AUTOMATED ANALYSIS OF SURVEILLANCE VIDEOS USING LATENT TOPIC MODELS

by

TANVEER AFZAL FARUQUIE

Department of Computer Science and Engineering

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute Of Technology Delhi

Hauz Khas, New Delhi-110016, India

July 2011

Dedicated to

everyone who has ever inspired me

Certificate

This is to certify that the thesis titled **"Automated Analysis of Surveillance Videos Using Latent Topic Models"**, which is being submitted by **Tanveer Afzal Faruquie** for the award of the degree of **Doctor of Philosophy in Computer Science and Engineering** to the **Indian Institute of Technology Delhi**, is a bona fide research work done under our guidance and supervision.

The thesis has reached the standard fulfilling the requirements of the regulations relating to the degree. The results obtained in the thesis have not been submitted to any other Institute for the award of any degree or diploma.

(Subhashis Banerjee)
Professor
Dept. of Computer Science and Engg.
Indian Institute of Technology Delhi
Hauz Khas, New Delhi-110016
India

(Prem K. Kalra)
Professor
Dept. of Computer Science and Engg.
Indian Institute of Technology Delhi
Hauz Khas, New Delhi-110016
India

Acknowledgements

PhD is a long journey; a journey that I feel blessed to have and thoroughly enjoyed. My thesis is about activity analysis in surveillance videos, but that is only a fraction of what I have learnt in this journey. Of course, I was not alone in this, I was lucky to have some wonderful people to guide and support me throughout.

It would be an understatement to say that this dissertation would not have been possible in its present form had it not been for the guidance, motivation and tremendous support from my supervisors. Prof. Subhashis Banerjee and Prof. Prem K. Kalra are incredible people whom I still have a lot to learn from. Their belief and constant encouragement made me strive for more and helped me achieve more than I could have.

I thank all my Student Research Committee members, Prof. S. K. Gupta, Prof. Santanu Chaudhury and Dr. Amit Kumar, for providing many valuable suggestions. I also thank Prof. S. N. Maheshwari and Dr. Amitabha Bagchi for their encouragement. I also thank the Dept. of Computer Science and Engineering and all the technical staff, particularly Mr. K. Kaushik, for their help on numerous occasions.

I had some wonderful friends to share this experience. My fellow students, Ayesha Chaudhury, Chetan Arora, Anant Vishnoi and Neeraj Goel brought vi-

brant energy and enthusiasm in the department. I also thank my colleagues at IBM, L. V. Subramaniam, Ullas Nambiar, Ashwin Srinivasan and Sumit Negi for their support and advice.

It is difficult to come up with words to thank one's family. Their constant support, unconditional love, sacrifice, and so much more. But, in this platform, I thank them, especially my mother for her sacrificial love and encouragement. In difficult times, I always find the strongest support in her. My brother Sameer and sister Lovi have always been my support and strength.

I want to thank my wonderful wife Sameena, for her constant love, support, help, encouragement and understanding. She sacrificed a lot for me and made this journey enjoyable. My achievement would not be possible without her unwavering support. With her by my side, I do not feel overwhelmed.

Finally, I thank God for giving us this life, this creation and its many unanswered questions.

Tanveer Afzal Faruquie

Abstract

Automated understanding and analysis of activities, a central task for visual surveillance, has become an important area of research. This is mainly due to the growth in surveillance of places such as large building complexes, railway stations, military facilities and public spaces. Computers can analyse motion patterns of objects to understand typical activities, learn semantically meaningful scene structures (such as paths commonly taken by objects), and detect deviant behaviour. As larger and more complex areas come under visual surveillance, manual analysis of activities becomes increasingly infeasible. At the same time, automated analysis of activities becomes more and more challenging. This has led to an increase in various aspects of automated analysis of surveillance videos that need to be addressed.

In this thesis, we develop several unsupervised frameworks using Bayesian network, called topic models, to automatically discover activities from complicated and crowded scenes using low level motion patterns. Using these unsupervised frameworks we address some aspects of automated analysis of visual surveillance, such as, discovering group activities in single view camera, learning global activities across multiple cameras, and finding the time dependent behaviour of activities.

Most of the existing approaches in visual surveillance for activity analysis are unable to model complex activities as they rely on simple probabilistic models, predefined rules, data specific heuristics, or elaborate feature extraction that fail for complicated scenes. In crowded scenes it is difficult to track objects and separate different types of co-occurring activities because of frequent occlusions. Our proposed frameworks do not rely on detection of objects, tracking of objects, or appearance features. Our frameworks can work with several low level features that can encode local motion patterns. Our topic models can structure dependency amongst a large number of variables to model complex activities. Any known knowledge about the data or scene can be incorporated into the framework as priors. We propose several topic models that are suitable for different aspects of surveillance and camera deployments.

We first propose a framework for single camera deployments that monitor fixed views, which are typically crowded. Crowded scenes are characterised by multiple co-occurring activities. It is difficult to separate individual activities and determine which of them co-occur frequently as group activities. We use a topic model that has a two layer hierarchical latent structure to correlate individual activities in lower layer with the group activity in the higher layer. Our model considers each scene to be composed of a mixture of group activities represented as a multinomial distribution. Each group activity is in turn composed as a mixture of individual activities represented using multinomial distributions. Each individual activity is represented as a distribution over local motion fea-

tures. We show that using this approach we are not only able to discover the usual activities present in a scene but can also extract the hidden association of these activities among themselves.

Next, we extend the above framework for multiple cameras deployment that monitors fixed views over a wide space. We assume that the topology of camera views is arbitrary and unknown. The field of views may have no overlap or any amount of overlap and the cameras need not be perfectly synchronised. We propose to use local motion features of individual camera views to jointly discover intra camera local activities and inter camera global activities without relying on any object appearances, finding camera correspondence or stitching tracks. We use a topic model having two level latent structure that considers global activities to be composed from camera specific local activities. It discovers activities localised to a camera as multinomial distributions over the local motion features of that camera. The higher level structure considers global activities as weighted mixture of local activities and uses a mixture of multinomials to model the composition. We experiment with multi-camera surveillance videos and show that our framework can discover camera specific local activities along with global activities across cameras without solving camera correspondence problem, knowing camera topology or camera calibration.

Next, we propose a framework for cameras with fixed views that also observe the time of occurrence of scenes. These type of deployments are common

in places like banks, malls etc., where normal day time activities are considered 'abnormal' during night. We use a topic model that not only discovers the activities but also their dependence over time. We propose to use agglomerative clustering on optical flow vectors to code direction and spatial information. In this model each activity is associated with not only a mixture distribution over these cluster occurrences but also on the distribution over time stamps of their occurrences. Our model can discover activity prominence at several scales ranging from few minutes to years. We demonstrate the effectiveness of our approach on multiple surveillance videos.

Finally, we conclude with some observations and suggestions for future work.

Contents