Human Motion Imagination and Prediction- A Survey

Wafaa Shihab Ahmed [1*], Abdul amir A. Karim [2]

[1]*Department of Computer Science, University of Technology, Baghdad, Iraq*
[2] *Department of Computer Science, University of Technology, Baghdad, Iraq*

[*]*Corresponding Author:* 111798@student.uotechnology.edu.iq

Received 28-9-2020 , Accepted 5-10-2020, published 2-1-2021

DOI: 10.18081/2226-3284/5-10/30-45

Abstract: Human motion generation and prediction is one of important subjects in computer vision, human robot interactions and animations. There are many methods have been used for human motion modelling. This paper illustrate a survey about human motion imagination and prediction and explain the types and methods of video generation and human motion modelling. In this paper the Recurrent Neural Network/ Long Short Term Memory (RNN/LSTM), Generative Adversarial Network (GAN) and Variational Auto Encoders (VAE) models have been introduced. These models have been used for generating the spatial-temporal cuboids or for predicting the intensity pixels trajectory in the scene and give good results with short term prediction. This paper also shows the perceptual quality metrics that used for computing the performance of the methods and techniques which has been used for modelling.

Keywords: Video generation, Human motion imagination, Human motion prediction, Perceptual quality metrics.

1. Introduction

Video understanding became one of the most critical activities in the field of computer vision. The temporal aspect of videos offers much better representations of the real world compared with still images, such as interactions between objects, human actions, and so on. The process of forecasting the future has recently gained expanded interest in the research community among the different tasks applied to videos. In this field, most previous studies focused on forecasting high-level semantics such as actions, events and motions in a video. Semantic forecasting offers information of what might occur in a video that is important for automating decision-making. However, the semantics forecasted are often limited to a specific task and offer only a partial explanation of the future [1]. Unfortunately, owing to the intrinsic instability of videos and shifts in different variables, like deformation, occlusion, object motion and background transformations, the prediction video based on pixel-level is very difficult. [2]. The prediction of human motion, which serves as one of the most critical components of robotic intelligence, allows rapid and high fidelity

responses to dynamic changes in the environment. A robot, for example, can quickly predict a void path collision by predicting the movement of nearby objects. [3]. There are types of methods and techniques used for generating and predicting human motion such as RNN/LSTM, Vartional Auto Encoder and Generative Adversarial Network (GAN).

2. Literature Survey

Below are some related works of the human motion imagination and prediction:

K. Fragkiadaki et al., in 2015 [5], they proposed a model of Encoder-Recurrent-Decoder (ERD) to recognize and predict the position of human body in video and in motion capture. The human motion temporal dynamic learned by a long short term memory (LSTM) model. They constructed a nonlinear transformation to encode the features of human pose and decode the LSTM output. They tested representations of ERD architectures to generate motion capture (mocap), labeling pose of body and predicted it in video. They tested this model on the dataset named H3.6M [4], which is consider largest dataset for video pose. **J. Martinez et al., in 2017** [6], they expanded Recurrent Neural Network (RNN) Through modeling the velocity of joints instead of explicitly calculating the body pose, and used a single linear layer for encoding pose features and decoding hidden states. This is achieved by proposing three modifications for the basic models of RNN usually used with motion of human, resulting an architecture of RNN in a salable and an easy which achieves state-of-the-art prediction of the human motion in efficient.. They suggested a residual architecture which modeling the 1st-order motion derivations that leading to prediction short term in smoothing and much precise. In the latter, they notice that zero-velocity poses have produced comparatively less error on mean angle distance, showing the utility of velocity modeling. Human 3.6 M (H3.6 M) dataset [4] used for testing. **P. Ghosh et al., in 2017** [7], Proposed a modern framework to learn the models of spatio-temporal motion prediction from data only. This approach, known as the Dropout Autoencoder LSTM (DAELSTM), will synthesize natural sequences of motion over long-term horizons1 without drastic drift or loss of motion. This Dropout Autoencoder (DAE) then is used by a 3-layer LSTM network to filter each expected pose, reducing the accumulation of associated errors and, subsequently, drifted over time. **J. B¨utepage et al., in 2017** [8], They suggested a deep learning method for capturing data of human motion that learns a generalized representations from a wide corpus data of (mocap) and good generalizing into unidentified and new motions. Human motion features representation extracted by learning encoding-decoding network to forecast 3D poses in future from latest experience. They presented three approaches, all the ideas based on the bottleneck encoding decoding from past to future frames.

Furthermore, they described three differences of the temporal model: symmetrical encoding, timescale encoding and structural encoding. Coding symmetrically. It fits the basic principle of automated encoders. Since the decoding process is a replicated representation of the encoding process, it can be interpreted as an approximation of the encoder's opposite. In Time Scale-encoding, filters of various sizes are convolved with input data. The convolutional layers output in an encoder decoder model is merged and then handled through the fully connected layers fashion. Hierarchy encoding: The body of human can be expressed as a tree attached to the nodes of the respective limbs in the body, that the nodes includes the joints of individual. All these models were trained with H3.6 M data-set videos. **H. Cai et al., in 2017** [9], They concentrated on human action videos and proposed a generic, two-stage deep system for producing videos of human action with no limitations or random numbers of limitations that systematically solve the three issues: video generation without any input frames, video prediction with the first frames and video completion with the first and final frames. They used a deep generative model for training in the first stage to generate a sequence of human pose from random noise. A network of skeleton-to-image is trained in the second stage and is employed to produce a video of human action using the full human pose sequence produced in the first stage. Human3.6m dataset [4] used for testing by this model. **B. Chen, et al., in 2017** [13], they proposed a new system by transformation generation, which generates imaginary videos from a single image. In a novel volumetric merging network, they applied the generated transformations on the original image to recreate frames in imaginary video. They also suggested a new RIQA metric assessment for measurement. In experiments, they used 3 datasets for testing, these data are Moving MNIST [10], 2D shape [11] and UCF101 [12]. **R. Villegas et al., in 2017** [1], proposed a deep neural network to predict future frames of realistic video sequences. To solve complicated development of pixels in video, they proposed decomposing motion and content, two main components producing dynamics in video. This model built for pixel level forecasting by the Encoder-Decoder Convolutional Neural Network and Convolutional LSTM, which separately identify the spatial structure of an image and the associated temporal dynamics. Trying to predict the next frame by separately modeling motion and content decreases the conversion the extracted features of content to the next frame content by the motion features defined, which simplifies the prediction job. They evaluated the proposed system on videos of human motion, using KTH, Weizmann action, and UCF-101 datasets. **C. Li et al., in 2018** [14], they presented a new approach built on convolutional neural

networks (CNN) for modelling human motion. The encoder of the long-term and encoder of the short-term have the same architecture, i.e. the CEM, which consist of three convolution layers and one fully connected layer. For each convolution layer the number of feature maps was 64, 128 and 128, and for fully connected layer the number of the output nodes was 512. A stride number for each convolution layer is set 2 to capture the long term correlations and enhance the accuracy of prediction. So they suggested a model of convolutional sequence-to sequence to predict human motions. They adjusted 2 types of convolutional encoders, the encoder of long-term and encoder of short-term, so that the information of the both distant and temporal motion used to predict the future. In the long term prediction this model outperform on state-of-the-art RNN models, in the testing, they used 2 datasets: the dataset named Human 3.6M [4] and dataset named Motion Capture CMU. **Y. Li et al., in 2018** [15], proposed a conditional variational autoencoder (cVAE) dependent on probabilistic models, for modeling the uncertainty. There are two unique attributes of their probabilistic model. Firstly, this model is a 3D-cVAE, i.e. the autoencoder is built in an architecture of spatial-temporal convolutions used to predict consecutive optical flows. Secondly, is the method of frame generation named the Flow2rgb model, the model will "imagine" the existence of the next frame by flow and start frame. A spatial temporal correlations and future uncertainty have been modelling in a 3D-cVAE model. For evaluating the model they testing their algorithm on 3 datasets. The KTH dataset, and 2 datasets the Waving Flag and Floating Cloud which collected form website. These 2 datasets represent dynamic texture videos. **Z. Huang et al., in 2018** [16 They proposed a network across space of human motion generation in video with two paths: a forward path that first samples / generates a series of motion vectors with low-dimension based on Gaussian Process (GP), which is combined with image of input person to form a moving human figure series; and a backward direction to re-extract the relevant latent motion representation dependent on the predicted human frames. KTH Dataset and Human3.6 M Dataset used for testing. **Y. Tang et al., in 2018** [3], they proposed a modified highway unit (MHU) to eliminate non-moving joints and estimate the next pose with the context of motion efficiently. Moreover, they improved the dynamic of motion through reducing the gram matrix error for predicting the long-term motions. The results of experiments illustrated the suggested system can promisingly predict the motion of human in future, which outperform over corresponding state-of the-art systems. They applied their experiments on the H3.6m mocap Dataset [4], **K. Xu et al., in 2018** [2], They proposed a novel edge guided for network of video predictions, that in the first modelling the

frame edges dynamic and forecast the frame edges in future, then the frames in future have been generated based on the guidance of future frame edges. This network includes of 2 modules the module of edge prediction based on the ConvLSTM and the frames of edge guided generation module. The experiments applied on KTH human action data and this model show the result was better than others especially with long term prediction. **L. Zhao et al., in 2018** [17], they suggested a mechanism for the generation consist of two-stages which video are produced from structure and afterwards modified through the temporal signals. The networks are training for learning the residual motions between the present frames and the future frames to model movements more effectively, so avoid learning movement-irrelevant data. They tested this method on two tasks to translate image-to-video these tasks are: retargeting the facial expressions and prediction of human pose. **J. N. Kundu et al**., **in 2019** [18], they proposed a novel probabilistic generative model called Bidirectional Human motion prediction – Generative Adversarial Network, or (BiHMP-GAN). They presented a novel strategy for recursive prediction. The architecture of the discriminator has been enhanced by allows to predict the intermediate part of pose sequence and used as conditioning to predict the latter part of the sequence. The BiHMP-GAN model applied on two available datasets Human 3.6M [4] and CMU MOCAP.

3. Understanding Human motion

Understanding human motion is focused on the interpretation of global patterns of motion, instead of the study of local characteristics such as hand movements or facial expressions. The analysis of human motion is attracting growing interest from researchers of computer vision. there are a wide variety of technologies, motivate this interest, such as athletic performance analysis, monitoring, human computer interfaces, storage and retrieval based on the content of image, and video conferencing [19].

Understanding of human motion has been a valuable aim for many researchers across various disciplines. Where every discipline has various aspect to the problem. Such achievements have been made in multiple fields with many different reasons for researchers to keep in mind. Together, combining these contributions gives one a deeper understanding of human motion [20].

Depending on the ambiguity inherent in the action, the process of analyzing human motion could be presented from different levels of depth. Human action modeling and identification involves the classification of motion understanding problem in term of motion taxonomy. [21].

The design of human motion has a challenging job, this is because the non-linear dynamic, high

dimensional, and random nature for the motion of human. [6].

4. Video Generation

Growing research attention has been devoted to video generation, specifically a video for human motion generation. Initial methods [22], [1] explicitly apply / extend traditional 2D GAN to 3D spatio-temporal video generation (i.e. using it to interact with 2D image generation). However, despite the high dimensional search space, these methods typically produce poor video quality (non-realistic appearance) [1]. To this aim, several recent methods [23],[24] have tried to restrict the generator to human skeleton details (e.g. skeleton diagrams or the position maps of the joint) and therefore to create more accurate human articulated motion. However, these approaches, still have essential restrictions. First, many of these algorithms need a corresponding particular skeleton sequence for image synthesis of each frame. In other words, to generate video, a sequence representation vectors of skeleton (or positions of joint) must be provided beforehand. In many instances, though, it is very difficult to have this data, which significantly restricts its use. Second, for supervised learning, these approaches often involve couples of picture frames with the same backdrop and similar persons. However, it is very costly to acquire such effective supervised training results, which in turn prevents increasing scale-up of algorithms training. [16].

Given previous observations, video generation is focused on creating future frames of high-fidelity by learning dynamic visual features from video. It is a good path for learning video representations because the model may have to learn to separate variety influences based on dynamic visual features, i.e. how objects moving and distort over time, how scenes changing as the camera moving, how the background changing while foreground objects moving, etc. [25].

Several works [26], [1] use adversarial learning to increase the efficiency of generation to enable the process of video prediction, i.e. inserting an adversarial loss [27] on the module for prediction [28]. Current methods of video generation primarily concentrate on two tasks:

1. Video prediction: i.e., from the observed frames sequence, the patterns of motion must be learned by the models and the next frames have been predicted/ generated. The RNN / LSTM 's have strong ability to model sequential data, these approaches are typically based on a repetitive structure (RNN of LSTM), they typically produce better results when using with short-term forecasts where the video is easy and quietly predicted. Although the effects of the prediction in long-term typically suffer from low quality of image, like blurring and deformations of the object. These current approaches use the

predicted frames recursively as inputs for more prediction in long-term predictions, so the values of error can accumulate and results in a sharp drop in the quality of predictions. [29], [10] and [30].

2. Video imagination: The second type of approach attempts to produce a series of frames directly depend on a single input [24] or a single type of scene [22]. As the motion patterns during the test process can no longer be detected, this task is more complex. These techniques use a model named GAN to produce cuboids spatio-temporal or use a Variational Autoencoders [31] for prediction the scene's intensity pixel trajectory. However, if no geometric constraints are provided for foreground object, the objects in the scene which shift randomly, resulting in great deformation of the created objects. In comparison, they find a common constraints across both types of approaches, i.e., the articulated mechanisms of the foreground motions of object (i.e. human) aren't very good modelled in the model of generation. Since prior generation approaches only take the entire appearances as inputs, if given no control, it would be hard to all the models for learning the relationship of structures between the articulated / partitions, leading to significant deformations during the motion. Restricted by this restriction, the video quality produced is far from satisfying. [23].

One of the strategies was using skeleton data to help create articulated motions, which is guided by the following observations. The articulated motions, on the one hand, is normally behind a strong structure / geometric restriction, which the skeletons can well represent.

In the other side, the skeleton (coordinates of the parts of body) acts as a very strong low dimensionality representation for motion of human compared to image of high dimensions. Therefore it may also be used to produce flexible posing as the underlying status parameters.

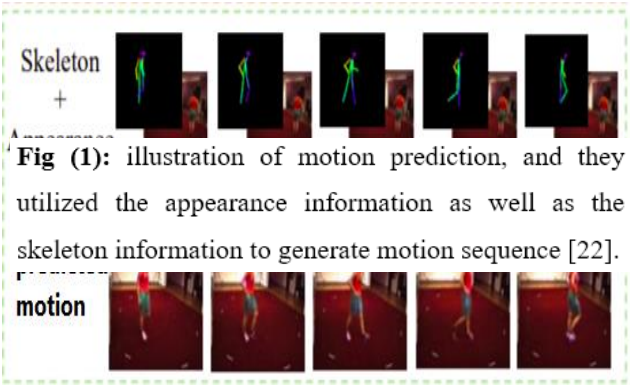

Fig (1): illustration of motion prediction, and they utilized the appearance information as well as the skeleton information to generate motion sequence [22].

Skeletons can also be converted to images one after the other, thereby eliminating the issue of long-term estimation shared by previous approaches. In addition, recent advances in human pose estimation techniques have made it easy to obtain skeleton data, thereby avoiding heavy human annotation [23]. Figure (1) illustrates (motion prediction).

5. Human Motion Imagination

Provided a static picture, humans will use their imagination to conceive about several scenes about what will happen next. For instance, given

the ballerina in Figure (2) (which can illustrate synthesizing multiple imaginary videos from one single image), one can clearly imagine the scene of the dancer leaping higher or landing gently. Video Imagination's role has been explained as an essential human capability for imagination of

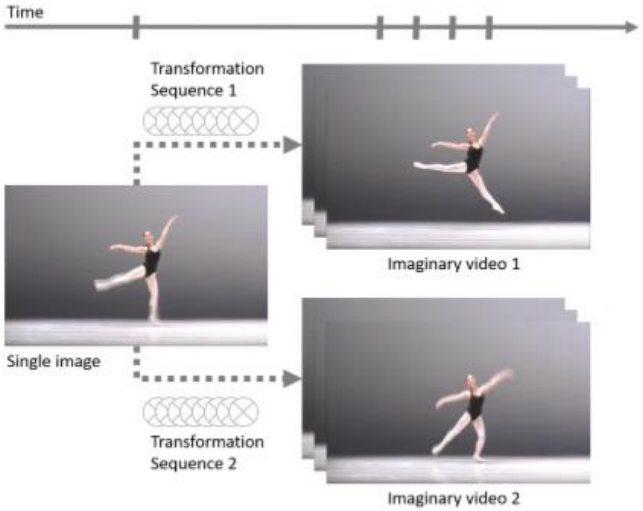

Fig (2): Synthesizing multiple imaginary videos from one single image. For instance, given an image of a dancing ballerina, the videos of the dancer jumping higher or landing softly are both plausible imaginary videos. Those videos can be synthesized through applying a sequence of transformations to the original image [13].

video by synthesizing fictional videos from one static image. This need to produce a diverse and realistic videos. There are more obstacles for imagination of video, screen preparation and prediction. In comparison to low-dimensional vectors in conceptual anticipation, visual creativity means generating true high-dimension pixel values. Furthermore, videos that are not similar to each other can all be rational, such as imaginary video 1 and imaginary video 2 in Figure (2) [13].

6. Human Motion Prediction

The prediction of human motion aims to generate frames of human motions in future and understanding a subject's behavior based on motion sequence observed [3].

The modelling of human motion is a classical challenge, with implementations covering the interaction between human and computer, the synthesis of motion, and virtual and augmented reality motion prediction. Recent works have focused on employing deep (RNNs) for modelling human motions, after the performance of deep learning approaches in many tasks of computer vision, with the objective of learning the representation of time-dependent which performing tasks like the prediction motions in short-term and long-term reconstruction of human motions. Figure (3) shows the motion prediction by using RNN [6].

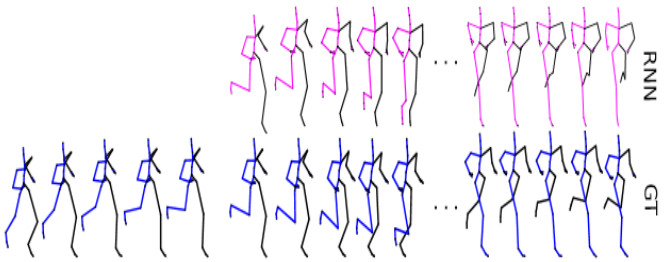

Fig (3): Frames on the left are the observations fed into the network. The middle part is the short-term prediction results for RNN (on the top) and our model (in the bottom). The right part is the long-term prediction results, in which RNN converge to a mean pose [14].

In the context of an assumptions of Markovian [32], [33], smoothing, or low-dimensionality embedding [34], conventional methods have generally placed specialist expertise on motion in their structures. A family of approaches focused on deep recurrent neural networks (RNNs) have recently demonstrated strong success on this job while attempting to be more neutral in their assumptions [6]. The approaches which based on Deep learning- have outperformed traditional methods on the problems that based on skeleton, such as 3D pose estimation [35] and actions recognition [36].

Predictions at the pixel level give a detailed and straightforward explanation of the visual world, and the current models of video recognition can be implemented to predict different meanings of the future on top of a predicted frame. The correlation between spatio-temporal in video provide self-supervision to predict frame, enabling a model to be purely unsupervised learning by the raw of video frames observing. Unfortunately, it is an incredibly difficult job to predict frames; not only because of the intrinsic ambiguity of the future, but also multiple variation variables in videos that contribute to complex dynamics in the values of raw pixels [1].

The human motion prediction based on pixel level as seen in Figure (4).

A main distinction between prediction of human motions and other sequence-to - sequence activities is that the motions of human is a highly restricted model of properties of the environment, properties of the human body and Newton's Laws [14].

Fig (4): Human Motion Prediction [37].

7. Perceptual Quality Metrics

These measures could be used either for a quantitative calculating of generated outcomes, or also as an error function by making some slight modifications to fulfill the necessary properties. These metrics were initially created to calculate the image compression codecs quality, such as JPEG15, in addition to neural network training. [38].

7.1 Mean Square Error (MSE)

Usage of the mathematical formula in equation (1), to measure the quantity of distortions in the videos [38].

$$MSE = \left(\frac{1}{W \times H}\right) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(P_{d(i,j)} - P_{s(i,j)}\right)^2 \qquad (1)$$

Where $p_{s(i,j)}$ represents the position of the pixel in (i,j) the original image,, $p_{d(i,j)}$ represents the

position of the pixel in (i,j) the destination image, where the rows and columns are represented by (x , y). MSE serves as a mathematical calculation tool used to calculate and quantify the difference between the source image and the resulted image. MSE metrics are of high quality where the importance of the variance is minimal. In practice, the importance of distortion for all pixels is measured, despite the fact that MSE does not take into account the human experience of visual content [38].

7.2 Peak Signal-to-Noise Ratio

The peak signal-to - noise ratio (PSNR) is a metric for assessing the similarity of the images produced. Is defined the ratio between the maximum potential picture strength and the corrupting noise that affects reconstruction accuracy. The PSNR value reflects by A logarithmic decibel scale, where a higher value means greater output. It is a rough estimate to determine reconstruction efficiency in terms of human experience, since its denominator is still depend on MSE. It is computed as follows:

$$PSNR(X,Y) = 10.log_{10}\left(\frac{{Y_{max}}^2}{MSE}\right) \qquad (2)$$

Where Y_{max} is represents he largest achievable intensity of any defined image with size W×H [39].

7.3 Structural Similarity (SSIM)

The structural similarity (SSIM) index found in [40] can be used as another performance parameter to predict the quality of the perceived image. This complete reference metric is an enhancement over PSNR as it is dependent on many human vision system theories. Therefore, dependent on luminance l(x, y), contrast c(x, y) and structural similarity s(x, y), both images are evaluated. These components are defined as follows [40]:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$
$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \qquad (3)$$
$$s(x,y) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

x, y = the original frames and distortion frames which used for comparison.

μx, μy= the intensity means of x, y.

σxy = The luminance of the x and y cross correlation.

σx, σy = the luminance variance.

C1, C2 = constants values.

It is possible to combine these terms to describe the SSIM index given by:

$$SSIM(x,y) = f\big(l(x,y).c(x,y).s(x,y)\big) \qquad (4)$$

It is possible to further simplify the terms for contrast and structure to cs(x, y) [40], resulting in:

$$SSIM(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5)$$
$$= l(x,y).cs(x,y)$$

It is possible to compute the SSIM index using a sliding window method [41]. A square kernel with specified size such as 11 × 11 and validated padding is then used, which moves pixel by pixel over the entire image. The index is then computed in each local area and at the end averaged to obtain the total image quality index for assessment. The metric value is in range SSIM(x,y)∈[0,1], where a maximum value refers to more similarity [39].

8. Application Domains

The prediction of motion is basic task for robotics of service, the vehicles of self-driving and advanced surveillance system [20]. Figure (5) shows these applications.

8.1 Robotics Mobile of Service: robotics of service work incrementally in open-ended home, industrial and urban areas connected with humans anticipating the motions of nearby objects is an essential requirement to secure and productive coordination of motions and contact between humans and robots. This is a daunting job due to insufficient on-board tools for computing and first-person sensing [20].

8.2 The Vehicles of Self-driving: The capability to predict motions of other road users is important in autonomous driving. As in the robotics of service domain, similar problems exist, but they are more prominent provided the higher vehicle masses and speeds and the corresponding greater damage that can potentially be incurred, especially to vulnerable road users (i.e. pedestrians and cyclists). In addition, vehicles needed to work in quickly moving, semantically rich outside traffic environments which involve difficult operational restrictions in real times. Awareness of traffic facilities (locations of lane, curbside, signage,

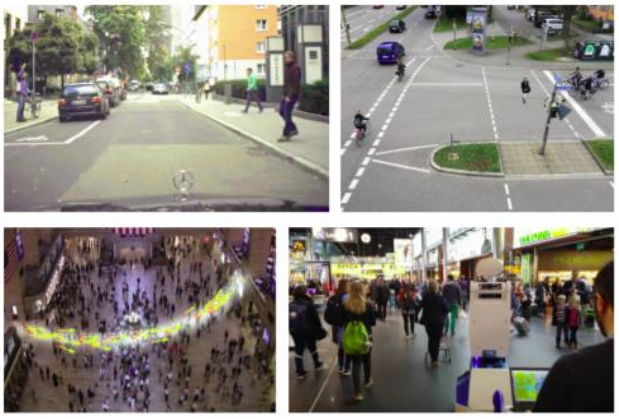

Fig (5). Application domains of human motion prediction. Top left: Will the pedestrian cross? Self-driving vehicles have to quickly reason about intentions and future locations of other traffic participants, such as pedestrians (Illustration from [42]). Top right: Advanced traffic surveillance systems can provide real-time alerts of pending collisions using communication technology. Bottom left: Advanced surveillance systems analyze human motion in public spaces for suspicious activity detection or crowd control (Illustration from [43]). Bottom right: Robot navigation in densely populated spaces requires accurate motion prediction of surrounding people to safely and efficiently move through crowds.

traffic signals, markers of other road, e.g. zebras) and traffic laws can assist in the motion predictions [20].

8.3 Surveillance Visual surveillance of vehicle traffics or humans crowd depends on the abilities to reliably tracking many targets through distributed network of stationary cameras. A range of surveillance activities such as individual recovery, perimeter security, traffic control, crowd management or retail analytics can be assisted by long-term motion prediction by further decreasing the number of false positive paths and tracking identifier changes, especially in dense crowds or through non-overlapping [43].

8.4 Computers Games the prediction of video is very helpful in teaching computer agents to play computer games [9].

8.5 The Animation Similar approaches have been used in animation to generate human position sequences [6].

9. Conclusion

In this paper the human motion imagination and prediction a survey has been presented that there are two methods for human motion prediction either based on skeleton of human to predict the next human pose or based on pixel level to predict the next frame. There are two tasks where the methods of video generation focused on them, the first one is video predictions the models require to learn the pattern of motions from the observed frames sequences and used them for prediction and generation the next frames. The second task is video imagination this task attempts to produce a series of frames directly depend on a single input or a single type of scene. There are two types of generated video the first is short term prediction and second is long term prediction. The RNN/LSTM has been used in these methods and achieved good results for short term prediction but the results in prediction of the long-term suffered from image with low quality, like blurring and objects deformations. The GAN or Variational Autoencoders model has been used with these methods for generating the spatial-temporal cuboids or for predicting the intensity pixels trajectory in the scene. At the end there are many applications for human motion prediction such as Service robots Mobile, Self-driving vehicles, Surveillance Visual, Computers Games and Animation. Future research directions include developing better models that can effectively train on multi-action data across all situations, particularly for long-term motion prediction.

References

[1] R. Villegas, J. Yang, S. Hong, X. Lin and H. Lee. "Decomposing Motion and Content for Natural Video Sequence Prediction," in ICLR 2017, pp. 1-22, 2017.

[2] K. Xu, G. Li, H. Xu, W. Zhang and Q. Huang, " Edge Guided Generation Network for Video Prediction," IEEE, 2018.

[3] Y. Tang, L. Ma, W. Liu and W. Zheng, "Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamic," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 935-941, 2018.

[4] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, "Human3.6m: Large Scale Datasets and Predictive Methods for 3d Human Sensing in Natural Enivoriments," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325-1339, 2014.

[5] K. Fragkiadaki, S. Levine, P. Felsen and J. Malik, " Recurrent Network Models for Human Dynamics," In Proceedings of the IEEE International Conference on Computer Vision, pp. 4346-4354, 2015.

[6] J. Martinez, M. J. Black and J. Romero, " On human motion prediction using recurrent neural networks," in IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv:1705.02445, pp. 2891-2900, 2017.

[7] P. Ghosh, J. Song, E. Aksan and O. Hilliges, " Learning Human Motion Models for Long-term Predictions," In 3D Vision (3DV), International Conference on IEEE, 2017.

[8] J. B¨utepage, M. J. Black, D. Kragic and H. Kjellstr¨om, "Deep representation learning for human motion prediction and classification,", In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[9] H. Cai, C. Bai, Y. Tai and C. Tang, " Deep Video Generation, Prediction and Completion of Human Action Sequences," arXiv: 1711.08682v3, 2017.

[10] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," In ICML., pp.843–852, 2015.

[11] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks,".In Advances in Neural Information Processing Systems. pp. 91–99, 2016.

[12] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[13] B. Chen, W. Wang, J. Wang and X. Chen, "Video Imagination from a Single Image with Transformation Generation," ACM Conference, Washington, DC, USA, 2017.

[14] C. Li, Z. Zhang, W. Sun, L. Gim and H. Lee, " Convolutional Sequence to Sequence Model for Human Dynamics," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5226-5234, 2018.

[15] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu and M. Yang, " Flow-Grounded Spatial-Temporal Video Prediction from Still Images," ECCV, Springer, pp. 1-16, 2018.

[16] Z. Huang, J. Xuand and B. Ni, " Human Motion Generation via Cross-Space Constrained Sampling," Proceedings of theTwenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 757-763, 2018.

[17] L. Zhao, X. Peng, Y. Tian, M. Kapadia and D. Metaxas, " Learning to Forecast and Refine Residual Motion for Image-to-Video Generation," In ECCV, Springer, 2018.

[18] J. N. Kundu, M. Gor and R.V. Babu, " BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN," Association for the Advancement of Artificial Intelligence, 2019.

[19] A. Rudenko and L. Palmieri, "Human motion trajectory prediction: a survey," the international journal of robotics research (ijrr), Vol. 39, no. 8, pp. 895-935, July 2020.

[20] A. N. Mohamed and M. M. Ali, " Human Motion Analysis, Recognition and Understanding in Computer Vision: A Review," Journal of Engineering Sciences, Assiut University, Faculty of Engineering, Vol. 41, no. 5, pp. 1928 – 1946, 2013.

[21] J. K. Aggarwal and S. Park, "Human Motion: Modeling and Recognition of Actions and Interactions," Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'04), IEEE, 2004.

[22] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," NIPS, pp. 613–621, 2016.

[23] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," In Proceedings of the 2017 ACMMM, pp. 199–207, 2017.

[24] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," In 2017 ICCV, pp. 3352–3361, 2017.

[25] Y. Jang, G. Kim and Y. Song, "Video Prediction with Appearance and Motion Conditions," Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[26] E. L. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 4417–4426, 2017.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, 2014.

[28] J. Xu, B. Ni and X. Yang, " Video Prediction via Selective Sampling," 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, pp. 1-11, 2018.

[29] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. P. Singh, "Action-conditional video prediction using deep networks in atari games," in NIPS, pp. 2863–2871, 2015.

[30] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video

prediction and unsupervised learning," CoRR, vol. abs/1605.08104, 2016.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," CoRR, vol. abs/1312.6114, 2013.

[32] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "A nonparametric bayesian network prior of human pose," In ICCV, 2013.

[33] P. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," In NIPS, 2000.

[34] J. Wang, A. Hertzmann, and D. M. Blei, "Gaussian process dynamical models,". In NIPS, 2005.

[35] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4948–4956, 2016.

[36] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition,". In European Conference on Computer Vision, pp. 816–833. Springer, 2016.

[37] C. Finn, I. Goodfellow, S. Levin, " Unsupervised Learning for Physical Interaction through Video Prediction," arXiv: 1605.07157v4 [cs.LG], pp.1-12, 2016.

[38] B. Sautermeister, "Deep Learning Approaches to Predict Future Frames in Videos," Master's Thesis in Computer Science, Departement of Informatics Technische Universitat Munchen, 2016.

[39] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," IEEE transactions on Image Processing, vol. 19, no. 6, pp. 1427-1441, 2010.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity." In: IEEE Trans. Image Processing 13.4, pp. 600–612, 2004.

[41] Z. Wang and A. C. Bovik, "A Univeral Image Quality Index," In: IEEE Signal Processing Letters 9, pp. 81–84, 2002.

[42] J.F.P. kooij, F. Flohr, E.A.I. Pool, and D.M. Gavrila, "Context-based path prediction for targets with switching dynamics," Int. J. of Comp. Vision (IJCV) : pp. 1–24, 2018.

[43] B. Zhou, X. Tang and X. Wang, "Learning collective crowd behaviors with dynamic pedestrian-agents," Int. J. of Comp. Vision (IJCV) vol. 111, no. 1, pp. 50–68, 2015.