Al-Qadisiyah Journal of Pure Science

Volume 26 | Number 4 Article 43

8-15-2021

Human Motion Prediction Using Wavelet Transform

Wafaa Shihab Ahmed
University of Technology, Baghdad, Iraq, 111798@student.uotechnology.edu.iq

Abdulamir A. Karim
University of Technology, Baghdad, Iraq, 110004@uotechnology.edu.iq

Recommended Citation

Ahmed, Wafaa Shihab and Karim, Abdulamir A. (2021) "Human Motion Prediction Using Wavelet Transform," *Al-Qadisiyah Journal of Pure Science*: Vol. 26: No. 4, Article 43.
DOI: 10.29350/qjps.2021.26.4.1354
Available at: https://qjps.researchcommons.org/home/vol26/iss4/43

Al-Qadisiyah Journal of Pure Science

QJPS

ISSN(Printed): 1997-2490 ISSN(Online): 2411-3514

DOI: /10.29350/jops.

http://qu.edu.iq/journalsc/index.php/JOPS

Human motion Prediction Using Wavelet Transform

Authors Names
a.Wafaa Shihab Ahmedr
b.Abdulamir A.Karim

Article History

Received on:14/6/2021
Revised on: 30/7/2021
Accepted on:18/8/2021

Keywords:
Wavelet Transform,
Convolution Neural Network,
Variational Auto Encoder (VAE),
Long Short Term Memory (LSTM).

DOI: https://doi.org/10.29350/ jops.2021.26. 4.1354

ABSTRACT

The goal of prediction human motion is to analyze a subject's behaviors based on observed sequences and produced future body motions. The results of conventional methods that predict the values of pixels produced blurring and artifacts frames and the accumulative error will lead to poor quality of prediction. So to obtain good video prediction, in this work the deep neural network has been employed and proposed using wavelet transform with CNN-VAE model to analyze the input data to multi scales and extract features using CNN to encode it by VAE model, LSTM model has been used to predict encoded data and decoded it by used CNN-decoder to produce the new predicted frames. The propose system achieved best results in PSNR, MSE and SSIM and made the time of training and testing (prediction) faster. The experiments have been applied on two dataset: KTH and Weizmann and generate realistic video of 1200 ms. The best accuracy was (0.97) with KTH dataset of waving motion.

1. Introduction

In the field of computer vision, motion video understanding has become one of the most important tasks. When opposed to static images, the temporal feature of video provides considerably better representations of the real world, such as interactions between objects, human behaviors, and so on. The procedure of predicting the future has gained the attention in the research field in a variety of fields [3].

Human motion prediction aims to analyze a subject's behaviors based on observed sequences and to produce future body motions. The approaches which based on Deep learning have outperformed traditional methods on pixel based problem and skeleton-based problem such as 3 Dimension poses estimation [24] and actions recognition [7], [11], [19].

The modelling of human motion is a classic problem at the confluence of graphics and computer vision, with applications ranging from human-computer communication to motion synthesizing to virtual and augmented reality motion prediction. Based on the success of deep learning techniques in a variety tasks of computer vision, researchers have recently focused on using deep recurrent neural networks (RNNs) to model human motion, with the aim of learning time-dependent representations that perform tasks like predicting short-term human motions and synthesizing long-term human's motions [15].

aUniversity of Technology, Baghdad, Iraq, E-Mail: 111798@student.uotechnology.edu.iq
bUniversity of Technology, Baghdad, Iraq, E-Mail:110004@uotechnology.edu.iq

One of the deep neural network methods which has been used in this paper is convolutional neural network with variational auto encoder (CNN-VAE) model and LSTM model. In this paper the wavelet transform has been used with CNN-VAE model to analyze the input data to multi Structure scales and to make the time of training and testing faster.

2. Related Work

Below are some related works clarify some methods used for predicting the human motion.

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.

P. Ghosh et al., in 2017 [6], 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.

R. Villegas et al., in 2017 [20], 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 [13], 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 spatialtemporal 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.

K. Xu et al., in 2018 [23], 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.

P. Liu, H. Zhang, W. Lian, and W. Zuo, in (2019) [12], they proposed a novel model called multi-level wavelet CNN (MWCNN), the proposed model achieved good trade-off between the size of receptive field and computational efficiency. This is done by embedded the wavelet transform into CNN structure which leads to minimize the resolution of feature map while at the same time, maximizing receptive fields. The proposed model improved the detailed filters and generalizing average, and can be used in image restoration processes. The results of experiments show the effectiveness of the novel model MWCNN for some functions like image denoising, single image super-resolution and removal artifacts in JPEG image and object classification.

3. The Preprocessing Data

In this stage the video has been framing and each frame has been processed by transforming the data from spatial domain to frequency domain using wavelet transform and the coefficients have been normalized from range [0,255] to range [0,1].

Discrete Wavelet Transform (DWT) decomposes the given image into one low-frequency sub-band and three high-frequency sub-bands using the property of dilations and translations by a single wavelet function called mother wavelet [22].

Haar wavelet transform is the oldest and most basic of the wavelet systems has constructed from the Haar basis function. The equations for forward Haar wavelet transforms and inverse Haar wavelet transform, are given by:

a) Forward Haar Wavelet Transform (FHWT)

Given an input sequence (xi) i=0…N-1, it is FHWT produce (Li) i=0…N/2-1 and (Hi) i=0…N/2-1 by using the following transform equations [22]:

1. If N is even

$$\left.\begin{aligned} L(i) &= \frac{x(2i)+x(2i+1)}{\sqrt{2}} \qquad ,i=0\ldots\left(\frac{N}{2}\right)-1 \\ H(i) &= \frac{x(2i)-x(2i+1)}{\sqrt{2}} \qquad ,i=0\ldots\left(\frac{N}{2}\right)-1 \end{aligned}\right] \qquad (1)$$

2. If N is odd

$$\left.\begin{aligned} L(i) &= \frac{x(2i)+x(2i+1)}{\sqrt{2}} \qquad ,i=0\ldots\left(\frac{N-1}{2}\right) \\ H(i) &= \frac{x(2i)-x(2i+1)}{\sqrt{2}} \qquad ,i=0\ldots\left(\frac{N-1}{2}\right) \\ L\left(\frac{N+1}{2}\right) &= x(N-1)\sqrt{2} \\ H\left(\frac{N+1}{2}\right) &= 0 \end{aligned}\right] \qquad (2)$$

b) Inverse Haar Wavelet Transform (IHWT)

The inverse one-dimensional HWT is simply the inverse to those applied in the FHWT; the IHWT equations are [22]:

1. If N is even

$$\left.\begin{aligned} x(2i) &= \frac{L(i)+H(i)}{\sqrt{2}} \quad ,i = 0 \ldots \frac{N}{2} - 1 \\ x(2i+1) &= \frac{L(i)-H(i)}{\sqrt{2}} \quad ,i = 0 \ldots \frac{N}{2} - 1 \end{aligned}\right\} \quad (3)$$

2. If N is odd

$$\left.\begin{aligned} x(2i) &= \frac{L(i)+H(i)}{\sqrt{2}} \quad ,i = 0 \ldots \frac{(N-1)}{2} \\ x(2i+1) &= \frac{L(i)-H(i)}{\sqrt{2}} \quad ,i = 0 \ldots \frac{(N-1)}{2} \\ x(N-1) &= L\left(\frac{N+1}{2}\right)\sqrt{2} \end{aligned}\right\} \quad (4)$$

4. Convolution Neural Network (CNN)

CNN is a very popular model for deep learning. These are especially appropriate for images as inputs, but they are often used in other tasks e.g. text, signals and other continual responds. The key distinction between CNN and other NN types is that the CNN input is an image, whereas the NN input is a numerical value (e.g. a feature vector). CNN includes three layers which are convolution layers, max-pooling or average-pooling layers, and fully-connected layers [1],[9],[8].

5. Variational Autoencoders

A variational autoencoder is an architectural that combines an encoder and a decoder and is trained to reduce the reconstructed errors between encoded-decoded data and the original data. However, in order to incorporate some regularization of the latent space, the encoding-decoding process has been somewhat modified: rather than encoding an input as a single point, it has been encoded as a series of points over latent space [17].

6. Long Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a more advanced version of recurrent neural network. Hochhreiter and Schmidhhuber (1997) proposed it as a solution to the vanishing gradient problems in the simple RNN. In many investigations, LSTM has been shown to be reliable and powerful for learning long-range dependency [18], [4], [16]. Fig.1. illustrate the structure of LSTM.

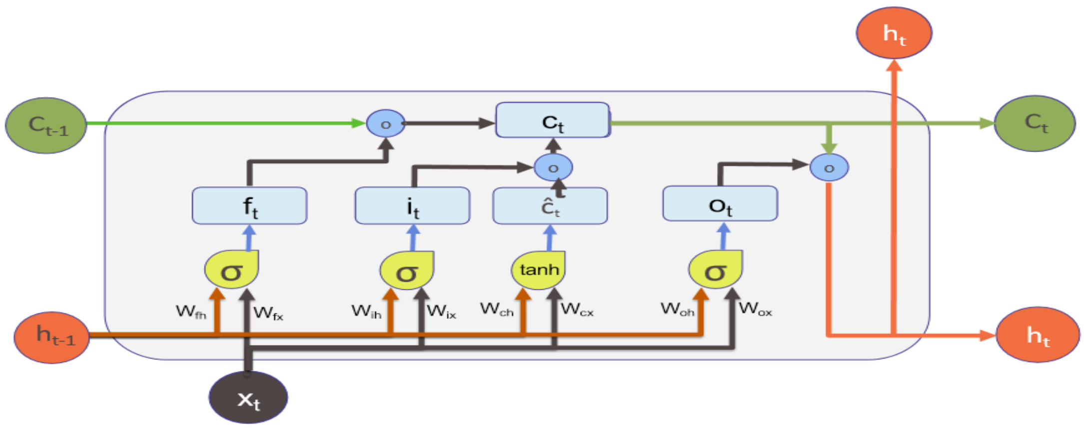

Fig 1. - LSTM block at time t [16]

7. The Proposed System

In the proposed system CNN-VAE model and LSTM model have been used to learn the representation of the input subband frames (LL) which acts the human motions such as (walking, boxing and waving). CNN-VAE model including cnn-encoder and cnn-decoder. The encoder receive the features from cnn and representing as latent variables by compute the variance (σ) and mean (μ) values and used in sampling operation using equation (5).

$$Sampling = \mu + \exp(0.5 * \sigma) * epsilon \qquad (5)$$

Where epsilon is random normal [0,1].

These latent variables have been learned by an encoder. The decoder part is the opposite of the encoder, where the sampled point is entered as input to the dense layer (fully connected) for decoding the values and the output of this layer will enter as input to the convolution layers. The result (Y) is compared with input image (X) by using equation (8) and compute the loss value by using equation (9) , the kullback divergence value has computed by equation (6 and 7) [9]:

$$KL\_Loss = 1 + (\sigma) - (\mu)2 - \exp(\sigma) \qquad (6)$$

$$KL_{Loss} = -0.5 * mean\ (KL_{Loss}) \qquad (7)$$

$$Reconstructed\_Loss = (X - Y)^2 \qquad (8)$$

$$Loss\ Value = \text{mean (Reconstructed } Loss + KL\ Loss) \ (9)$$

These steps repeated until reach to minimum loss value. The weights have been saved in file (vae-weights.h5).

CNN-VAE model has been illustrated in Fig.2. This model consist of two phases, training and prediction phases.

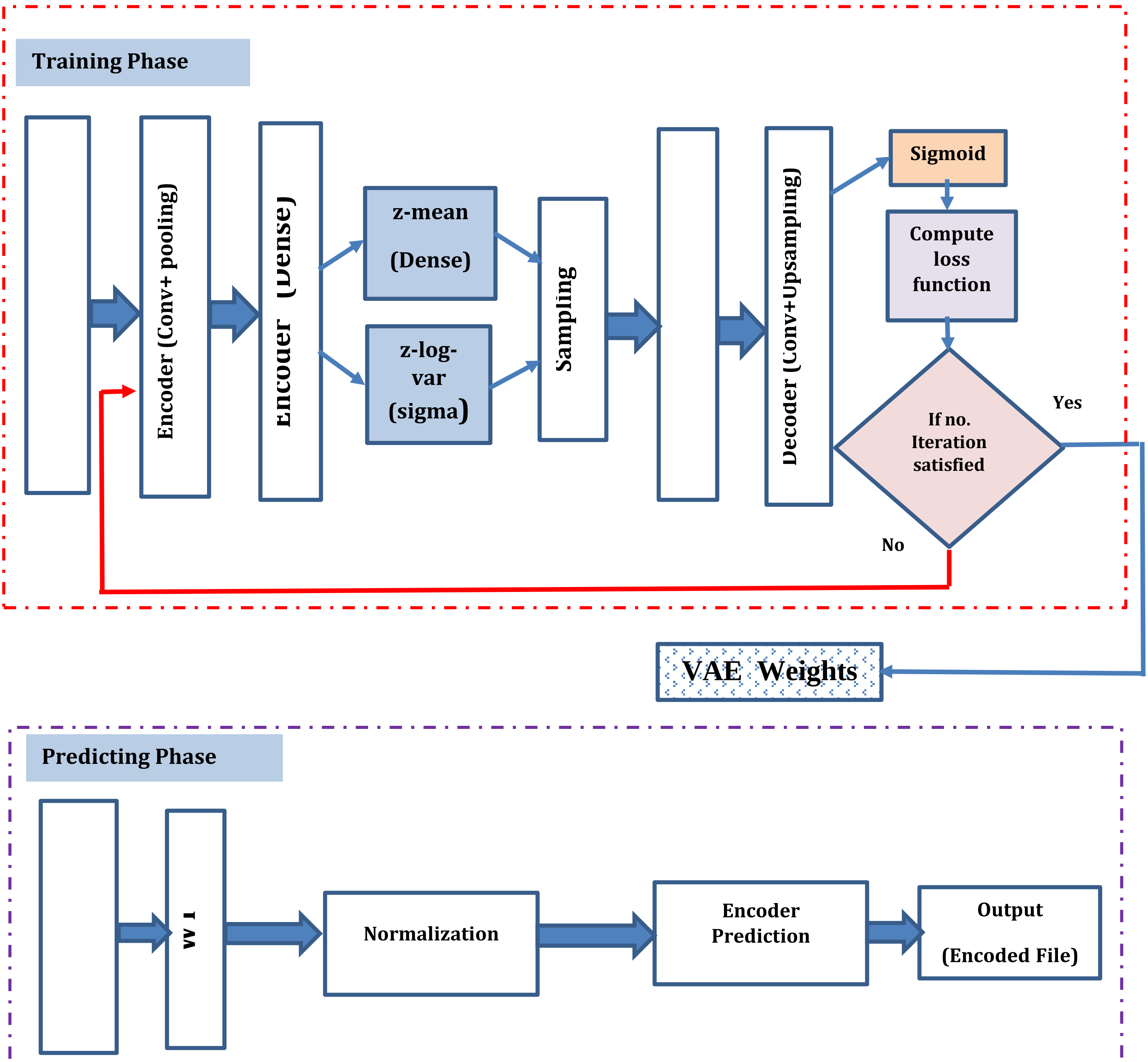

Fig. 2 - The Structure of proposed CNN-VAE Model

In the proposed system the process of generation is based on CNN-VAE model and on LSTM model. CNN_VAE model used to extract features and encoding it, after that the LSTM model has been used for training the encoding data (compressed data) to generate new frames. Fig. 3. Show the LSTM for training.

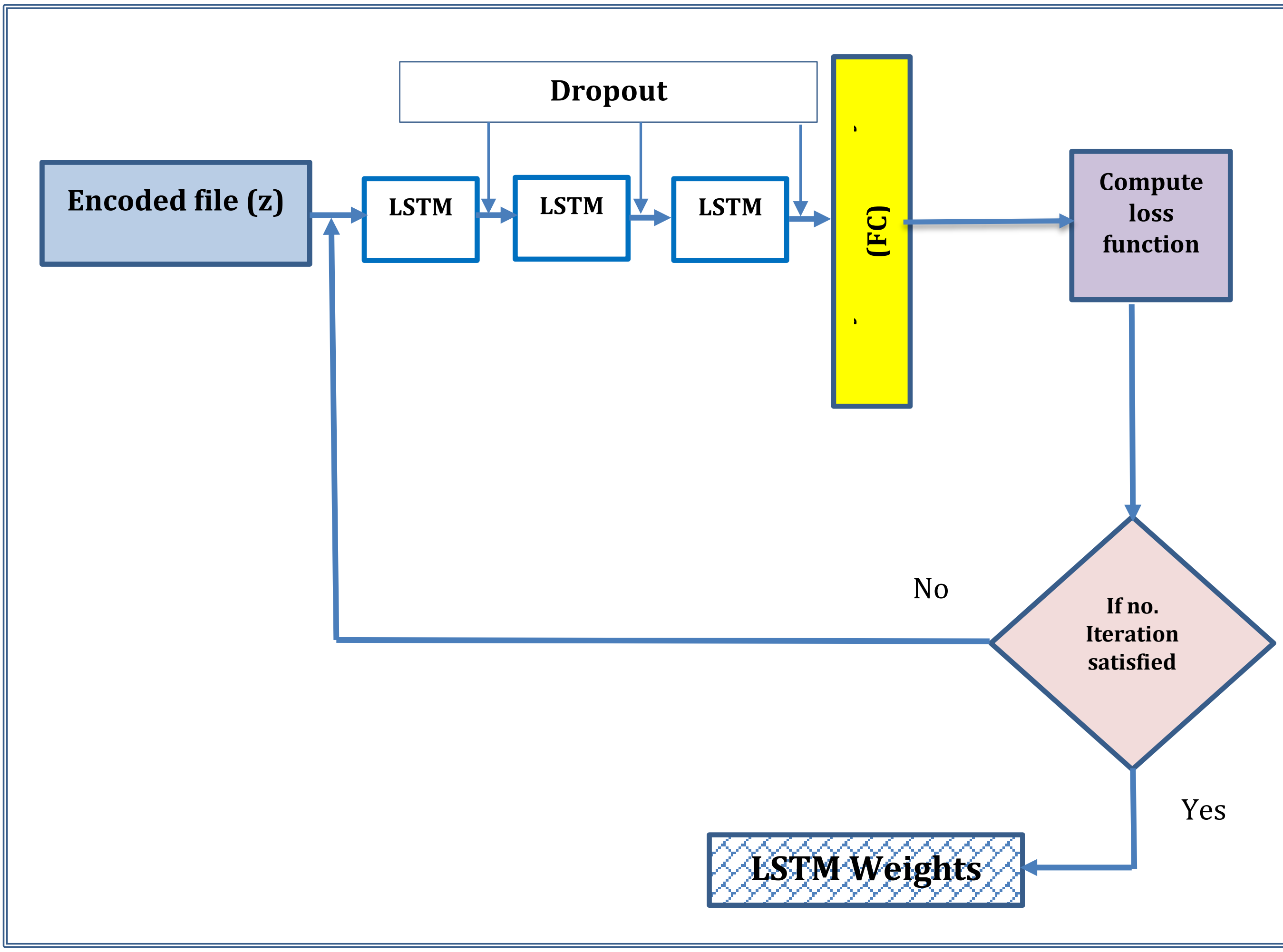

Fig. 3. - LSTM Structure for Training

7.1 The Training System

In the proposed system there are two training: training the CNN-VAE model and training the LSTM model. In the training CNN-VAE model the CNN-VAE encoder includes three convolutions layers with different number of filters (32, 64 and 128) and filter size 3×3. The number of max pooling layers are three and two dense layers (fully connected layer) with number of nodes 128. The activation function which has been used with each convolution layer is Relu. In the CNN-VAE decoder the number of convolution layers are four with three upsampling layers and two dense layers, sigmoid activation function used in the last convolution layer.

In the LSTM training the input to this model is encoded representation which stored in file and three LSTM layers have been used in the training, each layer have 512 nodes and after each layer the coefficients have been dropout. The output of lstm entered as input to the fully connected with number of nodes 1000. The error value between the input (Z) and output (Z ') has been computed by MSE measure. The weights of this network has been stored in file (lstm-weight.h5).

7.2 The Testing (Prediction) System

In the testing or prediction phase the future frames have been generated by using the weights of cnn-vae model and weights of lstm-mdn model.

In the process of generation as in Fig. 4. The input frame will transformed from spatial domain to frequency domain by using haar wavelet transform equations (1 and 2) and normalized, the coefficients will be encoded by using equation (5) and predicted LSTM model the result will be the new predicted encoded samples these samples have been decoded by using CNN-VAE decoder and the same result will back to LSTM model as input to predict the next encoded samples. The CNN-VAE decoder produced new reconstructed image this image has been normalized to original range. The result image represent the LL band from the original image. Each of remaining bands (LH, HL and HH) have been training and predicted as the same steps which the LL band has been trained and predicted. All reconstructed bands (LL, LH, HL and HH) have been concatenated to produce the new image. This new image enter to the inverse wavelet transform using equations (3 and 4) to produce the new frame, these new frames have been converted to a video.

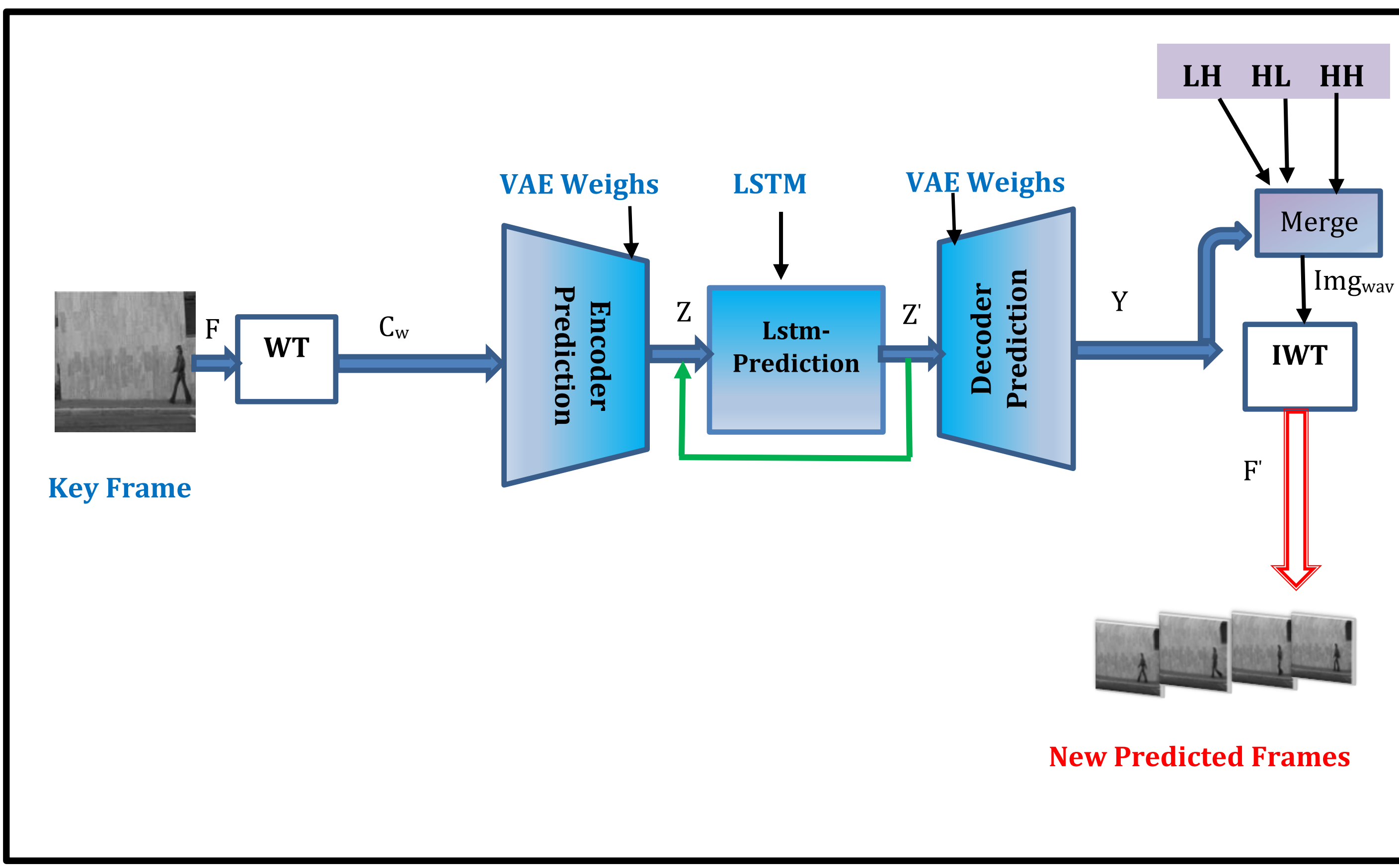

Fig. 4. - The Structure of the proposed Generation Phase.

8. The Results of Experiments

In this work the experiments have been implemented on two datasets Weizmann and KTH datasets. The motions which have been generated are walking and waving.

Below are the datasets which have been used in this work.

KTH dataset

This is includes 6 types of actions (boxing, hand clapping, hand waving, jogging, running and walking). This dataset contain on 599 action videos, these are taken by 25 various subjects with 4 scenarios (outdoors, outdoors with scale variations, outdoors with various cloths and indoors) [17]. This dataset is download from the website in reference [10], [2].

Weizmann dataset

This dataset consists of 10 classes of actions like "walking", "jogging", "waving" taken by 9 separate individuals to get a sum of 90 video clips. The video is shot with a static camera under a simple background [10], [3]. This database is downloaded from the website in reference [21].

In this work the wavelet transform has been implemented to transform the image from spatial domain to frequency domain the results of this implementation are show in Fig. 5:

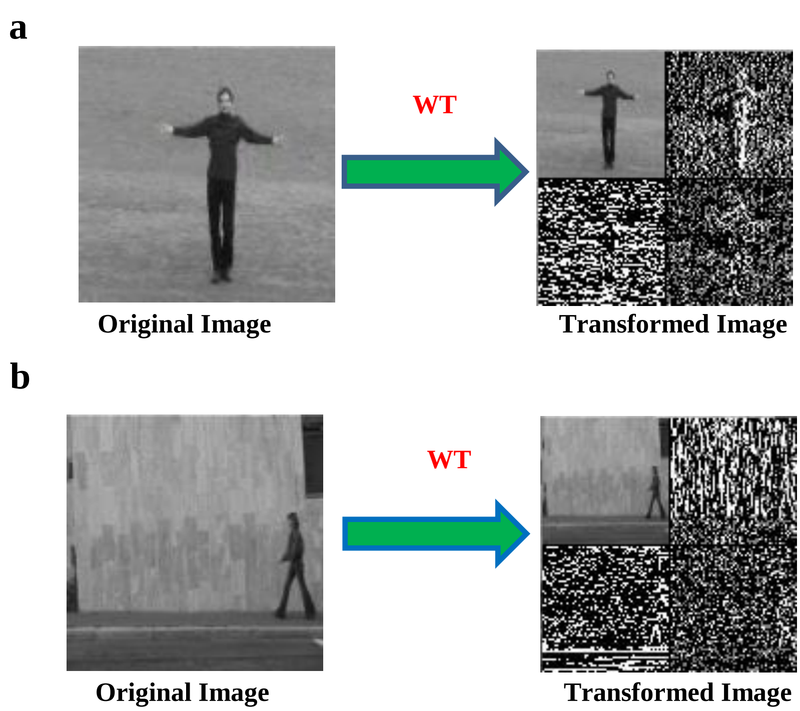

Fig. 5. – (a) KTH dataset for waving motion; (b) Weizmann dataset for walking motion.

In this work the experiments have been implemented on the subbands of transformed images in KTH dataset to generate new frames for waving motion and implemented on Weizmann dataset to generate walking motion by using the proposed system. The PSNR, MSE and similarity measures have been computed to measure the quality of new frames. Table 1. Show the quality measures, number of generated frames and time of each generated video measured in millisecond (ms). In this work we used 5 frames per second.

Dataset	Motion	Video	No. of generated Frames	MSE	PSNR	Similarity	Time of video in (ms)
Weizmann	Walking	1	25	Max=20.7 Min=7.46 Ave=9.51	Max=39.41 Min=34.97 Ave=38.46	Max=0.98 Min=0.96 Ave=0.974	5000 ms
Weizmann	Walking	2	17	Max=13.27 Min=7.67 Ave=9.08	Max=39.29 Min=36.90 Ave=38.61	Max=0.98 Min=0.97 Ave=0.972	3400 ms
Weizmann	Walking	3	17	Max=16.22 Min=11.59 Ave=13.99	Max=37.49 Min=36.03 Ave=36.69	Max=0.97 Min=0.97 Ave=0.97	3400 ms
KTH	Waving	3	11	Max=12.34 Min=3.85 Ave=5.74	Max=42.27 Min=37.22 Ave=32.97	Max=0.976 Min=0.958 Ave=0.971	1200 ms
KTH	Waving	4	11	Max=186.1 Min=5.61 Ave=60.61	Max=40.67 Min=25.43 Ave=40.84	Max=0.97 Min=0.86 Ave=0.93	1200 ms

Table 1 – The measures of quality frames

From Table 1, video 2 in Weizmann dataset gave best PSNR with best SSIM and low MSE value. Video 3 in KTH dataset gave best PSNR and SSIM with low MSE this is due to using wavelet transform which is reduced the dimension and removed the redundancy, so the prediction process has been implemented on each subband, this lead to reduce the training time.

The accuracy and loss value of CNN-VAE model training with number of epochs and batch size have been illustrated in Table 2 and in Fig. 6 and 7, the PSNR values for the two datasets have been illustrated in Fig. 8 and 9. Fig. 11. Show the qualitative comparison between the ground truth (original) frames and our proposed model.

dataset	No. of Training Frames	No. of epochs	Batch size	Learning rate	Accuracy of Training	Loss value
KTH	500	10000	10	0.0001	0.97	0.03
Weizmann	192	5000	5	0.0001	0.93	0.07

Table 2 – The accuracy and loss value for the CNN-VAE model training.

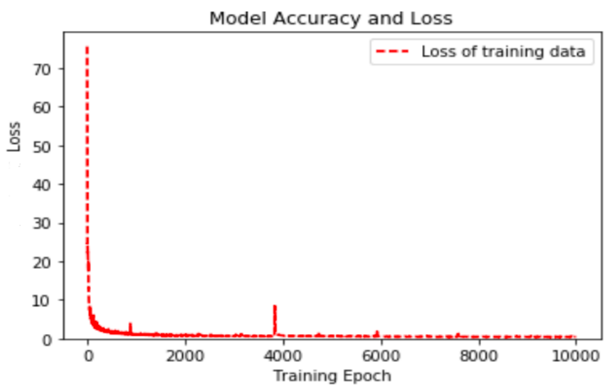

Fig.6. – The loss value of cnn-vae model of KTH dataset.

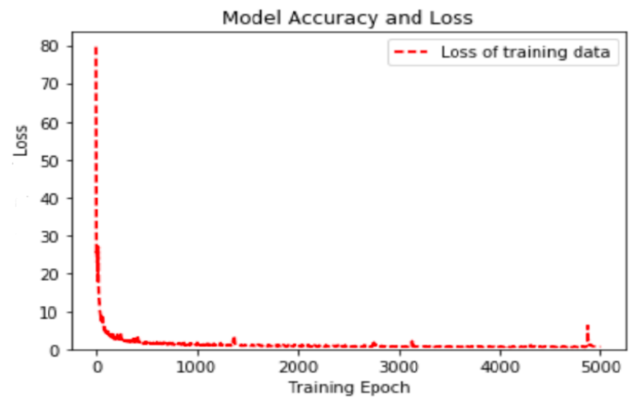

Fig.7. – The loss value of cnn-vae model of Weizmann dataset.

Fig.8. – The PSNR values for generated frames of Weizmann dataset.

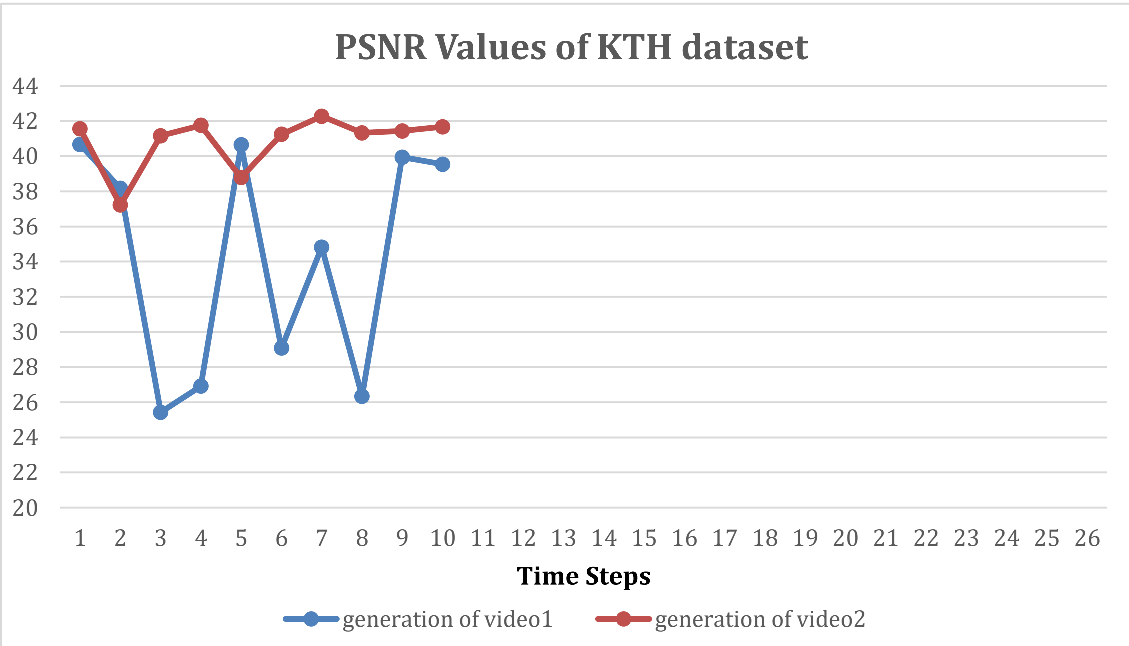

Fig.9. – The PSNR values for generated frames of KTH dataset.

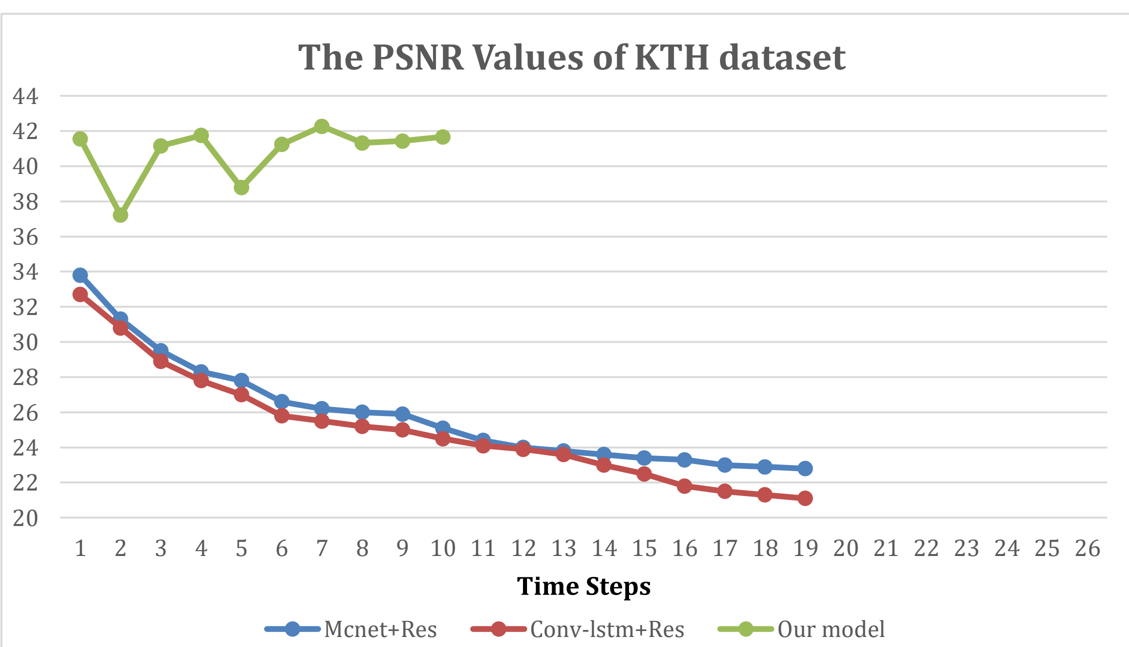

Fig.10. – The Comparison of PSNR values between conv lstm +res , mcnet+res and our model of KTH dataset.

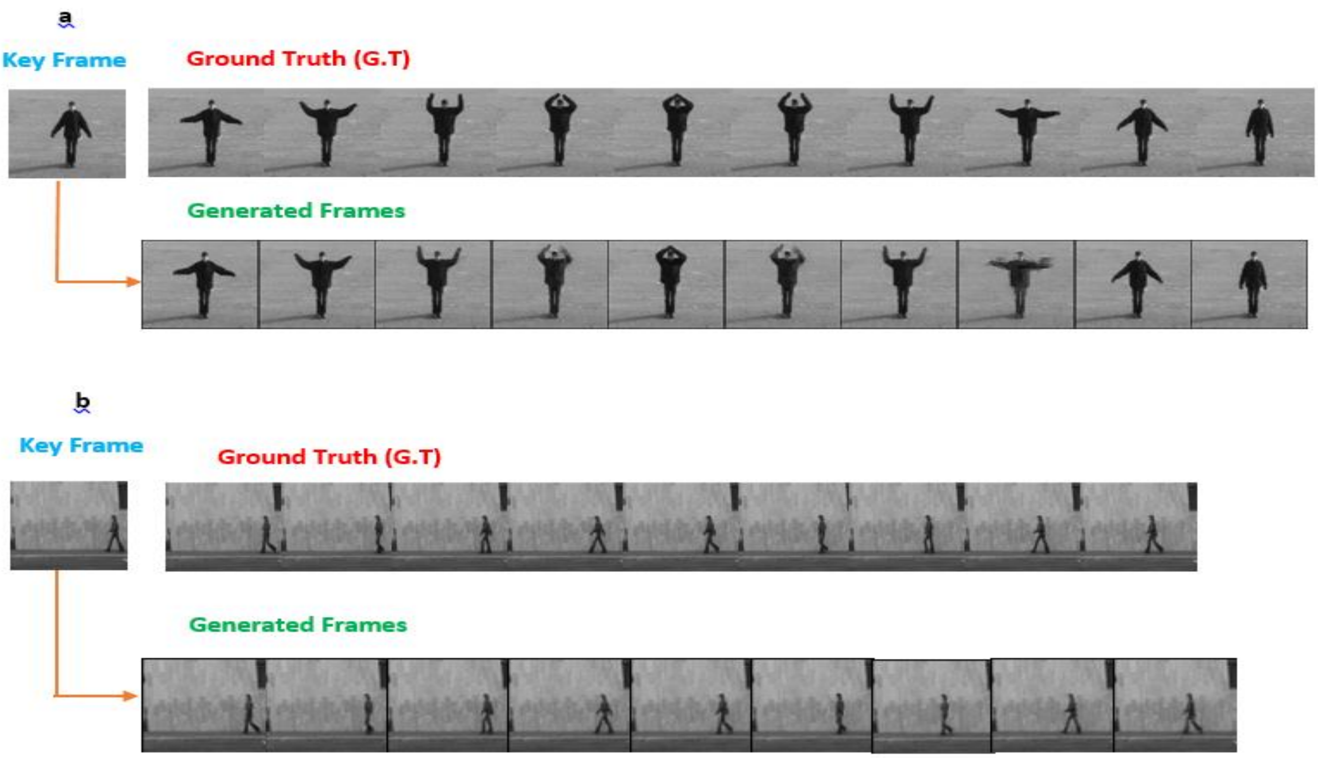

Fig. 11. – Qualitative comparison between our model and ground truth (a) KTH dataset for generation waving motion; (b) Weizmann dataset for generation walking motion.

Table 1, Fig. 8, 9 and 10 showed the system improved the PSNR values of the generated frames and the reconstructed frames have good quality comparison with other models results. In Fig. 10 the Conv LSTM+ res and MCnet models in the previous studies suffer from deformation over the time, while the proposed model maintained the quality of the frame and did not suffer from deformation over time.

Table 3. Illustrate some previous works comparison with our work.

Year	Authors	Method	Dataset	PSNR	SSIM
2017	Villegas et al.	MCnet model	KTH, Weizmann, UCF-101	PSNR of KTH First frame=38.0 last frame= 22.8, For Weizmann first frame=36.9 last frame=26.2	SSIM of KTH First rame=0.95 last frame= 0.75, for Weizmann first frame=0.97 last frame= 0.81

2017	Y. Wang, et al.	Spatio temporal LSTM with gate contoroller dual memory structure	Moving Minst, KTH and Radar echo	PSNR of KTH first frame=33.8, last frame=26.7	SSIM of KTH first frame=0.94 last frame=0.83
2018	K. Xu et al	CONV LSTM for edge guided prediction	KTH	PSNR of first frame=33.1 and last frame=24.9	
2018	Y. Li et al.	Optical flow with conditional VAE	KTH, Waving Flag and Floating Cloud		SSIM of KTH first frame= 0.97 and last frame=0.86
2021	**Our Proposed System**	Wavelet Transform with CNN-VAE model	KTH and Weizman	PSNR of KTH first frame=40.67 and last frame=39.54, for Weizmann first frame=39.29 and last frame= 37.92	SSIM of KTH first frame= 0.97 and last frame=0.92, for Weizmann first frame=0.98 and last frame=0.97

Table 3 – The Comparable between Previous Works and Our Work (Proposed System)

9. Conclusion

In this work the features have been extracted and encoded based on frequency domain by using the wavelet transform with CNN-VAE model. Haar wavelet transform reduced the dimensionality of the input image to the CNN VAE model and reduced the training time. The LSTM model used to train the encoded data and used the lstm weights for prediction the new encoded data and decoded it to produce new frames to construct video (long term prediction video). The proposed system achieved good results in PSNR, MSE and SSIM comparison with other models (Conv LSTM+res and MCnet) models. By the proposed system the time of video which has been generated is longer than generated video by the other models and the prediction video from the proposed system did not suffer from deformation and blurring over time.

References

[1] Ahmed, Wafaa Shihab. "Motion Classification Using CNN Based on Image Difference." In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), pp. 1-6. IEEE, 2020.
[2] Ahmed, Wafaa Shihab. "The Impact of Filter Size and Number of Filters on Classification Accuracy in CNN." In 2020 International Conference on Computer Science and Software Engineering (CSASE), pp. 88-93. IEEE, 2020.
[3] W. S. Ahmed and A. A. Karim, "Human Motion Imagination and Prediction- A Survey," MJPS, vol. 8, no. 2, pp. 30-45, 2021.
[4] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," in Journal of the American Medical Informatics Association, vol. 24(2), pp. 361–370, August (2016). ISSN 1527-974X. doi: 10.1093/jamia/ocw112.

[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, 2015, pp. 4346-4354.
[6] 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).
[7]] J.F. Hu,W. ShiZheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgbd activity recognition,". In Proceedings of the IEEE conference on computer vision and pattern recognition, (2015), pp. 5344–5352.
[8] Karpathy, G. Toderici, and S. Shetty, "Large-Scale Video Classification with Convolutional Neural Networks", Computer Vision and Pattern Recognition. IEEE, pp. 1725-1732, (2014).
[9] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", International Conference on Neural Information Processing Systems. Curran Associates Inc. pp. 1097-1105, (2012).
[10] KTH dataset http://www.nada.kth.se/cvap/actions/.
[11] 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, Springer, (2016), pp. 816–833.
[12] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level Wavelet Convolutional Neural Networks," (2019), pp. 1-12.
[13] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu and M. Yang, " Flow-Grounded Spatial Temporal Video Prediction from Still Images," ECCV, Springer, (2018), pp. 1-16.
[14] 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, (2018), pp. 5226-5234.
[15] J. Martinez, M. J. Black, and J. Romero, " on human motion prediction using recurrent neural network," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv:1705.02445, (2017), pp. 2891-2900.
[16] W. Mingkuan, "Sequential Images Prediction Using Convolutional LSTM with Application in Precipitation Nowcasting," master's thesis, University of Calgary, Calgary, (2019).
[17] J. Rocca, 2019. Understanding Variational Autoencoders (VAEs). https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73
[18] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," CoRR, abs/1402.1128, (2014).
[19] 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), (2018), pp. 935-941.
[20] 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. URL http://arxiv.org/abs/1402.1128.
[21] Weizmann dataset http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.
[22] W. Witwit, Y. Zhao, K. Jenkins and S. Addepalli, "Global motion based video super-resolution reconstruction using discrete wavelet transform," Multimed Tools Appl (2018) 77, pp. 27641–27660.
[23] K. Xu, G. Li, H. Xu, W. Zhang and Q. Huang, " Edge Guided Generation Network for Video Prediction," IEEE, (2018).
[24] 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, (2016), pp. 4948–4956.