IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Motion Capture Using Deep Learning

A.L Dhaigude , K.D . Dhamane. D.S Kale.

Final year student

Under guidance of Prof.P.S.Hanwante

Department of Artificial Intelligence & Data science

Anantrao Pawar College Of Engineering And Research, Pune

ABSTRACT

Generating animation from capture human motion. important problem in computer graphics. The creation of these virtual characters is different from traditional 3D animation but is based on real character movements and expressions. An overview of several mainstream motion capture systems in the field of motion capture is presented, and the application of motion capture technology in film and animation is explained in detail. The current motion capture technology is mainly based on complex human markers and sensors, which are costly, while deep-learning-based human pose estimation is becoming a new option. However, most existing methods are based on a single person or picture estimation, and there are many challenges for video multiples on estimation. The experimental results show that a simple design of the human motion capture system is achieved.

I. INTRODUCTION

As computational power has steadily risen over the years, so, too, has the realism and complexity of computer-generated scenes and animations. The complexity and computational Needs of techniques used by visual effects artists has kept pace with improvements in computation speed as summarized by Blink' s Law, which states " as technology advances, rendering Time remains constant" . For example, a single frame from a recent animated film could Range from several hours to several days to render on modern hardware.

In 1995, Pixar' s Original Toy Story required similar render times on hardware from that time. However, if Rendered on current-day machines, the film would take a fraction of the time to render. Although image rendering takes a significant portion of the computation time spent Generating a movie, other aspects of the film have grown in complexity as well.

Character Mesh deformations, for example, have also become more computationally demanding over The years. These mesh deformations are driven by character rigs, which controls how a Mesh is deformed according to a set of input parameters. As the detail and quality of mesh Deformations grow, so, too, does the complexity of the character rig.

At dream Works Animation, for example, character rigs were so complex that they were Unable to evaluate at interactive rates before the development of Libee and Premo, their Current in-house animation software. Previously, animators would enter numbers in a Spreadsheet and would wait for their workstation compute the deformed mesh and update A character on their screen at non-interactive rates.

To keep up with the growing complexity Of character rigs, they developed their current software to utilize multi-threaded hardware on High-end computing machines. As a result, animators are now able to adjust rig parameters And see the changes in the deformed character mesh in real-time, which can increase their Productivity. Despite these improvements, artists still require a significant amount of time to Produce a high-quality character animation. For example, one artist might spend a week of Effort to author 5 to 10 seconds of animation for a feature film.

METHODOLOGY

1 Processing Pipeline

Motion capturing is the process of recording the movement of objects or people. In recent years, many systems and approaches have been proposed for capturing human activities, and can be categorized in two types, as follows. Marker-based motion capture. Currently, marker-based motion capturing is a mature technique that has been used.successfully in many fields such as the motion picture industry and VR. However, this method is disadvantaged by the controller having to be worn as a marker suit with sensors such as optical mounted cameras [9]. Thus, this makes it impossible to capture the movement of people wearing ordinary clothes. Additionally, marker based motion capturing is sensitive to skin movement relative to the underlying bone [10]. Moreover, the exact placement of markers on anatomical landmarks is difficult to realize, and markers placed on the skin do not directly correspond to the

3D joint positions. Presently, many commercial automatic PyTorchic systems, developed by companies such as the Captury, Organic Motion, and Simi, can be used to investigate human motion. However, these system mostly employ multi-view cameras that can deduce the 3D position of the objects and body skeleton. Although the specifications of these systems have noticeable differences [11], the same underlying principles apply in terms of several points of interest being located in sequential images, converted to real-space coordinates, and used to infer the 3D pose of the underlying skeleton.

Markerlessmotion capture. A markerless system that can address the limitations and eliminate the need of body-worn sensors has attracted a substantial amount of attention in the field of computer vision and computer graphics, and has expanded the application of human motion capturing. Recently, this field has increasingly attracted more interest from researchers.The four major components of a markerless motion capturing system are (1) the camera systems being used, (2) the representation of the human body (the body model), (3) the image features being used, and (4) the algorithms used to determine the parameters (shape, pose, and location) of the body model [12]. There are two types of camera systems for markerless motion capturing, and can be distinguished according to whether or not a " depth map" is estimated. Currently, motion capturing systems based on depth-sensing are considered as an effective method of estimating a fullbody pose in real-time, for example, in [13], [14]. From an algorithmic viewpoint, markerless motion reconstruction can be classified into two main categories; discriminative approaches [15], [16], [17] and generative approaches [18], [19], [20] .

Discriminative approaches. In this field, the idea behind discriminative algorithms is to convert the motion capturing problem into a regression or pose classification problem, so as to achieve a much faster processing time, improve robustness, and reduce the dependence on initial guesses. Discriminative algorithms can be divided in two major groups. One approach is to discover a mapping directly from the image features to a description of the pose, such as by using machine-learning based regression [21], [22]. Alternatively, a dataset of pose examples can be created and then searched to discover the most similar known pose if the current image is given, as has been done in previous studies [23], [24]. However, discriminative algorithms have reduced accuracy and require a very large set of training data, which makes their application difficult. Therefore, the discriminative approaches are typically used as the initial guesses in generative approaches [26]. Generative approaches. In the generative motion capturing approaches, the pose and shape of the body are acquired by fitting the model to information extracted from images

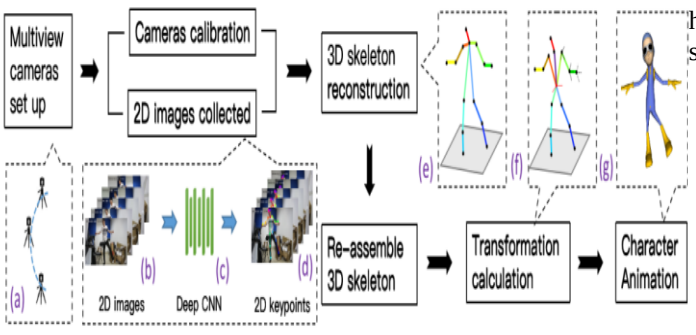


Fig 1.1. The processing pipeline. (a) multi-view cameras; (b) images captured; (c) 2D joint keypoints detector; (d) output of 2D joint keypoints; (e) 3D skeletion; (f) reassembled 3D skeleton and bone transformation; (g) animation of a 3D character.

This can generate a set of model parameters such as the body shape, bone length, and joint angle. Alternatively, the 3D body model can be compared against a 3D reconstruction, such as a visual hull, by minimizing the distances between the 3D vertices of the model and the 3D points of the visual hull [27] through a standard algorithm known as the iterative closest point. Contrary to discriminative algorithms, generative approaches are typically based on temporal information and can solve a tracking problem.

The majority of these methods parameterize the high dimensional human body and embed a low dimensional skeleton into the body model.

1 CAPTURING OF 2D IMAGES FROMCALIBRATED MULTIVIEW CAMERAS

First, the experimental environment is set up for body motion capturing. After their positions are fixed, all cameras are calibrated using fundamental matrix estimation for pairs of images. This is followed by bundle adjustment, which is a feature-based 3D reconstruction algorithm[40], we achieve the reconstruction of the 3D sample space, and obtain all of the camera parameters, including the intrinsic matrix and extrinsic matrix. Additionally, the above mentioned calibration procedure only has to be carried out once for each camera layout.

.2 2D JOINT DETECTION FROM MULTI-VIEW 2D IMAGES

Here, we describe our method of capturing the 2D human pose without landmarks, as shown in Fig. 2. In this part, a deep CNN detector from a state-of-the-art open-source 2D pose estimation library [17] is used to perform the detection of the 2D human pose keypoints. As is well known, as the number of cameras increases, the obtained reconstruction accuracy becomes higher. Additionally, using the Deep CNN method is time- and resource-consuming. Hence, using many cameras requires a substantial amount of computational resources. Therefore, to balance the calculation speed and quality of 3D joint reconstruction, a procedure of merging multi-view images into a single image to be fed into the deep CNN is designed to speed up the calculation and improve the synchronization of the output 2D poses. The details of the process are shown in Fig. 2. First, we down-sample all images and merge them into a single image. Next, by performing 2D pose detection, a series of 2D joints are obtained. Finally, the operation of upsampling and reordering consist of scaling the 2D joint into the original input image size and matching the corresponding camera id. After the above mentioned phases have been completed, we obtain the key joints of the 2D skeleton from all views.

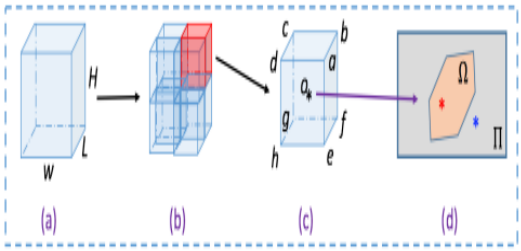


Fig. 1.2. Estimation of 3D joints using 2D joints. (a) a 3D sample space is initialized with width W, height H, and length L; (b) the space is subdivided; (c) one subspace is considered as a

3D sample unit; (d) the area Ω of the 3D sample unit is projected onto image Π.

Additionally, we defined a 3D sample cube with width w, height h, and length l as Cube{w,h,l}, and its center point with p 3 . project(p 3 , κ) represents the projection from p 3 to p 2 , where κ denotes the camera parameters. Thus, the projection area Ωi of the 3D sample cube in the i-th image is defined as follows:

$\Omega i=\{P^2|p^2 = \text{project}(P^3,K), P^3 \in \text{cube}\}$

Because the cube projection on the 2D plane is convex, its projection area can be calculated easily by its eight vertices {a, · · ·, g}. Accordingly, the foreground mask, known as a 2D joint keypoint, is the 2D projection of the corresponding 3D foreground object. Along with the cameras' viewing parameters, the 2D point p 2 defines a back-projected generalized line containing the actual 3D joint p 3 , as shown in Fig. 5.

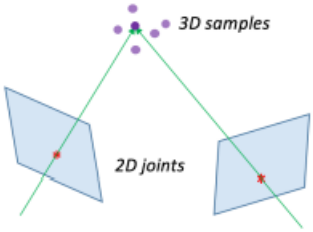
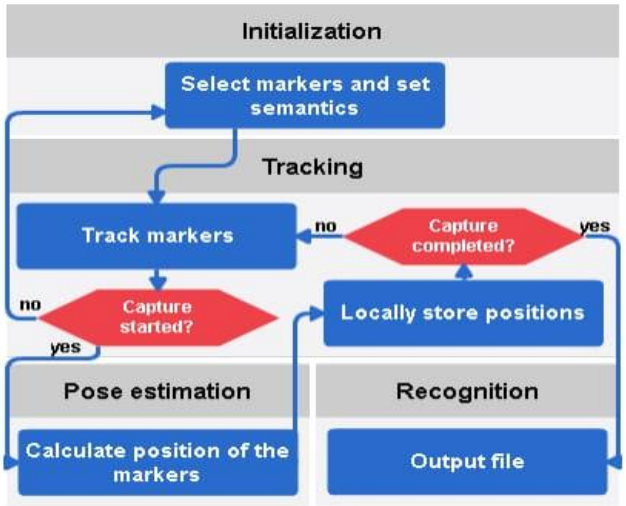


Fig.1.5. Projection that is a bounding geometry of the actual 3D point. Purple points are the 3D samples. Red stars are the 2D joints detected by the deep CNN

FLOWCHART



All the graphic user interface present in the project was implemented using the model of Interface Control and highlight input devices, present in the OpenCV library. Two windows are constructed in the beginning of the system, one called "Original", that show the frames obtained by the camera, and another called "Trackbar", with the controls of the reach of the image in color system HSV.

2 RESULTS

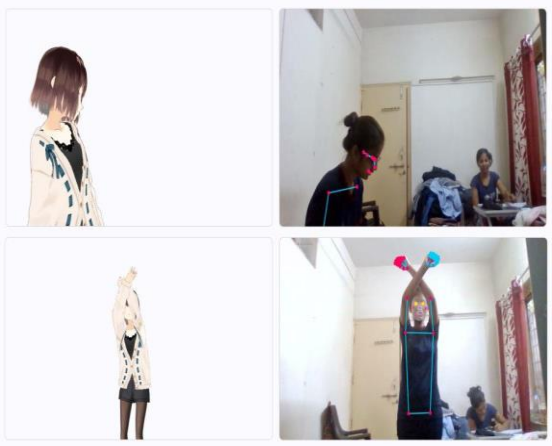
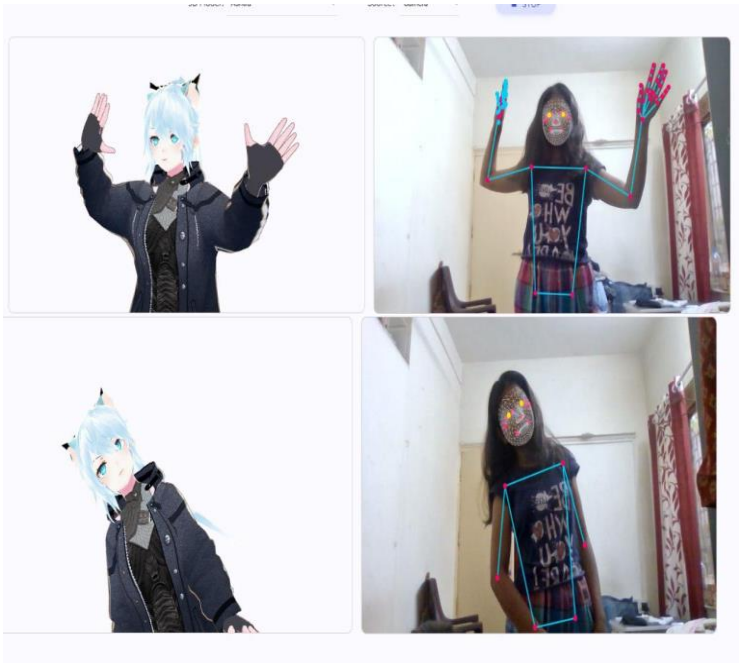


Fig. 1.4. Some experimental results on Humaneva-I dataset.

This paper proposes a new system for markerless human body motion capturing and animated character rigging using multi-view cameras. According to the experimental results, our system can produce accurate and robust 3D human body joints from multi-view camera images, which can then be used to rig 3D characters for animation. This system may be used in fields such as animation production, video game production, and VR game interaction. Thus, the production costs can be reduced and the human-machine interaction can be simplified considerably. However, the development of the proposed system is in the prototype stage, and there still exist issues that require further investigation, such as the stability of the calculated 3D joints, number and layout of the cameras, and efficiency of the entire system. In future work, we will focus on improving the reconstruction of the 3D joints and the capturing of body motion.The Avg 2D Err in multi-views. In general, the average 2D error of each view is kept within a small ranges. The very high intensity is the noise, when 3D points or 2D CNN points are not detected. Frames 1-4 represent the case that no 3D points are detected



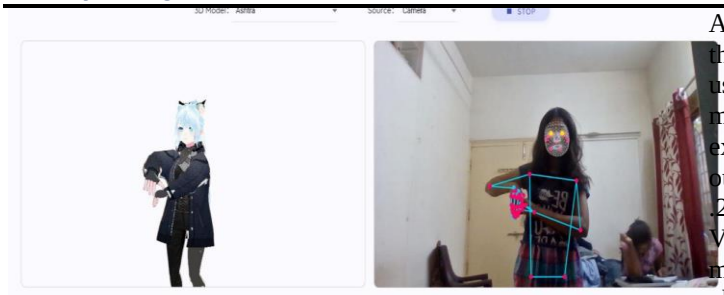
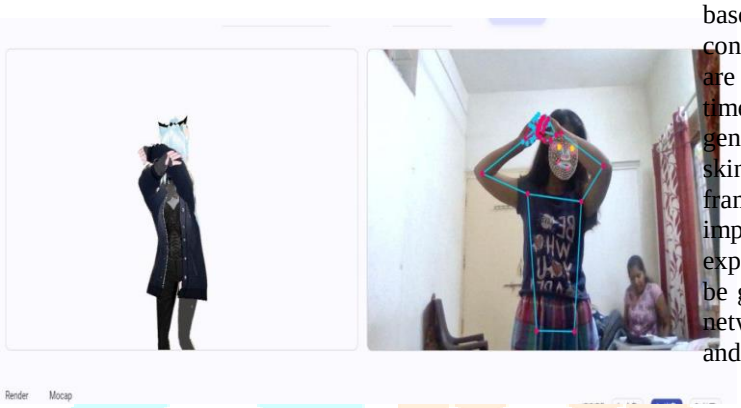


Fig. 1.6.. Xtion failure cases for which correct results were obtained by our system: (a) wrong bone transformation for right upper arm; (b) tracking failure (right upper arm); (c) tracking failure (with interference from other parts); (d) tracking failure (limb overlap)



Experimental analysis

In this part, we present more analysis about our system itself based on some additional experiments

Frame	View 0	View 1	View 2	View 3	View 4	Overall
10	19.7	11.8	13.8	13.2	24.3	16.5
20	17.9	12.6	12.4	9.7	23.3	15.1
30	28.1	18.0	18.6	16.2	23.7	20.9
mean	21.9	14.1	14.9	13.0	23.7	17.5

TABLE V

THE Avg 2D Err OF THE 2D REPROJECTION POINTS FROM THE 3D VOXEL AND 2D JOINTS WERE DETECTED BY THE CNN IN FIG. 15.

1) Accuracy analysis of our system: In order to evaluate the accuracy of our method, we define the average 2D error (Avg 2D Err) directly in the image in pixels (pix). The Avg 2D Err between the estimated pose x and the ground truth pose xˆ is expressed as the average Euclidean distance [43] between individual markers: d 2 (x, xˆ) = 1 n Xni=1 ||xi − xˆi ||, (3) where n is the number of joints, i = 1 . . . n, x is the 2D joint directly obtained from the 2D CNN detector, and xˆ is the 2D reprojected joint of the calculated 3D joint. Fig. 15 shows the comparison of the reprojected points between the calculated 3D joint points and the results directly obtained by the 2D CNN detector. Table V presents the results of calculating the Avg 2D Err to evaluate the effectiveness of the reconstruction joint in 3D space. The value (pixel distance) of each view is the average error of all 2D joints. Here, the image resolution is 1920×1080. The ratio of Avg 2D Err and image resolution is under 2%. The accuracy of the cameras' intrinsic and extrinsic parameters was not good, and this led to the Avg 2D Err ranges by view. From the comparison of one view, such as frame 10, view 0 in Fig. 15, it can be seen that the proposed 3D voxel joint estimation method can precisely reconstruct the key joints of the body. Additionally, the bone transformation calculation can obtain the correct bone pose in 3D space, and the results can be used to animate multiple 3D characters. Fig. 16 shows the multi-views Avg 2D Err value of consecutive frames in our example video sequence as shown in Fig. 15. It shows that our 3D joint estimation method is stable and reliable

2) Time consumption analysis of the whole pipeline: Table VI shows the time consumption of our pipeline in millisecond (ms) by 4 main phases. The image collection phase costs a lot for image preprocessing in a high frame rate (about 30fps). The second phase (2D joint detection) takes about 120ms, which is the maximum in all the phases, because the deep CNN is very time-consuming due to many large computational processes. Nevertheless, the 2D joint detection using CNN is still much faster than the method based on feature correspondence. Additionally, another time-consuming phase is 3D joint estimation, in which 18 joints are calculated one after another and the average processing time of each joint is about 5.6ms. The final phase is for generating character animation, including rigging and skinning, which usually takes less than 5ms. For now, the frame rate of our system can achieve 5fps by a preliminary implementation without any optimization in the experimental environment. In the future, the frame rate could be greatly improved by following strategies: optimizing the network structure of the deep CNN for 2D joints detection and adopting

3) The effect of the resolution of input images: In the process of 2D joint detection described in Section III-B, the input images captured by cameras are down-sampled and merged into a single image. In order to analyze the effect on

Merging & Down-sampling	No	Yes
Resolution (pix)	640 × 480	427 × 480
Time (ms)	633	**181**
Mean Abs 3D Err (mm)	38.6	52.1

TABLE VII

THE EXPERIMENTAL RESULTS ABOUT THE EFFECT OF DOWN-SAMPLINGAND MERGING INPUT IMAGES.

images separately, and the time consumption and accuracy of the results were compared with those of the proposed method. In the experiment, δ and σ were set to [20, 20, 20] and 4 respectively, and Box sequence of S1 in the HumanEva-I dataset was chosen as the test data. From the results listed in Table VII, it can be found that the input image resolution is responsible for the 3D joint estimation accuracy. Indeed, down-sampling operation will lead to a little increase in the 3D joint estimation error. However, merging input images into one big image makes the time consumption of the pipeline reduced dramatically. Generally speaking, it' s worth sacrificing a little accuracy to improve efficiency greatly.

4) More analysis of 3D joint estimation: The time consumption andresult quality (evaluated by Mean Abs 3D Err) of the 3D joint estimation phase were analyzed by a further experiment with different δ settings, in which the merged image resolution was set to 1280 × 960 and σ was set to 4. From the results listed in Table VIII, it can be found that the sample cube size δ plays an important role in terms of time consumption. When δ decreases, the time consumption increases dramatically. However, Mean Abs 3D Err is not sensitive to the change of δ. When δ alters, Mean Abs 3D Err doesn't change much. Therefore, it is not difficult to find a trade-off between computing precision and speed with appropriate parameters.

TABLE 1: Taxonomy of the state-of-the-art image-based 3D object reconstruction using deep learning.

Input	Training	1 vs. muli RGB, 3D ground truth, Segmentation.	One vs. multiple objects, Uniform vs. cluttered background.
	Testing	1 vs. muli RGB, Segmentation	
Output	Volumetric	High vs. low resolution	
	Surface	Parameterization, template deformation, Point cloud.	
	Direct vs. intermediating		
Network architec-ture	**Architecture at training**		**Architecture at testing**
	Encoder - Decoder TL-Net (Conditional) GAN		Encoder - Decoder
	3D-VAE-GAN		3D-VAE
Training	Degree of supervision	2D vs. 3D supervision. Weak supervision.	
		Loss functions.	
	Training procedure	Adversarial training. Joint 2D-3D embedding. Joint training with other tasks.	

The Image reconstruction table using deep

Conclusion

Due to its overwhelming characteristics, deep learning has been widely used in various research domains. From the current research trend, deep learning technology has very good application prospects. In this paper, we have considered the autonomous generation of facial expressions of 3D animated characters as the main research content along with improving the progress of deep learning. We have designed a method for generating facial expressions of animated characters, which is based on deep learning, and a localization experiment of 3D animation facial expression features is carried out to verify the operational superiority of the proposed model. Finally, through the matching experiments, it is proved that the 7proposed 3D animation facial expression generation method has a very good facial feature recognition effect. Additionally, generated facial expression features are more detailed, which has fully solved the problems associated with the traditional method and lay a solid foundation for the further development of the 3D animation design field. Recent advances in deep learning, and deep learning in particular, have provided new tools to apply to problems in character animation. To address growing complexity of filM quality character rigs, I have proposed methods to compress the computational cost of evaluating mesh deformations. Previously, these types of rigs have been specialized to individual films. In some film studios, these characters might even be inaccessible in future projects due to incompatibilities with updated animation software.

However, my proposed methods offer a solution to these common challenges with character rigs. First, my approach reduces the computational complexity of character rigs so that they can be evaluated realtime on low-powered, consumer-quality devices. As a result, my approach can increase the level of complexity of characters in games and interactive applications. Second, because these rig approximations are implemented as neural networks, character rigs can now be expressed as a fixedlength set of model parameters. This rep-presentation provides a common format in which any character can be expressed. Because deep learning libraries and packages are readily available, applications that evaluate these models can easily be written. Once trained, these approximation models no longer depend on the original animation software used to create the rig. The model parameters can also be used as an archival method for characters authored on outdated rigging software. As Another benefit, the models allow for character sharing between animation studios in cases where sharing their proprietary rigging software is an impracticality. Additionally, I have proposed tools for character control and assisted animation authoring. As digital characters continue to grow in complexity, animators continually spend more effort to control additional character details to achieve an ever-increasing level of realism and expressiveness. Although automated methods may never match the quality of artist-created animations, I have developed methods that allow artists to control coarse movements and deformations of a character so that they can focus their energy on finercharacters quickly through manipulation of a small set of control points rather than a long list of rig scale details that make an animation believable. The inverse kinematics methods allow artists to pose parameters.

Reference

1. Weiss M., Reich E., Grund S., Mülling C. K. W., Geiger S. M. Validation of 2 noninvasive, marker less reconstruction techniques in biplane high-speed fluoroscopy for 3-dimensional research of bovine distal limb kinematics. Journal of Dairy Science .2017;100(10):8372– 8384. doi: 10.3168/jds.201712563.

2. Zhang G., Lu R., Guo Q., Liu Y. Combining path-and-posture planning in a 3D environment. Tenth International Conference on Digital Image Processing (ICDIP 2018) .2018;22 doi: 10.1117/12.2503256.

Zhang L. Application research of automatic generation technology for 3D animation based on UE4 engine in marine animation. Journal of Coastal Research .2019;93(sp1):p. 652. doi: 10.2112/si93088.1.

Bian S., Zheng A., Gao L., et al. Zhang; Fully Automatic Facial Deformation Transfer. Symmetry .

Norris J., Creveling W., Porter E., Vassel E. Tracking individual targets for high density crowd scenes analysis. Current Urban Studies .2020;3

Chen S., Jiu Z. A method of stereoscopic display for dynamic 3D graphics on android platform. Journal of Web Engineering .2020;45 doi: 10.13052/jwe 1540-9589.195612.

Zhang X. U. H. U. I. A design method of group Animation fusion motion capturing data based on virtual reality technology. Proceedings of the 2021 13th International Conference on Measuring Technology And Mechatronics Automation; January 2021; Beihai, China.

Li X., Han Q., Zhang G. Large-size sprocket repairing based on robotic GMAW additive manufacturing. Welding in the World .2021;65(5):793– 805. doi: 10.1007/s40194-021-01080-9.

Bertiche H., Madadi M., Tylson E., Escalera S. DeePSD: Automatic deep skinning and pose space deformation for 3D garment animation. 2021..

Kim K., Park S., Lee J., et al. Large-scale Animation CelebFaces dataset via controllable 3D synthetic models. 2021.

Yin L., Yu T., Zhou L. Design of A Novel smart generation controller based on deep Q learning for large-scale interconnected power systems. Journal of Energy Engineering .2018;144(3) doi: 10.1061/(asce)ey.1943-7897.0000519.04018033