Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Next frame prediction using ConvLSTM

To cite this article: Padmashree Desai *et al* 2022 *J. Phys.: Conf. Ser.* **2161** 012024

View the article online for updates and enhancements.

doi:10.1088/1742-6596/2161/1/012024

Next frame prediction using ConvLSTM

Padmashree Desai[1], Sujatha C[2], Saumyajit Chakraborty[3], Saurav Ansuman[4], Sanika Bhandari[5], Sharan Kardiguddi[6]
School of Computer Science and Engineering, KLE Technological University, Hubballi, India
[1]padmahri@kletech.ac.in1, [2]sujata_c@kletech.ac.in, [3]saumyajit99@gmail.com,
[4]sarvkan@gmail.com, [5]sanikabhandari07@gmail.com [6]sharankardeguddi@gmail.com

Abstract. Intelligent decision-making systems require the potential for forecasting, foreseeing, and reasoning about future events. The issue of video frame prediction has aroused a lot of attention due to its usefulness in many computer vision applications such as autonomous vehicles and robots. Recent deep learning advances have significantly improved video prediction performance. Nevertheless, as top-performing systems attempt to foresee even more future frames, their predictions become increasingly foggy. We developed a method for predicting a future frame based on a series of prior frames that services the Convolutional Long-Short Term Memory (ConvLSTM) model. The input video is segmented into frames, fed to the ConvLSTM model to extract the features and forecast a future frame which can be beneficial in a variety of applications. We have used two metrics to measure the quality of the predicted frame: structural similarity index (SSIM) and perceptual distance, which help in understanding the difference between the actual frame and the predicted frame. The UCF101 data set is used for testing and training in the project. It is a data collection of realistic action videos taken from YouTube with 101 action categories for action detection. The ConvLSTM model is trained and tested for 24 categories from this dataset and a future frame is predicted which yields satisfactory results. We obtained SSIM as 0.95 and perceptual similarity as 24.28 for our system. The suggested work's results are also compared to those of state-of-the-art approaches, which are shown to be superior.

1. Introduction

Smart decision-making systems are capable of predicting, anticipating, and explaining future results. Due to its importance in a wide variety of computer vision applications, including autonomous vehicles and robotics, the problem of video frame prediction has become highly intriguing. Recent progress in deep learning has considerably increased video prediction performance. Nevertheless, their forecasts grow increasingly nebulous with top-performance systems trying to anticipate even farther frames.

Is it possible for an artificial intelligence system to forecast a photorealistic movie based on previous visual observations? An intelligent agent can plan its mobility based on the predicted video using an accurate video prediction model. We, as humans, solve this problem quickly and readily, but it is incredibly difficult for machines to do so. The challenge of producing future frames from a collection of successive frames is known as video prediction [1].

Predicting actions and movement in videos has been a source of active research. In disciplines like robotics, where an agent interacts with the outside environment and plans around what it considers to be the most likely future sequence of events, the ability to forecast the immediate future could be a substantial benefit [2]. Modelling and forecasting the future is critical for a variety of machine learning and computer vision applications, including human posture estimation and identification, pedestrian detection and tracking, and weather forecasting. Further- more, projection of future video frames is very challenging owing to the large number of variables that contribute to the changing dynamics of frames throughout time. Due to the uncertain nature of future events, the field is wide open for newer inventions to prosper.

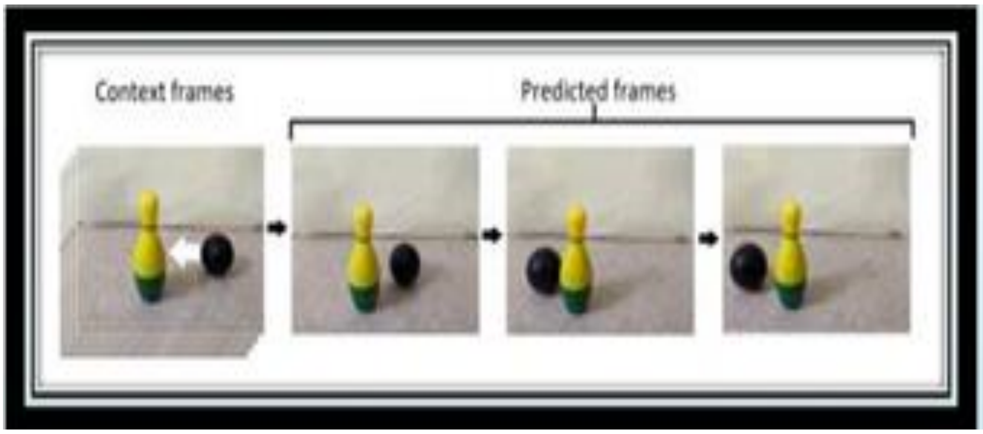

Figure 1. Frames Prediction: A ball thrown towards a pin

Like in the figure 1, context frames are given as input to the predicting system and future frames are generated. The ball is thrown towards the pin is given as input to the system and ball can be seen pass through the yellow pin as a predicted result.

Many real-life scenarios may be anticipated because they fulfil physical principles (e.g. inertia) such as the ping-pong robot ball parabola forecast [11]. The anticipation of moving items enables the system to decide in advance. Also, pictures can be anticipated in order to better comprehend the current environment [12]. There are several instances of predictive systems that benefit from next-frame predictions. Predicting future framework conditions, for example, permits self-employed people to make intelligent choices in different activities. Author of [14] demonstrated a video prediction system that aids robot decision-making by interpreting current images and anticipating the discrete raw pixel values that are jointly distributed across images. Here the system helps robots to decide by comprehending current pictures and by predicting the discreet raw pixel values jointly distributed among images. Oh et al. [15] integrated an artificial intelligence agent (AI) predictive with Q-learning deep algorithm to improve the efficiency in many games. Different techniques discussed in [16],[17] have offered a vehicle prediction visual system that forecasts that the photographer's future location will slow down or stop the cars. Klein et al. [18] forecast weather using next-frame prediction to forecast radar cloud pictures. In essence, next-frame prediction enables machine learning to gain a greater understanding of its surroundings and offers up endless possibilities for performing a range of activities that require predictive capacity.

In recent years, a slew of video prediction algorithms have been presented, following the success of deep neural networks (CNN), recurrent neural networks (RNN), generative adversarial networks (GAN), and Long Short-Term Memory (LSTM) [4]. Recent approaches to video prediction emphasise the separation of representations with varying temporal dynamics. Many spatiotemporal elements of variation exist in movies, making this a difficult topic for neural networks to model. To address this issue, a variety of approaches have been offered, the approach chosen by us for prediction of frame is the ConvLSTM model which It uses the inputs and previous states of its immediate neighbours to predict the future state of a cell in the grid [6]. LSTMs offer a wide range of parameters, including learning rates, input and output biases, and other factors, making it one of the most versatile prediction approaches.

In the proposed work, the quality of the forecast framework is measured through – Structural Similarity Index (SSIM) and perceptual similarity, which helps comprehend the difference between the current and the forecast frames. For project testing and training, the UCF101 data set is used. It is a series of realistic action videos that are collected from YouTube and have 101 categories of action. The ConvLSTM model has been trained and tested in 24 categories and a future framework is expected to provide successful results. We have received SSIM 0.95 and our system's perceptual similarity is 24.28. The outcomes of the work presented are also compared with the best-in-class approaches. Variants of other deep education methods, such as Generative Adversarial Network, can improve the proposed work (GAN).

1.1 Paper Organization

The remaining sections of this paper are organized as follows. Related research in future frame predictions is discussed in Section 2, while Section 3 discusses the proposed approach and implementation. The results, comparison and analysis are addressed in Section 4. Section 5 contains the conclusion as well as a discussion of possible research directions.

2. Related Work

In [1], the authors developed a method for forecasting future video frames using a prior sequence of continuous video frames. Rather than directly synthesising images, the goal here is to decouple the backdrop scene and moving objects in order to grasp the complicated scene dynamics. The future appearance of scene components is anticipated via non-rigid backdrop deformation and moving objects' affine transformation. In the future, the predicted appearances are integrated to form a reason- able video. When compared to other ways, our method produces significantly less ripping or distortion artefact. This work uses the Cityscapes and KITTI datasets, and they provide a separate-predict-composite model for future frame prediction [1].

Authors in [2] discussed that because of importance of many computer vision applications, the problem of video frame prediction has sparked a lot of interest. Video frame prediction algorithms that are supervised rely on labelled data, which isn’t always available. For video frame prediction, an unique self-supervised deep-learning algorithm dubbed Inception-based LSTM is utilised in this research. Inception networks are designed to implement larger networks rather than deeper networks. This network design has been found to improve picture categorization performance. Both Inception-v1 and Inception- v2 structures are used to test the suggested approach. The suggested Inception LSTM approaches are compared to convolutional LSTM when applied utilising the PredNet predictive coding framework for both the KITTI and KTH data sets. According to our findings, the Inception-based LSTM outperforms the convolutional LSTM. In terms of prediction, Inception LSTM outperforms Inception v2 LSTM. Inception v2 LSTM, on the other hand, has a lower computational cost than Inception LSTM [2].

The authors of [3] emphasised how difficult it is to anticipate the next frame given visual data. A recurrent convolutional neural network is trained to estimate depth from monocular video input, and it may then be used to compute the next frame, as well as the camera trajectory and current video picture. Unlike previous algorithms for next frame prediction, we use the scene geometry to our advantage and generate the next frame prediction using the predicted depth. Our method can generate detailed next-frame projections that account for each pixel's depth. This method also differs from others in that it estimates depth using a series of photos rather than a single still image (for example, in a movie). Using the KITTI dataset, the model is compared against state-of-the-art video prediction models, which are the best and most current models for next frame prediction.

Authors in [4] proposed a large-scale system that learns to anticipate future actions and objects by exploiting temporal structure in unlabelled video. The main idea behind the method utilised here is to train deep networks to predict how images will be represented in the future. They test this theory on two difficult “in the wild” video datasets, and the results show that learning with unlabelled films considerably aids in forecasting actions and anticipating objects. The assumption that freely available, unlabelled films are a good resource for learning this information [4].

Authors in [5] discussed about Convolutional networks and recurrent networks are used in video prediction models, and their mixtures frequently result in fuzzy forecasts. Blind spots, or a lack of access to all required prior knowledge for accurately forecasting the future, are a significant contributor to inaccurate predictions that has received less attention in the literature. To address this issue, we offer a totally context-aware architecture that utilises Parallel Multi-Dimensional LSTM units to gather all of a pixel's known historical context and then blends it using blending units. On three hard real-world video datasets: Human 3.6M, Caltech Pedestrian, and UCF-101, our model outperforms a robust baseline network of twenty recurrent convolutional layers and achieves state-of-the-art performance for next step prediction. It also accomplishes this with less variables than other recently proposed models, without the need of convolutional neural networks, multi-scale architectures, foreground and background modelling segregation, Motion flow learning, also known as adversarial training, is a technique for learning how to move in a certain. The findings demonstrate that for video prediction, a thorough understanding of the historical background is essential [5].

Authors in [6] discussed that forecasting, anticipating, and reasoning about future events is at the basis of intelligence, and it's one of the primary goals of decision-making systems like human-machine interaction, robotic navigation, and autonomous driving. Recently, two approaches have been explored, namely the usage of latent variable models to capture underlying stochasticity and adversarial trained models to produce crisper images. A novel multi-scale architecture encompassing both techniques is developed and deployed in this study to provide projections that appear more realistic and better span the spectrum of possible futures. Moving MNIST, UCF101, and Penn Action datasets are utilised to forecast a future frame by substituting some old techniques with advanced ones, resulting in a sharper image [6].

Authors in [7] discussed about the main goal is to create a unified generative adversarial network that can over time forecast accurate and temporally consistent future frames, even in difficult environments. A comparison is made with Prednet, which employs ten photos to forecast future frames, but the model constructed here only uses four images. One significant disadvantage of this method is that it is prone to inaccuracies due to lighting changes, occlusion, and sudden camera motion. The video pixel network, a probabilistic inference model comprising of resolution preserving CNN encoders and PixelCNN decoders, is proposed using the KITTI, Caltech pedestrian, UCF101, CUHK Avenue, and ShanghaiTech Campus datasets, which are trained and used in a standard format. To aggregate the above and forecast future frames, a convolutional LSTM is employed [7].

Authors in [8] proposed an architecture which makes use of the hierarchy of representations obtained from cascaded or multi-layer Convolutional Neural Networks. Separating and capturing the features at multiple levels of hierarchy is useful e.g., capturing high-level features makes the prediction task simpler, whereas capturing low-level features helps generating realistic frames. Thus here a design is proposed to use multiple ConvLSTMs, each of which is dedicated to model the temporal dynamics of each of multiple levels of CNN features [8].

Authors in [9] explored a topic, namely whether network topologies are necessary, and instead propose a new approach: the main goal is to find the video prediction algorithm with the least inductive bias while optimising network capacity. Finally, it is shown that by just boosting the capacity of a regular neural network, a high-quality video prediction may be produced even without the usage of the previously stated techniques (such as adversarial objectives, optical flows, and so on) [9].

Authors in [10] proposed an approach where the HMDB-51 video dataset is used to train an AlexNet convolutional neural network architecture to predict optical flow from static video frames using an unsupervised learning technique. Optical flow must be predicted in order to forecast future activities. Op- tical flow is a vector field that depicts each pixel's apparent mobility between two frames. In frames with unambiguous actions, the neural networks predicted optical flow surprisingly well, yet they made reasonable errors [10].

Numerous studies have explored the application of feature extraction, deep learning approaches, and picture recognition techniques that are applicable to image comprehension [20]-[24] by utilising multiple layers in the generator and descriptors. Authors in [25],[26] discuss feature extraction techniques which help extract features of any object of an image or video.

Each of the methods were on commonly used techniques for future prediction like Cyclic GAN, GAN, LSTM and other modified versions of these. KITTI dataset is used widely due to its high detail and multiple data for single instance circumstance. The other datasets used consists of Bouncing ball dataset, KTH dataset, pedestrian movement dataset. The most often utilised performance metrics are Mean Squared Error (MSE), Structural Similarity (SSIM), and Peak Signal to Noise Ratio (PSNR).The accuracy of one metric cannot be compared with another. Some models have good MSE value but lag in SSIM value and vice versa.

3. Methodology

Video segmentation is the first module, which converts the input video into video frames. The segmented frames are stored in an 2D array. LSTM networks are a subset of RNNs capable of establishing long-term connections. The video frames are sent one frame at a time to the second module i.e. ConvLSTM model and it extracts the features of first frame, stores it and takes the next input frame. It predicts the next frame using the features in LSTM memory cells. In third module, predicted frames are compared with the Ground Truth data and results are analyzed.

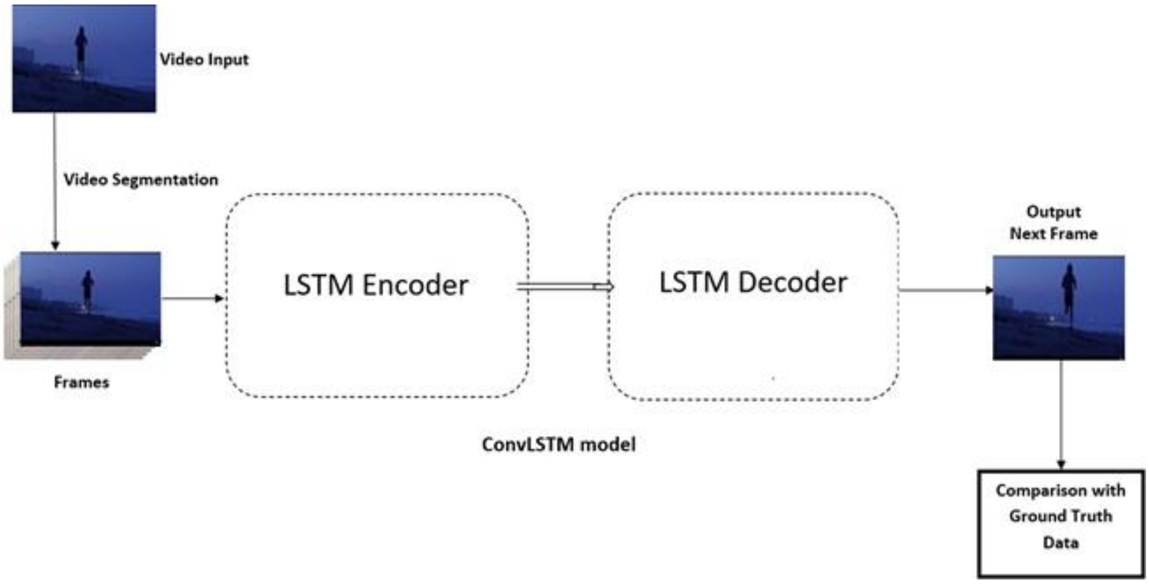

Figure 2. Architecture of the Proposed System

3.1 ConvLSTM Encoder-Decoder Architecture

The ConvLSTM model is divided into two sections: one that reads the input sequence and converts it to a fixed-length vector, and another that decodes the fixed-length vector and predicts the sequence. The encoder creates a two-dimensional output vector whose length is determined by the number of memory cells in the layer. The decoder is an LSTM layer that takes a 3D input of [samples, time steps, and features] and outputs a decoded sequence of the task's duration.

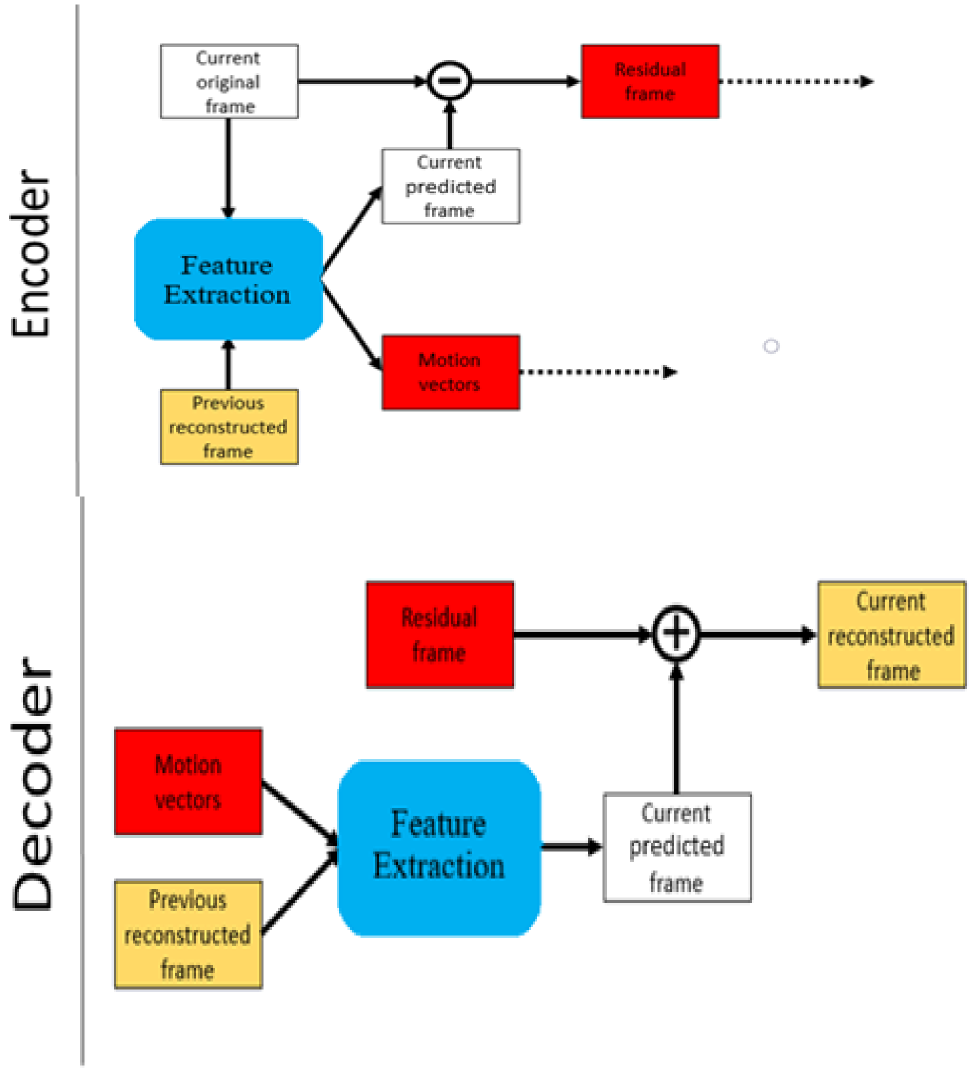

Figure 3. ConvLSTM Encoder-Decoder Architecture

A residual frame is created by subtracting the reference frame from the required frame in video compression methods. The error or residual frame is the term for this difference. Motion vector is a two-dimensional vector used for inter prediction that gives an offset from the decoded picture's coordinates to the reference picture's coordinates.

The ConvLSTM module's architecture is seen in Figure 3. The video frames are progressively delivered to the ConvLSTM model. In encoder of the ConvLSTM Architecture, the current original frame is pushed to feature extraction where it will form the motion vectors and a predicted frame with the help of previous reconstructed frame. A residual frame will be constructed from the original and anticipated frames. The Residual frame and Motion vectors are then given to the ConvLSTM model's decoder. In decoder, the motion vectors given by the encoder is used along with the previous reconstructed frame to extract the features to create a predicted frame. Then the residual frame of encoder and the predicted frame is combined to form a reconstructed frame. This frame is used as a memory cell for the ConvLSTM model which will be used in next encoder stage. After using five frames in the encoder-decoder phase, the predicted frame generated in the last decoder phase will be the output of the system.

To assess the prediction's accuracy, the system's output frame is compared to the ground truth data. The more similar the predicted and actual pictures are, the more precise the system. The model's performance is measured using the Perceptual Distance and Structural Similarity Index (SSIM).

3.2 Segmentation of video frames

First step in video prediction is to segment the video into frames. This is important as feature extraction of a video is to be done from these frames. Segmentation is referred as breaking down of video into individual frames which when coupled together form the video. The segmented frames are feed into the ConvLSTM module individually which stores the unique features to predict future frames of these input frames.

Suppose the video is of 5 seconds and the frames generated after each 0.5 seconds of difference, we get 10 frames which differ in their characteristics every 0.5 seconds. The threshold is not fixed and can vary based on our requirements and significance. The more the frames generated, more is the complexity for ConvLSTM modules. This increases the accuracy but it is not feasible as it requires high end machines to process such amount of data and takes lots of time. So, threshold must be ideal in order to process results with no such issues.

3.3 ConvLSTM Model Building and Training

In this module, ConvLSTM system model is built using different layers of LSTMs and trained from the dataset which is partitioned to training and testing directories. The image is sent through the Convolutional-LSTM model's convolutional layers, and the result is a set of features flattened to a 1D array. Padding is necessary prior to performing the convolution process to guarantee that the states have the same number of rows and columns as the inputs. At first the video frames are pre-processed by reshaping the images to lower dimensions and augmented using Gaussian Noise. For our model, combination of ConvLSTM2D and Conv2D layers are used. The height and width are configured before using in the model.

3.4 Next Frame Generation

In this module next frame is predicted using the trained ConvLSTM model. It consists of two parts. First, the generator model is called where the predicted 6th frame is added to the directory. Second is the prediction part where the model is called with testing directory.

3.4.1 Generator model

The Generator function concatenate the five frames from the input frames from the directory and after prediction, it adds the 6th frame to the same directory.

The generator function is used to add the 6*th* predicted frame using the ConvLSTM model and store in the directory. Here, it takes the five frames and concatenate each other and add a 6*th* frame which is generated in the prediction part. The frames are randomly shuffled for users to watch the predicted output randomly. At last, the train directory and test directory are added.

3.4.2 Prediction

The trained ConvLSTM model is called here with the generator model so that after prediction of 6th frame it is saved in the directory. The steps and epochs are added as arguments to the calling function.

In prediction part, the model is called with the generator function. The testing directory is included in the model. This function takes the configured batch size, the training directory, steps per epoch and configured number of epochs.

3.5 Comparison of the model's performance to the original data.

We have used two standard parameters to compare and analyze the predicted image of our system i.e. Perceptual Similarity and Structural Similarity Index (SSIM).

3.5.1 Perceptual distance

Perceptual similarity is the value of rate of change of colour perception between the actual output image and our predicted output.

Color perception is a complex and poorly understood topic that defines even the most sophisticated mathematical representations. Lesser the value of Perceptual similarity, better the image is predicted. C1 and C2 are colors of a pixel of predicted image and actual image respectively.

The distance between colors C1 and C2 (where each of the red, green and blue channels has a range of 0-255) is derived through equations (1) to (5):

$$r = \frac{\left(C_{1,R} + C_{2,R}\right)}{2} \tag{1}$$

$$\Delta R = C_{1,R} - C_{2,R} \tag{2}$$

$$\Delta G = C_{1,G} - C_{2,G} \tag{3}$$

$$\Delta B = C_{1,B} - C_{2,B} \tag{4}$$

$$\Delta C = \sqrt{\left(2 + \frac{r}{256}\right) x\, \Delta R^2 + 4\, x\, \Delta G^2 + \left(2 + \frac{255 - r}{256}\right) x\, \Delta B^2} \tag{5}$$

where r is the mean. C1, C2 are colors. ΔR, ΔG, ΔB are difference between C1 and C2 red channel, green channel and blue channel respectively.

3.5.2 SSIM

The Structural Similarity Index (SSIM) is a perceptual metric that measures image quality loss as a result of processing such as data compression or transmission losses. It's a full reference metric that necessitates the use of two photos from the same capture: a reference image and a processed image. Typically, the image after processing is compressed. The SSIM values range between 0 to 1 where 1 means a perfect match between the original image and the predicted one.

In our case, one is predicted image and one is actual image. SSIM Index quality assessment index is based on the computation of three factors; luminance (l), contrast (c) and structure (s). The overall index is a multiplicative combination of the three:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (6)$$

where μ_x, μ_y, σ_x, σ_y, and σ_{xy}, are the local means, standard deviations, and cross covariance for images x, y.

3.6 Dataset Description

UCF101 has 101 action classes and more than 13,000 clips and 27 hours of video footage, making it significantly larger than other datasets. The database is made up of actual user- submitted films with camera movements and a crowded backdrop. The majority of existing action recognition datasets suffer from two disadvantages:

i. In comparison to the variety of performed activities by humans in reality, such as KTH and Weizmann datasets, the number of their classes is often relatively low.

ii. The videos were shot in unnaturally controlled settings.

UCF101, on the other hand, addresses both issues because it is made up of web videos that were shot in unrestricted contexts and often feature camera motion, variable lighting conditions, partial occlusion, low quality frames, and other factors that make it realistic to use. We have used 24 categories from the dataset.

4. Results

The results obtained for different category are discussed in this section. The first five frames are the input frames taken from the videos segmented from the dataset and the sixth frame is the predicted next frame generated from proposed ConvLSTM model. Figure 4 to Figure 8 show the sample results for different categories Sports, Cars, Birds, Instruments and Exercises respectively.

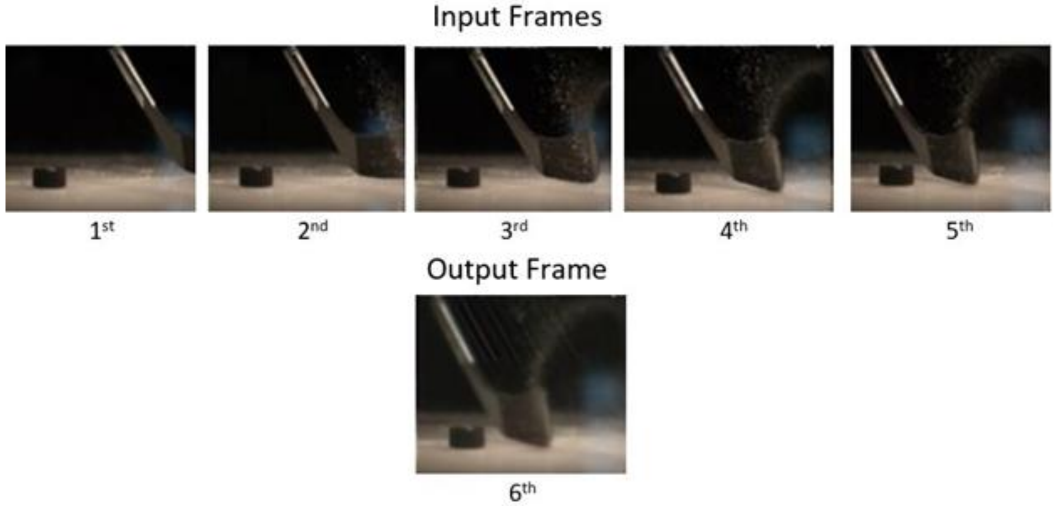

Figure 4. In 1st frame to 5th frame the hockey stick is gradually moving towards the puck and in predicted 6th frame, hockey stick is further moved forward.

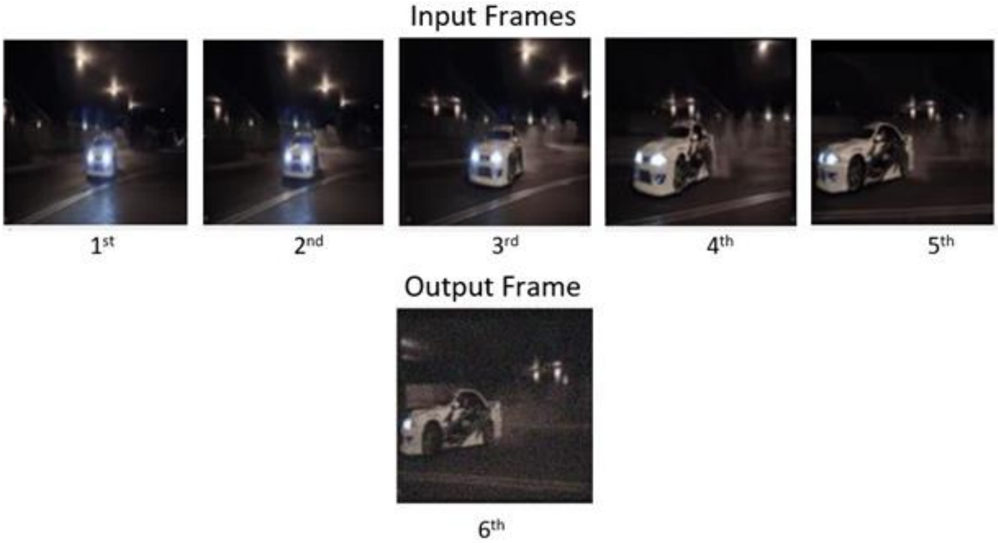

Figure 5. In 1st frame to 5th frame, the car is drifting on the road and in 6th frame car is further drifted.

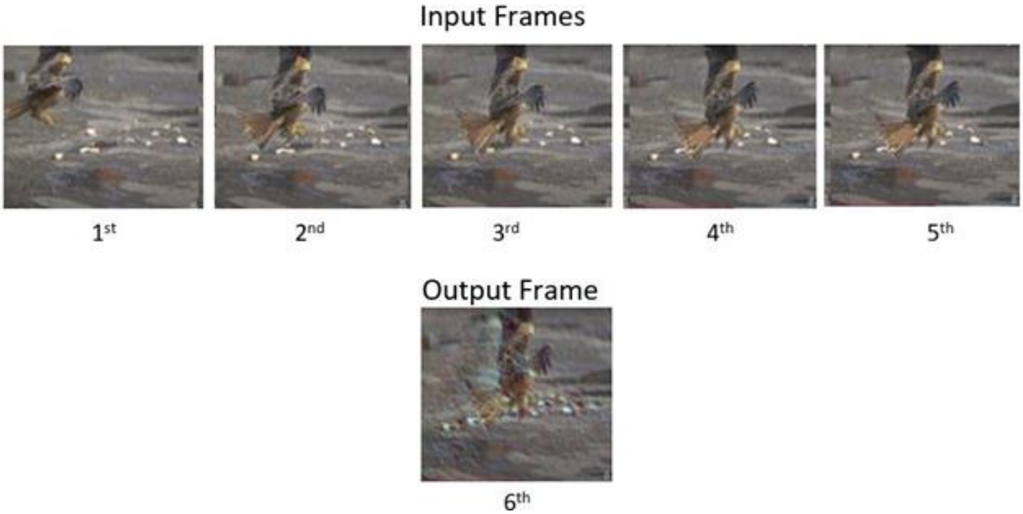

Figure 6: In 1st frame to 5th frame, a bird landing on ground and using claws to pick food and in 6th frame it further moved.

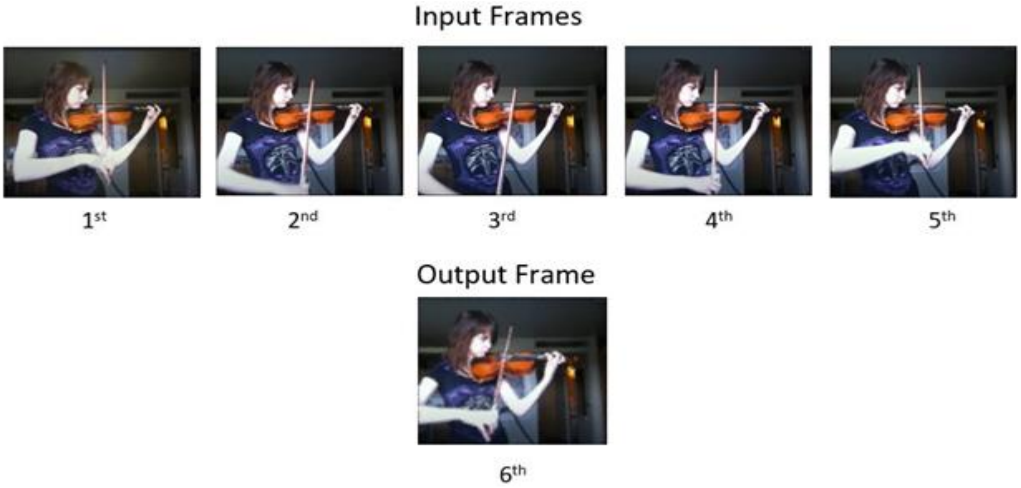

Figure 7. In 1st to 5th frame, a girl is playing violin and in predicted 6th frame the hand movement is changed.

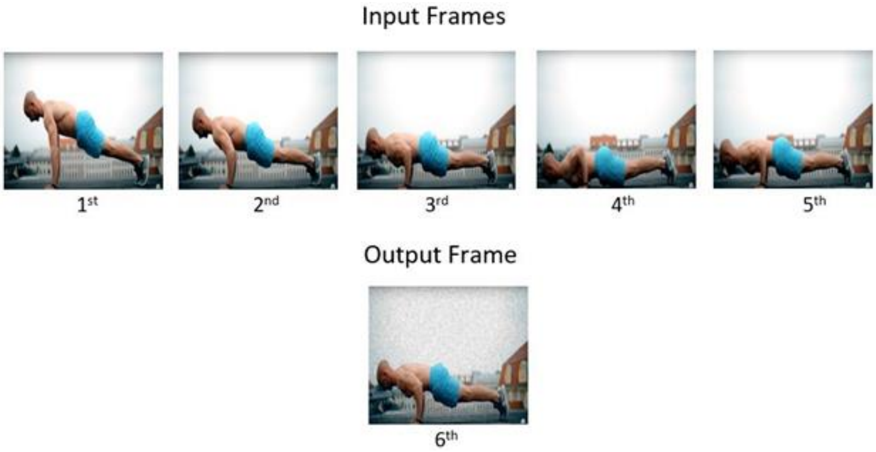

Figure 8. In 1st frame to 5th frame a guy doing Push-Up going downwards and then upwards and in the 6th frame the guy moved further upward.

4.1 Evaluation parameters

4.1.1 Perceptual Similarity

The perceptual similarity is the value of the rate of change of colour perception between the actual output image and our predicted output. Average perceptual similarity is measured by taking sample video from different categories and showed in table 1. The average Perceptual Distance of the system computed is 24.28 pu (perceptual units) taking five different categories. The lowest perceptual similarity is obtained for the category Exercises is 21.237.

Table 1. Perceptual similarity of different Categories

Category	Perceptual Similarity (p u)
Birds	25.786
Cars	26.638
Sports	24.997
Instruments	22.542
Exercises	21.437
Average	**24.28 p u**

4.1.2 Structural Similarity Index (SSIM)

SSIM is used to measure the similarity between two given images, i.e., predicted image and original image. SSIM for different categories of our model is shown in table 2. The average SSIM value is 0.95 of the system. For CARs category, SSIM value was foundout to be highest i.e., 0.97

Table 2. SSIM values for different Categories

Category	SSIM
Birds	0.95
Cars	0.97
Sports	0.94
Instruments	0.93
Exercises	0.96
Average	**0.95**

4.1.3 Comparison of SSIM with other models

The SSIM of our system is compared to available state-of-the-art methods in Table 3.

Table 3 SSIM values of different Models

Models	**SSIM**
SAVP [7]	0.86
HRPAE [9]	0.82
DVF [6]	0.94
ConvLSTM(Proposed System)	0.95

5. Conclusion

The suggested work makes a significant contribution by predicting the future frame using ConvLSTM. The prediction of the sixth frame is achieved by giving the input of the previous five frames. With 24 categories of videos from the UCF101 dataset, we achieved a mean perceptual similarity of 24.28 p u and a mean structural similarity index (SSIM) of 0.95 for the entire system. The SSIM value of our proposed system is found to be better when compared to the existing methods discussed in the result section. The ConvLSTM model predicts frames output more efficiently.

The models of Generative Adversarial Network (GAN) or its other variants can increase accuracy. Using better hardware, such as faster processors and more GPU cores,can also help to improve the accuracy of the predicted frame's resolution. The system can also be used to anticipate numerous frames and create a video output for future prediction.

6. References

[1] Wu, Y., Gao, R., Park, J. and Chen, Q., 2020. Future video synthesis with object motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5539-5548).

[2] Hosseini, M., Maida, A.S., Hosseini, M. and Raju, G., 2019. Inception-inspired lstm for next-frame video prediction. arXiv preprint arXiv:1909.05622.

[3] R. Mahjourian, M. Wicke and A. Angelova, "Geometry-based next frame prediction from monocular video," 2017 IEEE Intelligent Vehicles Symposium (IV), 2017, pp. 1700-1707, doi: 10.1109/IVS.2017.7995953.

[4] Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J. and Argyros, A., 2020. A review on deep learning techniques for video prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] Vondrick, C., Pirsiavash, H. and Torralba, A., 2016. Anticipating visual representations from unlabeled video. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 98-106).

[6] Byeon, W., Wang, Q., Srivastava, R.K. and Koumoutsakos, P., 2018. Contextvp: Fully context-aware video prediction. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 753-769).

[7] Kaur, J. and Das, S., 2020. Future Frame Prediction of a Video Sequence. arXiv preprint arXiv:2009.01689.

[8] Kwon, Y.H. and Park, M.G., 2019. Predicting future frames using retrospective cycle gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1811-1820).

[9] Fan, K., Joung, C. and Baek, S., 2020. Sequence-to-Sequence Video Prediction by Learning Hierarchical Representations. Applied Sciences, 10(22), p.8288.

[10] Villegas, R., Pathak, A., Kannan, H., Erhan, D., Le, Q.V. and Lee, H., 2019. High fidelity video prediction with large stochastic recurrent neural networks. Advances in Neural Information Processing Systems, 32, pp.81-91.

[11] Rosello, P., 2016. Predicting future optical flow from static video frames. Retrieved on: Jul, 18, p.2.

[12] Lin, H.I. and Huang, Y.C., 2019, August. Ball trajectory tracking and prediction for a ping-pong robot. In 2019 9th International Conference on Information Science and Technology (ICIST) (pp. 222-227). IEEE.

[13] Y. Miao, H. Dong, J. Al-Jaam, and A. El Saddik, 2019, A deep learning system for recognizing facial expression in real-time," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 15, no. 2, pp. 3301–3320.

[14] Kalchbrenner, N., Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A. and Kavukcuoglu, K., 2017, July. Video pixel networks. In International Conference on Machine

Learning (pp. 1771-1779). PMLR.

[15] Oh, J., Guo, X., Lee, H., Lewis, R. and Singh, S., 2015. Action-conditional video prediction using deep networks in atari games. arXiv preprint arXiv:1507.08750.

[16] Lotter, W., Kreiman, G. and Cox, D., 2016. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104.X.

[17] Liang, X., Lee, L., Dai, W. and Xing, E.P., 2017. Dual motion GAN for future-flow embedded video prediction. In proceedings of the IEEE international conference on computer vision (pp. 1744-1752).

[18] Klein, B., Wolf, L. and Afek, Y., 2015. A dynamic convolutional layer for short range weather prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4840-4848).

[19] Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H. and El Saddik, A., 2018. EVM-CNN: Real-time contactless heart rate estimation from facial video. IEEE transactions on multimedia, 21(7), pp.1778-1787.

[20] Desai, P., Pujari, J. and Kinnikar, A., 2016, August. Performance evaluation of image retrieval systems using shape feature based on wavelet transform. In 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-5). IEEE.

[21] Desai, P.D., Pujari, J. and Yaligar, N., 2012. Shape Based Features Extracted Using Wavelet Decomposition and Morphological Operators. International Journal of Advanced Research in Computer Science, 3(3).

[22] Desai, P., Pujari, J. and Sujatha, C., 2021. Impact of Multi-Feature Extraction on Image Retrieval and classification Using Machine Learning Technique. SN Computer Science, 2(3), pp.1-9.

[23] Desai, P., Pujari, J., Sujatha, C., Kamble, A. and Kambli, A., 2021. Hybrid Approach for Content-Based Image Retrieval using VGG16 Layered Architecture and SVM: An Application of Deep Learning. SN Computer Science, 2(3), pp.1-9.

[24] Desai, P., Sujatha, C., Shanbhag, R., Gotur, R., Hebbar, R. and Kurtkoti, P., 2021, June. Adversarial Network for Photographic Image Synthesis from Fine-grained Captions. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE.

[25] C Sujatha, Uma Mudenagudi, 2015, Gaussian mixture model for summarization of surveillance videos, pp. 1-4, National Conference on Computer Vision, Pattern Recog-nition, Image Processing and Graphics (NCVPRIPG).

[26] Nayak, U.T., Sujatha, C., Kamat, T.V. and Desai, P., 2021. Video Retrieval Using Residual Networks. In Advances in Computing and Network Communications (pp. 367-377). Springer, Singapore.