PREDICTING FUTURE ACTIVITIES OF PERSON AND HUMAN TRAJECTORY USING DEEP NEURAL NETWORK

PHASE II REPORT

Submitted by

HARSHA PRIYA G

in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANNA UNIVERSITY, CHENNAI

JANUARY 2021

ANNA UNIVERSITY, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled **"PREDICTING FUTURE ACTIVITIES OF PERSON AND HUMAN TRAJECTORY USING DEEP NEURAL NETWORK."** is the bonafide work of **HARSHA PRIYA G (2019207029)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. VALLI S
Professor and Head
Department of Computer Science and Engineering
Anna University
Chennai - 600 025

SIGNATURE

Dr. MARY ANITA RAJAM V
Professor
Department of Computer Science and Engineering
Anna University
Chennai - 600 025

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

CV	Computer Vision
ROI	Regions of Interest
ConvNet	Convolution Neural Networks
ConGD	Chalearn Gesture Dataset
3D	3 Dimension
LSTM	Long Short Term Memory
DANN	deleterious annotation of genetic variants using neural networks
MSR-VTT	Multimedia Search- Video to Text
TGIF	The Tumblr GIF
AV	Adhoc Video Search
ActEv	Activity Extended Video
NIST	National Institute of Standards and Technology
ADE	Average Displacement Error
FDE	Final Displacement Error

LIST OF FIGURES

CHAPTER 1

PROBLEM STATEMENT FOR PHASE II

In today's world we deal with humans and world via live incidents and map in our memory as videos. Video prediction has broad prospects in real-world scenarios, such as robots planning, autonomous driving, and anomaly detection in surveillance videos. Learning to predict future frames from a video sequence involves constructing an internal representation that models frame evolution accurately, including contents and dynamics.

Our project is the first on joint future path and activity prediction in streaming videos, and more importantly the first to demonstrate such joint modeling can considerably improve the future path prediction.

The problem statement and the objective of the project is to predict human trajectory and future activity using Computer vision techniques and deep learning architectures. An end-to-end, multi-task learning system utilizing rich visual features about human behavioral information and interaction with their surroundings is to be designed. To produce meaningful future activity prediction in addition to the path that benefits future path prediction. To facilitate the training, the network is learned with an auxiliary task of predicting future location in which the activity will happen.

CHAPTER 2

SUMMARY OF RELATED WORK FOR PHASE II

2.1 DIVERSE PLAUSIBLE FUTURE FRAMES PREDICTION

In the research work of Jinzhuo Wang *et al.*[7] presented a flexible and powerful video prediction system. Unlike popular video prediction methods that are performed at the global pixel level, they focused on ROIs and learn patterns of frame evolution at the transformations level. Given a sequence of frames or even a single frame,the system is able to accurately predict the next frame and long-term future frames. Moreover, it can produce diverse plausible future frames that preserve continuity and consistency with the input.

The merits of the work of Jinzhuo Wang *et al.*[7] are as follows:

- More accurate predictions is made by modeling the visual evolution.
- Focused on ROIs to avoid a heavy computational burden. Generated diverse plausible future frames.
- Performed video prediction conditioned on a single frame.

The demerits of of the work of Jinzhuo Wang *et al.* [7] are as follows:

- Video prediction is considered the challenging task of generating the future frames of a video.

2.2 GESTURE RECOGNITION USING CONVNET

In the research work of Z.Li *et al.*[5] uses a simple method for continuous gesture recognition using depth map sequences. A Depth sequences are first segmented so that each segmentation contains only one gesture. A ConvNet is used for feature extraction and classification.Experimental results on ChaLearn LAP ConGD dataset verified the effectiveness of the method. Extraction of the neutral pose and fusing different modalities to improve the accuracy is found challenging.

2.3 CRITICAL ANALYSIS OF HIGH LEVEL INTERPRETATION OF DETECTION OF IMPORTANT METHODS

In the research work of Borges *et al.* [3] provides a critical analysis of the major steps, from detection to high-level interpretation.Several important future directions, promising methods, generation and reconstruction of 3D observations, data-sets and annotation are highlighted theoretically. Anomalous behaviors are found difficult to model practically.

2.4 HUMAN TRAJECTORIES PREDICTION

In the research work of A. Alahi *et al.* [1] explains the LSTM-based model jointly reason across multiple individuals to predict human trajectories in a scene. It shows the pooling of entire hidden state to capture complex interactions in dense crowds. Multi-class settings where several objects such as bicycles, skateboards, carts, and pedestrians does not share the same space. Each object does not have its own label in the occupancy map. Human-space interaction is not modeled in this framework and does not allow modeling of human-human and human-space interactions.

2.5 DANN FOR ACTION RECOGNITION

In the research work of J.Wang *et al.* [8] A deep neural network called DANN for action recognition is introduced. To preserve motion structures in temporal domain, this method uses an adaptive method to determine the temporal size of network input and develop a metric pyramid pooling layer to resize the output before fully connected layers into fixed-size vector. The spatial size is still chosen in adhoc manner.

2.6 ADHOC VIDEO SEARCH (AVS) TASK

In the research work of G.Awad *et al.* [2] explains the motivation of mapping video embedding and language embedding into same semantic space. The models are trained on MSR-VTT and TGIF dataset with different visual and language architectures to achieve the best performance. Ad-hoc Video Search (AVS)task is only a task query video by natural language description in zero shot manner.

CHAPTER 3

HIGH LEVEL BLOCK DIAGRAM FOR PHASE II

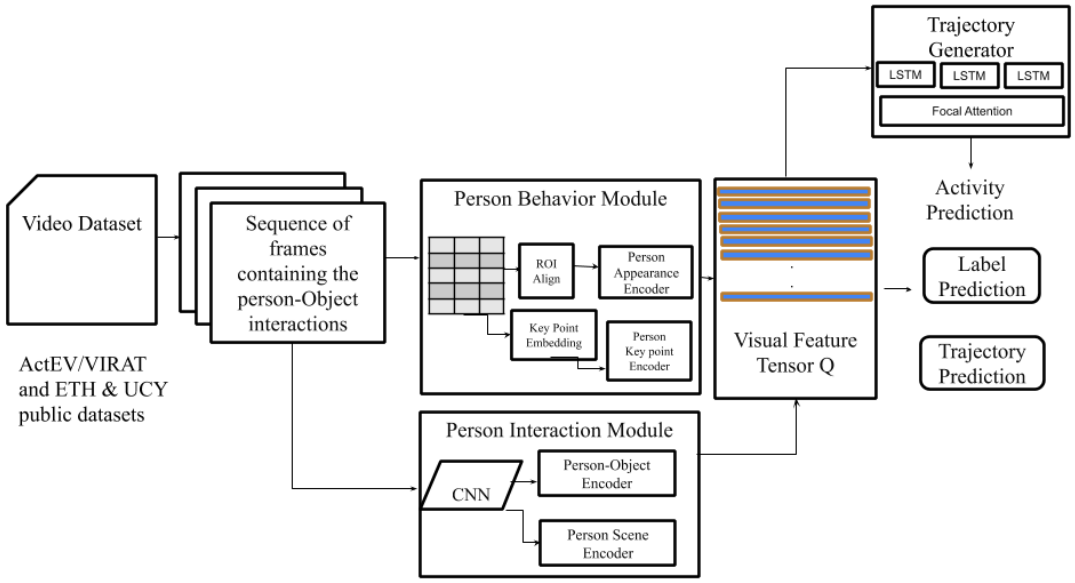

Figure 3.1: PREDICTING FUTURE ACTIVITIES OF PERSON AND HUMAN TRAJECTORY USING DEEP NEURAL NETWORK

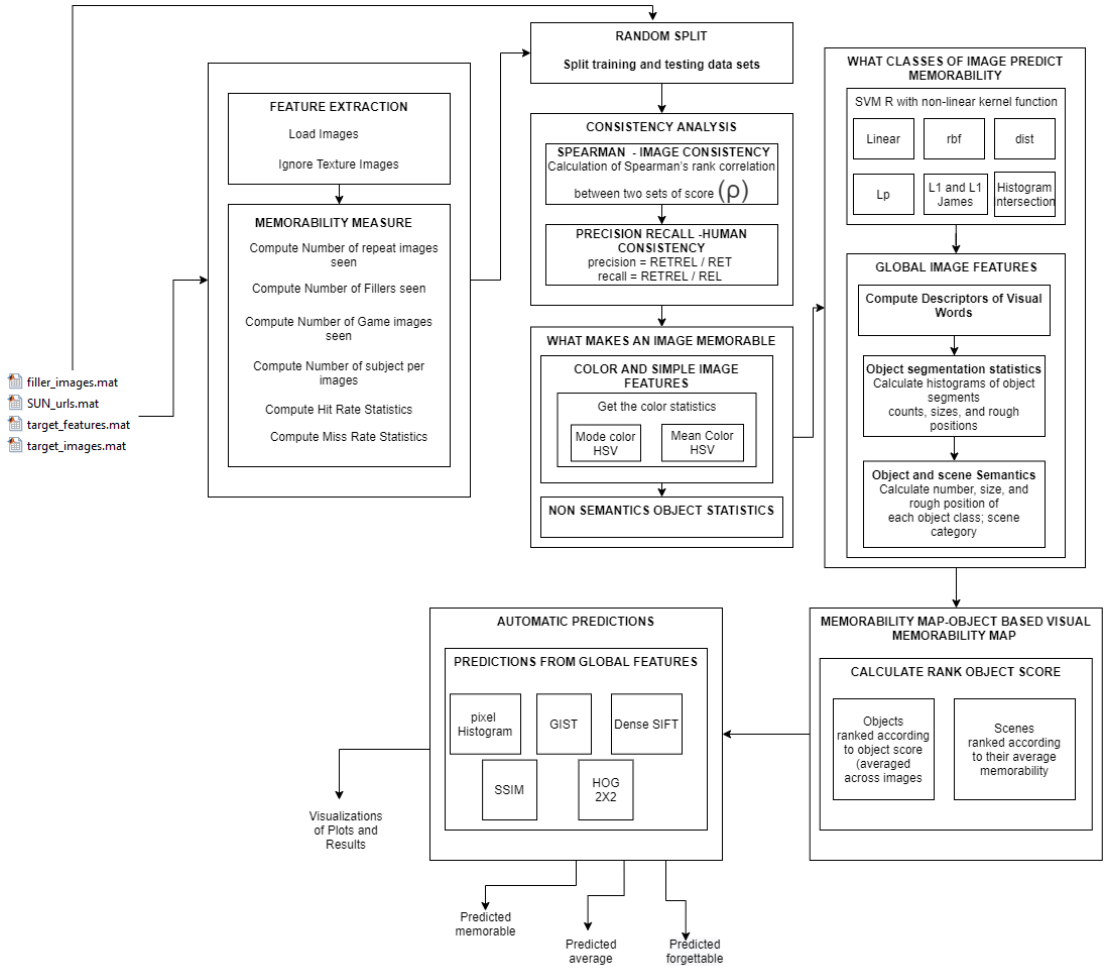

Figure 3.2: CLASSIFICATION OF IMAGES BASED ON MEMORABLE SCORE USING VISUAL MEMORY SCHEMA AND COMPUTER VISION TECHNIQUES

CHAPTER 4

FINAL DELIVERABLE OF PHASE II

4.1 FINAL OUTCOMES

A multi-task learning model which has prediction modules for learning future paths and future activities simultaneously is proposed in our project. As predicting future activity is challenging, we introduce two new techniques to address the issue.We encode a person through rich semantic features about visual appearance, body movement and interaction with the surroundings, motivated by the fact that humans derive such predictions by relying on similar visual cues.

Second, to facilitate the training, we introduce an auxiliary tasks for future activity prediction, i.e. activity location prediction. In the auxiliary task, we design a discrete grid which we call the Manhattan Grid as location prediction target for the system.

To the best of our knowledge, our work is the first on joint future path and activity prediction in streaming videos, and more importantly the first to demonstrate such joint modeling can considerably improve the future path prediction. We empirically validate our model on two benchmarks: ETH UCY, and ActEV/VIRAT.

Experimental results will show that our method outperforms state-of-the-art baselines, achieving the best-published result on two common benchmarks and producing additional prediction about the future activity.

CHAPTER 5

DATASET DESCRIPTION

- ActEV/VIRAT and ETH and UCY are public datasets released by NIST in 2018 for activity detection research in streaming video (https://actev.nist.gov/).
- This dataset is an improved version of VIRAT with more videos and annotations.
- It includes 455 videos at 30 fps from 12 scenes, more than 12 hours of recordings. Most of the videos have a high resolution of 1920x1080.
- The official training set for training and the official validation set for testing is used. The models observe 3.2 seconds (8 frames) of every person and predict the future 4.8 seconds (12 frames) of person trajectory.

CHAPTER 6

PERFORMANCE METRICS

Average Displacement Error(ADE)**:** The average Euclidean distance between the ground truth coordinates and the prediction coordinates over all time instants.

$$\textbf{ADE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_{pred}}||Y_t'^i - Y_t^i||}{NXT_{pred}} \tag{6.1}$$

Final Displacement Error (FDE): The euclidean distance between the predicted points and the ground truth point at the final prediction time instant Tpred.

$$\textbf{FDE} = \frac{\sum_{i=1}^{N}||Y_{T\,pred}'^{\,i} - Y_{T\,pred}^{\;\;i}||}{N} \tag{6.2}$$

REFERENCES

1. A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

2. G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. Smeaton, Y. Graham, et al. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. 2018.

3. P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology*, 23(11):1993–2008, 2013.

4. W. Choi and S. Savarese. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013.

5. Z. Li, W. Wang, N. Li, and J. Wang. Tube convnets: Better exploiting motion for action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3056–3060. IEEE, 2016.

6. W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

7. J. Wang, W. Wang, and W. Gao. Predicting diverse future frames with local transformation-guided masking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3531–3543, 2018.

8. J. Wang, W. Wang, R. Wang, W. Gao, et al. Deep alternative neural network: Exploring contexts as early as possible for action recognition. *Advances in Neural Information Processing Systems*, 29:811–819, 2016.