УДК 004.921

MAS MULTI-VIEW ACTION MEGRE 3D

Ding Aodi[1], Zhu Shuaiyu[2]

[1] Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, aoditing541@gmail.com
[2] Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, zhushuaiyu1001@gmail.com

In this experiment, the 2D human skeleton is detected from various viewpoints by reading in the data of athletes in 2D motion; then the geometric relationship is used to calculate the distance and match the human body from multiple viewpoints; finally, the 3D coordinates are calculated by using the 2D coordinates of the multiple viewpoints and the parameters of the camera, and the diffusion model trained on the 2D data using the MAS only requires multiple sets of different viewpoints to faster and more accurately build up the human body's motion image without any dead angle. The human motion image is constructed without any dead angle by using the diffusion model trained by MAS on 2D data.

Key words: 3D pose; MAS; human action image; multiple viewpoints; geometric.

Modelling principles and methods

Data preparation

HumanML3D is a 3D human movement language dataset derived from a combination of HumanAct12 and Amass datasets. It covers a wide range of human behaviours, such as daily activities, sports), acrobatics and artistic performances), the HumanML3D dataset consists of RBG images and spatial coordinates of human joints points, the dataset can be classified into a single-frame image dataset and a continuous frame video dataset according to whether the images are consecutive or not, and the dataset contains the joint vector data of the human motion sequences, the HumanML3D data follows the SMPL skeleton structure with 22 joints [6].

Principles of MAS Modelling

We learn a 2D motion diffusion model model from a set of videos, and then use the MAS algorithm to efficiently sample 3D motion from the learnt model. We extract valid 3D motion samples from the model. Our approach is based on the standard denoising cycle of ancestor sampling from diffusion models[5].MAS extends this concept to generate 3D motion by denoising multiple diffusion models simultaneously. Describing it. At each diffusion denoising step, all views are triangulated into individual 3D motions, which are then projected back to each view. This ensures multi-view consistency throughout the denoising process while conforming to a priori predictions. We further promote multi-view consistency by projecting 3D noise to each view whenever sampling from a 2D Gaussian distribution. We further encourage multi-view consistency by projecting a three-dimensional noise to each view whenever sampling from a Gaussian distribution during the two-dimensional ancestor sampling process [2].

The formula below models the projected motion prediction of X for all views, with the 2D sample X1 to Xv datasets denoised using G2D. At each iteration, we predict (P (X, 1), ... , P (X, V) as multi-view consistent motion [4].

$$X = \arg\min \sum_{v=1}^{V} || P(X', V) - X_0^V ||_2^2 \quad (1)$$

From the above equation we can see that an upsampling optimization x needs to be employed in order to speed up the convergence of the equation 3 Results of the experiment [1].

The following figure shows the three generated action animation software package puts the data sequence through the form of pictures to visualize the action, by filtering out the representative action behavior display.

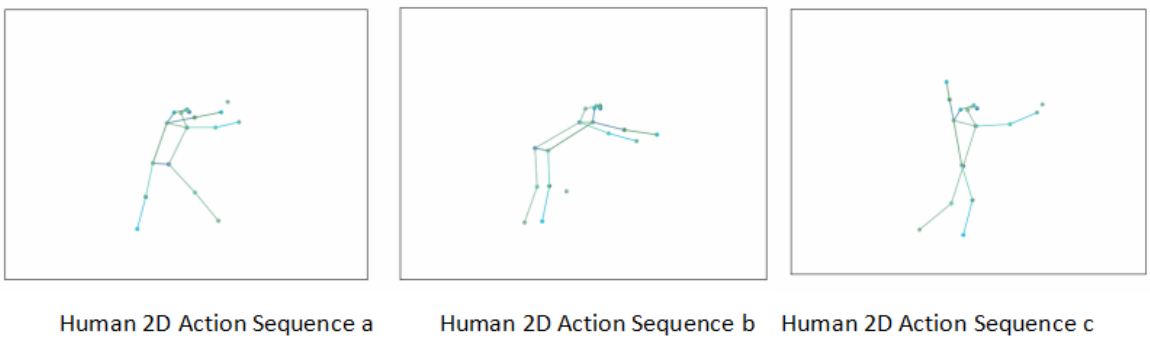

Fig. 1. Examples of three angles of gymnastic movement

From Figure 1, we can see three pictures of gymnastic movement examples, which record the coordinates of the gymnast's movements through different angles, and serve as the basic data for the next 3D movement synthesis using the MAS model.

The Figure2 below shows the results of the 3D motion sequences synthesized using the MAS model after removing the noise [3]. By looking at the results, it was found that the gymnast was in a full view of the 3D motion sequence showing the performance of the movement data from all angles.

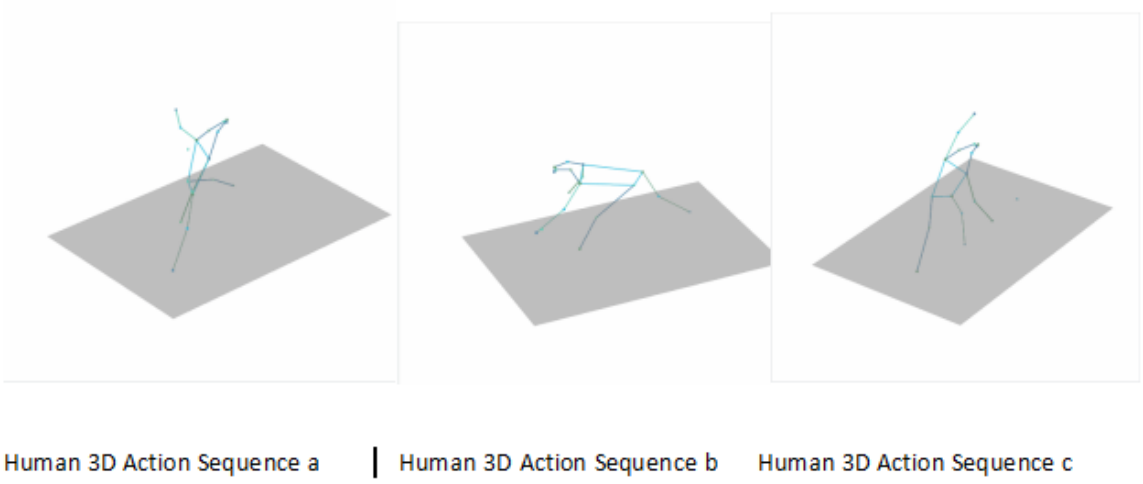

Fig. 2. 3D sequences after MAS modelling

From Figure 2 you can see the 3D video sequences generated by denoising the MAS model to complete the synthesized 3D video sequences, in order to show the video sequences I took three screenshots corresponding to the three angles of the example samples respectively [7].

Comparison of model results

In order to detect the MAS multi-angle synthesis 3D results, I added two additional models MotionBert and ElePose iterations performed 500 times, red represents this experimental research model MAS, green represents model ElePose model, blue model represents MotionBert, MotionBert is mostly used for human motion analysis, it is composed of a unified pre training phase and a task-specific optimization phase, which has recently been successfully practiced in natural language and computer vision.

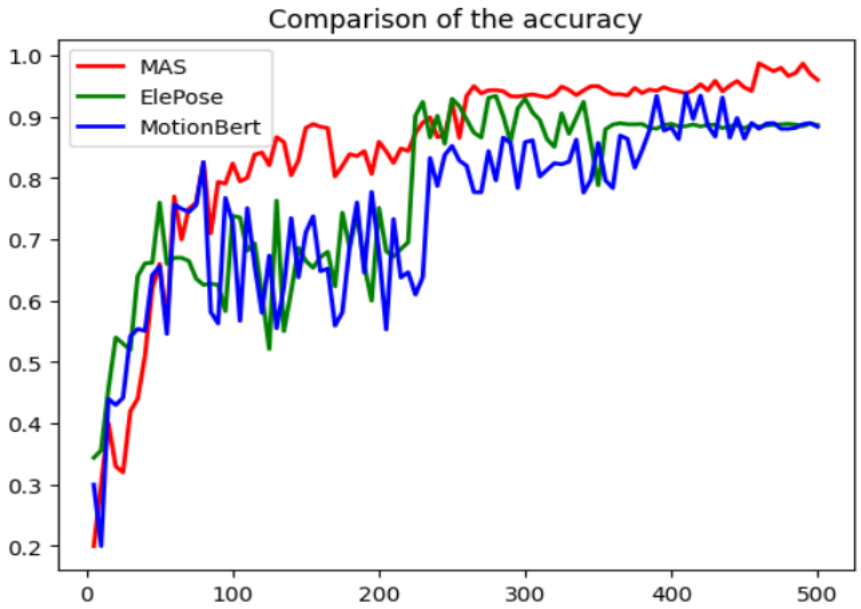

Fig. 3. Comparison of the accuracy of the three models

Figure 3, the red line segment MAS is in a stable state at iterations 300-500, with an accuracy of up to 97%, exceeding the models MotionBert and ElePose, so this experiment succeeded in achieving the task of synthesizing 3D motion from multiple angles quickly and with high accuracy.

Conclusion

In this experiment a generation method for 3D motion synthesis by MAS using multi-view 2D data is presented. By sampling multiple viewpoint diffusion models and denoising each viewpoint motion helps to generate 3D motion sequences quickly and accurately, and finally using a unique animation package to present a coherent video motion showing the results through text data. This experiment is based on the MAS method and developed an image data processing method, through the combination of 3D data animation and 2D data animation, cleverly achieve a three-dimensional and planar conversion, which focuses on the convergence of 3D sequences and 2D sequences processing and noise reduction processing. Provide an efficient and smooth method to achieve the combination of stereoscopic and planar video.

Reference

1. Learning diverse stochastic human-action generators by learning smooth latent transitions / Z. Wang [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence, v. 34, pp. 12281-12288, 2020. 2

2. Yamada T., Matsunaga H., Ogata T. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters, 3(4): pp. 3441-3448, 2018. 1, 2

3. Structure-aware human-action generation / P. Yu [et al.] // In European Conference on Computer Vision, pp. 18-34. Springer, 2020. 2

4. Memory-oriented decoder for light field salient object detection / M. Zhang [et al.] // NeurIPS, pp. 896-906, 2019. 2

5. On the continuity of rotation representations in neural networks / Y. Zhou [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5745-5753, 2019. 6

6. Kim Jh., Kim Js., Choi S. Flame: Freeform language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349, 2022. 2

7. Zero-1-to-3: Zero-shot one image to 3d object / R. Liu [et al.]. 2023. 3