HUMAN ACTION RECOGNITION FROM DEPTH MAPS AND SKELETON DATA

PARUL SHUKLA (*née* PANDEY)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2017

HUMAN ACTION RECOGNITION FROM DEPTH MAPS AND SKELETON DATA

by

PARUL SHUKLA (*née* PANDEY)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi

October 2017

Certificate

This is to certify that the thesis titled **Human Action Recognition from Depth Maps and Skeleton Data** submitted by **Ms. Parul Shukla (nee Pandey)** to the Indian Institute of Technology Delhi for the award of the degree of **Doctor of Philosophy**, is a record of bona-fide research work carried out by her under our guidance and supervision. The thesis, in our opinion, is worthy of consideration for the award of the degree of Doctor of Philosophy. To the best of our knowledge, the work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Dr. K. K. Biswas
Professor
Computer Science & Engineering Dept.
Indian Institute of Technology, Delhi

Dr. Prem K. Kalra
Professor
Computer Science & Engineering Dept.
Indian Institute of Technology, Delhi

Acknowledgments

The entire research expeience at IIT Delhi has been very enriching. I am indebted to all those who helped me in the realization of this dissertation. Firstly, I would like to express my sincere gratitude to my supervisors Prof. K. K. Biswas and Prof. P. K. Kalra for their valuable guidance and constant encouragement during the development and completion of this dissertation. I will always be thankful for their motivation and continuous support for allowing me the freedom to explore my interests, with understanding and patience. I am forever grateful to them for always taking the time to teach me the basics and providing valuable suggestions during the course of development of this dissertation. It has been a privilege to work under their guidance.

I would also like to thank my research committee members Prof. Subhashis Banerjee, Prof. Santanu Chaudhury and Prof. Parag Singla for their valuable guidance and support. Their words of encouragement helped in maintaining the required levels of motivation. Their insightful comments motivated me to explore the recent advances in the area of Recurrent Neural Networks, for which I will always be thankful to them. I would also like to thank Prof. S.K. Gupta, Prof. Saroj Kaushik and Prof. Amitabha Bagchi for their constant encouragement and support. Further, I take this opportunity to thank all the faculty members of Computer Science and Engineering who have encouraged me with their interactions over these years.

I thank my fellow labmates and friends Aditi Kapoor and Sonia Khetarpaul for their constant support, encouragement and motivation. I thank them for all those stimulating discussions and all the fun we have had in the past few years. The entire research experience would not have been the same without them. I am thankful to Yong Du and Lena Gorelick for providing me useful insights and clarifications. I would also like to thank Ierum, Mr. Bhateja, Avinash for helping me during the initial course work. Further, I would like to thank my colleagues Nisha Jain, Kanika Modi, Prachi Jain, Sunita Tiwari, Priti Jagwani, Yamuna Shukla, Shibashis Guha, Suvam Patra, Chinmay Narayan, Shashank Sharma for helping me during the different stages of development of this dissertation. I thank Noopur Arora for her enthusiasm and hard work during the development of chapter 5. I would also like to thank all the interns and M.Tech students who have worked with me during the past few years. I would also like to extend my gratitude towards all the staff members of Computer Science and Engineering department, especially Ms. Rekha Rathi for her continuous help throughout the processing of various administrative

tasks. I would also like to thank the staff of Database-AI Lab, Mr. M. Rathinam, Mr. S. S. Negi and Mr. Anil Sharma from Virtual Reality lab for taking care of all lab related technicalities. I would also like to thank the team of IIT Delhi involved in maintaining this institute to be such a peaceful, pleasing and research-oriented institute.

Last but not the least, I would like to express my gratitude towards my family for showing immense patience and for their constant support and encouragement. I thank my parents and brother for always believing in me and providing the much needed moral support and feedback. I am thankful to my entire family living nearby or far, for helping me in numerous ways, especially my grandfather Prof. K. N. Pathak for being an inspiration for all of us in the family. I would also like to thank my parents-in-law for always supporting and encouraging me to pursue my dreams. I thank my husband Ambika Prasad for providing me the much needed unconditional support throughout. This dissertation would not have been possible without his encouragement and support. The thesis would be incomplete without mentioning my daughter, Aadya. I am thankful to her for providing the reason to go that extra mile. I dedicate this thesis to all my family, friends and teachers who have helped and motivated me throughout life. Above all, I thank almighty for his blessings and kindness.

Parul Shukla

Abstract

In the recent years, rapid advancement in digital cameras and social media has resulted in widespread use and availability of images and videos. This, in-turn, has provided an impetus for automatic analysis of visual content in images and videos. Recognition of actions from videos has emerged as a promising field of research owing to many possible applications such as surveillance, human-computer interaction, content-based video retrieval, intelligent environment for assisted living etc. Further, recent advancements in sensing technology has resulted in an increased interest in the field of analysis of 3D data. Acquisition of 3D data has been greatly facilitated by devices such as Kinect. As a result, depth information about various objects and person in the scene can be easily obtained in the form of depth images. In addition to the depth maps, 3D position of body-joints is also available from Kinect. In this thesis, we address the issue of recognition of actions from sequence of depth maps and 3D skeleton data.

We propose a bag-of-feature based model for classification of actions from 3D skeleton data. To incorporate temporal information into the proposed model, we propose a method for constructing the model in a systematic and hierarchical manner. Depth maps or depth images are a richer source of 3D information since, the shape and geometric details of entire silhouette is available which may be beneficial particularly when the joints data is occluded or missing. We first study the applicability of widely used space-time interest points method. We explore the scenarios under which such a method is useful. In the literature, numerous methods have been proposed that analyse sequence of depth images by projecting them onto principal planes and accumulating the difference between consecutive projections. However, these methods only make partial observations from projected images as it involves considering the difference between consecutive projected frames. We propose to study the stacked silhouettes obtained after projection. We analyse the three 3D space-time action-shapes and obtain saliency and local orientation features from these action-shapes.

Depth images and 3D skeleton data each have their own merits. We propose a method based on fusion of information from both the modalities. Shape and motion features are extracted from depth images and skeleton data and their integration is done based on selecting the most relevant features to be used for classification.

Recent advancements in the field of neural network and their usage in numerous computer vision tasks has resulted in development of effective end-to-end systems for tasks such as object recognition, image captioning, scene understanding and action recognition. We explore one such model for developing an end-to-end system for recognition of actions from 3D skeleton data. We use Recurrent Neural network (RNN) and its variant Long Short-Term Memory (LSTM) for modelling temporal dependencies among body-joints' data. Our experimental study indicates that an efficient model with lesser number of trainable parameters can be developed by using only a subset of joints in comparison to the entire set of available joints.

Contents

List of Figures

List of Tables