2024 International Conference on Integration of Emerging Technologies for the Digital World (ICIETDW) | 979-8-3315-4117-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIETDW61607.2024.10939518

A Comprehensive Review on the Prediction of Video Frame and Motion Contents

Mohana Priya P
Department of Computer Science and Engineering
Vels Institute of Science Technology and Advanced Studies
Chennai, India
career.mohanapriya@gmail.com

Dr. Ulagapriya K
Department of Computer Science and Engineering
Vels Institute of Science Technology and Advanced Studies
Chennai, India
upriya.se@velsuniv.ac.in

***Abstract*— The capability of predicting, or analyzing the future outcomes of motion pictures using intelligent systems is the video frame prediction process. Video frame prediction is a significant and challenging aspect of Computer Vision (CV) applications. Video prediction plays a vital role in terms of robotics and Artificial Intelligence (AI) applications in detecting the future frames of a video, depending on which, the decision-making process can be carried out. The existing video frame prediction processes include pixel law based and motion-based predictions, which have been proven to be ambiguous. This leads to inaccurate predictions and indirectly deteriorates the performance of the real time applications. This paper performs a detailed review of the existing methodologies, their merits, pitfalls and challenges that create havoc in the video frame prediction process. In addition, this paper performs a detailed analysis on the existing datasets that assist in the video prediction process. This review summarizes various methodologies and datasets along with their performance analysis, which will open research possibilities in the field of video frame prediction.**

Keywords— Video frame prediction, Computer Vision, Robotics, Artificial Intelligence, Datasets.

I. INTRODUCTION

Prediction of the future motions [1] of the human body or any automated machines is the video frame prediction process, holds a vital role in real time intelligence applications. The video frame prediction [1] process holds good in varieties of real time applications namely, the robotics, video live streaming, accident control measures, smart city, industrial automations [2], etc. The prediction process is a real challenging process as it has influencing factors like the position of camera, changing of lighting aspect ratio, occlusions, speed of the moving object etc. Numerous techniques [3] have been proposed for the accurate prediction of the video frame. The Fig. 1 creates an anticipative question, "will the drone and ball collide each other?" The answer for this question requires prediction of the future frame. The prediction of the future video frames is technically referred to the x(n+1) frame for the video frame x(t). The scenario illustrated in the Fig 1 is composed of a video signal with the frames designated from $[X_{t-n},\ldots\ldots X_t]$ from which the future frames is to be predicted. The prediction of future frames of the video signal holds good in real time applications like, event anticipation, predicting the location of objects, video interpolation [4], long term planning, smart cities, predicting the pedestrian movement of normal and visually impaired people, autonomous vehicle etc. This article reviews the technologies employed in the prediction of future frames of the video signal. This review also analyzes the various datasets employed in the prediction process.

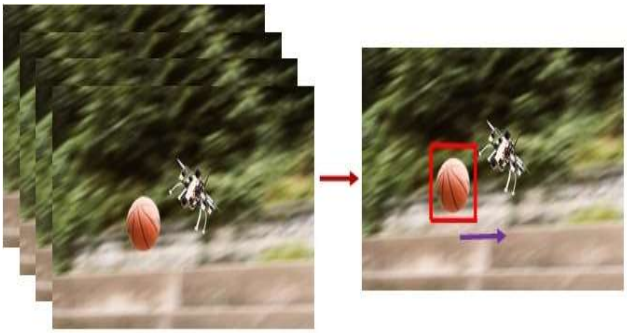

Fig. 1. Video signal x(t) considered for future frame prediction process

This review article is organized with a brief introduction in Section 1 and the various technological factors of video frame prediction in the Section 2. The Section 3 illustrates the various recent research results in the video frame prediction process along with the analysis of the various datasets associated with the process. A brief discussion is presented in the Section 4, illustrating the challenges and aspects for future research in the domain and the review is concluded in the Section 5.

II. ADVANCED TECHNOLOGIES IN PREDICTING VIDEO FRAMES

The capability of the system to predict or anticipate the future video frames is the major objective of the realtime decision making applications. Such a predictive nervous system is hard to be implemented in the machines, rather a peak performance shall be achieved by technologies like Deep Learning [4], Machine Learning and AI algorithms. The future frame of the video prediction process shall be classified into 4 types as depicted in Fig. 2. The video prediction process is broadly classified into four categories namely, the direct pixel synthesis, predicted space factorization, predicted space narrowing and uncertainty incorporation as depicted in Fig 2. This section describes how to perform the future frame prediction process in the video signal and the various methodologies involved in the prediction process. Our innate understanding of the universe comes from the conceptual

learning and foundational knowledge we acquired since we were young children.

Videos offer intricate changes and motion patterns in the temporal domain, in contrast to still images. Because of the temporal coherence, we can discover a wide range of local visually comparable deformations at fine resolution if we focus on a small region at the same geographical place throughout subsequent time steps Comparatively, if one were to examine the larger image, successive frames would appear visually distinct but maintain conceptual coherence. The primary causes of this variation in a video's visual look at various scales include, among other things, occlusions, adjustments to the lighting, and camera movement.

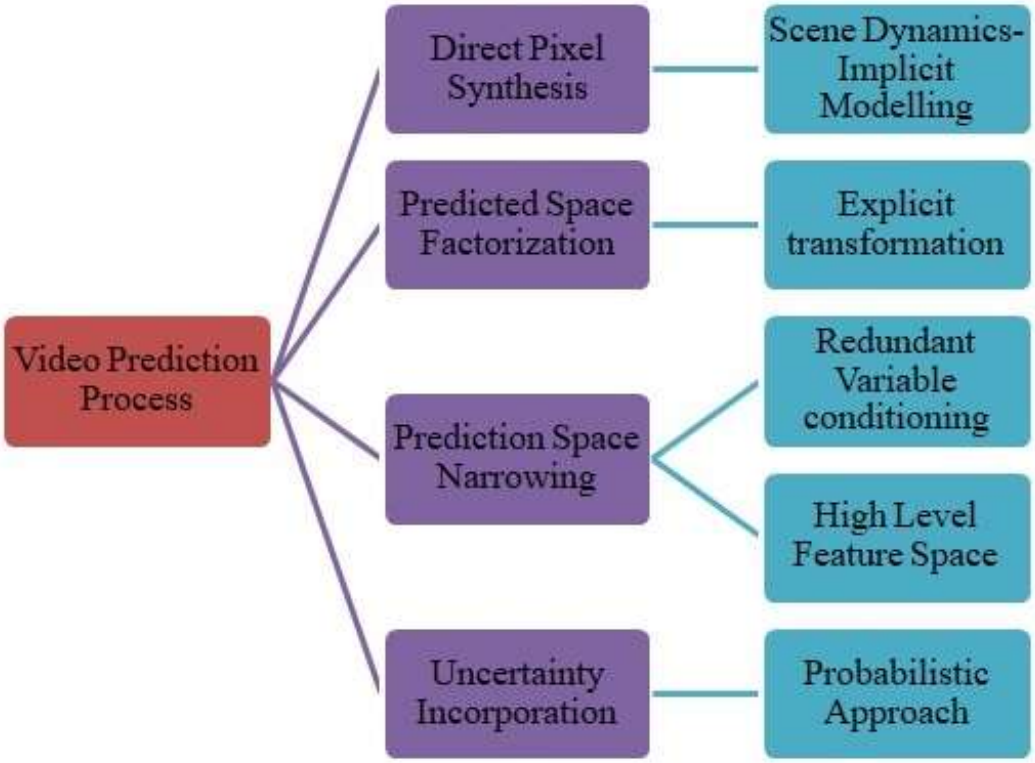

Fig. 2. Video Prediction process - Categories

A. Deep Learning Model in Future Frame Prediction

Convolutional Neural Networks (CNNs) are essential in predicting future frames, since they excel at capturing the spatial structure of pictures. These deep learning models have a significant function in this task. CNNs serve as the fundamental framework for visual prediction tasks, and researchers have proposed several improvements. These include incorporating additional convolutional layers, expanding kernel dimensions, integrating multiple scales such as in Laplacian pyramid reconstruction, employing dilated convolutions to capture extended spatial relationships, and enhancing receptive fields through pooling methods, albeit with some compromise in quality. Furthermore, Recurrent Neural Networks (RNNs) have shown outstanding proficiency in video prediction and other tasks involving sequence learning, such as language translation, voice recognition, and video/audio captioning. Generative models aim to understand the underlying distribution of individual classes, while discriminative models prioritize learning the decision boundaries between classes.

B. Machine Learning Model in Future Frame Prediction

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four variations of machine learning algorithms that may be classified according to the learning methodologies that they use. Direct pixel combination-based approaches and transform-based methods are the two primary categories that may be used to classify the techniques that are used for video forecasting which are based on machine learning. KNN-based video prediction approaches are among them, and they not only capture sequence dynamics, but they also take into consideration the long-term dependencies that are shown in movies.In addition to capturing sequence dynamics, KNN-based video prediction techniques also provide an explanation for long-term dependencies in videos.

C. *Datasets for Future Frame Prediction*

Self-supervised video prediction models often require sequences of videos as the data source. However, some video prediction systems rely on additional supervisory inputs, such as segmentation maps and individual poses. In this section, we will delve into the most pertinent datasets and delve into their advantages and disadvantages. These datasets have been categorized based on their primary purpose and have been succinctly summarized in Table I.

KTH: This dataset focuses on action recognition and consists of 2391 video sequences with an standard duration of 4 seconds. The backgrounds are uniform and the video clips are captured at a frame rate of 25 frames per second (fps) at an aspect ratio of 160 × 120 pixels.

UCF101: is an assortment of videos that showcase the art of action, sourced from the vast realm of YouTube. Every single one of these videos boasts a remarkable frame rate of 25 fps, capturing the essence of each action in exquisite detail, while maintaining a resolution of 320 × 240 pixels.

Penn Action Dataset: hails from the prestigious University of Pennsylvania, serving as a repository for both action and human pose recognition. Within its vast expanse lie 2326 captivating video sequences, each showcasing 15 distinct actions.

THUMOS-15: stands as a testament to the grandeur of action recognition challenges, its prominence reaching its zenith in the year 2015.

YouTube8M]: The Sports1M dataset, under the umbrella of YouTube8M, has been in existence since 2016. YouTube8M encompasses a wide range of videos, not just limited to sports.

Moving MNIST: The enchanting M-MNIST, also known as Moving MNIST, is a captivating collection of video predictions. The dataset may be synthetic type (S) or a real (R) type of data. In addition, the Table I represents whether the dataset is composed of videos Indoor(I) or Outdoor(O) area.

TABLE I. COMPARISON OF MACHINE LEARNING DATA ANALYSIS

Name of Dataset	No. of videos	No. of frames	S/R	Resolution of frames	Frame type	Nature of frames	Action classes	Year
KTH	2411	250000^2	R	160x120	RGB	O	6	2004
Weizmann	109	9000^2	R	180x160	RGB	O	10	2007
UCF101	13350	9000^2	R	240x240	RGB	I/O	15	2012
Penn Action	2526	1638^2	R	240x240	RGB	I/O	15	2013
THUMOS-15	18404	300000^2	S/R	320x240	RGB	I/O	101	2017
Youtube 8M	82000	NA	S/R	Var x Var	RGB	I/O	25	2017
CamVid	100	1820^2	S/R	480x270	RGB	I/O	15	2017
Apolloscape	404	2000^2	R	320x240	RGB	I/O	12	2018
CalTech Pedestrian	237	1000^2	R	320x240	RGB	I/O	12	2018
Moving MNIST	Var	Var	S/R	64x64	RGB	I/O	12	2020

III. VIDEO FUTURE FRAME PREDICTION- A REVIEW

This section describes the contributions made by various researchers in the field of future frame prediction in video signals using various methodologies. This section imparts a wide knowledge on how do the various technologies anticipates the future frames along with its efficiency and challenges. This sections acts as the major backbone of this research review article to understand about the future perspectives of research in this frame prediction domain.

M.Xu et al. (2019) proposed a Deep Reinforcement Learning (DRL) method [5] of identifying the future frames of the video signal. The authors employed Head Movement (HM) dataset for the detection of HM in both online and offline versions.

J.K.Lee et al. (2020) employed the CNN model [6] for the prediction of video frame in the Video Prediction Network (VPN). The performance analysis of the proposed work exhibited a gain range of -2.9% to -5.7%.

R.Yang et al. (2023) introduced an Advanced Learned Video Compression (ALVC) method [7] to anticipate the future frames of the visual signal. The in-loop prediction method employed end-to-end encryption method and by the efforts of the Recurrent Neural Networks (RNN) and Bi-Directional in-loop prediction mechanism, performs the anticipation process.

S.Li et al. (2021) constructed a Deep Multi-Branch Mask network (DMMNet) [8] for integrating the merits of the optical flow warping and pixel synthesis methods of RGB. The UCSD dataset has been used for the training and testing purpose and the accuracy has been achieved to 79.81%.

P.Kancharla et al. (2021) presented an efficient method of video frame prediction model [9] using the cognitive improvement model. The analysis of the proposed model outperforms well with reduced RMSE errors when compared with existing models.

L.Zhao et al. (2021) improved the performance of the future frame prediction of the video signal by using the Background Reference Frame (BRF) [10]. The reference frames were generated using the Surveillance Prediction Generative Adversarial Network (SP-GAN) model. The proposed work achieves a 5.8% gain during the prediction approach.

H.Choi et al. (2019) developed a novel frame prediction model using Deep Neural Network (DNN) [11], with an objective of enhancing the prediction efficiency. The DNN method detects both the uni-directional and bi-directional predictions of frames and reduced the luminance to 4.4%.

W. Lu et al. (2021) introduced a complete video prediction model that combines the optical flow prediction and pixel creation modules [12]. Utilizing the pedestrian data set UCF101, the suggested model demonstrated improved performance with a 30.9 PSNR value.

R.Szeto and et al. (2019) designed an integrative model using the bi-directional video prediction model [13] and temporal frame interpolation model for the prediction of future frames. The proposed work employed a unique data set and had exhibited an accuracy of 80.3%.

J.Wang et al. (2019) presented a novel future frame prediction model using the Recursive Pixel level prediction using the Region of Interests (ROI) [14] of the reference frames. The proposed work was tested with four different datasets and had proved to be efficient for real time applications.

TABLE II. SUMMARY OF LITERATURE REVIEW AND ITS PERFORMANCE

Ref	Technology Used	Application	Dataset Used	Performance
[5]	Deep Reinforcement Learning (DRL)	HM Prediction	HM dataset	Accuracy: 80.16%
[6]	Convolutional Neural Networks (CNN)	VPN	Youtube	Gain: -5.6%
[7]	Recurring Neural Networks (RNN)	In loop prediction	Own dataset	Accuracy: 76.17%
[8]	Deep Multi-Branch Mask network (DMMNet)	Video Prediction	UCSD dataset	Accuracy: 79.81%
[9]	Perceptual Strengthening Model	Video Prediction	Own dataset	RMSE: 12.96%
[10]	Background Reference Frame (BRF)	Video Surveillance	Reference Pic	Gain: 5.8%
[11]	Deep Neural Network (DNN)	Video Surveillance	Own dataset	Luminance: 4.4%
[12]	Optical Flow Estimation module with the Pixel Generation	Video Prediction	UCF101	PSNR 30.9%
[13]	Bi-Directional Video Prediction model	Video Prediction	Own dataset	Accuracy: 80.3%
[14]	Recursive Pixel level prediction	Video Prediction	Four dataset	Accuracy: 81.04%

The Table II presents the summarized abstract of the findings performed through the literature review performed in this section.

IV. DISCUSSION

This section summarizes the findings from the review performed in the previous section. Despite the contributions of numerous Deep learning algorithms in the video frame prediction process, there exists a room for performance improvement in the prediction process. The following shortfalls have been identified from the review process. Short-term horizons are the only limitations of current approaches. Although frames for the near future can be projected with good accuracy, the prediction problem naturally becomes multimodal over the long term horizon. The predictions were first conditioned on frames that had already been forecasted. However, these autoregressive models have a propensity to accrue prediction errors, which cause the generated prediction to gradually deviate from the anticipated result. This has a direct bearing on the pixel-wise loss functions, which train the system to prioritize visual attractiveness. Selecting the loss function is a research open issue that directly affects the prediction quality. Ultimately, another possible unresolved issue in the qualitative assessment of video prediction is the absence of valid and equitable assessment models.

A. Future Research Perspectives

This review paper highlights some potential future research fields depending on the analyzed literature identifying the most advanced video prediction systems.

- Considering loss functions
- Usage of Synthetic Database
- Updation of Evaluation Metrics

V. CONCLUSION

The significance of video future frame prediction processes has increased due to the increase in real-time

monitoring applications. In this regard, this paper presents an in-depth review of the available approaches and demonstrates that the deep learning algorithm plays an important part in the prediction process, specifically through the use of Neural Networks (NN). This study's primary focus is on the analysis

and categorization of over 30 different methodologies together with the datasets they have employed. Three aspects of the methods were examined: their description, their advancement over earlier research, and their performance outcomes. Additionally, a proposed taxonomy based on their primary contribution has been used to classify them. To help the reader quickly detect minor details, we have also included a tabular summary of the datasets and methodologies that compare them. In summary, video prediction is a potential way to enable rich spatio-temporal correlations to be learned via self-supervision and to give existing intelligent decision-making system prediction capabilities. Even with these impressive advances, deep learning approaches for video prediction remain unsatisfactory.

REFERENCES

[1] M. Jubran, A. Abbas and Y. Andreopoulos, "Sequence-Level Reference Frames in Video Coding," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1578-1591, March 2022, doi: 10.1109/TCSVT.2021.3070423.

[2] H. Kataoka, T. Suzuki, K. Nakashima, Y. Satoh and Y. Aoki, "Joint Pedestrian Detection and Risk-level Prediction with Motion-Representation-by-Detection," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 1021-1027, doi: 10.1109/ICRA40945.2020.9197399.

[3] J. Jeong, M. Hong, J. W. Kim and S. Kim, "A Fast 4K Video Frame Interpolation based on StepWise Optical Flow Computation and Video Spatial Interpolation," 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2021, pp. 1140-1143, doi: 10.1109/ICTC52510.2021.9621048.

[4] S. Kannan, Prabakaran. D, Dhenesh Kumar. S and Shivaram. S, "A Deep Learning-Based Convolution Neural Networks to Forecast Wind Energy," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111917.

[5] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo and Z. Wang, "Predicting Head Movement in Panoramic Video: A Deep Reinforcement Learning Approach," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, pp. 2693-2708, 1 Nov. 2019, doi: 10.1109/TPAMI.2018.2858783.

[6] J. -K. Lee, N. Kim, S. Cho and J. -W. Kang, "Deep Video Prediction Network-Based Inter-Frame Coding in HEVC," in IEEE Access, vol. 8, pp. 95906-95917, 2020, doi: 10.1109/ACCESS.2020.2993566.

[7] R. Yang, R. Timofte and L. Van Gool, "Advancing Learned Video Compression With In-Loop Frame Prediction," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 5, pp. 2410-2423, May 2023, doi: 10.1109/TCSVT.2022.3222418.

[8] S. Li, J. Fang, H. Xu and J. Xue, "Video Frame Prediction by Deep Multi-Branch Mask Network," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 4, pp. 1283-1295, April 2021, doi: 10.1109/TCSVT.2020.2984783.

[9] P. Kancharla and S. S. Channappayya, "Improving the Visual Quality of Video Frame Prediction Models Using the Perceptual Straightening Hypothesis," in IEEE Signal Processing Letters, vol. 28, pp. 2167-2171, 2021, doi: 10.1109/LSP.2021.3118639.

[10] L. Zhao, S. Wang, S. Wang, Y. Ye, S. Ma and W. Gao, "Enhanced Surveillance Video Compression With Dual Reference Frames Generation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1592-1606, March 2022, doi: 10.1109/TCSVT.2021.3073114.

[11] H. Choi and I. V. Bajić, "Deep Frame Prediction for Video Coding," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 7, pp. 1843-1855, July 2020, doi: 10.1109/TCSVT.2019.2924657.

[12] W. Lu, J. Cui, Y. Chang and L. Zhang, "A Video Prediction Method Based on Optical Flow Estimation and Pixel Generation," in IEEE Access, vol. 9, pp. 100395-100406, 2021, doi: 10.1109/ACCESS.2021.3096788.

[13] R. Szeto, X. Sun, K. Lu and J. J. Corso, "A Temporally-Aware Interpolation Network for Video Frame Inpainting," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1053-1068, 1 May 2020, doi: 10.1109/TPAMI.2019.2951667.

[14] J. Wang, W. Wang and W. Gao, "Predicting Diverse Future Frames With Local Transformation-Guided Masking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 12, pp. 3531-3543, Dec. 2019, doi: 10.1109/TCSVT.2018.2882061.