Half-Yearly Progress Report for Jan-May 2020 and July-Nov 2020

Data Sheet for Ph.D Scholars

Name: Sonam Gupta

Registration No: CS18D005

Department: Computer Science and Engineering

Date of Joining: 16th July 2018

Date of Upgradation (if any):

Specialization / Stream: Computer Vision

Area of Research work: Future Frame Prediction in Video

Category of Admission: HTRA

Guide: Prof. Sukhendu Das

Co-Guide(s):

Date of DC meetings:

Description	Event	Date
1st DC meeting	Comprehensive Viva	Attempt 1: 20th August 2019
2nd DC meeting	Research Proposal Seminar (1st Seminar)	
3rd DC meeting	Mid-Term Review DC meeting (3-3.5 years from the date of joining)	
4th DC meeting	Research Colloquium (2st Seminar)	
Six monthly DC meeting	After five years from the date of registration, upto maximum period of the programme	SMD 1: SMD 2: SMD 3: SMD 4:
Final DC meeting	Synopsis at Dean AR Office	

Details of Course Work

S.No	Course No	Course Title	Sem/Year	Credit	Grade
Core Courses					
1.	CS6015	Linear Algebra and Random Processes	1	12	S
2.	CS5691	Pattern Recognition and Machine Learning	1	15	B
Elective Courses					
1.	CS5020	Nonlinear Optimization: Theory and Algorithms	1	12	A
2.	CS5800	Advanced Data Structures & Algorithms	1	12	A
3.	CS6350	Computer Vision	2	12	S
4.	CS7015	Deep Learning	2	12	A
5.	CS6870	Digital Video Processing	2	12	B
6.	CS6777	Optimization Methods for Computer Vision Applications	3	12	S
Compulsory Courses / Optional Courses					
1.	ID6020	Introduction to Research (Institute Module)	1	0	Pass
2.	CS6021	Introduction to Research	1	0	Pass

Signature of Scholar

Signature of Guide

Contents

i) Title of Research Work
Future Frame Prediction in Videos

ii) Problem Definition / Research Objectives
Much work in video prediction tasks have been focused on generation of a deterministic future. The output of such models generally represents either a deterministic or the average of the possible futures, and is also blurred. Also, the future prediction is done only upto a few frames as it gets blurry for longer durations. The blurriness arises due to the model's inability to predict and retain high frequency information for frames farther in time.

During my PhD research work, I am trying to design new Deep Learning methods, built on top of existing models to address the above challenges in video prediction. In particular, I will work on the following aspects of the problem:

1. Predicting frames so that both high and low frequency information is retained.
2. Exploring appropriate evaluation metric for evaluating the results of multiple future predictions.
3. Proposing a new problem statement for predicting an abstract concept/representation in future.

iii) Summary of Work Done before Review (From Date of Admission till now)
During the first semester, I worked mostly upon improving/gaining knowledge in foundation courses that will help in research. I took Linear Algebra and Random Processes (LARP), Non-Linear Optimization (NLO) courses in which I learnt about Matrices and their properties, Eigen values and Eigen Vectors, Singular Value Decomposition, Convex Sets, Convex Functions, Primal Dual Formulations. These mathematical tools will enable me to understand and formulate the mathematical explanations better. As Machine Learning and Deep Learning are prevalent in Computer Vision community, I took Pattern Recognition and Machine Learning course wherein I have learnt about Regression, Gaussian Mixture Models, Hidden Markov Models, Support Vector Machine through assignments and class lectures. Also, to brush up the Data Structures and Algorithms concepts, I took Advanced Data Structure and Algorithms course. The course covered advanced topics like Max Flow algorithms along with other Graph Algorithms. Many problems in domain of Computer Vision can be mapped to graphs. I believe that the fundamentals built by learning these courses will help me to better understand the literature in the area. With these four courses, I completed the coursework requirement of PhD.

In second semester I primarily focussed upon building domain knowledge through courses and literature. I did three courses namely Deep Learning, Computer Vision and Digital Video Processing. The concepts learnt in these courses would help me to understand the state of art methods in the literature. Computer Vision course exposed me to different problems that exists in literature like Depth Estimation, Segmentation, Object Detection and Recognition. I also learnt some of the possible ways in which these problems can be solved (partially or fully) using classical (Non-deep) methods. The course also covered fundamentals of Computer Vision and various handcrafted features like edges, corners, SIFT, SURF which have been a great success in the past and are still used successfully in many applications. As part of the course, I did a term project, in which I worked on the problem of finding the Depth from a monocular RGB Image. For this project, after following the literature, I have used an existing Deep Learning method based on
Transfer learning [1]. The methods followed an Encoder-Decoder based architecture, where both Encoder and Decoder were CNN based models. I also got to explore different loss functions for the same architecture.

The Deep learning course covered topics like Multi-Layer Feed Forward Neural Networks, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long short term memory (LSTM), Auto Encoders. I implemented these concepts for tasks like Image classification, Function Approximation, Image Annotation, Image Captioning.
Digital Video Processing course, covered the seminal papers for Background Subtraction, Edge Detection, Corner Detection, Markov Random Fields based methods for Segmentation, Stereo Matching, Shape based matching etc. This course introduced me to various challenges like texture, illumination, scale, rotation that may come when proposing an algorithm.

In third semester, I carried out literature survey for the task of Future Frame Prediction so that I can scope out the problem to work upon. I explored the current papers in some of the top conferences as CVPR, NIPS, ECCV. Based on the literature review I attempted and qualified the comprehensive viva in August 2019. During third semester, I appeared for comprehensive viva. I presented the literature review covering some of the state-of-the-art methods [2, 3, 4, 5, 6, 7, 8] for Future Frame Prediction. I explained the methods of [2, 6] in detail. While carrying out the literature survey I read paper on different problems including Object Detection, Segmentation, Video Summarization along with Future Frame Prediction. At present, I am working on the problem of Future Frame Instance Segmentation Prediction. The motivation for this comes from methods proposed in [9]. Recently, Luc et. al [9] have shown that predicting future frames at the RGB level is more complicated than predicting frames at an abstract level such as semantic segmentation which is generally sufficient for many applications. They have empirically shown that predicting the semantic segmentation of future frames is more efficient than predicting RGB frames and then segmenting those. Subsequently, in paper [10], Luc et. al have proposed the novel task of future instance segmentation prediction. One of the underlying challenges here is that the number of instances is not constant across frames in a video sequence. Hence, the authors have proposed to predict the future frames at fixed-size convolutional feature level. The evaluation metrics (Average Precision and Mean Intersection over Union) suggests that there is a lot of room for improvement of this method. Also, the method takes about 5 days to be trained on 2 GPUs. Thus, coming up with a smaller and efficient architecture will be of help. Currently, I am working on both of the above aspects of the problem.

Among other things, I attended the course on “Optimization methods for Computer Vision Applications”. I also attended 1-week Summer School on Computer Vision held at IIIT Hyderabad in July 2019. These exposures helped me to get a broader understanding of both the old and recent trends in Computer Vision.

iv) Work Done during Review (Odd Semester 2020 and Even Semester 2020)

On exploring the future instance prediction task, the running time of the method seemed to be a bottleneck. There are four levels to be trained with each taking 2 to 5 days for training on 2 GPU system. Thus, I decided to work with a smaller model. One of the limitations of existing methods is the blurriness of frames when predicting farther time steps in future. For the first paper, I decided to address this challenge using various ideas which are explained below.

A frame becomes blurred when the edges in the generated frames are not sharp. Thus, I explored the idea of predicting both future Edges as well as Future RGB frames from a model. The predicted edges were further merged with the RGB frames using compact bilinear pooling to make the predictions sharper. MCNET [11] was chosen as the base paper. After trying few modifications like co-attention, fusion, the results were better than the base paper but could not achieve the state-of-the-art (SOTA) performance when compared with 2020 future frame prediction papers. Motivated by Domain adaptation literature, we also tried three encoder and two decoder approach where the

first encoder encoded the edge information, the second encoder encoded the RGB information and the third encoder encoded the information shared between edges and RGB frames. The results of this approach were also subpar compare to the SOTA (2020) papers.

There were parallel works which were trying to modify the LSTM itself (rather than just the CNN architectures) for the task of future frame prediction. But the existing modified LSTMs also does not pay attention to different frequencies of the signals and uses same states for both high and low frequency components of the signals. Thus, I tried to modify ST-LSTM [12] so that it uses a separate state for high and low spatio-temporal frequency content.

I am further exploring the use of Discrete wavelet transform with CNN architecture to generate future frames while capturing the frequency information (both spatial and temporal) in a video robustly. Currently, I am working on use of DWT and multi-frequency LSTMs. I first explored the effect of predicting future frames in wavelet domain using ST-LSTM [12] based recurrent neural network. One of the key observations was that the sparsity of wavelet domain leads to faster convergence of model. Hence, I have proposed a wavelet residual network which is efficient to train and also have faster inference time while maintaining the state-of-the-art performance. Currently, I have submitted a paper on this in ICIP 2021.

v) Issues affecting Research Progress, if any
None

vi) Future Plans, with proposed timeline
For the next semester, I plan to do the following:

1. Focus on proposing a new problem.
2. Submit papers to two conferences and one Journal.

vii) Visible Research Output:

(a) Paper(s) Published in Journals
None

(b) Full Paper(s) Published in Conference Proceedings
None

(c) Seminars/Workshops/Conferences/Exchange programs attended and Papers Presented

- Attended 11th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2018).
- Attended Summer School on Computer Vision at IIIT Hyderabad from July 1, 2019 to July 7, 2019. The summer school curriculum consisted of a series of lectures and demo/lab sessions. Experts talks were delivered focussing on recent advances in the Computer Vision.

(d) Awards/Honours, if any

- I was awarded a cash prize for being among the top 20 participants in the Computer Vision Summer School. I am eligible to apply for travel grant of upto 1 lakh for CVPR 2020 or ECCV 2020 oral submission.
- Awarded Alation Ph.D. Fellowship for 3 years (Jan 2020 – Dec 2021).

viii) References:

[1] I. Alhashim, P. Wonka, “High Quality Monocular Depth Estimation via Transfer Learning by”, in arXiv preprint 1812.11941, 2019.

[2] H. Cai and C. Bai, “Deep Video Generation, Prediction and Completion of Human Action Sequences,” in Proceedings of the European Conference on Computer Vision, 2018.

[3] J. Xu, B. Ni and X. Yang, “Video Prediction via Selective Sampling,” in Advances in Neural Information Processing Systems, 2018.

[4] P. Bhattacharjee and S. Das, “Predicting Video Frames using Feature Based Locally Guided Objectives,” in Asian Conference on Computer Vision, 2018.

[5] Walker, Jacob, et al. The pose knows: Video forecasting by generating pose futures. ICCV, 2017.

[6] P. Bhattacharjee and S. Das, “Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage Generative,” in Advances in Neural Information Processing Systems, 2017.

[7] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using LSTMs,” in International Conference on Machine Learning, 2015.

[8] M. Michael, C. Couprie, and Y. LeCun, “Deep multi-scale Video Prediction beyond Mean Square Error,” in International Conference on Learning Representations, 2016.

[9] P. Luc and et al., “Predicting Deeper into the Future of Semantic Segmentation,” in International Conference on Computer Vision, 2017.

[10] P. Luc, C. Couprie, Y. LeCun and J. Verbeek, “Predicting Future Instance Segmentation by Forecasting Convolutional Features,” in the Conference on Computer Vision and Pattern Recognition, 2017.

[11] R. Villegas, J. Yang, S. Hong, X. Lin and H. Lee, “Decomposing motion and content for natural video sequence prediction”, in the International Conference on Learning Representations, 2017.

[12] Y. Wang, M. Long, J. Wang, Z. Gao and P. S. Yu, “Predrnn: Recurrent neural networks for predictive learning using spatiotemporal LSTMs”, in Advances in Neural Information Processing Systems, 2017.