkey, a data set based on the body keys of monkeys in the coco dataset format, which contains 1,428 consecutive images from eight videos. The authors use QMonkey to train a model that recognizes monkey body movements. And the authors propose a new non-standing posture normalization method for motion transfer between monkeys and humans. Finally, the authors utilize motion transfer and structural similarity contrast algorithms to provide a reliable evaluation method for animal exercise courses, eliminating the subjective influence of teachers on scoring and providing experience in the combination of artificial intelligence and drama education.

1. Introduction

Animal Exercise [1] is a method of training acting created by the Soviet dramatist Stanislavsky, and it is now mostly seen in the basic courses of acting majors. Usually, in their freshman year, students take a 16-week Animal Exercise course. The learning content of the Animal Exercise course is to observe and imitate the actions of animals such as "walking, eating, sleeping, hunting", etc. During the exercises, students will imitate a large number of animals, such as clever monkeys, ferocious tigers, aggressive roosters, etc. [2]. Through the vivid imitation of animals, the flexibility of the limbs is exercised, and the body and mind can be relaxed on the stage. Some students with poor physical shape cannot meet the requirements, and it is difficult to accurately express the external characteristics of animals. At this time, a lot of physical exercises and the guidance of teachers are needed to make the animal images created by the students realistic and credible on the stage.

At present, the teaching of Animal Exercise courses is mainly based on teachers passing on the course con-

*Corresponding Author:
Yu Qi,
The Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Japan;
Email: d21200029a@edu.teu.ac.jp

DOI: https://doi.org/10.30564/jcsr.v4i2.4759

tent to students according to the theories in the textbooks and their own teaching experience. This kind of teaching method based on oral teaching will inevitably bring about teaching deviations. And due to teachers' personal preferences, there is also the problem of unfair scoring [3]. When students practice, the guidance given by different teachers is different, which will make students' thinking confused. Therefore, we need to provide a unified evaluation standard for the course, so that different teachers have a unified guideline in teaching, and provide students with an introspection environment, students can analyze the difference between their own imitation and animals, so as to get the ability to promote. We use motion transfer [4] to achieve this, transferring the original video motions of the animals to the students' own target videos.

Motion transfer technology refers to transferring the motion of the initial object in the initial motion video to the target object to produce the target motion video. Berkeley researchers have proposed a method to transfer human actions in different videos, which requires only two simple given videos, the target person, we want to synthesize his performance; the other is the original video, we want to combine his actions Transfer to the target person. However, these motion transfer are not suitable for quadrupeds, and they are all carried out in a standing posture.

In this study, we aim to provide an objective evaluation criterion for Animal Exercise courses for acting professional learners through the image similarity metric in motion transfer. Migration between humans and animals is very challenging due to several problems during the study. First of all, let's choose monkeys, which have obvious characteristics and appear relatively frequently in animal practice courses as the subject of the experiment. We utilize the coco dataset [5] to detect keypoints for students, but in the object detection dataset, we did not find monkey-related datasets. Second, in human-to-human action transfer, the mismatch of the scale of the source and target characters can cause the target characters to appear large relative to the background or surrounding objects or appear to be suspended. To address this issue, Chan et al. [6]. Devised a method for global pose normalization. They use the four coordinates of the source video character's near point, far point, nose, and ankle as the benchmark to correct the position of the target video character to make the generated target video more realistic. However, this method is only suitable for standing human posture specification, and the normalization effect for other movements such as crawling is not ideal. To solve the above two problems, we found 8 monkey videos on the open-source video website, marked 1428 monkey keypoint pictures, and created a monkey keypoint dataset named QMonkey. In order to deal with the second difficulty, we use the method of image scale filling to standardize the size of the target scale.

We summarize our contributions as follows: 1) We have created a dataset of monkey keypoints, which can provide a data basis for target detection for subsequent researchers. 2) Our experiments demonstrate that motion transfer between humans and quadrupeds is also possible. 3) We provide an evaluation standard for performance teachers when conducting animal practice tutoring.

2. Related Works

2.1 Motion Transfer

Human motion transfer refers to transferring the motion of objects in the source motion video to the target object and generating the target motion video [6]. At present, the most effective motion transfer technologies are inseparable from four steps: human pose estimation [7], training the source image generation model [8], posture normalization, and using the model to generate the person's movement in the target image. Based on Chan et al. [6], this research carried out an innovation suitable for the motion transfer between humans and monkeys.

2.1.1 Human Posture Estimation

Human posture estimation (HPE) refers to obtaining the posture of the human body from given sensor input. [9] We use OpenPose [7] to estimate the pose of monkeys and human limbs. OpenPose is a bottom-up pose estimation method. It first detects all the keypoints in the image and then uses PAF (Partial Affinity Fields) model to associate the keypoints with obtaining the correct image of the keypoints of the limbs. This method does not need to detect the object first and is very suitable for detecting keypoints of the monkey's limbs.

2.1.2 Generative Model

Since the creation of GANs by Goodfellow et al. [10], many interesting variants have emerged, some of them can transfer the style of two images [11], and some can generate high-resolution images from low-resolution images [12]. And pix2pixHD GANs [8] can generate source images with target actions. It is a variant of Conditional GANs [13] and redesigns the generation network based on pix2pix GANs [14], enabling the algorithm to complete high-quality image conversion.

2.1.3 Posture Normalization

In the human motion transfer, when the scale of the

person in the sources image and the person in the target image do not match, the target person may appear large relative to the background or surrounding objects or appear to be floating. To solve this problem, Chan et al. [6] designed a method of global posture normalization. They use the four coordinates of the source video person's near point, far point, nose, and ankle as benchmarks to correct the position of the target video person to make the generated target video more realistic. However, this method is unsuitable for non-standing monkeys, so we propose a new posture normalization method.

2.2 Dataset

COCO is large-scale object detection, segmentation, and captioning dataset [5]. It has 25,000 annotated human images, including labels for 17 keypoints such as nose, wrist, and knee. We use the COCO dataset to train the openpose model to recognize human poses. Since we also need to recognize the monkey's pose, we created a miniature monkey dataset for this experiment.

2.3 Image Similarity Index

LPIPS [15] uses depth features as a perceptual metric to judge the similarity of images. It is different from the widely used SSIM [16] and PSNR [17] evaluation indicators. It can evaluate the similarity of two images in a more human-like perceptual way. Since animal exercise is a way to improve performance, it will eventually be shown to the audience in a performance. At this time in our research, the visual similarity is significant, so we choose LPIPS as the evaluation benchmark for animal exercise.

3. Method and Experiment

3.1 Method

This experiment uses the same OpenPose parameters as Cao et al. [7] and selects the vgg19 [18] network structure. This combination is widely used in human posture detection and has a good detection effect. We use a pre-trained model for human pose detection to save experiment time. Since the QMonkey dataset (details are described in 3.2) is small, it is prone to overfitting when training monkey posture detection. We will terminate the training when the loss value reaches 0.006, taking 20,000 iterates and about 70 hours. When training the motion transfer model, parameters are the same as those of Chan et al. [6]. All training processes are performed on the Ubuntu16.04 operating system and a TITAN RTX graphics card. When using LPIPS for image similarity comparison, we choose the vgg19 network structure as the comparison parameter, which is consistent with the network structure of the OpenPose.

The evaluation system process is as follows:

1) First, use the COCO dataset to train an OpenPose model that can recognize human poses. On the other hand, use the QMonkey dataset to train an OpenPose model that can recognize monkey poses.

2) Perform posture estimation recognition on human videos and monkey videos to obtain images and posture images. We use monkey images and its posture images as the source dataset and human images and its posture images as the target dataset.

3) Train the generative model using the monkey images and its posture images.

4) Using the monkey posture images as the standard, normalization the human posture images. Simultaneously process human images.

5) Using normalized human posture images and generative models, generate human-action-based videos of monkey movements.

6) Comparing the structural similarity between the generated image and the posture image, respectively, to obtain an evaluation.

3.2 Monkey Posture Dataset

It is well known that OpenPose is very mature and accurate for human attitude recognition. Although the model trained on the human dataset can detect the pose of a few monkeys, the probability of such detection is too low to meet the needs of this experiment. Therefore, we created a new dataset of monkeys in an effective way to solve monkey pose recognition. We created a mini dataset QMonkey for monkeys. It is described in the format of the COCO dataset but only contains human-like pose information. The data from 8 different monkey videos with 1,428 images, as shown in Figure 1. Since the current dataset is small, only a few monkeys can be effectively identified, we will make it public after expanding the dataset.

3.3 Target Scale Normalization

We found from the existing animal videos that the distance between the animal and the camera is uncontrollable. When filming, the animals don't make the movements we want and don't walk within our designed range. Smaller animals may fill the screen, and larger animals may be very far from the camera. The method in [6], the everybody method, can no longer satisfy this experiment, so for the problem of mismatched target size ratios, we propose a target scale normalization method applicable in both standing and non-standing situations.

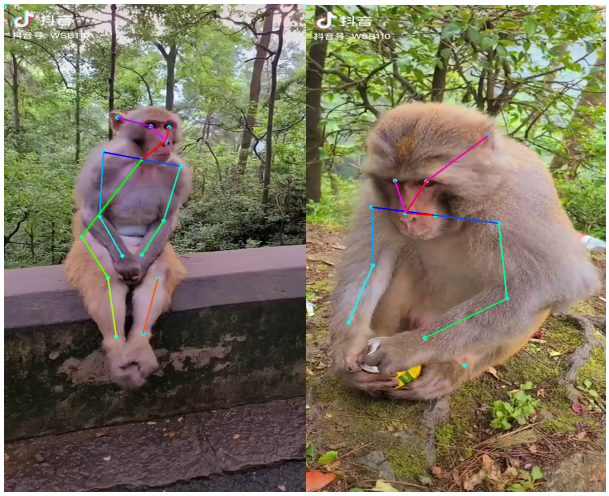

Figure 1. An Example of Images in the Qmonkey dataset

As shown in Figure 2, we record the four maximum points on the monkey posture image frame's top, bottom, left, and right. Then calculate the average width and height of the monkey posture, which are recorded as Mwidth and Mhigh, respectively. Similarly, the average width and height of the pose in the human pose map are calculated and recorded as Pwidth and Phigh, respectively. We calculate the ratio that needs to be filled or enlarged.

$$Wscale = (Pwidth / Mwidth) - 1$$

$$Hscale = (Phigh / Mhigh) - 1$$

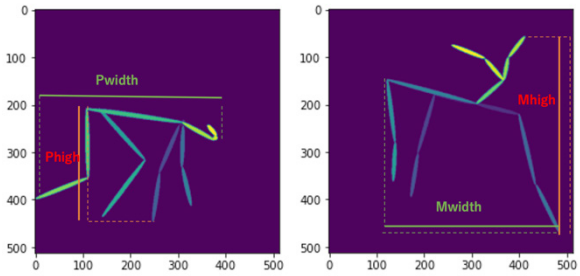

Figure 2. Human (left) and monkey (right) posture image, and their width and high

If neither Wscale nor Hscale is negative, select the larger value as the fill scale of the human posture image. If both Wscale and Hscale are negative numbers, we choose the smaller value and take the absolute value as the reduction ratio of the human posture image. When you use human images as source data and monkey images as target data to train the generative model, in that case, you can swap the human width and height with the monkey width and height in the formula.

3.4 Animal Exercise Evaluation

Here we construct the evaluation method using two sets of comparison graphs. The first group is the posture image of humans and monkeys, as shown in Figure 2, which can accurately reflect the direction of each limb and whether the degree of joint bending is similar. However, since humans and monkeys have different body proportions, we also need a second set of comparison images. Figure 3 consists of the original image of the monkey and the generated image of the human imitating the monkey. We use LPIPS to compare the two sets of images' similarity and then sum and average the scores to obtain the reference value.

Figure 3. Original image (left) and generated image of the monkey (right)

3.5 Animal Exercise Evaluation

3.5.1 Posture Normalization

We compare our scale normalization method and the everybody method of Chan et al. [6] with human images as source and monkey images as target data. The images in Figure 4 are respectively the posture image of a human imitating monkey, monkey posture image, monkey posture image normalized by our method, and monkey posture image normalized by the everybody method.

Our method downscales the monkey pose map to bring the pose scale closer to the pose images of humans imitating monkeys. The everybody method chooses the enlargement process, so that the scale of the monkey pose map and the human-imitation monkey pose image become larger, and part of the pose information is lost. It can be seen that our method is more suitable for standardized processing in non-standing postures.

3.5.2 Evaluation System

We selected two images from a video imitating a monkey for comparison in the evaluation experiments. During the monkey's walking, the hand and leg on the same side have two states, the leg moves forward close to the arm,

and the hand moves forward away from the leg.

In Figure 5, the experimenter on the left paid attention to the walking order of the limbs when imitating the monkey's walking, which was basically the same as the monkey's walking posture. The experimenter on the right walks clumsily when imitating the monkey's walking, which is not the same as the monkey's posture. After comparing the structural similarity between the two, the average scores of the generated map and the pose map are 0.309 and 0.366, respectively.

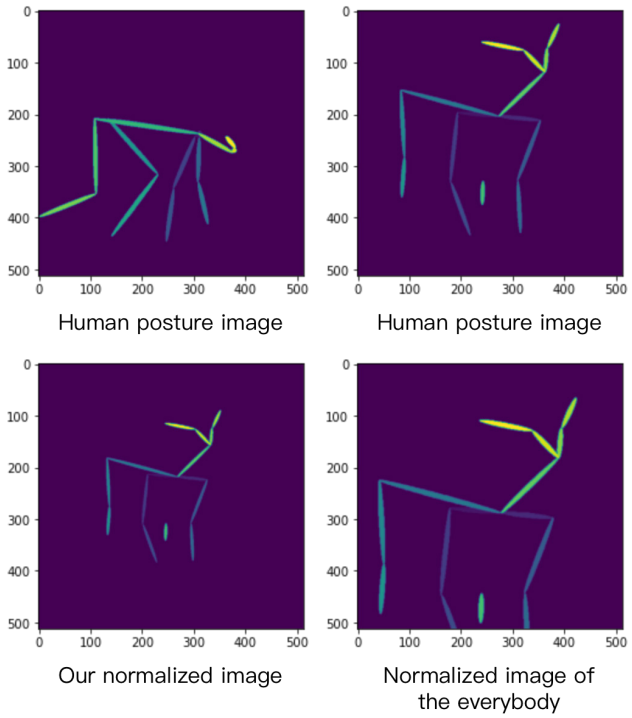

Figure 4. Comparison between our method and the everybody method

Figure 5. A comparison image of imitating the same monkey action

We surveyed 20 drama education teachers and asked them to rate 50 sets of comparison images. The teachers range in age from 28 to 70 years old, and the teaching age ranges from 1 to 41 years. Each set of comparison images contains one monkey image and ten images of humans imitating monkeys for 510 images. The evaluation scores are S, A, B, C, D from high to low. By analyzing the survey results, we came up with the evaluation criteria in Table 1.

Table 1. Scope of the proposed evaluation scope derived from the survey results

Numerical value	Evaluation
<0.3	S
0.3-0.33	A
0.34-0.37	B
0.37-0.4	C
>0.4	D

4. Results and Discussion

We found that the ability of a generative network to generate realistic images depends not only on whether the actions are mimicked the same but also on whether the limb proportions between the source and target data are similar. Although the generative network has a specific anti-interference ability and can generate images by lengthening or shortening limbs of different sizes, the movement transfer of different species increases the difficulty of generating realistic images. This research uses the target scale normalization method suitable for non-standing posture, normalizes the original data and target data to an approximate scale, and minimizes the influence of different limb scales on the model. Since we still cannot overcome the problem of limb scale changes due to camera angle, we compare the pose map and the generated map simultaneously to improve the reliability of the comparison results.

The Animal Exercise evaluation system provides teachers with a standard reference benchmark for teaching or examination, so we have adopted a grading method after research. Of course, since there is currently no action transfer between humans and animals that performs well, we cannot yet use the scores as a direct reference benchmark.

We found that whether the generation network can generate real images depends not only on whether the actions imitated are the same but also on whether the proportions of the limbs between the targets are similar. The generation network has a certain degree of anti-interference. It can stretch or shorten limbs of different sizes to generate, but this will also affect the quality of the generated image. Here we use the method of normalizing the target proportion to

make the target in the same size, try to eliminate the influence of different body proportions. However, we cannot overcome the problem of body proportion changes caused by the camera angle, as shown in Figure 5. Therefore, we compare the key point map and the generated map together to improve the reliability of the comparison result.

5. Conclusions

At present, the combination of drama education and artificial intelligence is still in its infancy, and there are a large number of technical blank areas. Many technologies can only be used from research in other fields and cannot completely solve the existing problems in drama education. The evaluation system of Animal Exercise courses fills the shortage of uneven teaching levels among teachers and too subjective teaching evaluation for Animal Exercise courses. To solve the problems, we proposed a new AI-based posture evaluation method. The Animal Exercise course evaluation system only provides teachers with an evaluation range for teaching and examination and does not entirely solve the problem of students' introspection. This is also our future work direction. Next, we will focus on developing the motion transfer algorithm between humans and monkeys to make the motion generated by the model more realistic, simplify the workflow, reduce the generation time, and facilitate the use in teaching.

Author Contributions

Yu Qi: Conceptualization; Methodology; Writing the initial draft.

Chongyang Zhang: Resources; Data Curation; Writing - Review & Editing.

Hiroyuki Kameda: Supervision.

Conflict of Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service, or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "Animal Exercise: A New Evaluation Method".

Funding

This research received no external funding.

References

[1] Stanislavski, C., 1989. An actor prepares, Routledge.

[2] Adler, S., 2000. The Art of Acting, ed. Howard Kissel (New York: Applause, 2000).

[3] Qi, Y., Zhang, C., Kameda, H., 2021. Historical summary and future development analysis of animal exercise. ICERI2021 Proceedings, 14th annual International Conference of Education, Research and Innovation. pp. 8529-8538, IATED.

[4] Aberman, K., Wu, R., Lischinski, D., et al., 2019. Learning character-agnostic motion for motion retargeting in 2d. arXiv preprint arXiv:1905.01680.

[5] Lin, T.Y., Maire, M., Belongie, S., et al., 2014. Microsoft coco: Common objects in context. European conference on computer vision. pp.740-755.

[6] Chan, C., Ginosar, S., Zhou, T., et al., 2019. Everybody dance now. Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5933-5942.

[7] Cao, Z., Simon, T., Wei, S.E., et al., 2017. Realtime-multi-person2d pose estimation using part affinity fields. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291-7299.

[8] Wang, T.C., Liu, M.Y., Zhu, J.Y., et al., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798-8807.

[9] Andriluka, M., Pishchulin, L., Gehler, P., et al., 2014. 2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on computer Vision and PatternRecognition. pp. 3686-3693.

[10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial nets. Advances in neural information processing systems. 27.

[11] Yoo, D., Kim, N., Park, S., et al., 2016. Pixel-level domain transfer. European conference on computer vision, Springer. pp. 517-532.

[12] Ledig, C., Theis, L., Huszár, F., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681-4690.

[13] Mirzaand, M., Osindero, S., 2014. Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784.

[14] Isola, P., Zhu, J.Y., Zhou, T., et al., 2017. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125-1134.

[15] Zhang, R., Isola, P., Efros, A.A., et al., 2018. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586-595.

[16] Wang, Z., Bovik, A.C., Sheikh, H.R., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 13(4), 600-612.

[17] Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. Ssim. 2010 20th international conference on pattern recognition. IEEE. pp. 2366-2369.

[18] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.