Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230073

A review of motion generation technology

Zhe Yang

Shanghai DianJi University, Shanghai, China, 201306

yz91570@163.com

Abstract. Nowadays, deep learning and neural network-related research play a very important role in the widely use of artificial intelligence -related technologies, Among them, the hot development in the direction of generative adversarial networks (GAN) has given birth to many generation-related techniques. For example, MoCoGAN is based on the implementation principle of GAN, which enables video generation of different actions of the same character or the same action of different characters, through the method that decompose video into actions and content. This paper introduces the history and principles of MoCoGAN, starting from the prospect of using MoCoGAN in artificial intelligence (AI) industry and the technical challenges that need to be overcome in the future application of action generation. Besides, this paper also discusses the two main issues of how to improve the quality of video generation using MoCoGAN and the input conditions that are the most central problem to GAN networks. By summarizing the optimization solutions of other researchers in these two areas in recent years, this paper searches the core problems need to be solved and propose a broad prospect for future video generation techniques that can be implemented by using MoCoGAN in human-computer interaction (HCI) area.

Keywords: Artificial Intelligence, Deep Learning, MoCoGAN, Generate Motion, HCI.

1. Introduction

With the rapid development of artificial intelligence (AI) industry nowadays, deep learning and neural network related applications are constantly being developed. After generative adversarial networks (GAN) networks [1] was proposed in 2014, its application in the field of generation has gradually become one of the hottest directions of deep learning in recent years. Many scholars have used the principles of GAN networks to study and modify them to fit their own experimental directions. Initially, GAN was mostly used for data and image generation, such as pix2pix, cycleGAN, etc. Subsequently, applications in the direction of video were gradually developed. Caroline Chan's "everybody dance now", published in 2018, uses the principles of GAN to achieve the effect of motion migration [2], unlike the migration of actions, MoCoGAN implements the generation of actions. The technique of generating movements is of great importance in the field of human posture estimation. If it is possible to generate a variety of movements without any apparent contradiction, the generated actions can then be used for the creation of motion datasets. This can effectively solve the current problem of small variety of samples and large production cost of motion datasets in the field of human posture estimation. It is even possible to use video generation technology in conjunction with human-computer interaction (HCI) and industries that are in the early stages of development, such as Virtual Reality (VR) or Mixed Reality (MR). This will broaden the idea of the development of new technologies and show more clearly

DOI: 10.54254/2755-2721/30/20230073

the interaction between man and machine that can be improved in operation. In a word, video generation technology will have a very significant impact on this area.

This article reviews the development process from the origin of GAN networks to MoCoGAN, introduces the principle of MoCoGAN, its application direction, and the effect of its implementation so far; summarizes the current defects of MoCoGAN and how they have been optimized in recent researches; this paper also discusses the importance of solving these deficiencies and the prospect of application afterwards.

2. History of GAN network development

2.1. The history of GAN networks to MoCoGAN networks

Since the concept of GAN networks was first proposed in 2014, the field of generative adversarial networks has rapidly become one of the hottest research directions in recent years. By making the generators and discriminators trained against, it is possible to use GAN networks to generate many items, characters, scenes, actions, etc. that would not otherwise exist. This technology has made many remarkable contributions to the artificial intelligence area.

Due to the powerful functionality of GAN networks, their applications in several research directions in artificial intelligence, such as image, video, and speech generation, are considered to have good prospects for development. Although GAN networks have been proven to have excellent practical value in static object generation, such as the convenience of AI painting for designers, and the migration of painting style to provide more possibilities for pre-trained materials needed for autonomous driving, there are still many problems that need to be optimized in the generation of dynamic video, action and other samples. Therefore, this paper focuses on the application of GAN networks in video generation.

As shown in Table 1, this paper summarizes some of the variants of GAN networks for video generation in recent years. In 2016 Carl Vondrick et al. [3] proposed a VGAN that generates coupling by dividing the video into foreground and background, i.e., Decompose the video into motion foreground and static background. The study of VGAN for both static and dynamic parts of video scenes also provides important ideas for MoCoGAN, which is the main focus of this paper. In 2016, Yipin Zhou et al. [4] studied RNN-GAN, which takes the time series of video as the pointcut, and uses the temporal modeling capability of Recurrent Neural Network (RNN) to predict and generate objects in the future state. The concept of RNN-GAN can also be used to develop in Natural Language Processing (NLP), music generation, etc. as well. Also using time series as an entry point is TGAN proposed by Masaki Saito in 2017 [5]. Combining the concepts of VGAN and TGAN, in 2018 Sergey Tulyakov et al. [6] proposed the MoCoGAN network for generating videos. With reference to the idea of VGAN and TGAN mentioned above decomposes the video into two parts, motion and content, and also imports the concept of time series. With these processes, videos with the same content and different actions or the same action and different content can be successfully generated.

Table 1. GAN network development in the direction of video generation.

Name	Year
VGAN	2016
RNN-GAN	2016
TGAN	2017
MoCoGAN	2018

DOI: 10.54254/2755-2721/30/20230073

2.2. The theory of MoCoGAN network and its effect

The framework of MoCoGAN network is shown in Figure 1 below, which mainly consists of four sub-networks: the recurrent neural network RM, The image generator GI, The image generator DI and The video discriminator DV. DI and DV will act as discriminator to judge the videos generated by GI and RM. The core idea of MoCoGAN is to divide the potential space of an image into two potential spaces of content and motion, and by sampling different points in the content space or different motion trajectories in the motion space, it can generate videos of the same object performing different movements or videos of different objects doing the same action. The videos generated by using MoCoGAN can be used not only for generating actions but also for changing facial expressions, etc. Despite the partial shortcomings and deficiencies of this technique in the action generation area, it has already been tested on the Taiji dataset and it has been shown to successfully generate simple action videos. Also, it has better results in changing human facial expressions.

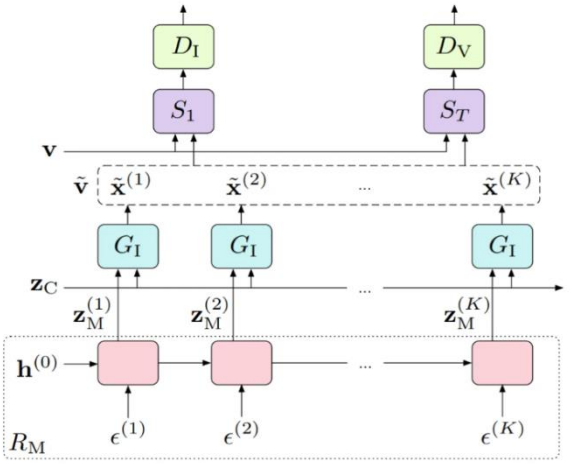

Figure 1. The MoCoGAN framework for video generation [6].

3. Deficiencies of MoCoGAN and optimization methods

This paper summarizes the deficiencies or shortcomings of MoCoGAN network at the present phase, and by comparing the optimization methods proposed by other scholars for MoCoGAN video generation, this paper summarizes the improvement directions that can be attempted in the future and look into what problems can be solved by using MoCoGAN to generate the video or which field the MoCoGAN video generation technique can be applied the future.

3.1. Deficiencies of MoCoGAN

Although MoCoGAN has definitely evolved in video generation, showing better test results on datasets such as Taiji, Moving MNIST, etc., there are still some deficiencies exist in the field of video generation using MoCoGAN, and the main problems exist in these following directions.

(1) Unstable video quality. This problem is the most serious problem of MoCoGAN, the generated video may be blurred, distorted or even incoherent, which is a very fatal problem for video, before discussing how to generate more complex videos with more elements, in the first place, this problem will affect the visual experience of users. So, one of the most important optimization directions so far is to solve the basic video quality issues like improve the clarity of the generated video, solve the frame drops problems reduce the sense of dissonance and so on.

(2) Weak semantic consistency. The main problem is that the generated video may violate the rules of real-world physics, and its impact is particularly severe in the area of human posture estimation. The generated movements may have misaligned joints, reversed joints, etc., which can be a serious violation. This is a problem that must be solved if the videos generated by MoCoGAN want to be used in the

DOI: 10.54254/2755-2721/30/20230073

future. Same as the basic properties of video such as sharpness and frame rate, the generated videos must conform to the physical rules, which also is a key factor to improve the video quality.

(3) Sensitivity to input conditions. Because GAN networks are based on the principle of generators and discriminators against generation, they are very sensitive to the input conditions. If the input conditions are not accurate or sufficient, the generated videos may lack diversity or unable to meet the expected requirements. In recent years, in addition to studying how to improve the input conditions in order to generate videos that meet the requirements, some researchers have also put emphasis on the pursuit of convenience by focusing on zero-shot learning and few-shot learning, i.e., not relying on samples or using as few samples as possible to generate videos that meet their requirements.

(4) High time and resource consumption. Because of the underlying framework of MoCoGAN, which is a deep learning model evolved from a complex generative adversarial network, a large amount of computational resources and training time are required to generate videos that can meet user requirements, which means if one uses normal device specifications to generate a high quality video, it takes a lot of time to generate and the computer heats up, and on a practical level, generating video in real time is not very likely.

3.2. Optimization methods designed by other researchers in recent years

Up to now, many researchers have proposed methods to optimize different aspects of MoCoGAN, and due to the above mentioned shortcomings and deficiencies in hardware issues are not the focus of the discussion, so this paper summarizes the most important directions, i.e., the quality of the generated video and the output conditions on which the GAN network is most dependent, and what researchers have done in recent years to optimize the MoCoGAN neural network framework to produce meaningful effects on the generated video. Yatin Dandi et al. [7] in 2020 proposed a modeling technique for joint training of image and video generation that can improve the quality of videos by pre-training a video generation model allowing videos to share information flexibly across frames, by applying this approach to MoCoGAN, which effectively reduces the FVD score [8] of generated videos at 2AFC, proving that it can effectively improve the quality of the generated videos. Kawamoto et al. [9] started from the perspective of input conditions by incorporating the learning method of CGAN [10] into MoCoGAN, subdividing the potential space into more action-related classes and using condition-MoCoGAN to try to reach zero-sample learning for video generation. The results using this condition-MoCoGAN on the Weizmann action database demonstrate that although the zero-sample generated videos are not satisfactory in terms of quality obtained, there is feasibility, the feasibility of implementation still exists. In addition, the generation with class-conditional has a better balance than the video generated by MoCoGAN alone. Others like G3AN proposed by Yaohui Wang in 2020 [11]. Although is different from the implementation method by using MoCoGAN, they borrow the decomposition concept of MoCoGAN, use a three-stream Generator to model appearance and motion separately, and uses a novel spatio-temporal fusion method to generate videos that are more realistic and can handle high resolution video generation tasks. There is also a deep learning framework proposed by Siarohin A. et al. in 2019 [12], which consists by three parts, an unsupervisely trained keypoint Detector, a dense-motion prediction network, and a Motion Transfer Network. It enables image to video conversion generation in the form of motion prediction. Very positive results were achieved on the Taiji dataset. The two mentioned above are also two typical cases in the direction of video generation, from which we can conclude that the current optimization direction in the field of video generation is still focused on these two main issues of video quality and compliance with the laws of physics. Though GAN networks have excellent performance in the field of object generation, due to the issues involved in video generation regarding the decomposition of elements in the video and time sequences, it is necessary to continuously make targeted modifications to various deep learning neural network frameworks in order to achieve breakthrough research results.

DOI: 10.54254/2755-2721/30/20230073

4. Research Prospects

MoCoGAN, as one of the most important research results in the field of video generation in recent years, has a broad application prospect. Up to now, the technology of video generation is considered to be most applicable to the field of human pose estimation, and the most important problem in this field is the lack of dataset, which will require huge labor and scene cost to produce a dataset, and may even have the problems of action diversity, insufficient samples, and angle limitation. If MoCoGAN can solve the above-mentioned deficiencies, video generation will bring great convenience to the production of human action-related datasets, which can not only effectively solve the problems of space and human resources, but also can generate a series of more complex actions to involve in more detailed and elaborate action detection tasks. In addition, the value of video generation technology for virtual reality, games, and movie special effects is also enormous. In the field of film, in recent years, Japanese Tokusatsu film and TV series have begun to experiment with real-time synthesis of virtual backgrounds and real characters to achieve a more ambitious scene layout, but the basic action still requires the actors themselves, if the action generation technology can be perfectly improved in the future, it will allow the actors to take the general action, and then through the posture-generation, to achieve a more smooth, more ornamental action scenes. In the popular Mixed Reality (MR) field, the video generation based on MoCoGAN will also provide more interaction diversity, and the generated actions can provide action guidance for the real users. Mixed Reality technology itself has so far been used more for instructional purposes in areas such as mechanical operations and surgical training, where the generated actions can be superimposed on the person to more intuitively represent the main points of operations in these areas and to guide the subject of the operations, creating a more complete instructional system by generating actions and operational tips. In the field of virtual reality (VR), video generation technology can also be applied to improve the game experience of game players, the generated movements will make the non-player character in the game look more realistic, which not only reduces the pressure of the modeler of frame-by-frame modeling, but also brings a better game experience for players. In addition virtual reality technology can also be applied in the field of physical rehabilitation. If patients are given a space with great immersion, they will have better rehabilitation results due to the feedback information given by the brain in the virtual space, and as mentioned above, the combination of motion generation and human posture estimation can help patients better improve their rehabilitation efficiency.

Therefore, video generation technology has great significance for many fields such as entertainment, teaching and even medical care, and research in recent years has shown that how to improve the quality of video generation and reduce the sense of violation has been a key research topic in the field of video generation. It is believed that in the future, when there is a major breakthrough in these problems, the field of video generation can really be put into use in a large number of industries, bringing more creative research results to our life.

5. Conclusion

This paper compiles the development process of MoCoGAN, a video generation technology, from the establishment of the idea to the proposed method, and summarizes the technical deficiencies that need to be improved in the field of video generation in general, starting from the viewpoint of the basic elements of video, and discusses the deficiencies in the most basic video quality, i.e., clarity, frame rate, etc., of the generated video as a video. Then, identifies the deficiencies in the physical rules that need to be addressed in order to maintain the perception, i.e., the semantic consistency of the elements in the video. Next, beginning with the structure and properties of GAN networks, discusses the input condition problem, which is the most important for generative adversarial networks. Hardware issues are also analyzed, but because the main focus of this paper is on optimizing the video quality and the structure of the neural network, hardware-related issues are not analyzed too much. For the most important requirements of video quality and input conditions of the GAN network, this paper summarizes the optimization solutions proposed by several researchers in this area, and studies the basic quality of the video by pre-training the video material to effectively improve the video quality. As for the input conditions of GAN networks, the improvement of video balance by condition-MoCoGAN for generating

DOI: 10.54254/2755-2721/30/20230073

human actions is investigated and the possibility of using condition-MoCoGAN under zero-shot learning conditions is also explored. Finally, the main improvement directions of this technique are proposed for future development, and the future improvement of the optimized video generation technique using MoCoGAN in combination with the hottest fields of Mixed Reality (MR) and Virtual Reality (VR) in the fields of entertainment, training, and medical care are presented. All these studies demonstrate the unlimited potential of video generation and its promising applications prospects. Although the field of video generation has not yet made significant research progress or shown mature application situations, it is believed that in the near future, video generation technology will slowly enter people's view from the dataset aspect with the continuous optimization of existing models, and will continue to produce phenomenal research results in combination with new industries or the most advanced research fields.

References

[1] Goodfellow I., Pouget-Abadie J., Mirza M., et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] Chan C., Ginosar S., Zhou T., et al. Everybody dance now [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2019: 5933-5942.

[3] Vondrick C., Pirsiavash H., Torralba A. Generating videos with scene dynamics [J]. Advances in neural information processing systems, 2016, 29.

[4] Zhou Y., Berg T. L. Learning temporal transformations from time-lapse videos [C]// Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 262-277.

[5] Saito M., Matsumoto E., Saito S. Temporal generative adversarial nets with singular value clipping [C]// Proceedings of the IEEE international conference on computer vision. 2017: 2830-2839.

[6] Tulyakov S., Liu M. Y., Yang X., et al. Mocogan: Decomposing motion and content for video generation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1526-1535.

[7] Dandi Y., Das A., Singhal S., et al. Jointly trained image and video generation using residual vectors [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 3028-3042.

[8] Unterthiner T., van Steenkiste S., Kurach K., et al. FVD: A new metric for video generation [J]. 2019.

[9] Kimura S., Kawamoto K. Conditional mocogan for zero-shot video generation [J]. arXiv preprint arXiv:2109.05864, 2021.

[10] Mirza M., Osindero S. Conditional generative adversarial nets [J]. arXiv preprint arXiv:1411.1784, 2014.

[11] Wang Y., Bilinski P., Bremond F., et al. G3AN: Disentangling appearance and motion for video generation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5264-5273.

[12] Siarohin A., Lathuilière S., Tulyakov S., et al. Animating arbitrary objects via deep motion transfer [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2377-2386.