UNIVERSITY OF CALIFORNIA
RIVERSIDE

Exploration of Contextual Relationships for Robust Video Analysis: Applications in Camera Networks, Bio-image Analysis and Activity Forecasting

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Anirban Chakraborty

August 2014

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Ertem Tuncel
Dr. Stefano Lonardi

The Dissertation of Anirban Chakraborty is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The last five years of my life had been a fascinating journey in pursuit of knowledge and I am grateful to all the fantastic people who were instrumental in making it a successful one.

I am deeply indebted to my advisor Dr. Amit K. Roy-Chowdhury, who provided the vision, encouragement and all the advices I needed to proceed through my doctoral studies and complete my dissertation. He has always been a constant source of encouragement and inspirations throughout the ups and downs of my life as a researcher at UC Riverside. This dissertation would not have been possible without his guidance and active support.

I also express my heartfelt gratitude to my dissertation committee members, Dr. Ertem Tuncel and Dr. Stefano Lonardi. Dr. Tuncel has been a great teacher and mentor for me since the very early days of my student life at UCR. Attending his lectures and working with him as a teaching assistant have been invaluable experiences for me, and I cannot thank him enough for that. Special thanks to Dr. Stefano Lonardi for giving me thoughtful feedback and constructive comments on my oral exam, thesis and final defense. I would also like to thank Dr. G. Venugopala Reddy for providing me with several live-imaging datasets to work with and for all his advice on my research from biological perspective.

I owe a lot to all my fellow researchers in the Video Computing Group at UCR. I would like to thank Abir Das, Mahmudul Hasan, Katya Mkrtchyan, Nandita Nayak, Dr. Min Liu and Dr. Ahmed Tashrif Kamal for giving me some of the most valuable ideas for my research through the numerous brainstorming sessions in the lab. I learned a lot and developed as a researcher by interacting and collaborating with them. I convey

my special thanks to Elliot Staudt, Utkarsh Gaur and Dr. Chong Ding. I shall always cherish our long intellectual discussions on the most diverse topics imaginable. Research and lab-work never felt boring in our group.

I met so many wonderful persons in the Bengali community in and around Riverside and some of them have become my friends for life. I sincerely thank my two house-mates Dr. Tapan Sarkar and Tasneem Raihan for their tireless support and encouragements. They were like my family so far away from home who never let me feel homesick. I am also thankful to Dr. Tusar Saha, Anindya Ganguly, Dr. Monobina and Ambarish Mukherjee, Jayashree and Dr. Pradip Bag for their gift of so many precious and joyful moments. I am ever indebted to my aunt and uncle Sumita and Dr. Animesh Ray. Whenever I needed any help, I always knew that they were the people to bother. Without their active help and support, I couldn't have come this far.

Words cannot express the depth of my respect and gratitude to my mother Sanchita Chakraborti and my father Narayan Chandra Chakraborti for their lifelong guidance, teachings, and above all, unconditional love. Since my childhood, my father inculcated me with the idea of 'simple living, high thinking'. He was the one who motivated me to pursue an academic career. Among many other things, my mother taught me how to withstand all adversities in life through power of patience and perseverance. All credits go to these two persons for whatever I have achieved so far. I am also immensely grateful to my fiancée Arpita Bhattacharjee, whose love and support made it easy to trudge through these five long years of graduate studies.

I would also like to thank the National Science Foundation for their grant (IIS-0712253) to Dr. Amit K. Roy-Chowdhury, which partially supported my research.

Acknowledgment of previously published or accepted materials: The text of this dissertation, in part or in full, is a reprint of the material as is appears in three previ-

ously published or accepted papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all three publications, directed and supervised the research which forms the basis for this dissertation. The papers are, as follows.

1. 'Adaptive Geometric Tessellation for 3D Reconstruction of Anisotropically Developing Cells in Multilayer Tissues from Sparse Volumetric Microscopy Images' (Published in PLOS ONE, August 5, 2013). The co-authors Dr. Mariano M. Perales and Dr. G. Venugopala Reddy contributed with data-sets used in this work and technical expertise.

2. 'A Conditional Random Field Model For Tracking In Densely Packed Cell Structures' (Accepted for publication in International Conference on Image Processing, October 2014).

3. 'Consistent Re-identification in a Camera Network' (Accepted and to be published in European Conference on Computer Vision, September, 2014). The co-first author Abir Das contributed to experimentation, analysis and writing.

To my parents.

ABSTRACT OF THE DISSERTATION

Exploration of Contextual Relationships for Robust Video Analysis: Applications in Camera Networks, Bio-image Analysis and Activity Forecasting

by

Anirban Chakraborty

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, August 2014
Dr. Amit K. Roy-Chowdhury, Chairperson

Recently, there has been a surge of interest in modeling contextual information for various computer vision applications. Use of inter-object/inter-activity context has ushered in significant performance improvement over most classical video analysis approaches that independently estimate multiple variables of interest, which are, in fact, interrelated. In this work, we explore the problems of multi-camera data association, spatio-temporal cell tracking and activity forecasting, where the contextual relationship models could be extremely useful but are little studied in literature.

Most existing data association techniques focus on sequentially matching pairs of data-point sets and then repeating this process along space-time to achieve long-term correspondences. However, in multi-camera problems, simply combining the local pairwise association results often leads to inconsistencies over the global space-time horizons. We present an optimization framework to combine all pairwise data-association results over a network, which not only establishes global consistency but also improves pairwise accuracies. The proposed 'Network Consistent Data Association' (NCDA) method is also capable of handling the problem of variable number of data-points across different sets of instances in the network and its application is shown in the classic person

re-identification problem.

In spatio-temporal cell tracking problem, 3D cells within tightly clustered tissues are imaged over time at various depths of the tissue. The objective is to associate these 2D cellular projections that often lack good feature quality. We exploit the tight spatial topology of the cells in a CRF model and obtain robust pairwise similarity measures between the 2D cell segments, which are further combined together via the NCDA method to yield consistent and accurate cell lineages. The estimated lineages are utilized in the proposed tessellation based reconstruction method for generating 3D structures of individual cells.

We also explore the activity forecasting problem in continuous videos. We model the simultaneous and/or sequential nature of human activities as contextual information on a graph and combine that with the interrelationship between static scene cues and dynamic target trajectories. The forecasting problem is then solved as inference problem on an MRF model defined on this graph and high accuracy is observed throughout our experiments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many computer vision problems attempt to recognize, associate or predict the identity of a set of objects or actions from collected videos. However, most often these estimation tasks are performed independently for every variable of interest. Let us consider the case of human activity recognition as an example. Majority of the classical approaches to activity recognition model activities in videos individually, though activities are often related to one another in space and/or time and rarely occur independently. Again, in multi target tracking problems, data association and estimation of target states are mostly performed independently for each individual target, even though motions of different targets are often highly correlated. These observations, in recent years, have motivated computer vision researchers to understand the importance of modeling the inter-object/inter-activity relationships, often broadly termed as 'contextual information', in various video analysis problems. Information from the spatial co-occurrence statistics of activities in natural videos as well as their temporally sequential patterns have been incorporated as context into the traditional independent activity recognition models to improve their performance by a large margin [73, 101, 102, 18]. Similarly, in

multi target tracking, interrelationship between targets and other scene points having strong motion correlation with the targets is exploited to accurately estimate target states even under occlusion [39]. These contextual relationship models have also been adopted and shown significant promise in many other computer vision problems. One such application area is the object categorization, where co-occurrence of multiple objects in a scene is exploited through probabilistic graphical models to improve object recognition performances [82, 35].

In this thesis, we explore a number of video analysis problems where the contextual relationship models could be extremely useful but little studied in the literature. In the first chapter, we investigate the applicability and importance of such models in multi-camera data association problems with special focus on person re-identification. Existing data association techniques mostly focus on sequentially matching pairs of data-point sets and then repeating this process along space-time to achieve long term correspondences. However, in many such data association problems, especially in a multi-camera setting, a set of data-points may be observed at multiple spatio-temporal locations and/or by multiple agents in a network and simply combining the local pairwise association results between sets of data-points often leads to inconsistencies over the global space-time horizons. For example, most existing person re-identification methods focus on finding similarities between persons between pairs of cameras (camera pairwise re-identification) and this often leads to infeasible associations when results from different camera pairs are combined. In Chapter 2, we present a framework to combine all pairwise data-association results over a multi-camera network that not only establishes consistency in the global association results but also improves pairwise data association accuracies. We name it 'Network Consistent Data Association' (NCDA), which is formulated as an optimization problem. The proposed NCDA can be solved as a binary

integer program leading to a globally optimal solution and is capable of handling the challenging data-association scenario where the number of data-points varies across different sets of instances in the network. We have tested NCDA in two application areas, viz., person re-identification and spatio-temporal cell tracking and observed consistent and highly accurate data association results in both the cases.

Spatio-temporal cell tracking is another application area where an inter-object contextual relationship network can be very useful. Using confocal microscopes, multicellular biological tissues are often imaged at multiple time points to observe the growth of hundreds of individual cells tightly packed in the tissue. At each time point, 3D cells within the tissue are imaged at various confocal planes resulting in a spatio-temporal network of 2D cellular projections. The cell tracker aims to find correspondences between cell image slices along both 'z' (depth of the tissue) and time. Because of the multi-dimensional nature of this tracking problem, spatial and temporal correspondences obtained by choosing the most similar candidate for each cell independently do not guarantee consistent results automatically. Moreover, because of the low SNR in the confocal images and stereotype in features extracted from the 2D cells' projections, the estimated pairwise similarity scores are often misleading. To overcome this problem, in Chapter 3, we present a framework where the tight spatial topology of neighboring cells in a multicellular field is modeled using conditional random fields (CRF) and robust pairwise similarity scores are estimated through inference on these CRFs. These similarity scores are further combined together via the method proposed in Chapter 2 to generate consistent and highly accurate spatio-temporal cell lineages. We present results on (3D+t) confocal image stacks of Arabidopsis shoot meristem and show that the method is capable of handling many visual analysis challenges associated with such cell tracking problems, viz. poor feature quality of individual cells, low SNR in parts of images,

variable number of cells across slices and cell division detection.

The estimated cell lineages can be utilized to obtain important cell growth and division statistics. Towards that objective, we propose a novel spatial context based cell resolution 3D reconstruction method in Chapter 4. The tight packing of the cells in a tissue enables us to estimate the 3D structures of individual cells using the slice information of the cell as well as that of its nearest neighbors through an adaptive geometric tessellation model. The proposed method, named as the 'Adaptive Quadratic Voronoi Tessellation' (AQVT), is capable of handling both the sparsity problem and the non-uniformity in cell shapes by estimating the tessellation parameters for each cell from the sparse data-points on its boundaries. We have tested the proposed 3D reconstruction method on time-lapse CLSM image stacks of the Arabidopsis Shoot Apical Meristem (SAM) and have shown that the AQVT based reconstruction method can correctly estimate the 3D shapes of a large number of SAM cells.

Finally in Chapter 5, we explore an emerging problem in video analysis. Ability to successfully forecast activities that are yet to be observed is a very important video understanding problem, and is starting to receive attention in the computer vision literature. It can be observed that the series of consecutive activities performed by an actor often follows a fixed temporal sequence. Also, for collective activities, actions of the involved actors are strongly synchronized with each other within a spatio-temporal window. Motivated by these observations, we model the simultaneous and/or sequential nature of human activities on a graph as contextual information and combine that with the interrelationship between static scene cues and dynamic target trajectories, termed together as the 'activity and scene context'. The forecasting problem is then posed as an inference problem on a MRF model defined on the graph. We perform experiments on the publicly available challenging VIRAT ground dataset and obtain high forecasting

accuracy for most of the activities, as evidenced by the results.

1.1 Related Work

In this section, we provide detailed literature surveys on each of the application areas we explore in this thesis.

1.1.1 Network Consistent Data Association

Network level consistency has not been dealt with in data association problems and to the best of our knowledge, our work on person re-identification in [21] is the first work that introduced this idea in multi-camera data association. However, there have been a few correspondence methods proposed in recent years in different computer vision areas, *e.g.*, point correspondence in multiple frames and multi target tracking that utilized contextual relationship to obtain better data association results. In one of the early works [87], finding point correspondences in monocular image sequences is formulated as finding a graph cover and solved using a greedy method. A suboptimal greedy solution strategy was used in [91] to track multiple targets by finding a maximum cover path of a graph of detections where multiple features like color, position, direction and size determined the edge weights. In [9], the authors linked detections in a tracking scenario across frames by solving a constrained flow optimization. The resulting convex formulation of finding k-shortest node-disjoint paths guaranteed the global optima. However, this method does not actively use appearance features into the data association process which might lead to ID switches among different pairs of cameras resulting in inconsistency. An extension of the work using sparse appearance preserving tracklets was proposed in [5]. With known flow direction, a flow formulation of a data-

association problem will yield consistent results. But in data-association problems with no temporal or spatial layout information, the flow directions are not natural and thus the performance may widely vary with different choices of temporal or spatial flow.

Person re-identification is an important class of multi-camera data association problem, where associating targets across pairs of camera FoVs individually often leads to globally inconsistent results when the pairwise associations are fused over the network of cameras. However, this problem of network inconsistency is never studied in the past in re-identification literature. Here we briefly go over the major research directions in person re-identification problem to date. The proposed approaches addressing the pairwise re-identification problem across non-overlapping cameras can be roughly divided into 3 categories, (i) discriminative signature based methods [4, 6, 63, 66], (ii) metric learning based methods [7, 1, 96], and (iii) transformation learning based methods [43, 80]. Person specific discriminative signatures are computed using multiple local features (color, shape and texture) [6, 63, 66, 54] or salient features learned in an unsupervised framework [99]. Metric learning based methods try to improve the re-identification performance by learning optimal non-Euclidean metric defined on pairs of true and wrong matches [24, 61] or by maintaining redundancy in colorspace using a local Fisher discriminant analysis based metric [78]. Works exploring transformation of features between cameras tried to learn a brightness transfer function (BTF) between appearance features [80], a subspace of the computed BTFs [43], linear color variations model [36], or a Cumulative BTF [81] between cameras. Some of these works [36, 43] learned space-time probabilities of moving targets between cameras which may be unreliable if camera FoVs are significantly non-overlapping.

As the above methods do not take consistency into account, applying them to a camera network does not give globally feasible re-identification results. In a very

recent paper [21], we have introduced the network consistency in solving the person re-identification problem. In [21], the presentation of the method and the constraints used in the integer program are specific to that particular problem (re-identification). However, in this thesis, we provide a generalized formulation for solving any network level data association problem and further show how the generalized constraints can be simplified for problems in specific application areas. Besides the person re-identification problem, we also show applications of this data association method in the spatio-temporal (3D+t) cell tracking problem and how the generalized constraints can be translated into their cell tracking problem specific form.

1.1.2 Spatio-temporal Cell Tracking

There has been some work on automated tracking and segmentation of cells in time-lapse images, for both plants and animals. One of the well-known approaches for segmenting and tracking cells is based on evolution of active contours [28, 59, 58, 76, 27]. However, this method is not suitable for tracking where all the cells are in close contact with each other and share very similar physical features, nor is there any reported result on spatial correspondence. In fact, in spatio-temporal image stacks where the cells are arranged in compact multilayer structure, slice of a new cell can legitimately appear at the exact same spatial location as that of a different cell located in the layer just above it. This characteristic, along with the fact that these tightly packed cells are mostly stationary can force the active contour based tracker to generate false spatial tracks.

The Softassign method uses the information on point location to simultaneously solve both the problem of global correspondence as well as the problem of affine transformation between two time instants iteratively [19, 37]. However, these methods are more suitable for aligning global features than finding correspondences between

non-uniformly growing individual cells. Although [37] presents a sample result on a shoot meristem without validating against ground truth, it is not enough to evaluate the accuracy of this method on a typical 4D confocal data.

Besides the aforementioned approaches, tracking based on association between detections such as [44, 50] has shown good performance on time-lapse images. In [10], the authors proposed a cell tracking method on phase contrast time-lapse images that performs a global association of tracklets generated by frame-by-frame detection based tracking. Many other algorithms that have been successfully applied to single molecule localization and 2D movement tracking have been reviewed in [45]. [22, 46] describes probabilistic framework for joint detection and tracking of melanosomes. In [62], the authors have proposed a multiple hypothesis based framework that can be applied to solve particle tracking and 3D cell segmentation problems, which include splitting and merging. In [95], the authors presented a method for tracking large number of particles undergoing dense motion by integrating motion models at particle, local and global levels. However, these methods perform well when the feature quality or the underlying motion model is reliable. In fact, for many applications such as the one presented in this paper, there is no motion information available and hence it cannot be exploited for tracking.

We are looking at a more challenging problem, where the features extracted from each cell may not be reliable enough for accurate data association. As an example, in this paper the experiments are performed on confocal time lapse image stacks of plant shoot apical meristem, where hundreds of cells are tightly clustered in a multi-layered architecture and only the boundary of each cell is visible. Thus the features extracted for each cell could only be the shape and area, which could often be non-discriminating between cells even from a local neighborhood. In such cases, for tracking and data

associations in absence of very reliable features one can use the states of objects or points other than the target, that have strong spatio-temporal correlations with the states of the target. These correlations are utilized to rectify/estimate the target states in absence of reliable measurements for the target. Such secondary 'contextual information' in the visual tracking literature and have resulted in significant improvements in tracking accuracy. For example, in [39], feature points from the scene that are not on the target but have strong motion correlation with the target are used to estimate the target state under occlusion.

In [65, 64], a spatio-temporal tracking algorithm for Arabidopsis SAM was proposed, where relative positional information of neighboring cells were used to generate unique features for each cell. The best cell pair in two different image slices across space or time is found based on the computed features and the correspondence is grown sequentially outwards from these 'seed cell points' using a local search mechanism to find match between the rest of the cells. However, the location of this spatial search window depends on the position of the last tracked cell and hence this method tends to accumulate error that can throw the tracker off for cells spatially distant from the 'seed'. Therefore, in [15], we posed the spatio-temporal cell tracking problem as an inference problem on a conditional random field where the relative positional information of a cell with respect to its neighbors are utilized to generate robust associations between cells in two spatially/temporally consecutive image slices.

However, most of these methods have focused on slice to slice/pairwise cell tracking. The method in [65] utilizes indirect *paths* between any two slices to improve the pairwise tracking accuracy. However, this method does not ensure spatio-temporally consistent association results. Also, the tracking in the (3D+t) stack is done in a sequential manner and a globally optimal solution is not achieved. We utilize the NCDA

method presented in Chapter 2 that yields globally optimal and consistent correspondences between 2D cell slices when built on top of any method that can generate similarity scores between cells, such as [15].

1.1.3 Cell Resolution 3D Reconstruction

There are several methods of shape and size estimations for individual cells such as impedance method [72] and light microscopy methods [31]. Methods such as [47] are used to study changes in cell sizes in cell monolayers. In live plant tissues, a number of work focussed on the surface reconstruction [55, 93]. But we are looking at a much more challenging problem where the subject of study is a dense cluster of cells arranged in multiple spatial layers. In such cases, one of the popular practices is to use Confocal Laser Scanning Microscopy (CLSM) to image cell or nucleus slices at a very high spatial resolution and then reconstruct the 3D volume of the cells from those serial optical slices which has been shown to be reasonably accurate [29, 33, 100]. However, performance of the current imaging based 3D reconstruction techniques depends heavily on the availability of a large number of very thin optical slices of a cell and the performance rapidly deteriorates in the cases where the number of cell slices becomes limited. This problem is very common especially in CLSM based *live cell imaging* when the time gap between successive observations is small. In order to keep the cells alive and growing for a longer period of time and obtain frequent observations, a cell cannot be imaged in more than 2-4 slices, i.e., high depth-resolution and time-resolution cannot be achieved simultaneously.

A very recent method [33] accurately reconstructs the Shoot Apical Meristem of Arabidopsis. This method uses a dataset containing fine slice images acquired from 3 different angles, each at a Z-resolution of 1 μm. They have reported 24 hours as

the time resolution in imaging. But, for analyzing the growth dynamics of cell clusters where the time gap between successive cell divisions is in the range of 30 to 36 hours, we need a much higher time resolution in imaging in order to capture the exact growth dynamics. To obtain longer cell lineages at high time resolution we may have to sacrifice the spatial or depth resolution and hence the number of image slices in which a cell is present can be really small. With such a limited amount of image data, the existing 3-D reconstruction/ segmentation techniques cannot yield a good estimate of cell shape. Other recent approaches that attempted the problem of cell resolution 3D reconstruction from sparse collection of slices to estimate approximate cell volumes are presented in [13, 16].

Motivated by the rigid, tightly packed physical structure of SAM, we propose the cell resolution 3D reconstruction in a geometric tessellation framework. A tessellation is a partition of a space into closed geometric regions with no overlap or gap among these regions. In case of the SAM tissue, each cell is represented by such a closed region and any point in the 3D SAM structure must be the part of one and only one cell. In fact, there are some recent works in the literature as [67] which predicted that the 3D structures of Arabidopsis SAM cells could be represented by convex polyhedrons forming a 3D 'Voronoi' tessellation pattern.

A Voronoi tessellation is one of the simplest form of partitioning of the metric space, where the boundaries between two adjacent partitions are equidistant from a point inside each of these regions, also known as the 'sites'. In [67, 38], these sites are the approximate locations of the center of the cell nuclei about which the tissue is tessellated into individual cells. However, this work used a dataset where both the plasma membrane as well as the nucleus of each cell is marked with fluorescent protein, whereas, in the dataset under our study, only the plasma membrane is visible under

the confocal microscope. In this thesis, we present a cell resolution 3D reconstruction method based on a quadratic distance based anisotropic Voronoi tessellation, where the distance metric for each cell is estimated from the segmented and tracked sparse data-points for the cell [14].

1.1.4 Activity Forecasting

In computer vision research, majority of the work related to human activity in video has focused on the task of recognition of simple to more complex activities [79]. Many existing work exploring context focus on spatio-temporal relationship of features [84, 40], interactions of objects and actions/activities [97, 18, 56], AND-OR graph based scene representation [41, 88]. Methods such as [102, 101, 73, 8, 92] studied spatio-temporal relationship between activities in a scene. The method in [73] takes a probabilistic approach and model the spatial and temporal relationship between activities into the potential functions of a Markov Random Field model. Besides these, spatio-temporal context in structured videos with manually defined rules has been modeled using 'Markov Logic Networks' in [70]. We are strongly motivated by these methods to model the spatio-temporal context between activities and scene context between actors' motion, objects in the scene and the activities on an 'activity graph'. However, none of these methods can be applied to solve the activity forecasting problem as they assume that observations from all the activity regions are available throughout the recognition process.

There have been some recent work on the emerging topic of early recognition of ongoing activities. The method in [83] approached this problem by representing an activity as an integral histogram of spatio-temporal features and subsequently used a novel dynamic bag-of-words approach to model how these feature distributions change

over time. Authors in [98] developed a 'spatial-temporal implicit shape model' which characterizes the space time structure of the sparse activity features extracted from a video and the early recognition is done using a random forest structure. The authors in [42] proposed a max-margin framework based on structured SVM to recognize partially observed events. However, these methods rely on the availability of a partial set of information for the ongoing activity where a typical activity forecasting problem should be able to forecast probable future activities well before the start of the activity segments.

Very recently, the activity forecasting problem was introduced in [51]. The authors combined semantic scene labeling with inverse optimal control to forecast probable actor trajectories, which in turn helps predict destinations and future actions. However, there are a number of differences between our method and [51]. [51] investigates the effect of the static scene environment on future activities, whereas we use both static cues from the scene and dynamic cues from target trajectories and model their interrelationships for forecasting future activities. Unlike a pure trajectory based approach in [51], we combine the target trajectory information with the motion based activity recognition methods in a dynamical model. Finally, we show results on the recent release of VIRAT dataset containing 11 diverse activities where [51] tested their forecasting method on a dataset of three activities.

1.2 Organization of the Thesis

The rest of the thesis is organized as follows. We present the generalized theory of network consistent data association (NCDA) in Chapter 2 along with its application in the person re-identification problem. Spatio-temporal cell tracking problem, the second application area of the NCDA, is presented in Chapter 3. The conditional random

field based pairwise cell tracking module is also detailed in this chapter. In Chapter 4, we propose the cell resolution 3D reconstruction pipeline for dense multilayer tissues based on an adaptive quadratic Voronoi tessellation. Finally, the problem of activity forecasting and its solution strategy based on a spatio-temporal context model are given in Chapter 5. We conclude the thesis in Chapter 6 with an outline of the future direction of research.

Chapter 2

Network Consistent Data Association

2.1 Introduction

In many computer vision problems such as tracking, re-identification *etc.*, associating detected targets across space and/or time is of utmost importance. Most data association approaches are sequential in nature, *i.e.*, they try to find correspondences between pairs of instances of a set of data-points and repeat this process along space/time to obtain long term correspondences. However, this local approach for finding correspondences may lead to inconsistencies over the global space-time horizons. The goal of this chapter is to show how globally consistent correspondence results can be obtained by enforcing suitable network-level constraints over the entire set of observation data-points. We explain the problem more precisely through two examples below.

Consider the well studied person re-identification problem where the objective is to associate targets across cameras with non overlapping field-of-views (FoVs). Most

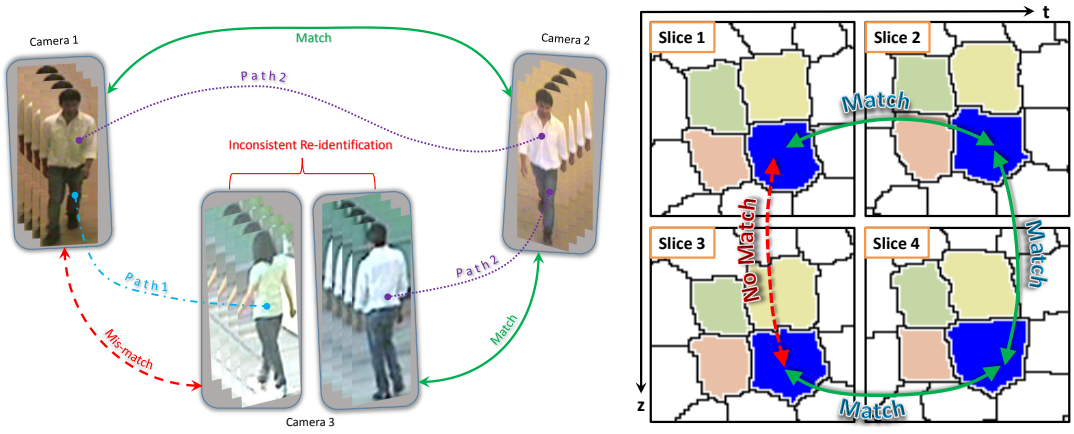


Figure 2.1: Example of network inconsistency in data association. (A) A person re-identification case. Among the 3 possible re-identification results, 2 are correct. The match of the target from camera 1 to camera 3 can be found in two ways. The first one is the direct pairwise re-identification result between cameras 1 and 3 (shown as 'Path 1'), and the second one is the indirect re-identification result in camera 3 given via the matched person in camera 2 (shown as 'Path 2'). The two outcomes do not match and thus the overall associations of the target across 3 cameras are not consistent. (B) A similar case of network inconsistency in spatio-temporal cell tracking problem. In this schematic, association results between 2D projections of the same 3D cell on four spatio-temporal image planes are analyzed. The pairwise associations need to be consistent across the loop over the four image slices. This consistency can be used to obtain correspondences when there are no direct pairwise matches or to correct wrong ones. For example, the correspondence between the same cell in image slice 1 and slice 3 (broken arrow) is established via an indirect path (solid arrows) through slices 2 and 4 to restore network consistency.

widely used approaches focus on pairwise re-identification, *i.e.*, association between two camera FoVs. Even if the re-identification accuracy for each camera pair is high, it might contain many global association inconsistencies over the entire network if three or more cameras are considered. Matches between targets given independently by every pair of cameras might not conform to one another and, in turn, may lead to inconsistent mappings. Thus, in person re-identification across a camera network, multiple paths of correspondences may exist between targets from any two cameras, but ultimately all these paths must point to the same correspondence maps for each target in each camera. An example scenario is shown in Fig. 2.1(A). Even though camera pairs 1-2

and 2-3 have correct re-identification of the target, the false match between the targets in camera pair 1-3 makes the overall re-identification across the triplet inconsistent. It can be noted that the error in re-identification manifests itself through inconsistency across the network, and hence by enforcing consistency the pairwise accuracies can be improved as well.

Spatio-temporal cell tracking is another application area where consistent data association is important. Using confocal microscopes, multicellular biological tissues are often imaged at multiple time points to observe the growth of hundreds of individual cells in the tissue. At each time point, cells within the tissue are imaged at various confocal planes, thus resulting in a four dimensional (3D + t) stack of images. Each cell, therefore, may have projections on different spatio-temporal planes. The spatio-temporal cell tracking aims to find correspondences between cell image slices along both 'z' (depth of the tissue) and time. Because of the multi-dimensional nature of this tracking problem, spatial and temporal correspondences obtained by choosing the most similar candidate for each cell independently do not guarantee consistent results automatically. Note that, as in the case of re-identification, a 2D cell segment in any spatio-temporal image slice must not have more than one match in any other spatio-temporal image and if at least one spatio-temporal path exists in the network that associates two cell slices, they must be projections of the same cell onto two image planes. Example of network-level inconsistent data association results in the spatio-temporal cell tracking problem is presented in Fig. 2.1(B).

The network inconsistency problem in data-association is not only observed in the person re-identification or the cell tracking tasks, but is also visible in other correspondence problems of similar nature, *e.g.* optical flow computation, feature tracking etc. In fact, a data-association approach that ensures network level consistency will be

valuable for any problem where sets of data-points observed at multiple spatio-temporal locations need to be associated with each other.

2.1.1 Contributions of the Present Work:

Motivated by such scenarios, we propose a novel consistent data association scheme over a network. When the same set of data-points is observed by different sensors (such as different camera FoVs) or observed by the same sensor repeatedly at different spatial and/or temporal locations (such as at different spatio-temporal imaging planes in confocal microscopy stacks), the possible associations between such projections of the same set of observations construct a network with the observed data-points as nodes. To achieve a consistent and optimal data association, we pose the problem as an optimization problem that minimizes the global cost of associating pairs of targets on the entire network constrained by a set of consistency criteria (as mentioned before). Since consistency across the network is the motivation as well as the building block of the proposed method, we term this as the *Network Consistent Data Association (NCDA)* strategy.

We start by computing the pairwise similarity scores between sets of targets which are the input to the proposed method. Unlike assigning a match for which the similarity score is maximum among a set of probable candidates, our formulation picks the assignments for which the total similarity of all matches is the maximum while maintaining the constraint that there is no inconsistency in the assignment among any two sets of targets given the assignments between all other sets of targets across the network. The resulting optimization problem is translated into a binary integer program that can be solved using standard branch and cut, branch and bound or dynamic search algorithms based methods available in [85].

The proposed NCDA method is further generalized to a more challenging scenario in data association where the number of targets may vary across different sets of instances in the network. For example, in person re-identification problems all persons may not appear in all the cameras or in case of cell tracking, 2D projections of new cells appear as we image deeper into a tissue and hence the number of 2D cell slices widely vary across imaging planes. The objective function and the constraints are, therefore, modified to incorporate probable one-to-none mappings without jeopardizing the network consistency. We also provide a proof showing that the one-to-one NCDA can be directly derived as a special case of this more generalized formulation.

We show the general applicability of the proposed method by testing it in two previously mentioned computer vision application domains in Chapters 2 and 3. In this chapter (Chapter 2), we provide a detailed mathematical treatment on the proposed NCDA method and experimental results on the problem of person re-identification [21]. In the next chapter (Chapter 3), we introduce the challenging problem of spatio-temporal cell tracking in a live tissue and show how the NCDA method can be applied to this problem to improve the tracking performance. Throughout these two chapters, we describe how each of the various computer vision challenges can be mapped to the exact same generalized NCDA problem, which can then be solved to generate unambiguous and more accurate data association results.

2.2 The Network Consistent Data Association Problem

In this section we describe the proposed approach in details. The Network Consistent Data Association (NCDA) method starts with the pairwise similarity scores between the targets. First we describe the notations and define the terminologies as-

sociated to this problem that would be used throughout the rest of the paper before delving deeper into the problem formulation.

2.2.1 Notations and Terminologies

1. Node: A 'node' is a data-point/target that needs to be associated with other data-points via NCDA. For person re-identification problems, a node represents a target in the FoV of a camera, whereas, in cell tracking problem, a node is a 2D segmented cell (at any given spatio-temporal location).

2. Group: A 'group' is a collection of nodes. A node can never be associated with any other node from the same group it belongs to. For example, in a typical person re-identification problem, the set of all targets appearing in the FoV of the same camera is a group and for spatio-temporal tracking, the collection of 2D cell segments in one image slice can be assumed a group. Thus, a node is a member of a group. Let the i^{th} node in the group g be denoted as $\mathcal{P}_i^g$.

3. Similarity score matrix: This is a matrix data structure containing feature similarity scores between nodes belonging to two different groups. Therefore, for each pair of groups in a network there is one such matrix. Let $\mathbf{C}^{(p,q)}$ denote the similarity score matrix between groups p and q. Then $(i,j)^{th}$ element in $\mathbf{C}^{(p,q)}$ denotes the similarity score between the nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$.

4. Assignment matrix: We need to know whether the nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ are associated or not, $\forall i,j = \{1,\cdots n\}$ and $\forall p,q = \{1,\cdots m\}$. The associations between targets across groups can be represented using 'Assignment matrices', one for each pair of groups. Each element $x_{i,j}^{p,q}$ of the assignment matrix $\mathbf{X}^{(p,q)}$ between the group pair

(p, q) is defined as follows,

$$x_{i,j}^{p,q} = \begin{cases} 1 & \text{if } \mathcal{P}_i^p \text{ and } \mathcal{P}_j^q \text{ are the same targets} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

If the number of nodes is the same in all groups, then $\mathbf{X}^{(p,q)}$ is a permutation matrix, *i.e.*, only one element per row and per column is 1, all the others are 0. Mathematically, $\forall\, x_{i,j}^{p,q} \in \{0, 1\}$

$$\begin{aligned} \sum_{j=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall i = 1 \text{ to } n \\ \sum_{i=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall j = 1 \text{ to } n \end{aligned} \tag{2.2}$$

5. Edge: An 'edge' between two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ from two different groups of nodes is constructed between the i^{th} node in group p and the j^{th} node in group q. It should be noted that there will be no edge between the nodes of the same group. There are two attributes connected to each edge. They are the similarity score $c_{i,j}^{p,q}$ and the association value $x_{i,j}^{p,q}$.

6. Path: A 'path' between two nodes $(\mathcal{P}_i^p, \mathcal{P}_j^q)$ is a set of edges that connect the nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ without traveling through a node twice. Moreover, each node on a path belongs to a different group. A path between $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ can be represented as the set of edges $e(\mathcal{P}_i^p, \mathcal{P}_j^q) = \{(\mathcal{P}_i^p, \mathcal{P}_a^r), (\mathcal{P}_a^r, \mathcal{P}_b^s), \cdots\ (\mathcal{P}_c^t, \mathcal{P}_j^q)\}$, where $\{\mathcal{P}_a^r, \mathcal{P}_b^s, \cdots \mathcal{P}_c^t\}$ are the set of intermediate nodes on the path between $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$. The set of association values on all the edges between the nodes is denoted as $\mathcal{L}$, *i.e.* $x_{i,j}^{p,q} \in \mathcal{L}$, $\forall i, j = [1, \cdots, n]$, $\forall p, q = [1, \cdots, m]$ and $p < q$. Finally, the set of all paths between any two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ is represented as $\mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$ and the z^{th} path is $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$.

2.2.2 The NCDA Objective Function

For the pair of groups (p, q), the sum of the similarity scores of association is given by $\sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q}$. Summing over all possible pairs of groups, the global similarity score can be written as

$$\mathbf{C} = \sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q} \tag{2.3}$$

2.2.3 Identification of Constraints

Let us first discuss the problem of one-to-one data association where the number of data-points per group is constant and each data-point from one group would have exactly one match in another group. This type of data association problem is often relevant to the person re-identification datasets, where the same set of persons appears across the FoVs of all the cameras in the network. Later, we shall present a more generalized version of NCDA where number of data-points belonging to different groups may vary and therefore a data-point may or may not have a match in another group.

The set of constraints are as follows.

1. Pairwise association constraint: For the one-to-one association scenario, a data-point from the group p can have only one match from another group q. This is mathematically expressed by the set of Eqns. 2.2. This is true for all possible pairs of data groups and can be expressed as,

$$\begin{aligned} \sum_{j=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall i = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q \\ \sum_{i=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall j = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q \end{aligned} \tag{2.4}$$

2. Loop constraint: This constraint comes from the consistency requirement. If two nodes are indirectly associated via nodes in other groups, then these two nodes must also be directly associated. Therefore, given two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$, it can be noted

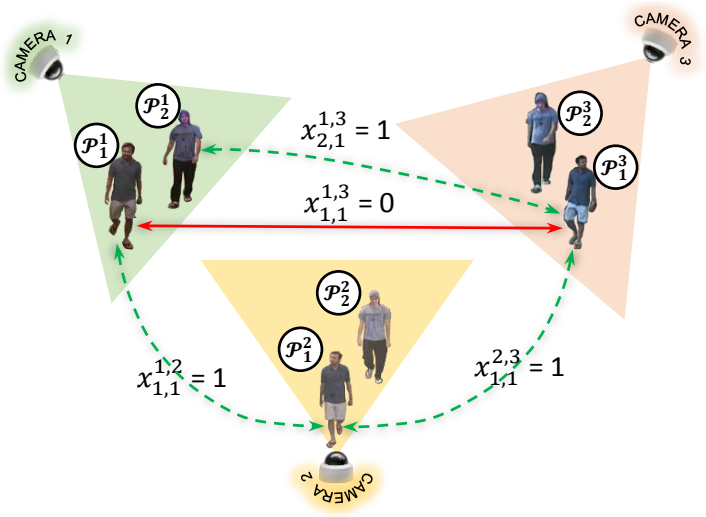


Figure 2.2: An illustrative example showing the importance of the loop constraint in a data-association problem. It presents a simple person re-identification scenario in a camera network involving 2 persons (data points) in 3 cameras (groups).

that for consistency, a logical 'AND' relationship between the association value $x_{i,j}^{p,q}$ and the set of association values $\{x_{i,a}^{p,r}, x_{a,b}^{r,s}, \cdots x_{c,j}^{t,q}\}$ of any possible path between the nodes has to be maintained. The association value between the two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ has to be 1 if the association values corresponding to all the edges of any possible path between these two nodes are 1. Keeping the binary nature of the association variables and the pairwise association constraint in mind the relationship can be compactly expressed as,

$$x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1 \tag{2.5}$$

$\forall$ paths $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$, where $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$ denotes the cardinality of the path $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$, *i.e.* the number of edges in the path. The relationship holds true for all i and all j. For the case of a triplet of cameras the constraint in Eqn. 2.5 simplifies to,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 2 + 1 = x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \tag{2.6}$$

An example from person re-identification involving 3 cameras and 2 persons is illustrated with the help of Fig. 2.2. Say, the raw similarity score between pairs of

targets across cameras suggests associations between $(\mathcal{P}_1^1, \mathcal{P}_1^2), (\mathcal{P}_1^2, \mathcal{P}_1^3)$ and $(\mathcal{P}_2^1, \mathcal{P}_1^3)$ independently. However, when these associations are combined together over the entire network, it leads to an infeasible scenario - $\mathcal{P}_1^1$ and $\mathcal{P}_1^2$ are the same person. This infeasibility is also correctly captured through the constraint in Eqn. 2.6, *i.e.*, $x_{1,1}^{1,3} = 0$ but $x_{1,1}^{1,2} + x_{1,1}^{2,3} - 1 = 1$, thus violating the constraint.

For a generic scenario involving a large number of groups of nodes where similarity scores between every pair of groups may not be available, the loop constraint equations (*i.e.* Eqn. 2.5) have to hold for every possible triplet, quartet, quintet (and so on) of groups. On the other hand, if the similarity scores between all nodes for every possible pair of groups are available, the loop constraints on quartets and higher order loops are not necessary. If loop constraint is satisfied for every triplet of groups then it automatically ensures consistency for every possible combination of groups taking 3 or more of them. So, in such a case, the loop constraint for the network can be written as,

$$\begin{gathered} x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\ \forall\, i,j,k = [1, \cdots n],\ \forall\, p,q,r = [1, \cdots m],\ \text{and}\ p < r < q \end{gathered} \tag{2.7}$$

Unlike the person re-identification case, formation of triplets is not possible for the cell tracking problem because of the structure of the data. However, it can be shown that the entire spatio-temporal cell tracking network can be exhaustively partitioned into quartets of cell slices and the loop constraints are expressed accordingly (discussed in the next chapter).

2.2.4 Overall Optimization Problem For One-to-One Associations

By combining the objective function in Eqn. 2.3 with the constraints in Eqn. 2.4 and Eqn. 2.5, we pose the overall optimization problem for the case of one-to-one map-

ping between groups as,

$$\underset{\substack{x_{i,j}^{p,q} \\ i,j=[1,\cdots,n] \\ p,q=[1,\cdots,m]}}{\text{argmax}} \left(\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q} \right)$$

$$\begin{aligned}
\text{subject to } & \sum_{j=1}^{n} x_{i,j}^{p,q} = 1 \ \forall i = [1,\cdots,n] \ \forall p,q = [1,\cdots,m],\ p<q \\
& \sum_{i=1}^{n} x_{i,j}^{p,q} = 1 \ \forall j = [1,\cdots,n] \ \forall p,q = [1,\cdots,m],\ p<q \\
& x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1 \\
& \forall\, i,j = [1,\cdots n],\ \forall\, p,q = [1,\cdots m],\ \text{and } p<q \\
& \forall \text{ paths } e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q) \\
& x_{i,j}^{p,q} \in \{0,1\} \ \forall i,j = [1,\cdots,n],\ \forall p,q = 1 \text{ to } m,\ p<q
\end{aligned} \tag{2.8}$$

The above optimization problem for optimal and consistent data association is a binary integer program.

2.3 NCDA for Variable Number of Data-points In Each Group

As explained in the previous sub-section, network consistent data association can be achieved by solving the binary IP formulated in Eqn. 2.8. However, the assumption of one-to-one association between targets across groups may not be valid in many practical scenarios, especially when there are unequal numbers of data-points in different groups. For re-identification in a camera network, there may be situations when every person does not go through the FoV of every camera. For the spatio-temporal cell tracking problems, there could be variable number of segmented cell slices on the

images at different spatio-temporal locations. In such cases, a data-point may not have association with any other data-point from another group and hence the values of assignment variables in every row or column of the assignment matrix can all be 0. However, a one-to-many association is still infeasible as before. For re-identification example, a person from any camera p can have *at most* one match from another camera q. As a result, the pairwise association constraints now change from equalities to inequalities as follows,

$$\begin{aligned} \sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1, \cdots, n_p] \ \forall p, q = [1, \cdots, m], \ p < q \\ \sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1, \cdots, n_q] \ \forall p, q = 1 \text{ to } m, \ p < q \end{aligned} \tag{2.9}$$

where, n_p snd n_p are the number of nodes (datapoints) in groups p and q respectively.

However, with this generalization, it is easy to see that the objective function (ref. Eqn. 2.8) is no longer valid. Even though the provision of 'no match' is now available, the optimal solution will try to get as many associations as possible across the network. This is due to the fact that the current objective function assigns reward to both true positive (correctly associating a datapoint across groups) and false positive associations. Thus the optimal solution may contain many false positive associations. This situation can be avoided by incorporating a modification in the objective function as follows,

$$\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \tag{2.10}$$

where k is any value in the range of the similarity scores. This modification leverages upon the idea that, typically, similarity scores for most of the true positive matches in the data would be much larger than majority of the false positive matches. In the new cost function, instead of rewarding all positive associations we give reward to most of the true positives, but impose penalties on the false positives. As the rewards for all true

positive (TP) matches are discounted by the same amount k and as there is penalty for false positive (FP) associations, the new cost function gives us optimal results for both 'match' and 'no-match' cases. The choice of the parameter k depends on the similarity scores generated by the chosen method, and thus can vary from one pairwise similarity score generating method to another. Ideally, the distributions of similarity scores of the TPs and FPs are non-overlapping and k can be any real number from the region separating these two distributions. However, for practical scenarios where TP and FP scores overlap, an optimal k can be learned from training data. A simple method to choose k could be running NCDA for different values of k over the training data and choosing the one giving the maximum accuracy on the cross validation data. So, for this more generalized case, the NCDA problem can be formulated as follows,

$$
\begin{aligned}
&\underset{\substack{x_{i,j}^{p,q} \\ i=[1,\cdots,n_p] \\ j=[1,\cdots,n_q] \\ p,q=[1,\cdots,m]}}{\operatorname{argmax}} \left(\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \right) \\
&\text{subject to } \sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1,\cdots,n_p] \ \forall p,q = [1,\cdots,m],\ p<q \\
&\sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1,\cdots,n_q] \ \forall p,q = [1,\cdots,m],\ p<q \\
&x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1 \\
&\forall\, i = [1,\cdots,n_p],\ j = [1,\cdots,n_q],\ \forall\, p,q = [1,\cdots m],\ \text{and } p<q \\
&\forall \text{ paths } e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q) \\
&x_{i,j}^{p,q} \in \{0,1\} \ \forall i = [1,\cdots,n_p],\ j = [1,\cdots,n_q], \\
&\forall p,q = [1,\cdots,m],\ p<q
\end{aligned}
\tag{2.11}
$$

2.4 Equivalence Between One-to-One NCDA (Eqn. 2.8) and The Generalized NCDA (Eqn. 2.11)

If the similarity score matrix and the assignment matrix are vectorized, one can rewrite the problems in Eqn. 2.8 and Eqn. 2.11 in standard binary integer program form. The one-to-one NCDA problem in Eqn. 2.8 can be rewritten as

$$\begin{gathered} \underset{\underline{\mathbf{X}}}{\operatorname{argmax}} \ \underline{\mathbf{C}}^T \underline{\mathbf{X}} \\ \text{subject to } \mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}},\ \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}} \\ \underline{\mathbf{X}} \text{ is composed of binary variables.} \end{gathered} \tag{2.12}$$

where $\mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}$ is the pairwise association constraint (same as Eqn. 2.4) and $\mathbf{B}\underline{\mathbf{X}} \leq \underline{\mathbf{1}}$ is the rewritten loop constraint (same as Eqn. 2.5). The value of d is 1 in case of person re-identification problems where the loop constraints are expressed on triplets of groups. However, for the cell tracking problem, $d = 2$ as here the IP is written using quartet based loop constraints (see Eqn. 3.27 for more details).

The generalized form of NCDA (Eqn. 2.11) can, similarly be rewritten as,

$$\begin{gathered} \underset{\underline{\mathbf{X}}}{\operatorname{argmax}} \ (\underline{\mathbf{C}}^T - k\underline{\mathbf{1}}^T)\underline{\mathbf{X}} \\ \text{subject to } \mathbf{A}\underline{\mathbf{X}} \ \leq \underline{\mathbf{1}},\ \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}} \\ \underline{\mathbf{X}} \text{ is composed of binary variables.} \end{gathered} \tag{2.13}$$

Now let us prove that the problem expressed by Eqn. 2.13 is equivalent to the problem expressed by Eqn. 2.12 under the condition that the number of data-points/targets is constant and there exists a one-to-one mapping between targets across groups. Let $\underline{\mathbf{X}}^*$ be the optimal solution to the problem expressed by Eqn. 2.13. To prove the equivalence, we have to show that $\underline{\mathbf{X}}^*$ also maximizes the problem expressed by Eqn. 2.12.

Since $\underline{\mathbf{X}}^*$ maximizes the objective function under the constraints as expressed by Eqn. 2.13, we can write,

$$(\underline{\mathbf{C}}^T - k\underline{\mathbf{1}}^T)\underline{\mathbf{X}}^* \geq (\underline{\mathbf{C}}^T - k\underline{\mathbf{1}}^T)\underline{\mathbf{X}} \text{ for } \{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} \leq \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\} \tag{2.14}$$

where both $\underline{\mathbf{X}}^*$ and $\underline{\mathbf{X}}$ are composed of binary variables.

Since $\{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\} \subset \{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} \leq \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\}$, the relation (2.14) holds true for the feasible set of Eqn. 2.12, *i.e.*,

$$\begin{gathered}(\underline{\mathbf{C}}^T - k\underline{\mathbf{1}}^T)\underline{\mathbf{X}}^* \geq (\underline{\mathbf{C}}^T - k\underline{\mathbf{1}}^T)\underline{\mathbf{X}} \text{ for } \{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\} \\ \implies \underline{\mathbf{C}}^T\underline{\mathbf{X}}^* - k\underline{\mathbf{1}}^T\underline{\mathbf{X}}^* \geq \underline{\mathbf{C}}^T\underline{\mathbf{X}} - k\underline{\mathbf{1}}^T\underline{\mathbf{X}} \text{ for } \{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\} \\ \text{with both } \underline{\mathbf{X}}^* \text{ and } \underline{\mathbf{X}} \text{ composed of binary variables.}\end{gathered} \tag{2.15}$$

Now for all $\underline{\mathbf{X}}$ and $\underline{\mathbf{X}}^*$ that satisfy $\mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}$ (*i.e.*, for the case when the same set of n targets appear in all m groups),

$$\underline{\mathbf{1}}^T\underline{\mathbf{X}}^* = \underline{\mathbf{1}}^T\underline{\mathbf{X}} = \text{Num. of group pairs} \times \text{Num. of targets} \tag{2.16}$$

This is because, each row and column of the assignment matrix for pair of groups contains exactly one 1, resulting in the sum of all elements of the assignment matrices being n.

Using the above relation in Eqn. 2.15 we get,

$$\begin{gathered}\underline{\mathbf{C}}^T\underline{\mathbf{X}}^* \geq \underline{\mathbf{C}}^T\underline{\mathbf{X}} \text{ for } \{\underline{\mathbf{X}} : \mathbf{A}\underline{\mathbf{X}} = \underline{\mathbf{1}}, \mathbf{B}\underline{\mathbf{X}} \leq d\underline{\mathbf{1}}\} \\ \text{with both } \underline{\mathbf{X}}^* \text{ and } \underline{\mathbf{X}} \text{ composed of binary variables.}\end{gathered} \tag{2.17}$$

Therefore, $\underline{\mathbf{X}}^*$ also maximizes the problem 2.12, thus proving the equivalence.

2.5 Experiments and Results

In this section, we evaluate the NCDA method on the person re-identification problem.

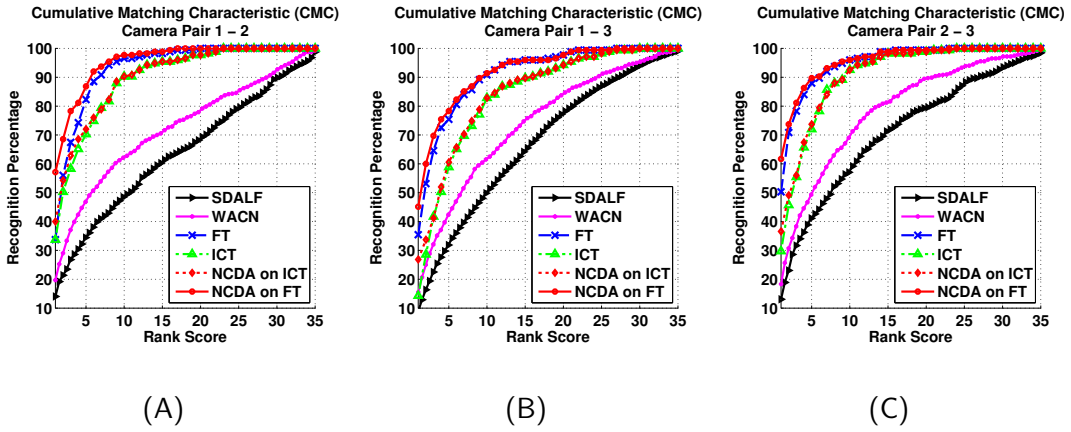
Datasets and Performance Measures: We performed experiments on two benchmark datasets - WARD [66] and one new dataset RAiD introduced in [21]. Though state-of-the-art methods for person re-identification *e.g.*, [17,6,60] evaluate their performances using other datasets too (*e.g.*, ETHZ, CAVIAR4REID, CUHK) these do not fit our purposes since these are either two camera datasets or several sequences of different two camera datasets. WARD is a 3 camera dataset with 70 people while RAiD has been collected across 4 cameras with 43 persons walking through them. Results are shown in terms of recognition rate as Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values, as is the common practice in the literature. The CMC curve is a plot of the recognition percentage versus the ranking score and represents the expectation of finding the correct match inside top t matches. nAUC gives an overall score of how well a re-identification method performs irrespective of the dataset size. In the case where every person is not present in all cameras, we show the accuracy as total number of true positives (true matches) and true negatives (true non matches) divided by the total number of unique people present. All the results used for comparison were either taken from the corresponding works or by running codes which are publicly available or obtained from the authors on datasets for which reported results could not be obtained.

Pairwise Similarity Score Generation: The camera pairwise similarity score generation starts with extracting appearance features in the form of HSV color histogram from the images of the targets. Before computing these features, the foreground is segmented out to extract the silhouette. Three salient regions (head, torso and legs) are extracted from the silhouette as proposed in [6]. The head region S^H is discarded, since it often consists of a few and less informative pixels. We additionally divide both body and torso into two horizontal sub-regions based on the intuition that people can wear

shorts or long pants, and short or long sleeves tops.

Given the extracted features, we generate the similarity scores by learning the way features get transformed between cameras in a similar approach as [80, 43]. Instead of using feature correlation matrix or the feature histogram values directly, we capture the feature transformation by warping the feature space in a nonlinear fashion inspired by the principle of Dynamic Time Warping (DTW). The feature bin number axis is warped to reduce the mismatch between feature values of two feature histograms from two cameras. Considering two non-overlapping cameras, a pair of images of the same target is a feasible pair, while a pair of images between two different targets is an infeasible pair. Given the feasible and infeasible transformation functions from the training examples, a Random Forest (RF) [12] classifier is trained on these two sets. The camera pairwise similarity scores between targets are obtained from the probability given by the trained classifier of a test transformation function as belonging to either the set of feasible or infeasible transformation functions. In addition to the feature transformation based method, similarity scores are also generated using the publicly available code of a recent work - ICT [4] where pairwise re-identification was posed as a classification problem in the feature space formed of concatenated features of persons viewed in two different cameras.

Experimental Setup: In our implementation we used the following settings: 1. To be consistent with the evaluations carried out by state-of-the-art methods, images were normalized to 128×64. The H, S and V color histograms extracted from the body parts were quantized using 10 bins each, 2. image pairs of the same or different person(s) in different cameras were randomly picked to compute the feasible and infeasible transformation functions respectively, 3. all the experiments are conducted using a multi-shot strategy where 10 images per person is taken for both training and testing,



(A) (B) (C)

Figure 2.3: CMC curves for the WARD dataset. Results and comparisons in (A), (B) and (C) are shown for the camera pairs 1-2, 1-3, and 2-3 respectively.

4. the RF parameters such as the number of trees, the number of features to consider when looking for the best split, etc. were selected using 4-fold cross validation, and 5. for each test we ran 5 independent trials and report the average results.

2.5.1 WARD Dataset

The WARD dataset [66] has 4786 images of 70 different people acquired in a real surveillance scenario in three non-overlapping cameras. This dataset has a huge illumination variation apart from resolution and pose changes. The cameras here are denoted as camera $1, 2$ and 3. Fig. 2.3(A), (B) and (C) compare the performance for camera pairs $1-2, 1-3$, and $2-3$ respectively. The 70 people in this dataset are equally divided into training and test sets of 35 persons each. The proposed approach is compared with the methods SDALF [6], ICT [4] and WACN [66]. The legends 'NCDA on FT' and 'NCDA on ICT' imply that the NCDA algorithm is applied on similarity scores generated by learning the feature transformation and by ICT respectively. For all 3 camera pairs the proposed method outperforms the rest with rank 1 recognition percentage as high as 61.71% for the camera pair 2-3. The next runner up is the method

Table 2.1: Comparison of NCDA with state-of-the-art methods on the WARD dataset in terms of the nAUC values.

Camera pair	SDALF	WACN	ICT	FT	NCDA on ICT	NCDA on FT
1-2	0.6487	0.7328	0.8780	0.9136	0.8835	**0.9317**
1-3	0.6825	0.7496	0.8240	0.8905	0.8299	**0.8981**
2-3	0.7206	0.7966	0.8881	0.9278	0.8910	**0.9330**

applying only feature transformation which has the recognition percentage of 50.29% for rank 1. Comparison of different re-identification methods with the NCDA based ones also presented in terms of nAUC values in Table 2.1. As can be observed, NCDA on FT performs the best on the WARD dataset for all the camera pairs. NCDA on ICT also performs better than ICT only.

To show how the proposed method yields consistent re-identification results where pairwise method fails, two example cases are provided in Fig. 2.4. At first, re-identification is performed on 3 camera pairs independently on the WARD data by FT method. In the first example, though the camera pairs $1-2$ and $2-3$ gave correct association (red dashed lines) for both the targets, the incorrect associations between camera pair $1-3$ (red dashed line) make the re-identification across the 3 cameras inconsistent. Similarly, in the second example, incorrect associations between targets across camera pair $1-2$ make the overall re-identification results inconsistent. However, in both the case, the NCDA exploits the consistency requirement and makes the resultant re-identification across 3 cameras correct, which are shown using green arrows.

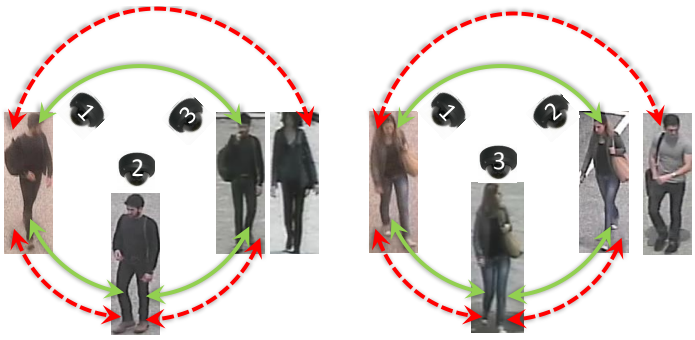
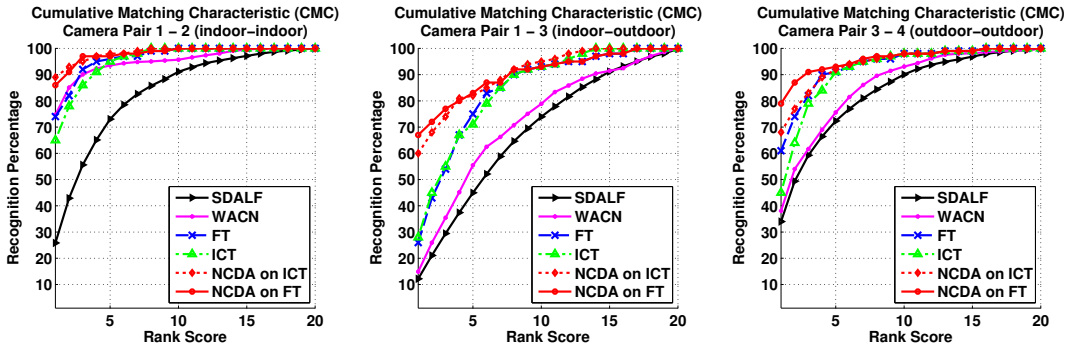


Figure 2.4: Two examples of correction of inconsistent re-identification from WARD dataset. The red dashed lines denote re-identifications performed on 3 camera pairs independently by FT method. The green solid lines show the re-identification results on application of NCDA on FT. The NCDA algorithm exploits the consistency requirement and makes the resultant re-identification across 3 cameras correct.

2.5.2 RAiD Dataset

We collected this dataset [21] to test the proposed method on a larger network. The dataset was collected using 2 indoor (cameras 1 and 2) and 2 outdoor (cameras 3 and 4) cameras. It has large illumination variation that is not present in most of the publicly available benchmark datasets. 41 subjects were asked to walk through the FoVs of these 4 cameras and the dataset contains a total of 6920 images of these 41 persons.

The proposed NCDA is compared with the same methods used for the WARD dataset. 21 persons were used for training while the rest 20 were used for testing. Figs. 2.5(A) - (C) compare the performance for camera pairs 1-2, 1-3 and 3-4 respectively. We see that the proposed method performs better than the rest for both the cases when there is not much appearance variation (for camera pair 1-2 where both cameras are indoor and for camera pair 3-4 where both cameras are outdoor) and when there is significant lighting variation (for the camera pair 1-3, where camera 1's FoV is indoor and camera 3's FoV is outdoor). Expectedly, for camera pairs 1-2 and 3-4 the performance of the proposed method is the best. For the indoor camera pair 1-2 the proposed method applied on similarity scores generated by feature transformation (NCDA on FT) and on

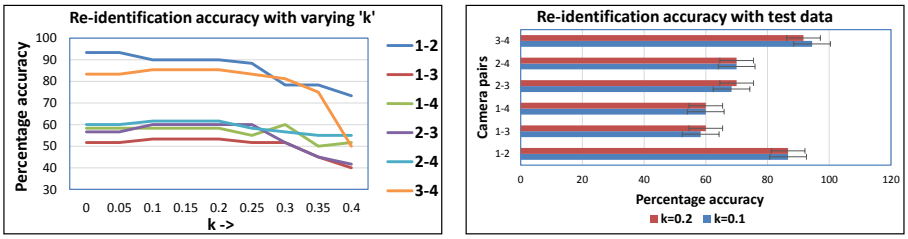


(A) (B) (C)

Figure 2.5: CMC curves for RAiD dataset. In (A), (B), (C) comparisons are shown for the camera pairs 1-2 (both indoor), 1-3 (indoor-outdoor) and 3-4 (both outdoor) respectively.

the similarity scores by ICT (NCDA on ICT) achieve 86% and 89% rank 1 performance respectively. For the outdoor camera pair 3-4 the same two methods achieve 79% and 68% rank 1 performance respectively.

In all the camera pairs, the top two performances come from the NCDA method applied on two different camera pairwise similarity scores generating methods. It can further be seen that for camera pairs with large illumination variation (*i.e.* 1-3) the performance improvement is significantly large. For camera pair 1-3 the rank 1 performance shoots up to 67% and 60% on application of NCDA algorithm to FT and ICT compared to their original rank 1 performance of 26% and 28% respectively. Clearly, imposing consistency improves the overall performance with the best absolute accuracy achieved for camera pairs consisting of only indoor or only outdoor cameras. On the other hand, the relative improvement is significantly large in case of large illumination variation between the two cameras.



(A) (B)

Figure 2.6: Performance of the NCDA algorithm after removing 40% of the people from both cameras 3 and 4 in the RAiD dataset. In (A) re-identification accuracy on the training data is shown for every camera pair by varying the parameter k after removing 40% of the training persons. (B) shows the re-identification accuracy on the test data for the chosen values of $k = 0.1$ and 0.2 when 40% of the test people were not present.

2.5.3 Re-identification with Variable Number of Persons

Next we evaluate the performance of the proposed method for the generalized setting when all the people may not be present in all cameras. For this purpose, from the RAiD dataset we chose two cameras (namely camera 3 and 4) and removed 8 (40% out of the test set containing 20 people) randomly chosen people. No change is made in cameras 1 and 2. For this experiment the accuracy of the proposed method is shown with similarity scores as obtained by learning the feature transformation between the camera pairs. The accuracy is calculated by taking both true positive and true negative matches into account and it is expressed as $\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique people in the testset}}$.

Since the existing methods do not report re-identification results on variable number of persons nor is the code available which we can modify easily to incorporate such a scenario, we can not provide a comparison of performance here. However we show the performance of the proposed method for different values of k. The value of k is learnt using 2 random partitions of the training data in the same scenario (*i.e.*, removing 40% of the people from camera 3 and 4). The average accuracy over these two random partitions for varying k for all the 6 cameras are shown in Fig. 2.6(A).

As shown, the accuracy remains more or less constant till $k = 0.25$. After that, the accuracy for camera pairs having the same people (namely camera pairs 1-2 and 3-4) falls rapidly, but for the rest of the cameras where the number of people are variable remains significantly constant. This is due to the fact that the reward for 'no match' increases with the value of k and for camera pair 1-2 and 3-4 there is no 'no match' case. So, for these two camera pairs, the optimization problem (in Eqn. 2.11) reaches the global maxima at the cost of assigning 0 labels to some of the true associations (for which the similarity scores are on the lower side). So any value of k in the range $(0 - 0.25)$ will be a reasonable choice. The accuracy of all the 6 pairs of cameras for $k = 0.1$ and 0.2 is shown in Fig. 2.6(B), where it can be seen that the performance is significantly high and does not vary much with different values of k.

2.6 Conclusion

When the same set of data-points are observed by multiple agents and/or at multiple spatio-temporal locations, pairwise data-association may often lead to infeasible scenarios over the network of agents and the global space-time horizon. In this chapter, we have proposed a generalized data-association method that not only maintains consistency across the network of agents or amongst observations across spatio-temporal locations, but also improves the data-association accuracy. This global data-association technique, termed as the 'Network Consistent Data Association' (NCDA), is posed as a binary integer program on a graph with multiple network-level constraints and a globally optimal solution can be found using standard optimization techniques such as branch and bound, dynamic search etc. We have shown that the proposed NCDA method is also capable of handling the challenging data-association scenario where the number of

data-points varies across different sets of instances in the network.

In this chapter, the performance of the NCDA method is studied by applying it to the classic multi-camera person re-identification problem. Analysis of the experimental results indicates that the proposed method improves both network level consistency and pairwise re-identification accuracy. In Chapter 3, we shall show how the same generalized NCDA can be applied to the challenging spatio-temporal cell tracking problem in densely packed multilayer biological tissues.

Chapter 3

Context Aware Spatio-temporal Cell Tracking In Densely Packed Multilayer Tissues

3.1 Introduction

In developmental biology, the causal relationship between cell growth patterns and gene expression dynamics has been one of the major topics of interest. A proper quantitative analysis of the cell growth and division patterns in both the plant and the animal tissues has remained mostly elusive so far. Information such as rates and patterns of cell expansion and cell division play a critical role in understanding morphogenesis in a tissue. The need for quantifying the cellular parameters such as average rate of cell divisions, cell cycle lengths, cell growth rates etc. and observing their time evolution is, therefore, extremely important.

Towards this goal, with the advancements in microscopy and other imaging

techniques, time lapse videos are being collected to quantify the behavior of hundreds of cells in a tissue over multiple days. For visualizing the cells over time within a densely packed multilayer tissue, one such in-vivo time-lapse microscopy technique is confocal laser scanning microscopy (CLSM) based *Live Cell Imaging*. With this technique, optical cross sections of the cells in the tissue are taken over multiple observational time points to generate spatio-temporal image stacks. For high-throughput analysis of these large volumes of image data, development of fully automated image analysis pipelines are becoming necessities, thereby giving rise to many new automated visual analysis challenges.

Automated cell tracking with cell division detection is one of the major components of all such pipelines (such as [33]) that analyzes the live cell imaging data. A review of current cell tracking imaging methodologies can be obtained in [49]. The computational challenges related to a robust design of cell tracker come from multiple sources such as variable number of cells in the field of view (FoV), deformation of cell shapes, complex topologies of cell clusters, low SNR in the images, etc. In this chapter, we present an automated visual tracker for cells tightly packed in developing multilayer tissues. This calls for developing strategies for *temporal* associations of the cells. Moreover, since at every time point of observation a cell could be imaged across multiple spatial images, the tracking method must be capable of finding correspondences in the *spatial* direction as well. Beyond these, the tracker has to be able to detect cell divisions, detect new cells as the deeper layers of the tissues are imaged, differentiate between cells in a close neighborhood sharing similar physical features and generate correct matches in presence of low SNR. These challenges are evident in the sample CLSM image stack of a live Arabidopsis shoot meristem, as shown in Fig. 3.1.

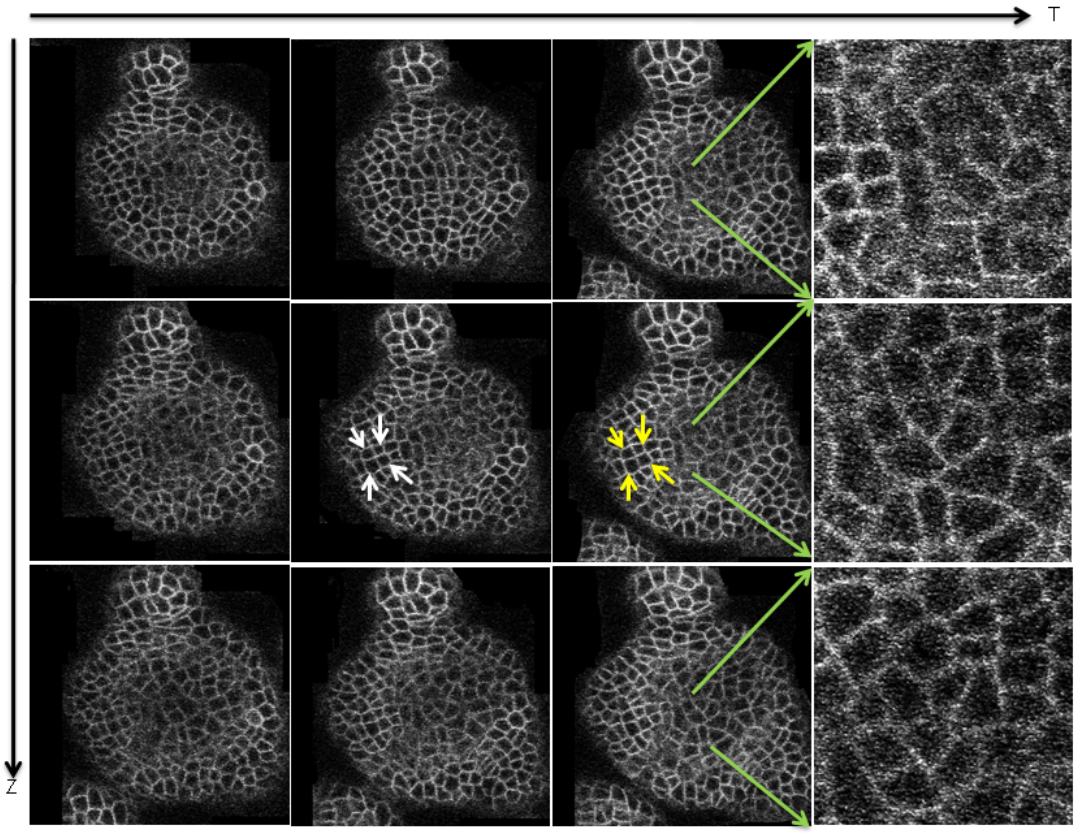


Figure 3.1: A typical 4D (X-Y-Z-T) live-imaging data. A live Arabidopsis shoot meristem tissue is imaged using a confocal laser scanning microscope at multiple time points. The plasma membranes of the cells are stained with fluorescent proteins and that is why the cell walls are the only visible parts. Each of the first three columns of images presents Z stack of image slices, i.e., the cross sections of the tissue imaged at various depths of it. When such images are collected over time to capture the growth of the tissue along with that of individual cells in it, it forms a 4D image stack. As can be seen from the figure, there are various challenges associated with the problem, viz., growth/deformation of the cells in the tissue, stereotypical cell shapes in the tissue and hence less discriminative physical features (as an example, 4 cells from a close neighborhood are marked with white and yellow arrows respectively in two consecutive time points which have very similar shapes and sizes), minor shifts between images and low SNRs in the central regions of the tissue. We have zoomed into these low SNR regions in the 4^{th} column of the figure. As seen, it is really difficult to even manually mark the boundaries of a number of cells in these regions.

In this work, we propose to solve the spatio-temporal tracking problem as a graph inference problem. All pairs of images which are either spatially or temporally consecutive are first analyzed to obtain cell slice to cell slice similarity measures exploiting context information. For every such pair of images, we build a graph on one of the images with individual cells as the nodes and neighboring nodes sharing an undirected edge between them. We further define a Conditional Random Field (CRF) on the graph, the probable states of each node being the candidate cell correspondences from the next image. A distance defined on the physical features extracted from a cell and that of each of its candidate matches is used to constitute the node potential. The spatial context is modeled on each of the edges based on the relative location of the cell and its neighbors by utilizing the tight spatial topology of the cell clusters. The approximate marginal for each node is obtained by a Loopy Belief Propagation scheme. Treating these marginals as similarity measures between a cell slice to its spatial/temporal candidates, we further apply the NCDA method introduced in the previous chapter to generate globally consistent 4D spatio-temporal correspondences. Unlike the person re-identification problem, similarities between cells are computed only between immediate spatio-temporally neighboring image slices and hence, the loop constraints are simplified accordingly. The overall tracking pipeline is shown in Fig. 3.2.

3.2 Overview of the Proposed Method

As mentioned earlier, many animal and plant tissues (such as the shoot meristem of a plant, epithelial tissues in animals) are a collection of tightly packed small cells arranged in clonally different layers forming a solid 3D structure. To visualize the internal parts of these 3D structures we employ imaging techniques such as Confocal

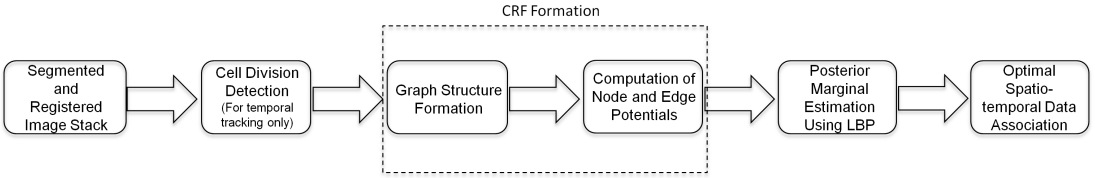


Figure 3.2: Proposed cell tracking framework - different sequential components in the proposed method. The input to the method is a Watershed segmented and registered 3D or 4D image stack. For temporal tracking only, the next stage is detection of possible cell division events. The tracking is done sequentially on pairs of spatially or temporally consecutive slices. For any of such pairs, once the cell divisions are detected, we remove the parent and children cells from the respective segmented images and build a graph on one of the images of the pair based on neighborhood structure around each cell with individual cells as nodes in the graph. The candidate matches for each cell is found from the other image in the pair under consideration (for details, see Sec. 3.3.1). The graph is then represented as an CRF. The node and edge potentials are computed using methods described in Sec. 3.3.5 and Sec. 3.3.6 and finally the marginal posteriors on the states of individual nodes are estimated using loopy belief propagation. These steps are repeated for every sequential pairs of images in the stack (along 'z' and 't'). Finally, for a 4D image stack, optimal spatio-temporal correspondences in the entire stack are obtained using these computed marginals in the proposed NCDA method (Sec. 3.4).

Laser Scanning Microscopy (CLSM) that generates serial optical cross sections of the tissue at various focal planes, thereby generating a 3D stack of images, each containing tightly packed 2D cross sections of 3D cells. In case of time-lapse 'live cell imaging', the same tissue is imaged at successive time points resulting in a collection of a number of such 3D stacks. 2D segmentation techniques (such as Watershed) are employed to segment out individual 2D cell cross sections on each of the confocal slices. The problem of finding correspondences between such 2D cell slices along the depth of the tissue is called 'spatial tracking', analogous to the 'temporal tracking' problem where such correspondences are estimated between slices of the same cell at successive observational time points. The objective of this work is to provide a solution strategy to the general spatio-temporal tracking problem in a 4D confocal image stack.

The 2D slices of the cells in the tissue are already registered in one 3D stack. We can also register cell slices across time between any two image slices (Fig. 3.1) of the

tissue using methods such as [68]. This ability to register cell slices across space and time, along with the fact that the relative positions of the centroids of two neighboring cells in the tissue do not vary substantially across both time and space motivate us to pose the problems of spatio-temporal cell tracking as a graph-based feature matching problem.

3.2.1 Graph Structure

As can be seen in Fig. 3.3, a graph can be built on top of every slice image in the tissue. The nodes of the graph would be the 2D segmented cell cross-sections and each of the immediately neighboring cells would share a link/edge between them. In spatial tracking, each of these cells can either have a correspondence to one of the cell slices in the next z-slice or they can have no correspondence - in case the cell ends in the present image slice and not imaged in the deeper slice. Also, in case of temporal tracking, a cell might be out of Field-of-View (FoV) or not detected because of noise in the image. Thus, a candidate set of cell slices from the subsequent image can be estimated for each cell in the image slice on which the current graph is built and this candidate set can be considered as the set of all possible states/labels for a certain node in the graph. An additional state, corresponding to the case that the cell is not imaged in the next confocal plane or next time point needs to be included in this set. For temporal tracking, we first detect the cell division events across the two images (Fig. 3.4) and then build the graph with the rest of the cell slices in the first image as the nodes. The details on how the graph is formed and the set of states/labels for each node is ascertained are given in Sec. 3.3.1.

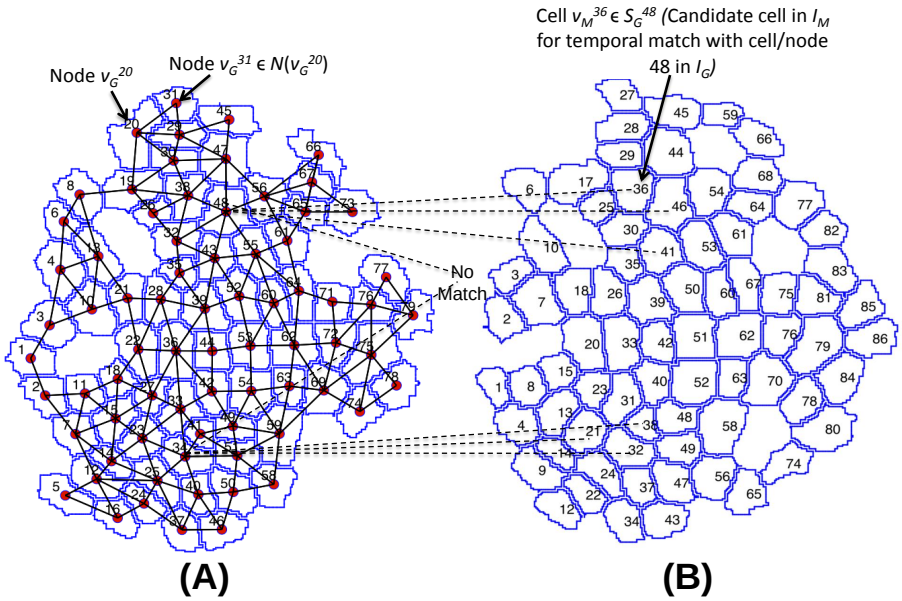


Figure 3.3: Graph Structure. (A) For tracking cells between two spatially and temporally consecutive image slices, a graph is built on one of the images, where the nodes of the graph are the segmented cells and two neighboring cells share an edge between them. For temporal tracking, the cells undergoing division are set aside before constructing the graph. (B) From the next image slice, the candidate matches for each cell in A are estimated. Again, for temporal tracking, the children cells after division are also removed from the image and the candidate set of best 'K' states for each node in A is estimated through a search in B in a spatial window around the location of each of the nodes in A. A 'K+1'th state is added to each of the candidate sets corresponding to the case that the cell is not imaged or poorly imaged in B, referred to as the 'No Match' state in the figure. Now the graph is expressed as a CRF, where the node potentials are computed based on feature distances between each node and its candidates (see Sec. 3.3.5) and the edge potentials are computed based on the relative locations of the neighboring nodes in A and the same between any two cells from within their respective candidate sets in B (Sec. 3.3.6).

3.2.2 Computation of Potential Functions

As we have mentioned earlier, the relative positions of the centroids of the neighboring cell slices do not vary a lot in short time intervals or along z and hence the knowledge of the most probable state for any cell can substantially aid in estimating the maximum likely state for its neighboring cells. We consider the graph just formed as a 'Conditional Random Field'. The node potentials in the CRF are computed based on the shape similarity between a cell slice and each of its candidates along with their relative centroid locations. The edge potentials are obtained based on similarity between the relative positions of two neighboring nodes and that of their any two candidate cells' centroids in the successive slice or time point. The details of the edge and node potential computations are described in Sec. 3.3.5 and Sec. 3.3.6.

3.2.3 Computation of marginal posteriors: Pairwise similarities between cell slices

Once the graph is formed and the necessary potential functions computed, the next step is to design a strategy on this graph to estimate the marginal posteriors for each of the nodes - how likely it is for a node to be associated to each of its candidates. We employ a 'Loopy Belief Propagation' (LBP) scheme (based on the well known 'Sum-Product' algorithm [52]) for this purpose. In Sec. 3.3.7, we show the iterative parallel inter-node message updation strategy in the traditional sum-product scheme and compute the marginals.

3.2.4 Complete spatio-temporal cell tracking: Network Consistency

Once the marginal for each node (cell) is estimated between every pair of spatially/temporally consecutive images in the stack, the next problem is to generate asso-

ciations between all these 2D cell slices without affecting spatio-temporal consistency. Similarity measure between any two such 2D cell slices is obtained from the computed marginals. A 4D graph is built by combining all the pairwise graphs (as discussed in Sec. 3.3.1) and removing links between cells in the same image slice. The spatio-temporal tracking problem is now posed as the problem of optimally putting labels (1 - correspondence, 0 - no correspondence) on the edges of this graph and the previously introduced NCDA method is used for this purpose. As the number of 2D cell segments may vary across space and time, the generalized NCDA, presented in Eqn. 2.11, is used at this stage. The final set of labels, obtained by solving the NCDA integer program, is the spatio-temporal tracking result in the 4D stack.

3.3 Graphical Model Design and Inference

3.3.1 Graph Formation on 2D Segmentations

Let us define the problem to be to find correspondences between the cells in two segmented confocal image slices I_G and I_M. The Watershed segmentation of I_G and I_M produces two sets of cell segments Ω_G and Ω_M respectively. Thus, the set of observations is given as

$$\mathcal{O} \;=\; \Omega_G \cup \Omega_M, \tag{3.1}$$

which comprises of 2D Watershed segmentations of both I_G and I_M. However, for temporal tracking, we first detect if some cells form I_G have divided into pairs of cells in I_M following the method described in Sec. 3.3.3 and remove the parent cells that has undergone division from Ω_G and the divided children from Ω_M. The graph and the candidate states of each node of the graph are thereafter formed using the remaining subsets of cells V_G and V_M containing N_G and N_M cells respectively, i.e. the remaining

cells

$$v_G^1, v_G^2, \cdots v_G^{N_G} \ \in \ V_G \subseteq \Omega_G$$
$$v_M^1, v_M^2, \cdots v_M^{N_M} \ \in \ V_M \subseteq \Omega_M \tag{3.2}$$

The graph is built on I_G and the set of nodes V_G is same as the set of segmented cells. Any two nodes v_G^i and v_G^j will have an edge between them if v_G^i and v_G^j are spatial neighbors. For tightly packed cluster of cells, v_G^i and v_G^j are neighbors if they share a common boundary and thus the set of all neighbors of a cell v_G^i would be

$$N(v_G^i) \ = \ \{v_G^j \ s.t. \ v_G^i \text{ and } v_G^j \text{ share common boundary}\}. \tag{3.3}$$

Thus, we can represent the graph g_G on I_G as an adjacency matrix A_G between the nodes,

$$\begin{aligned} A_G\left(i,j\right) &= 1 \text{ iff } v_G^j \in N\left(v_G^i\right), \\ &= 0, \text{ otherwise} \end{aligned} \tag{3.4}$$

3.3.2 Determination of Candidate States For Every Node

For finding correspondences between cells across two segmented slices I_G and I_M, the graph is built on the slice I_G following Sec. 3.3.1. Each node in the graph, corresponding to each cell slice v_G^i represents a random variable x^i that can take a label from the set S_G^i which is the set of K closest segments in the slice I_M around the point $\mathbf{c}_G^i$, the centroid of v_G^i on I_G. Therefore,

$$S_G^i \ = \ \{s_1^i, s_2^i, \cdots s_K^i\} \tag{3.5}$$

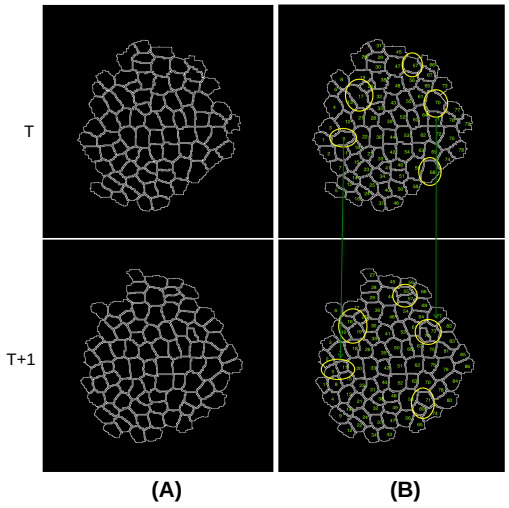


Figure 3.4: Cell Division Detection. (A) Two segmented image slices one time point apart. (B) The ellipses in image at T mark the parent cell that have undergone divisions between time points T and T+1 and those at time point T+1 mark the children cells after division.

where, $s_k^i \in V_M \ \forall k = 1, 2, \cdots K$ and

$$||\mathbf{c}_M^{s_1^i} - \mathbf{c}_G^i|| \leq ||\mathbf{c}_M^{s_2^i} - \mathbf{c}_G^i|| \cdots ||\mathbf{c}_M^{s_K^i} - \mathbf{c}_G^i||$$

$$\leq ||\mathbf{c}_M^{s_j^i} - \mathbf{c}_G^i|| \ \forall \ j \in \{1, 2, \cdots N_M\}, \ j \notin S_G^i \tag{3.6}$$

We can safely assume that the actual tracked cell slice in I_M would be amongst the K closest cells, as I_G and I_M are already registered.

Now, we add an additional label s_0^i to the candidate set S_G^i that represents the case where the cell slice v_G^i is not imaged in the slice I_M. Thus, the complete set of candidate states becomes

$$S_G^i \ = \ \{s_0^i, s_1^i, \cdots s_K^i\}. \tag{3.7}$$

For the datasets under study in this thesis, cells are tightly packed and the neighboring cells share common boundaries. However, for experimentation with other datasets, where the cells are generally not compactly arranged, this set can be represented as

$$N(v_G^i) \ = \ \{v_G^j \ s.t. \ ||\mathbf{c}_G^i \ - \mathbf{c}_G^j||_2 \leq th\}, \tag{3.8}$$

where c_G^i and c_G^j are the centroids of v_G^i and v_G^j respectively. th is a displacement threshold that can be learned from a set of training images as a constant (say 1.5) times the maximum displacement observed between a cell and its match in pairs of images. Note that this value can vary across different datasets.

3.3.3 Cell Division Detection

To detect cell divisions before forming the graph g_G in temporal tracking, we first compute the candidate sets C_G^i in I_M for a segmented cell slice $\omega_G^i \in \Omega_G$ following similar method as in Eqn. 3.6. Next we form all possible pairs of the candidate cells from C_G^i that share a boundary as in

$$D_G^i \;=\; \left\{(cd_p^i,\; cd_q^i)\; s.t.\; cd_p^i \in N(cd_q^i) \text{ and } cd_p^i, cd_q^i \in C_G^i\right\}. \tag{3.9}$$

Now, if the cell ω_G^i has divided into two children cells cd_p^i and cd_q^i, then ideally the shape of ω_G^i should be very similar to the combined shape of cd_p^i and cd_q^i, taken together (i.e. to the shape of $cd_p^i \cup cd_q^i$) and each of cd_p^i and cd_q^i would be approximately half the size of ω_G^i. Motivated by this physical property associated with cell division, we compute a Modified Hausdorff Distance (MHD) metric to estimate the shape similarity between $b(\omega_G^i)$ and $b(cd_p^i \cup cd_q^i)$, where b is the set of boundary points on a shape, when the point coordinates are recomputed with respect to the shape centroid. With these, we compute a set of distances as

$$d(\omega_G^i,\; D_G^i) \;=\; \frac{1}{t_1} MHD(b(\omega_G^i), b(cd_p^i \cup cd_q^i)) \;+ \frac{1}{t_2}\left[\left|\frac{1}{2} - \frac{area(cd_p^i)}{area(\omega_G^i)}\right| \;+\; \left|\frac{1}{2} - \frac{area(cd_q^i)}{area(\omega_G^i)}\right|\right] \tag{3.10}$$

If $\min d(\omega_G^i,\; D_G^i) \;\leq\; 1$, then it is inferred that the cell ω_G^i has divided into a pair of cells (cd_p^i, cd_q^i) for which this minimum is obtained. The values of the parameters t_1 and t_2 are learnt from a training image set and the details of parameter learning is

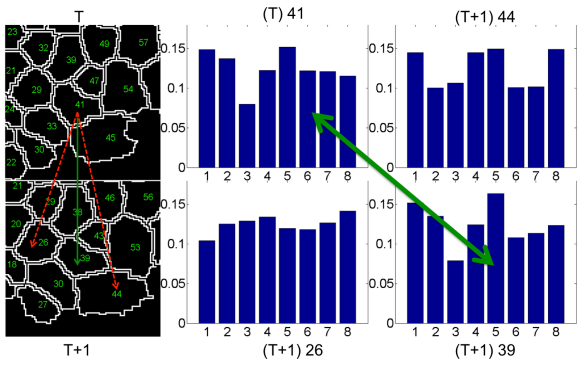


Figure 3.5: Shape descriptor for individual cells. Cell 41 at time point T has cells 26, 39 and 44 as candidates for correspondence at time T+1. The correct correspondence for 41 at T is 39 at T+1. In the left column of figure, the correct correspondence is shown in green arrow whereas the the incorrect ones are shown in red. Shape histogram descriptors are computed for each cell following the method described in Sec. 3.3.5. As expected, the histogram for 41 at T is similar to that of 39 at T+1 and the descriptors for the other two candidate cells are very different.

described in Sec. 3.5.4.

Once the cell division events are detected for one or more cells in I_G, the graph g_G is constructed using the methods described in Sec. 3.3.1 and 3.3.2 after eliminating the parents undergoing division and the divided children cells from Ω_G and Ω_M respectively and forming V_G and V_M.

3.3.4 Conditional Random Field Modeling

Let the set of random variables associated with v_G^i be $X = \{x^1, x^2, \cdots x^{N_G}\}$, which are to be estimated given the observation I_M. These random variables correspond to the state of each node in the graph and the support for each of these variables is the candidate set as discussed in Sec. 3.3.2.

Then the overall CRF is expressed as

$$\begin{aligned} P(X;\mathcal{O}) &= \exp\left(-E(X;\mathcal{O})\right)/Z \\ &= \exp\left\{-\sum_{c\in clq(X)} E_c(X;\mathcal{O})\right\}/Z, \end{aligned} \tag{3.11}$$

where Z is the partition function and E is the energy function defined on all the cliques of the graph, which can be further split into individual nodes and edges as

$$E(X;\mathcal{O}) = \sum_{i=1}^{N_G} E_i\left(x_i;\mathcal{O}\right) + \sum_{i=1}^{N_G} \sum_{j:v_G^j\in N(v_G^i)} E_{ij}(x_i,x_j;\mathcal{O}). \tag{3.12}$$

Then,

$$\begin{aligned} P(X;\mathcal{O}) &= \frac{1}{Z}\prod_{i=1}^{N_G}\exp\left(-E_i(x_i;\mathcal{O})\right) \cdot \\ &\quad \prod_{\substack{(i,j) \\ : v_G^j\in N(v_G^i)}} \exp\left(-E_{i,j}(x_i,x_j;\mathcal{O})\right) \\ &= \frac{1}{Z}\prod_{i=1}^{N_G}\phi_i(x_i;\mathcal{O}) \cdot \\ &\quad \prod_{\substack{(i,j) \\ : v_G^j\in N(v_G^i)}} \psi_{i,j}(x_i,x_j;\mathcal{O}) \end{aligned} \tag{3.13}$$

Here ϕ_i represents the node potential of any node v_G^i in g_G, and ψ_{ij} is the edge potential from node v_G^i to node v_G^j. If we are only interested in tracking cells between any two image slices, we have to maximize $P(X;\mathcal{O})$ to estimate the optimal states for every node. Towards that objective, we first estimate the approximate marginal distributions $P(x_i;\mathcal{O})$ at each node using belief-propagation scheme as described later. The optimal states that maximize the posterior distribution could be then estimated by maximizing the marginals independently.

3.3.5 Computation of Observation/Node Potential:

The node potential is defined on every node of the graph, which is the likelihood on the label taken by a node belonging to V_G, given the observation $\mathcal{O}$. It is analogous to the probability distribution of any node v_G^i being assigned to each of its candidate states. This distribution is computed independently for each node based on its shape similarities and proximities in location of its centroid from each of its candidates.

For measuring similarities between cell shapes, we generate a shape histogram descriptor for each of the cells, which is very similar to one of the methods described in [2]. First we recompute co-ordinates of a cell's peripheral points by shifting the origin to the cell's centroid. Next, we partition the x-y plane into 8 angular sectors centered at the origin and compute the mean Euclidean distances of the peripheral points falling into each of these partitions from the origin. The set of these distances forms a 8 bin histogram descriptor for the shape of a cell, the angular sectors being sorted counter clockwise from x-axis. Note that, unlike the classical shape histograms of this sort [2], we compute mean distances from each sector instead of counting the number of points, as the latter gives us scale invariance and may lead to a high match score between a legitimate cell and a small region generated by over-segmentation on noisy images. Some sample descriptors of a cell and its candidates for correspondence are given in Fig. 3.5.

Let the shape histogram associated with the cell slice v_G^i be h_G^i and that with the candidate slice s_j^i be h_M^j (as $s_j^i \in V_M$). We computed the K-L divergence (KLD) between h_G^i and h_M^j which gives us a distance measure between these two cell slices and suppose it is represented as $d_1^i\left(v_G^i, s_j^i\right)$,

$$d_1^i\left(v_G^i,\ s_j^i\right)\ =\ \mathrm{KLD}\left(h_G^i,\ h_M^j\right)\ . \tag{3.14}$$

We also compute the distances between the centroids of a cell slice in I_G and

each of its candidates in I_M and the distance is given by,

$$d_2^i\left(v_G^i,\ s_j^i\right) \ = \ \|\mathbf{c}_M^{s_j^i} - \mathbf{c}_G^i\|_2 \ . \tag{3.15}$$

Hence, the overall distance between a cell slice v_G^i and one of its candidates $v_M^{s_j^i}$ is expressed as a combination of normalized d_1 and d_2 as

$$d^i\left(v_G^i,\ s_j^i\right) \ = \ w\frac{d_1^i}{\lambda_1} \ + \ (1-w)\frac{d_2^i}{\lambda_2}, \ \ 0 \leq w \leq 1 \ . \tag{3.16}$$

The corresponding node potential for each node is

$$\phi_i\left(x_i = s_j^i; \mathcal{O}\right) \ = \ \exp\left(-d^i(v_G^i,\ s_j^i)\right) \ \forall j = 1, 2 \cdots K \tag{3.17}$$

and

$$\phi_i\left(x_i = s_0^i; \mathcal{O}\right) \ = \ 1 - \max_j \left\{\phi_i\left(x_i = s_j^i; \mathcal{O}\right), \ \ j = 1, 2, \cdots K\right\} \tag{3.18}$$

The normalization parameters λ_1 and λ_2 (in Eqn. 3.16) are learnt from a training dataset. See Sec. 3.5.4 for details on parameter estimation.

3.3.6 Computation of Spatial Context/Edge Potential:

This potential function is defined on edges connecting pairs of neighboring nodes and is representative of the conditional distribution $P(x_j|x_i, \mathcal{O})$. The computation of the potential function depends on the fact that if two neighboring cells v_G^i and v_G^j are tracked to two cell slices v_M^p and v_M^q, then the relative position of v_G^j with respect to v_G^i should be very similar to that of v_M^q and v_M^p. As a result, if v_G^i is tracked to v_M^p then the probability that v_G^j corresponds to v_M^q gets boosted if

$$\mathbf{c}_G^j - \mathbf{c}_G^i \ \approx \ \mathbf{c}_M^q - \mathbf{c}_M^p \ , \tag{3.19}$$

where $\mathbf{c}_G^i, \mathbf{c}_G^j, \mathbf{c}_M^p, \mathbf{c}_M^q$ be the centroids of $v_G^i, v_G^j, v_M^p, v_M^q$ respectively.

Clearly, the additional evidences for matching two cell slices in I_G and I_M comes in the form of local neighbourhood structure based contextual information.

Thus, the contextual transition potentials between any two nodes v_G^i and v_G^j taking non-zero states can be expressed as a function of the shift between the relative positions of those nodes

$$\psi_{i,j}\left(x_i = s_p^i, x_j = s_q^j; \mathcal{O}\right) = \exp\left\{-\gamma\|(\mathbf{c}_G^j - \mathbf{c}_G^i) - (\mathbf{c}_M^{s_q^j} - \mathbf{c}_M^{s_p^i})\|_2\right\} \tag{3.20}$$

$\forall p, q = 1, 2, \cdots K$, where $s_p^i \in S_G^i, s_q^j \in S_G^j$ and $i, j \neq 0$.

Now, in both spatial and temporal tracking, there is one more state s_0^i for every node i that corresponds to the case that the particular cell is not imaged in the successive slice (spatial or temporal). Thus, the transition potentials must also incorporate the case where one of the cells is not tracked and its neighboring cell is matched to one of the cells in the next slice or not matched to any cell and vice versa. Incorporating these values, the complete edge potential function between any two neighboring nodes v_G^i and v_G^j would be

$$\psi_{i,j}\left(x_i = s_0^i,\ x_j = s_q^j; \mathcal{O}\right) \ = \ \frac{1}{K+1}\ \forall q = 0, 1, \cdots K\ . \tag{3.21}$$

This corresponds to the case when v_G^i is not matched to any cell in I_M. When both the cells v_G^i and v_G^j have correspondences in the subsequent spatial or temporal image I_M,

$$\psi_{i,j}\left(x_i = s_p^i,\ x_j = s_q^j; \mathcal{O}\right) \ = \ \exp\left\{-\gamma\|(\mathbf{c}_G^j - \mathbf{c}_G^i) \ - \ (\mathbf{c}_M^{s_q^j} - \mathbf{c}_M^{s_p^i})\|_2\right\}\ , \tag{3.22}$$

for $p, q \neq 0$.

Finally, when v_G^i has a match in the next spatial or temporal image slice I_M, but its neighbour v_G^j does not, then the corresponding edge potential entries become

$$\psi_{i,j}\left(x_i = s_p^i,\ x_j = s_0^j; \mathcal{O}\right) = 1 - \max_q\left\{\psi_{i,j}\left(x_i = s_p^i, x_j = s_q^j; \mathcal{O}\right),\ q = 1, 2, \cdots K\right\} \tag{3.23}$$

for $p \neq 0$.

3.3.7 Loopy Belief Propagation: Estimation of Marginals

The next step involves the computation of the marginal probability distributions for the states x_i of each node $v_G^i \in V_G$, given the observations $\mathcal{O}$. For computation of the marginals at each node, we choose to use a very popular local *message-passing* algorithm known as *Belief Propagation* (BP) [77]. Since there are many loops or cycles in our graph, the algorithm is called a *Loopy Belief Propagation* (LBP). This is an iterative algorithm and at l^{th} iteration, each node v_G^i computes a message to be sent to each of its neighbors and the message sent to $v_G^j \in N(v_G^i)$, according to the popular *Sum-Product* algorithm [52], is,

$$m_{i,j}^{(l)}(x_j) = \alpha \sum_{x_i} \left\{ \psi_{i,j}(x_i, x_j; \mathcal{O}) \phi_i(x_i; \mathcal{O}) \prod_{x_k : v_G^k \in N(v_G^i) \setminus v_G^j} m_{k,i}^{(l-1)}(x_i) \right\} \tag{3.24}$$

where α is a normalizing constant. Note that the updation strategy employed here is parallel, i.e. all the edges in the CRF are updated simultaneously in each iteration.

Also, at each iteration l, each node v_G^i produces an approximate marginal distribution

$$P^{(l)}(x_i; \mathcal{O}) \;=\; \alpha \phi_i(x_i; \mathcal{O}) \prod_{x_j : v_G^j \in N(v_G^i)} m_{j,i}^{(l)}(x_i) \tag{3.25}$$

For a tree type graph, these approximate marginal distributions are guaranteed to converge to the true marginals, but for a graph as ours that contains multiple loops there is no guarantee of convergence of the LBP [90]. However, in literature, such as [71], LBP has shown very good empirical performance and in most of our experiments the method converged very quickly.

3.4 Optimal Data Association: Combining Spatial and Temporal Cell Tracking and Resolving Association Ambiguities Through NCDA

Once the marginal posterior distribution for every cell is computed via LBP, the next step is to infer the optimal states for individual cells (candidates) from these computed marginals. If tracking is performed along either of the two dimensions 'z' or 't', the MAP estimates on these marginals would give us the optimal correspondences, i.e.

$$\hat{x_i} \;=\; \arg_{x_i} \max\, P^{(L)}(x_i; \mathcal{O}) \tag{3.26}$$

where L being the iteration when LBP converges and this optimum state corresponds to either the 'no-match' case or a specific cell in I_M.

However, when the tracking problem is multidimensional in nature, optimal or unambiguous tracking result may not be attained by simply maximizing the marginals between every spatially/temporally consecutive image slices. Spatial and temporal correspondences obtained by choosing the most similar candidate for each cell independently between every pair of images might not conform to one another and in turn, can lead to infeasible spatio-temporal mappings. It can be noted that in spatio-temporal tracking, multiple paths of correspondences may exist between cells from any two image slices and all these paths must point to the same correspondence maps between individual cells. Therefore, the objective is to obtain network consistent associations between the 2D cell slices in the entire spatio-temporal image stack using the similarity scores generated via previous method and the NCDA method, presented in the last chapter, is utilized for that purpose.

Following the NCDA problem formulation, each 2D image slice (containing a cluster of tightly packed cell slices) is treated as a 'group' and individual 2D cells on these slices are the nodes. Also, for any given image slice, similarity scores are computed only to its immediate spatio-temporal neighboring slices (i.e. slice above, slice below, slice at same 'z' at previous time point and the same at next time point). This architecture yields a network of image slices (groups) that can be exhaustively covered using quartets of groups. Fig. 3.9(A) shows one such quartet in a large network. Following similar notations for nodes and groups as those in the last chapter, the loop constraints for this problem are,

$$\begin{gathered} x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,l}^{r,s} + x_{l,j}^{s,q} - 2 \\ \forall\, i = [1, \cdots n_p],\, \forall\, j = [1, \cdots n_q] \\ \forall\, k = [1, \cdots n_r],\, \forall\, l = [1, \cdots n_s] \\ \forall\, p, q, r, s = [1, \cdots m], \quad \text{and } p < r < s < q \end{gathered} \tag{3.27}$$

Here, i, j, k, l are 2D cell segments on the p^{th}, q^{th}, r^{th} and s^{th} image in the stack respectively. Using the marginal posteriors as similarity scores between a cell slice to its spatial/temporal candidates, we further run the NCDA for generating complete optimal 4D spatio-temporal correspondences between 2D cell slices. As the number of 2D cell slices widely vary across spatial/temporal cross-sections, the generalized NCDA problem in Eqn. 2.11 is used to generate the tracking results.

3.5 Experiments and Results

3.5.1 Data Collection and Preprocessing

For the experiments performed in the present study, the 3D structures of the tissues are imaged using single-photon confocal laser scanning microscope and we have specially dealt with the 'Shoot Apical Meristem' (SAM) of the plants that showcase all the challenges associated with any spatio-temporal cell tracking problem in a tightly packed multilayer tissue. The SAM of *Arabidopsis Thaliana* consists of approximately 500 cells and they are organized into multiple cell layers that are clonally distinct from one another. By changing the depth of the focal plane, CLSM can provide in-focus images from various depths of the specimen. To make the cells visible under laser, fluorescent dyes are used. The set of images, thus obtained at each time point, constitute a 3-D stack, also known as the 'Z-stack'. Each Z-stack is imaged at a certain time interval (e.g. 3-6 hours between successive observations) and it is comprised of a series of optical cross sections of SAMs that are separated by approx. 1.5-2 μm. A standard shoot apical meristematic cell has a diameter of about 5 - 6 μm and hence in most cases, a single cell is not visible in more than 3-4 slices when the tissue is sparsely imaged at the aforementioned z-resolution to avoid photodynamic damage to the cells.

Each 2D image slice in the 4D confocal image stack is further segmented into individual cell slices. The choice of the 2D segmentation algorithm is largely data-specific. For our experiments on the SAM tissues, we use an adaptive Watershed segmentation method [69] that learns the 'h-minima' threshold directly from the image data so that a uniformity in cell sizes is maintained as a result of the segmentation. This method works satisfactorily for SAM cells as, in general, all SAM cells on a 2D confocal slice have similar sizes. This 2D segmentation method is also robust to over and under-segmentation

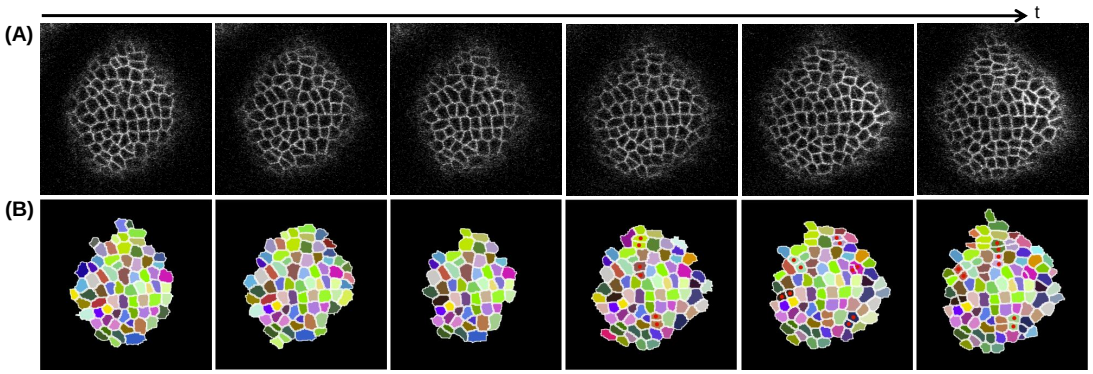


Figure 3.6: Results on the temporal tracking on Arabidopsis SAM live imaging dataset with time resolution of 3 hours. (A) Raw confocal image slices at 3 μm deep into the tissue imaged every 3 hours from 3^{rd} hour of observation to 18^{th} hour. (B) Temporal tracking result shown by color coding the cells. The same cells are marked with the same color. After cell division, the children cells are marked with the same color as their parent, also a red dot is put at the center of each of the children.

errors to a large extent.

The image slices in one single 3D confocal stack is already registered because of minimal movement of the tissue specimen during imaging at any given time point of observation. However, during successive observations the specimen is moved in and out of the imaging setup which causes rotation and shift of the imaged 3D stack from that at the previous time point. Thus, the image slices in successive time points have to be registered prior to the cell-tracking. We register the cell slices across time between any two confocal images of the tissue using a 'local graph' based registration technique [68]. This is a fully automated landmark based registration method that finds out correspondences between the two image slices and utilizes these correspondences (landmarks) to register one image to the other.

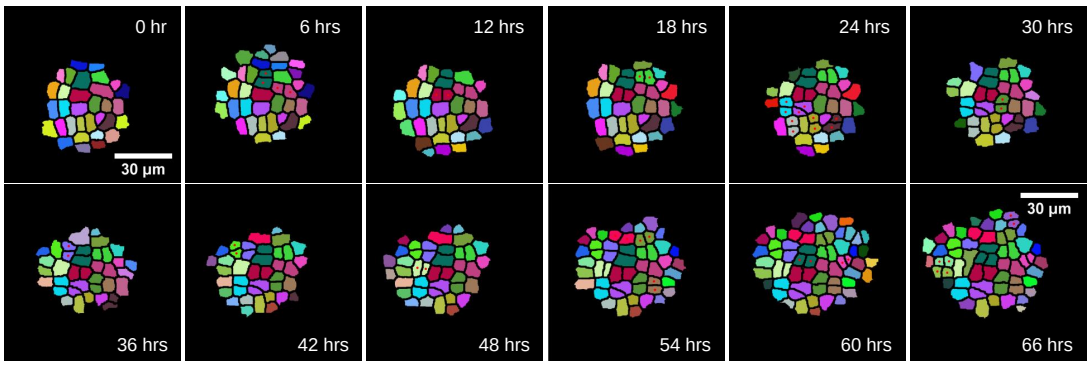


Figure 3.7: Results on the temporal tracking on Arabidopsis SAM live imaging dataset where the images are collected through three days with time resolution of 6 hours between successive observations. Tracking result is shown by color coding the cells. The cell division events are also detected with perfect accuracy and are displayed on this figure the same way as Fig. 3.6.

3.5.2 Pairwise/Slice to Slice Tracking Results

In the first part of the results, we analyze the performance of the CRF based similarity score generation method on pairs of images. Note that, the message passing scheme yields marginal posteriors for every cell over its candidate set. Now, even before applying NCDA for generating 4D tracking results, MAP estimates on these posteriors would give us tracks of cells if only spatial or temporal tracking is performed.

Fig. 3.6 shows a typically obtained result for temporal tracking. Fig. 3.6(A) shows raw confocal image slices at a depth of 3μm from the tip of SAM through six consecutive time points (3^{rd} to 18^{th} hours), each observed every three hours. Although the images are registered, because of the growth of the cells in the tissue there are local shifts in the cells' positions, which make the task of temporal tracking more challenging. The segmentation and tracking results for these slices are shown in Fig. 3.6(B), where the slices of one cell across different time points are marked with the same color. We have also marked the 12 cell division events detected by the tracker on the same images. The children cells are marked with red dots and they share the same color with their

Table 3.1: Tracking Result Summary

	TP	FP	TN	FN
Spatial	86%	0.25%	12.13%	1.62%
Temporal	83%	0%	14.66%	2.34%
Division	31/33	0	-	2/33

parent cell.

To evaluate the accuracy of the tracker in the situations where the temporal resolution is small, i.e., the 3D stacks are imaged after large time gaps, we tested the proposed tracker on a temporal stack where the imaging is done every 6 hours. With a longer gap between observations, the deformation of the cells in the tissue is even more visible, which results in larger shifts between centroid locations of the same cell in successive time points. Moreover, there are more number of cell division events which makes the temporal tracking problem even more challenging. Because of the robustness of the proposed method, we obtain highly accurate temporal tracking results as seen in Fig. 3.7. We also show that our method is capable of maintaining tracks for long duration (66 hours as shown in the figure) and it detected all legitimate cell division events.

The pairwise tracking result on an Arabidopsis SAM dataset (12 time points, each 3 hours apart and 7 slices at each time point) is summarized quantitatively in Table 3.1. TP corresponds to the cases where two cell slices are correctly matched either in space or time. When cell slices from two different cells are incorrectly matched together, it falls under FP. When the tracker fails to pick up a correct correspondence, it is represented by FN and its opposite case is tabulated under TN.

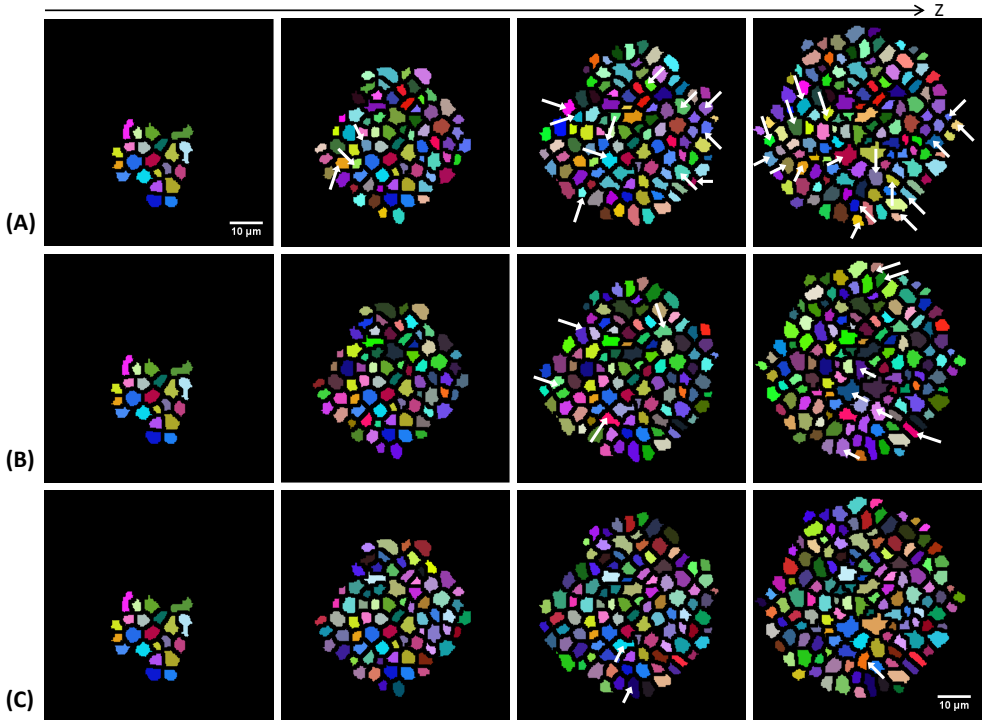


Figure 3.8: Comparison of the spatial tracking results as obtained from the proposed method with the results from [65] and a baseline tracker. The results are shown on a set of four spatially sampled image slices from a 3D image stack of Arabidopsis SAM. The tracking results are shown using similar color-coding as in the previous figures and the locations of errors in tracking are marked by white arrows. (A) The results obtained by using the baseline tracker contain many errors as it is designed on local cell shape features and the cell shapes even from a close neighborhood can be very stereotypical. (B) Results obtained by using [65] are much better in accuracy but still contain a number of FP, FN and switched tracks. (C) The proposed method performs the best out of these three with very few errors and no track switching.

We have compared the CRF based pairwise tracking method with the 'local-graph' based cell tracker [65] and also with a baseline tracker. In order to show the improvements in tracking accuracy using contextual information, we designed the baseline tracker on the same local cell shape features as used to compute the node potentials in Sec. 3.3.5 and the tracker associates cell slices across images using 'Hungarian algorithm'. Also, if any associated pair of cells have a feature distance larger than a predefined threshold, the track is terminated and re-initialized. Fig. 3.8(A) shows the tracking result by using this baseline tracker on four spatially sampled image slices from a 3D image stack. A number of wrong associations are marked by white arrows. Fig. 3.8(B) and Fig. 3.8(C) shows tracking results on the same images for [65] and the CRF based method respectively. The baseline tracker generates many wrong associations because the cell shapes are often very similar even in a close neighborhood. The tracker proposed in [65] performs much better than the baseline tracker and the errors comprise of both false-positives and false-negatives along with a number of switched tracks. The pairwise tracking method, presented in this chapter, however, performs the best as the errors obtained are much fewer in numbers than both [65] and the baseline and therefore validates the fact that the local contextual information indeed aids spatio-temporal cell tracking for tightly packed multilayer tissues.

3.5.3 Analysis of 4D Tracking on Spatio-temporal Image Stack: Effect of NCDA

The effect of NCDA towards improvement of spatio-temporal tracking results is shown in Fig. 3.9. In Fig. 3.9(A), a sample 2X2 block of images of Arabidopsis SAM is shown, which contains two spatially neighboring image slices at each of two

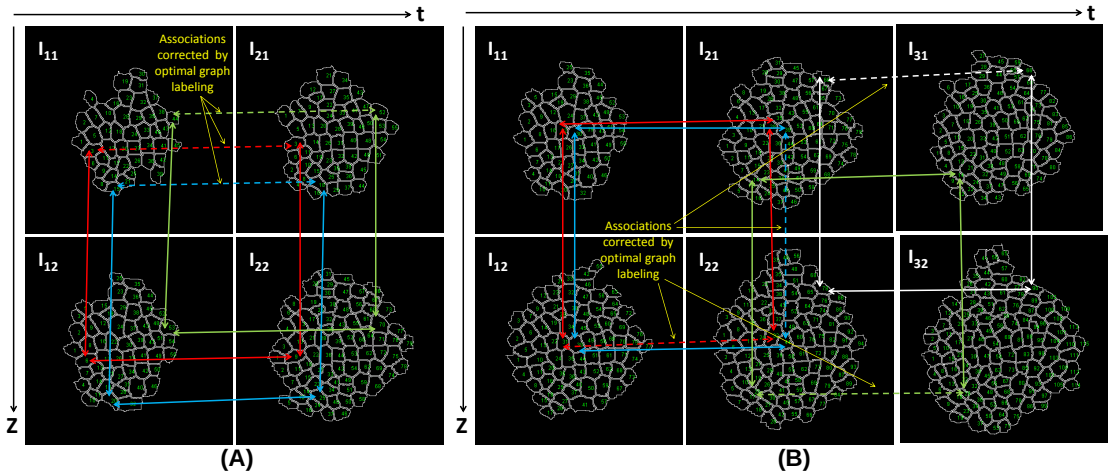


Figure 3.9: Effect of the NCDA towards improvement of spatio-temporal tracking results. (A) The figure shows a spatio-temporal 2X2 block of confocal images. Pairwise assignments between cells in spatial or temporal pairs of images are obtained by performing MAP inference on graphs formed on every image slice. Infeasible 4D assignments are observed when these pairwise associations are combined over the stack. Examples of such infeasibilities are shown for three cell slices. The solid arrows represent correct associations between cell slices and the broken arrows depict no association which is incorrect and cause the infeasibility. Our proposed data association approach establishes consistency in association and corrects these errors. (B) Similar results are observed in a 2X3 confocal stack. False negatives in pairwise spatial or temporal tracking results are rectified using NCDA.

consecutive time points of observation. Pairs of image slices are chosen and CRFs are formed for each of the pairs ($I_{11} - I_{12}, I_{12} - I_{22}, I_{21} - I_{22}$ and $I_{11} - I_{21}$). Now, marginal posteriors are estimated using LBP and MAP inferences are drawn to generate pairwise correspondences. When these pairwise associations are combined together, spatio-temporally infeasible associations are observed for a number of cells. For example, correct associations are found between cell 15 in I_{11} and cell 20 in I_{12}, cell 20 in I_{12} and cell 25 in I_{22}, cell 25 in I_{22} and cell 18 in I_{21}. Therefore, for spatio-temporal feasibility, cell 15 in I_{11} and cell 18 in I_{21} must also be associated. However, according to the aforementioned MAP inference, no associations for cell cell 15 from I_{11} is found in I_{21}. Similar infeasibilities are observed for cells 3 and 44 in I_{11}. The network consistent data association technique, when applied on the previously computed marginal posteriors for pairs of images, corrects these infeasibilities and establishes the associations.

Fig. 3.9(B) shows similar results on a 2X3 confocal image stack. As before, correct associations obtained using MAP inference on the graphs are shown in solid arrows. The false negatives are shown using broken arrows, which are further corrected using the application of NCDA (by enforcing the loop constraints on the corresponding quartets). Note that the NCDA method can correct a false negative irrespective of its appearance in spatial or temporal tracking in the 4D stack.

Although the number of network inconsistencies may seem a small (3 in the 2X2 stack and 4 in 2X3 stack) percentage of the total number of cells, it is of utmost importance that each such error is rectified. A few inconsistencies per slice may add up to a large number of errors in a typical confocal stack consisting of thousands of 2D cell slices. Moreover, a tracking error not only affects the corresponding cell lineage, but may also affect the tracking accuracies for a number of its neighbors in the tightly

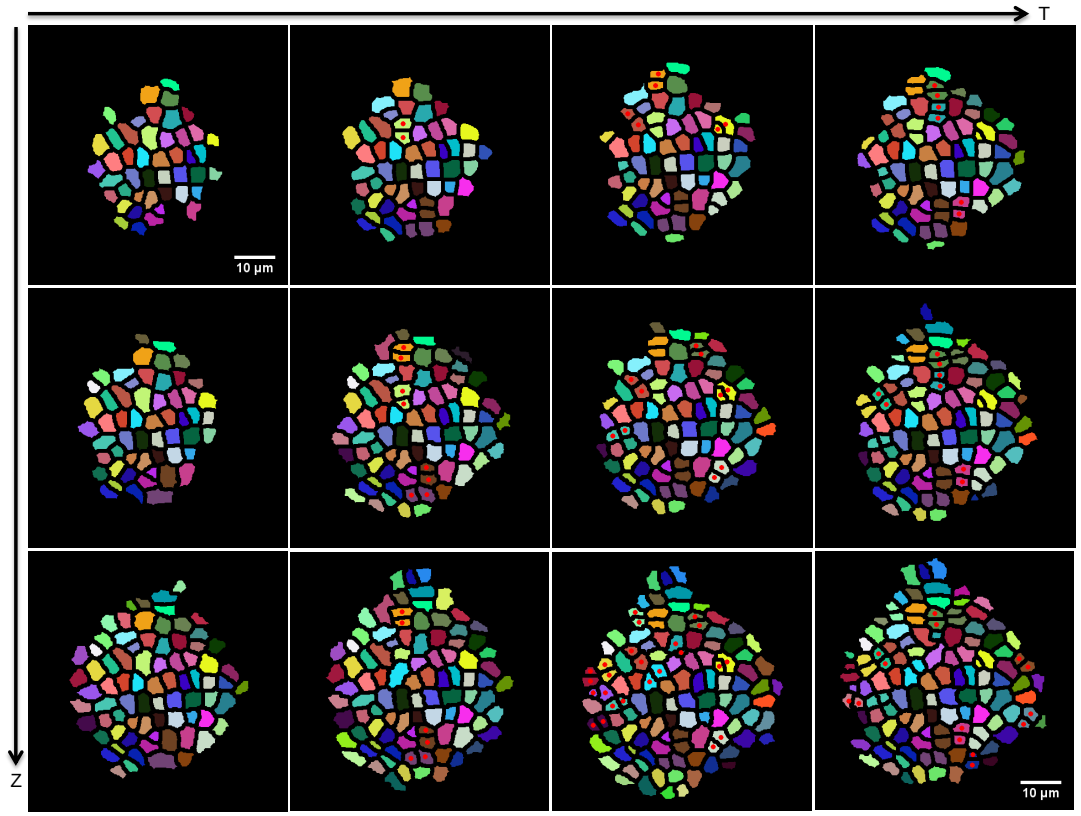


Figure 3.10: Results showing combined spatio-temporal tracking on Arabidopsis SAM dataset. A number of cells are tracked across four time points of observation ($12^{th}, 15^{th}, 18^{th}$ and 21^{st} hours). Three image slices are sampled from the 3D stack at each time point (at 3 μm, 4.5 μm and 6 μm respectively). Cell slices corresponding to the same cell across space and time are marked with the same color. Cell divisions are also detected and the children cells having the same colors as their parents are marked with red dots.

packed multilayer tissue.

Results on complete spatio-temporal tracking are shown in Fig. 3.10. We sample three consecutive spatial slices from confocal stacks at 4 different time points (at $12^{th}, 15^{th}, 18^{th}$ and 21^{st} hours of observation) and the tracking result for them are shown using color coding. It can be observed that slices of new cells appear as we go deeper into the tissue and as expected, they are not matched to any cell from the slice above.

3.5.4 Learning the Model Parameters

In this section, we describe how are the different parameters associated with the pairwise cell tracking model set. We use manually ground-truthed correspondences in a subset of the data as our training set to learn the best set of values for different parameters and the same set of learned parameter values are used for all the experimental results shown in this chapter. The parameters used in the cell division detection method (see Sec. 3.3.3) are learned independently from the CRF parameters. For each of the parameters, we first choose a range for the parameter value and then uniformly sample a number of values from within that range. Now, we generate combined sets of parameter values using every possible combination. Finally, on the training dataset, the best set of parameters out of all such candidates is selected using a 5-fold cross validation. The aforementioned ranges of different parameter values are fixed by observing the variation in cell division detection and tracking errors when each of these individual parameters are changed and then obtaining an optimal range for each parameter.

Fig. 3.11 shows the variations of cell division detection error with the parameters t_1 and t_2 (Eqn. 3.10) on a training image set. In Fig. 3.11(A) the cell division area parameter (t_2 in Eqn. 3.10) is varied when the shape parameter t_1 is kept constant at

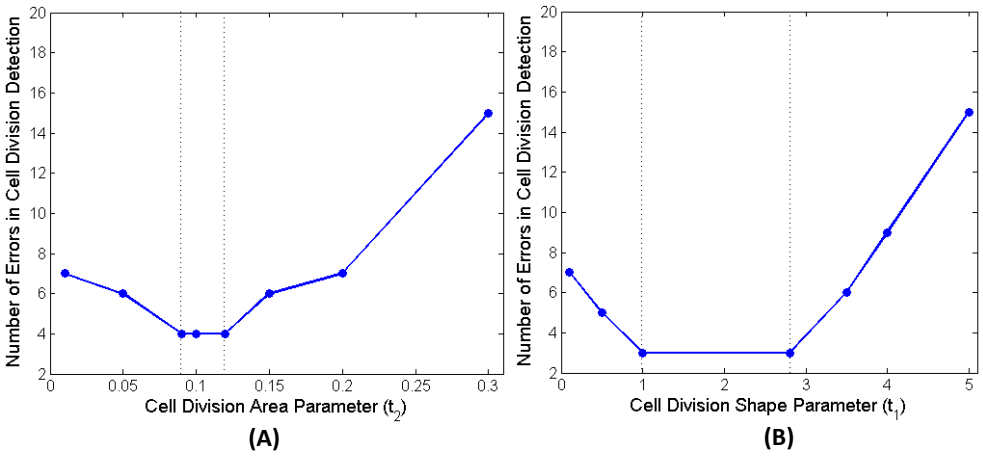


Figure 3.11: Variation of cell division detection error with the parameters t_1 and t_2 (Eqn. 3.10) on a training image set. (A) The cell division area parameter (t_2 in Eqn. 3.10) is varied with the shape parameter t_1 is kept constant at a large value (10^6). The optimal region corresponding to the lowest cell division detection error is within the two vertical lines. (B) With the parameter t_2 being fixed at a large value 10^6, the shape parameter t_1 (see Eqn. 3.10) is varied independently over a range. As before, the optimal range is within the two vertical lines on the figure.

a large value (10^6) to minimize the effect of the cell shape distance measure over cell division detection accuracy. There are 8 legitimate cell division events in this training dataset. The optimal region corresponding to the lowest cell division detection error is within the two vertical lines shown in the figure and any value within this region would yield the lowest cell division detection error. If the parameter t_2 is decreased below this optimal range, the number of detected cell divisions decreases. This leads to increase in false negatives. For choice of parameters greater than those in the optimal region, false positive cell division detection error increases. Similarly, with the parameter t_2 being fixed at a large value 10^6 (to minimize the effect of the cell area based distance measure), the shape parameter t_1 (see Eqn. 3.10) is varied independently over a range and the dynamics of cell division detection errors is observed in Fig. 3.11(B). As before, the optimal range is within the two vertical lines on the figure. The variation of cell division detection error with t_1 shows a very similar trend to that of t_2, as with decrease

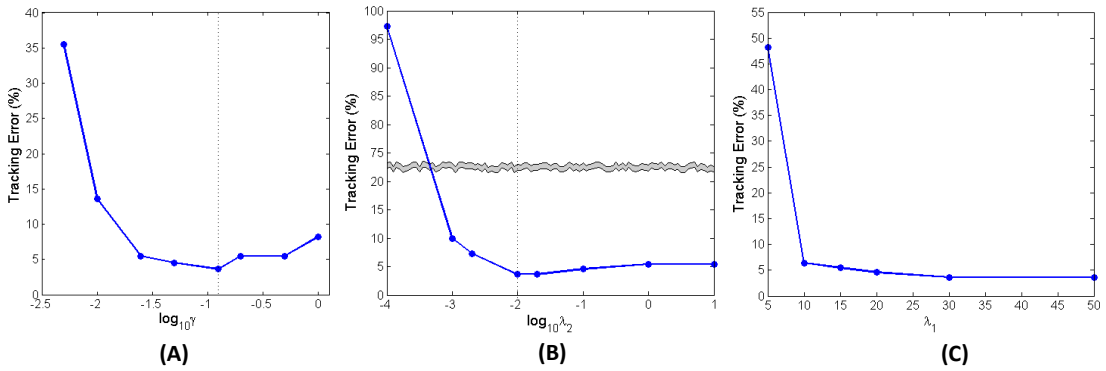


Figure 3.12: Variation of tracking errors with edge and node potential parameters on a training dataset. (A) Tracking error rate is plotted against $log_{10}\gamma$ (edge potential) and an optimal choice of γ can be $\frac{1}{8}$. (B) Variation of tracking error with node potential parameter λ_2 is shown. The optimal range of λ_2 can be chosen as [0.01 10]. (C) Variation of tracking error with node potential parameter λ_1 is observed when λ_2 is fixed at 0.01.

in t_1 below the optimal region, the number of false negatives increases and with increase in t_1 above the optimal values, more and more false cell division events are detected, thereby rapidly increasing the cell division detection error.

Fig. 3.12(A) shows the variation in cell tracking error in a training image stack with change in the edge potential parameter γ. As γ is varied over a very large range, the tracking error is plotted against $log_{10}\gamma$. As γ is decreased rapidly (in the range of 0.01 to 0.001), edge potential values get closer to 1 uniformly for both true and false associations (Eqn. 3.20), as well as the edge potential value for the no-match case becomes close to 0 (Eqn. 3.23). This results in a large increase in false positives (both ID switches and forced matching different cells). For large values of γ, the potential function value for the no-match case can get close to 1, even when there are legal matches available and this results in an increase in false negatives. Between these two extremes, the optimal range of values for γ can be found (around 0.1, as seen in Fig. 3.12(A)).

Variation in tracking error with change in the node potential parameters is shown in Fig. 3.12(B-C). λ_2 is varied over a wide range and the variation in tracking

error is observed in Fig. 3.12(B). For a very low value(in the range of 10^{-3} to 10^{-4}) of λ_2, the node potential for the no-match state can be as high as 1 and resultingly, most of the cells will not have any association to the target image slice. This will rapidly increase the number of false negatives. Again, with a large value of λ_2, the number of false positives will be large. The optimal range of λ_2, therefore, lies within these extreme values (such as 0.01 to 10 in Fig. 3.12(B)). If λ_2 is fixed at an optimal value (0.01), then the variation of tracking error with λ_1 can be observed in Fig. 3.12(C). Note that, now the node potential weight w (Eqn. 3.16) is chosen as 0.5. The optimal range for λ_1 can be chosen in the range of ten (say, 10 to 40).

Once the optimal ranges for the parameters are estimated, the best parameter values can be obtained via cross validation over multiple sets of parameters sampled from these optimal ranges.

3.5.5 Discussion on the Limitations of the Proposed Method

The accuracy of the proposed cell tracking method depends on spatio-temporal registration of the image slices in the 4D stack. The parameters of the tracker can be tuned in order to account for some error in registration but for major transformations across image slices, the tracker would not work satisfactorily. As we have shown in our experiments, the tracker can handle moderate deformations of the growing cells. However, if the deformation changes both the nominal shape of individual cells as well as the topology of their local neighborhood, it becomes more challenging and in some cases leads to failure of the tracker. Likewise, this present tracking algorithm is not designed to handle large displacements or motions of individual cells, but it can still provide good tracking accuracy as long as the local neighborhood structure around a cell is not jeopardized.

3.6 Conclusion

In this chapter, we have presented a method for automatically tracking individual cells in spatio-temporal image stacks of closely packed developing multilayer tissues. We observed that cells in a close cluster in the tissue can have very similar image features and hence we leveraged upon the local spatial geometric structure and topology of the relative positions of the neighboring cells to generate robust similarity measures between 2D cell segments in different spatio-temporal image slices in presence of imaging noise. We have also shown how to detect cell divisions prior to temporal tracking in order to find out the proper terminating point of individual cell lineages. Finally, we have provided a strategy to optimally and feasibly combine the spatial and temporal similarity measures between cell slices via the previously introduced NCDA method to generate a complete 4D spatio-temporal tracking result. Experiments were conducted on confocal live-imaging stacks Arabidopsis SAM and the results indicate the high tracking accuracy obtained through the method presented. We also show how errors in tracking can be corrected by establishing network consistency using the proposed NCDA.

Chapter 4

Adaptive Geometric Tessellation for Cell-resolution 3D Reconstruction

4.1 Introduction

For quantitative understanding of morphogenesis in a plant or animal tissue, statistics on rates and patterns of cellular deformation could play a critical role. As explained in the previous chapter, novel cell resolution imaging techniques and image analysis tools are being developed towards this goal. Live cell imaging is a class of microscopy, where the same living cells are observed and imaged at regular time intervals over several hours to monitor their motion or displacement and to visualize cell growth and division dynamics. A typical image processing and analysis pipeline for high throughput, quantitative analysis of these large volumes of data consists of a number of fundamental components. We have presented one such image analysis component, viz.,

the spatio-temporal cell tracker, in Chapter 3. In the present chapter, we shall describe a cell resolution 3D reconstruction method for densely packed multilayer tissues that exploits the cellular lineages estimated through the tracking module. Estimation of cell shape and volumes as a function of time is most fundamental to understanding of the growth process. Due to the large quantity of data collected during the growth of a tissue, automated computational methods for robust estimation of 3D cell structures and cell volumes are absolutely necessary in order to obtain statistically significant results of these growth parameters.

Inspite of the extreme usefulness of CLSM based live cell imaging for analysing such tissue structures, there are number of technical challenges associated with this imaging technique that make the problem of cell shape estimation non-trivial. Some of these challenges were discussed in the last chapter while presenting the cell tracking framework. To keep the cells alive and growing, we have to limit the laser radiation exposure to the specimen, i.e., if dense samples in one time point are collected, it is highly unlikely that we will be able to get time lapse images as the specimen will not continue to grow in time due to high radiation exposure. Therefore the number of slices in which a cell is imaged is often very low (2-4 slices per cell). Again, the fluorescent signal fades as we image the deeper layers of the tissue, thereby bringing in the problem of very low SNR in parts of the confocal image stack.

Please note that in some cases, a two-photon excitation microscopy or light sheet microscopy can be better choices for live-cell imaging for more efficient light detection and less photo-bleaching effect. But, a large number of data sets exist that are imaged using CLSM or exhibit the characteristic of our data and our method can be useful in analyzing them. We have found that two photon excitation is toxic to SAM cells than the single photon CLSM and since the SAM is surrounded by several devel-

oping flower buds, the side ward excitation may not be possible. Also, by designing an image-analysis method that is capable of handling the worse quality data, we can ensure that same or better accuracy can be achieved on a data-set having superior image quality and resolution. Thus, from an image analysis perspective, we are looking at a very challenging problem where we want to obtain a 3D surface reconstruction of arbitrary cell shapes from a set of very sparsely sampled data points in presence of unavoidable imaging noise. Also, the reconstruction pipeline must be fully automated. In most cases, manual analysis (which has been the trend) is usually extremely tedious and, often, only provides qualitative trends in the data rather than precise quantitative models.

4.1.1 Contributions of the Present Work:

In this study, we have looked at the problem of 3D reconstruction of a tightly packed multi-layer tissue from its Z-sparse confocal image slices. As a special example, in this paper we have proposed a novel, fully automated cell resolution 3D reconstruction framework for Shoot Apical Meristem (SAM) of Arabidopsis Thaliana. SAM, also referred to as the stem cell niche, is a very important part of a plant body plan because it supplies cells for all the above ground plant parts such as leaves, branches and stem. A typical Arabidopsis SAM is a densely packed multi layered cell cluster consisting of about five hundred cells where the cell layers are clonally distinct from one another. The tight tessellation of cells in SAM enabled us to estimate the 3D structure of individual cells using the slice information of the cell as well as that of its nearest neighbors. The 3D estimation is based on prior geometrical tessellation models, the parameters for which are estimated from the sparse image data to hand and then this model is used to partition the 3D SAM structure into individual cellular regions.

Being motivated by the methods in [67, 38], we first assume a 'Voronoi' tessellation model based on Euclidean distance metric to segment/reconstruct the 3D cell shapes. Through the results obtained on 3D sparse confocal stacks of SAM images we show that this model yields a good approximation of cell shapes where the shapes and sizes of the cells are uniform along all three axes of the cells and the major axes of growth of the neighbouring cells are isotropic. But, in practice, this is not always the case and the cells can have very anisotropic shape and growth, even in a close neighborhood. In such cases, the Voronoi tessellation using the Euclidean distance fails to generate accurate enough cell walls.

We, therefore, propose an anisotropic Voronoi tessellation defined on a quadratic distance metric to capture the growth anisotropy of individual cells along all of their axes. We show that the parameters of this metric can be estimated from the sparse set of confocal image slices of the individual cells and the tessellation based on this metric can provide very accurate 3D cell shapes as quadratic surfaces even in the case of non-uniform cell shapes, sizes or growth along different cell axes. The tessellation, named as the 'Adaptive Quadratic Voronoi Tessellation' (AQVT) and the 3D reconstruction technique based on it, presented in this chapter, provide accurate enough 3D reconstructed cell shapes and sizes in the SAM as validated through the experiments in Sec. 4.4.2. We also show that the proposed anisotropic Voronoi tessellation (AQVT) can also be applied on tissues, where the cell shapes in the tissue follow a standard Euclidean distance based Voronoi tessellation.

4.2 Overview of the Proposed Method

As mentioned earlier, the input to the cell resolution 3D reconstruction pipeline is a set of segmented and spatially associated 2D cell slices from the confocal image stack. The details on CLSM imaging technique and the resulting image data structure are described in Sec. 3.5.1. In principle, any segmentation and tracking method can be used for preprocessing the data, as the reconstruction method is independent of the preprocessing steps we choose to employ. For segmentation we have preferred an adaptive Watershed method [69] as it produces very accurate and realistic cell boundaries for the SAM confocal data, even in presence of imaging noise. Please note that the contribution of our method lies in the post segmentation and tracking stage though a better segmented data is guaranteed to improve the performance of both the subsequent tracking and 3D reconstruction modules. In order to find a cell's correspondence across multiple z-slices, we have used our context-aware cell tracker [15] presented in the last chapter.

Once the sparse image slices are segmented and tracked to generate the initial clustering of individual cell slices, the objective is to obtain full 3D reconstructions of these cells. As explained above, for live imaging with frequent observations in time, the number of slices in which a particular cell can be present is very small (e.g. 2-4 slices/cell). Unfortunately, the existing 3-D reconstruction methods are not capable of handling such sparsity in data. Motivated by the physical structures of the cells, we handle this issue by assuming a prior geometrical 3D tessellation model for the tissue, the parameters of which is estimated to fit the given sparse set of segmented cross sectional images.

In [67, 38], the authors have used a standard Voronoi tessellation technique

to estimate the cell boundaries from the known information of the nucleus location for individual cells. Motivated by their work, we first show how a Voronoi tessellation can be fitted to our CLSM dataset which only have the partial cell wall information (Sec. 4.3.1.1, Sec. 4.3.1.2). Unlike the dataset used in [67,38], we do not have the cell nuclei marked in our dataset. The standard affine Voronoi tessellation is not always accurate enough to reconstruct the cells in SAM as different cells in the tissue can have very diverse sizes as well as the neighbouring cells might not have isotropic growth directions. This motivates us to propose an anisotropic Voronoi tessellation model (Sec. 4.3.2) for the tissue, which is also a generalization of the standard Affine Voronoi tessellation. We show how to estimate the parameters (Sec. 4.3.2.1) of an anisotropic or quadratic distance function for individual cells from the sparse data-points on the boundaries of these cells. The proposed quadratic Voronoi tessellation approach, termed as the 'Adaptive Quadratic Voronoi Tessellation' (AQVT) is then used to cluster a dense point cloud obtained from within the estimated 3D surface of the SAM and thereby, to generate the final 3D shapes of each individual cell (Sec. 4.3.2.2).

4.3 Detailed Methods: The 3D Reconstruction Framework

4.3.1 Standard Voronoi Tessellation Based 3D Reconstruction

In this section, we introduce the standard Voronoi tessellation and briefly describe its properties. Then we present how the 3D cell structures can be reconstructed as Euclidean distance based Voronoi regions.

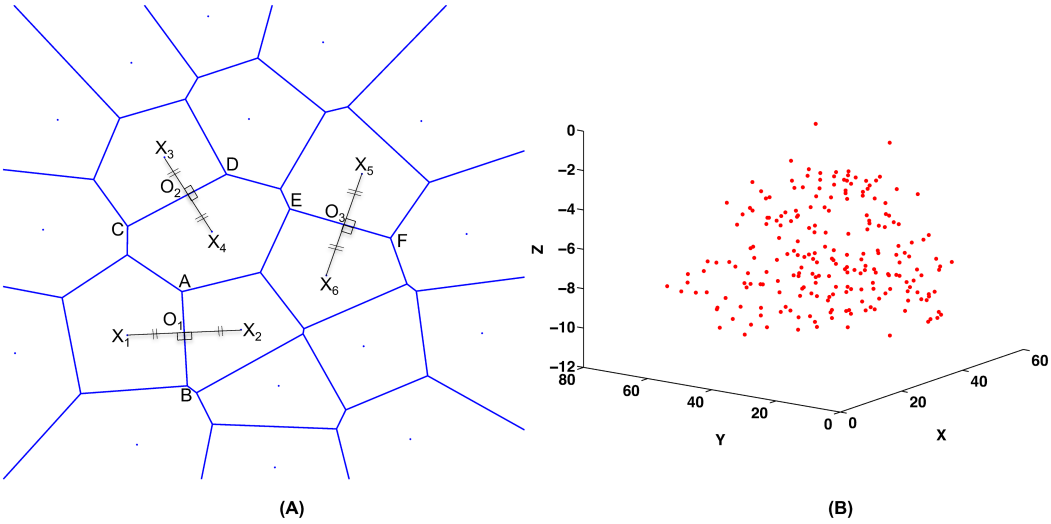


Figure 4.1: A schematic of Voronoi Tessellation and Estimated SAM cell centroids as Voronoi sites. (A) A Voronoi diagram based on the Euclidean distance metric for twenty one sites in 2D. The figures also show that the Voronoi edges are perpendicular to the line joining any two neighbouring sites. $\overline{\text{AB}}$, $\overline{\text{CD}}$, $\overline{\text{EF}}$ are three of the Voronoi edges and they are the perpendicular bisectors of $\overline{\text{X}_1\text{X}_2}$, $\overline{\text{X}_3\text{X}_4}$, $\overline{\text{X}_5\text{X}_6}$ respectively. (B) Centroids are estimated for around two hundred cells in a SAM tissue, which are also the sites of Euclidean distance based Voronoi tessellation.

4.3.1.1 A Brief Overview of Voronoi Tessellation and its Properties

'Voronoi Diagram' is a geometric minimization diagram that splits its embedding space into different non-overlapping regions. Each of these regions is characterized by a generating point or an object also known as the 'site'. All other points in each of these regions are closer to the site in its region than to any other site in the entire embedding space. The closeness of the points to the sites is computed using a distance metric. There can be different types of sites ranging from a point, a line to any complex geometric shape. Depending on the type of sites, the distance metric or the embedding space, several different variations of the Voronoi diagram can be defined. Detailed discussions on many of such variants can be found in [75, 3, 11].

Based on the characteristic of site and distance function, the locus of the points equidistant from two neighboring sites (also termed as the 'bisectors' or 'edges') can be

hyperplanes or higher order hypersurfaces. We call the Voronoi diagrams with hyperplane bisectors as the 'Affine Voronoi' diagrams. The most common example of such an affine diagram is the Voronoi diagram of points based on the Euclidean distance metric.

Let there be n point sites in a space $\mathbb{R}^d$, and the set of all sites $\mathcal{S}$ be $\{\mathbf{s}_1,\ \mathbf{s}_2,\ \cdots \mathbf{s}_n\}$. The Voronoi regions associated with these sites are represented as $V(\mathbf{s}_1),\ V(\mathbf{s}_2),\ \cdots\ V(\mathbf{s}_n)$.

$$V(\mathbf{s}_i)\ =\ \bigcap_{\mathbf{s}_j \in \mathcal{S}, j \neq i} H(\mathbf{s}_i,\ \mathbf{s}_j) \tag{4.1}$$

where $H(\mathbf{s}_i,\ \mathbf{s}_j)$ is the half-plane defined by,

$$H(\mathbf{s}_i,\ \mathbf{s}_j)\ =\ \Big\{\mathbf{x} \in \mathbb{R}^d \mid d(\mathbf{x}, \mathbf{s}_i)\ <\ d(\mathbf{x}, \mathbf{s}_j)\Big\} \tag{4.2}$$

The distance $d(\mathbf{x}, \mathbf{s}_i)$ for a standard Voronoi diagram is the Euclidean distance defined by

$$d(\mathbf{x}, \mathbf{s}_i)\ =\ \|\mathbf{x}\ -\ \mathbf{s}_i\|_2$$

Again, the set of all points on the bisector between two Voronoi regions $V(\mathbf{s}_i)$ and $V(\mathbf{s}_j)$ is given by,

$$B(\mathbf{s}_i,\ \mathbf{s}_j)\ =\ \Big\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x} - \mathbf{s}_i\|_2 = \|\mathbf{x} - \mathbf{s}_j\|_2\Big\} \tag{4.3}$$

Some of the properties of the Euclidean distance based Voronoi diagram with point sites are as follows:

(1) The line joining any two Voronoi sites $\mathbf{s}_i$ and $\mathbf{s}_j$ is always perpendicular to the Voronoi bisector/edge $B(\mathbf{s}_i,\ \mathbf{s}_j)$,

(2) The perpendicular distances from $\mathbf{s}_i$ and $\mathbf{s}_j$ to $B(\mathbf{s}_i,\ \mathbf{s}_j)$ are equal to one another,

(3) Any point $\mathbf{x} \in V(\mathbf{s}_i)$ would satisfy

$$\|\mathbf{x} - \mathbf{s}_i\|_2\ \leq\ \|\mathbf{x} - \mathbf{s}_j\|_2,\ \forall j \in \{1, 2, \cdots\ n\}$$

(4) The Voronoi regions, thus produced, are convex polyhedrons and can be expressed as intersection of a finite number of open or closed half spaces.

In Fig. 4.1(A), we have shown some of the properties for a 2D Voronoi tessellation with twenty one sites (X_1, X_2, X_3 etc.). It can be observed that each of these Voronoi regions is a convex polygon.

4.3.1.2 Voronoi Sites Estimation From Sparse Data

After segmentation and tracking, we are given a sparse set of 3D data-points that lie on the boundary between the neighbouring cells. Our objective is to fit a Voronoi tessellation to these data-points to obtain complete structures of individual cells, represented as Voronoi polyhedrons in 3D. Therefore, ideally the sparse data-points should lie on the bisectors between the neighbouring Voronoi regions. Given a very sparse set of data-points on the bisectors as in the case of 'Live cell imaging', the only way to reconstruct the Voronoi diagram is to first estimate the approximate locations of Voronoi sites, from which the Voronoi edges/bisectors can then be computed.

Given a set of generating sites, the construction of the Voronoi diagram can be done through several methods, the most popular of which being Fortune's 'sweep line' algorithm. The inverse problem, i.e. to obtain the locations of the sites given the Voronoi bisector, is, however, less studied in the literature. In [30], Evans et al. proposed a linear least-square technique to estimate the Voronoi sites by fitting a Voronoi diagram over a given tessellation pattern. Again, in [34], a number of algorithms have been proposed to obtain the Voronoi sites given the vertices of the Voronoi polygons. But neither of these methods is applicable to the sparse data that we have. These methods require the knowledge of the complete structures of the Voronoi polyhedrons, which is precisely the output that we are after. In fact, because of the extreme sparsity in our dataset, it is

rather impossible to obtain unique estimates of the Voronoi sites for individual 3D cells.

In this situation, the knowledge about the physical structure of the SAM stem cells helps us in obtaining approximate locations of the Voronoi sites, where each cell is represented as a Voronoi region. In [38], the authors observed that the SAM cells can be represented as Voronoi regions where the sites are located at the approximate centers of the cell nuclei. This observation, along with the fact that the nucleus, located in the central region of the cell, contributes to the majority of the size of a SAM cell motivates us to devise a simple strategy to estimate the approximate site locations, even when the nuclei are not imaged in the confocal image stack.

Given that the sparse set of points on the segmented and tracked slices of a cell c are $P^{(c)}_{sparse}$ (the 3D data-point set $\{(x^{(c)}_{sparse},\ y^{(c)}_{sparse},\ z^{(c)}_{sparse})\}$), the approximate centroid location of the cell would be $\hat{\mathbf{s}}_c\ =\ [\bar{x}^{(c)}_{sparse},\ \bar{y}^{(c)}_{sparse},\ \bar{z}^{(c)}_{sparse}]$, where the elements are the arithmetic means of $\{x^{(c)}_{sparse}\},\ \{y^{(c)}_{sparse}\}$ and $\{z^{(c)}_{sparse}\}$ respectively. Thus, $\hat{\mathbf{s}}_c$ is also the estimated approximate location of the site corresponding to the Voronoi region representing the cell c. The centroids for approximately 200 cells from a SAM tissue is shown in Fig. 4.1(B).

Now, the next step would be to generate a dense point cloud sampled from within the SAM structure and to cluster this dense set of 3D data-points into different Voronoi regions (representing individual cells) based on the estimated locations of the sites $\hat{\mathbf{s}}_c,\ c = 1, 2, \cdots C$.

4.3.1.3 Generation of Dense Point Cloud to be Partitioned Into Cells: Global Shape of SAM

At this stage, we estimate the 3D structure of the SAM by fitting a smooth surface to its segmented contours. The surface fitting is done in two steps. In step

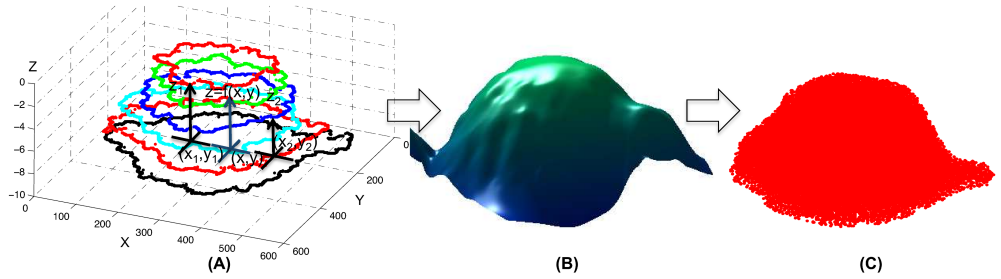


Figure 4.2: Generation of the dense point cloud from within the reconstructed SAM surface. (A) The SAM contours extracted from the confocal image stack using Level-Set segmentation, (B) The SAM surface is reconstructed using linear interpolation on a local neighbourhood of points on the SAM contours, (C) A very dense point cloud is extracted from within the reconstructed SAM surface which is clustered using the proposed reconstruction technique into individual cells.

one, the SAM boundary in every image slice is extracted using the 'Level Set' method (Fig. 4.2(A)). A level set is a collection of points over which a function takes on a constant value. We initialize a level set at the boundary of the image slice for each SAM cross section, which behaves like an active contour and gradually shrinks towards the boundary of the SAM. Let the set of points on the segmented SAM contours be P^{SAM} $(\{x^{SAM}, y^{SAM}, z^{SAM}\})$.

In the second step, we fit a surface on the segmented points P^{SAM}. Assuming that the surface can be represented in the form $z = f(x, y)$ (where the function f is unknown), our objective is to predict z at every point (x, y) on a densely sampled rectangular grid of points bounded by $\left[x_{min}^{SAM}, y_{min}^{SAM}, x_{max}^{SAM}, y_{max}^{SAM}\right]$. As the segmented set of data points are extremely sparse, this prediction is done using a linear interpolation on a local set of points on the grid around the point (x, y). As the value (z) for the point (x, y) is approximated by a linear combination of the values at a few neighboring points on the grid, the interpolation problem can be posed as a linear least-square estimation problem. We also impose a smoothness constraint in this estimation by forcing the first partial derivatives of the surface evaluated at neighboring points to be as close as

possible. A MATLAB visualization [23] of the surface is shown in Fig. 4.2(B).

Once the SAM surface (S^{SAM}) is constructed, we uniformly sample a dense set of 3D points (P_{dense}, a visualization can be found in Fig. 4.2(C)) such that every point in P_{dense} must lie inside S^{SAM}. Thus, $P_{dense} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_N\}$ and the required output from the proposed algorithm is a clustering of these dense data points into C cells/clusters such that $P_{dense} = \{\hat{P}^{(1)}_{dense}, \hat{P}^{(2)}_{dense}, \cdots \hat{P}^{(C)}_{dense}\}$ starting from the sparse set of segmented and tracked points $P_{sparse} = \{P^{(1)}_{sparse}, P^{(2)}_{sparse}, \cdots P^{(C)}_{sparse}\}$ obtained from the confocal slice images of individual cells.

4.3.1.4 Segmentation of the Dense Point Cloud Into Voronoi Cells

In this step, we partition the dense point cloud P_{dense}, obtained in the last step, into C clusters based on the site locations $\hat{\mathcal{S}} = \{\hat{\mathbf{s}}_c\},\ c = 1, 2, \cdots C$, estimated in Sec. 4.3.1.2.

The c^{th} cell can be represented as a collection of dense data-points belonging to the c^{th} Voronoi region as

$$\hat{P}^{(c)}_{dense} = \{\mathbf{x} \in P_{dense} \mid \|\mathbf{x} - \hat{\mathbf{s}}_c\|_2 \leq \|\mathbf{x} - \hat{\mathbf{s}}_k\|_2\ \forall k \in \{1, 2, \cdots C\}\} \tag{4.4}$$

Once the dense point cloud belonging to each Voronoi region is obtained, we can construct convex polyhedrons with each of these dense point clusters ($\hat{P}^{(1)}_{dense}, \cdots \hat{P}^{(C)}_{dense}$) to obtain the cell resolution 3D reconstruction of SAM.

4.3.2 An Adaptive Quadratic Voronoi Tessellation (AQVT) for Non-uniform Cell Sizes and Cell Growth Anisotropy

In a tissue like SAM, cells do not grow uniformly along all three axes (X, Y, Z). In fact, most of the cells show a specific direction of growth. Again, neighboring cells

in SAM, especially in the central region (CZ), are not likely to grow along the same direction. Thus, even if a tessellation is initially an affine Voronoi diagram, it is not likely to remain so after a few stages of growth. Such cases of non-uniform cell sizes and anisotropic growth can be captured in a more generalized non-affine Voronoi tessellation called the 'Anisotropic Voronoi Diagrams'. In the most general form of such diagram for point sites, the distance metric has a quadratic form with an additive weight [11].

Following similar notations used in previous sections, for a set of anisotropic sites $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots \mathbf{s}_n\}$ in $\mathbb{R}^d$, the anisotropic Voronoi region for a site $\mathbf{s}_i$ is given by,

$$V_A(\mathbf{s}_i) = \Big\{\mathbf{x} \in \mathbb{R}^d \mid d_A(\mathbf{x}, \mathbf{s}_i) \leq d_A(\mathbf{x}, \mathbf{s}_j) \; \forall j \in \{1, 2, \cdots n\}\Big\} \tag{4.5}$$

where

$$d_A(\mathbf{x}, \mathbf{s}_i) = (\mathbf{x} - \mathbf{s}_i)^T \Sigma_i (\mathbf{x} - \mathbf{s}_i) - \omega_i \tag{4.6}$$

Σ_i is a d x d positive definite symmetric matrix associated with the site $\mathbf{s}_i$ and $\omega_i \in \mathbb{R}$. Thus each of the anisotropic Voronoi regions is parameterized by the triplet $(\mathbf{s}_i, \Sigma_i, \omega_i)$. Further assuming $\omega_i = \omega_j \; \forall i, j \in \{1, 2, \cdots n\}$, the distance function becomes

$$d_Q(\mathbf{x}, \mathbf{s}_i) = (\mathbf{x} - \mathbf{s}_i)^T \Sigma_i (\mathbf{x} - \mathbf{s}_i) \tag{4.7}$$

As the bisectors of such a Voronoi diagram are quadratic hypersurfaces, these diagrams are called 'Quadratic Voronoi Diagrams', wherein every Voronoi cell i is parameterized by $(\mathbf{s}_i, \Sigma_i)$ pairs.

$$V_Q(\mathbf{s}_i) = \Big\{\mathbf{x} \in \mathbb{R}^d \mid d_Q(\mathbf{x}, \mathbf{s}_i) \leq d_Q(\mathbf{x}, \mathbf{s}_j) \; \forall j \in \{1, 2, \cdots n\}\Big\} \tag{4.8}$$

From Eqn. 4.7, it can be observed that Σ_i is essentially a weighting factor that non-uniformly weights distances in every Voronoi regions along every dimension. When all the Voronoi regions are equally and uniformly weighted along every axis,

$\Sigma_i = \mathrm{I}_{dxd}\ \forall i = 1, 2, \cdots n$ and the resulting diagram for point sites becomes an Euclidean distance based Voronoi diagram.

4.3.2.1 Estimating the Distance Metric From Sparse Data: Minimum Volume Enclosing Ellipsoid

Now, the problem at hand is to estimate the parameter pair for each cell/quadratic Voronoi regions from the sparse data-points, as obtained from the segmented and tracked slices, that belongs to the boundary of each cell. Given the extreme sparsity of the data, there is no available method that would provide Σ_is for each region. We, in this work, propose an alternative way of estimating $(\mathbf{s}_i,\ \Sigma_i)$ pairs directly from the sparse data-points.

An ellipsoidal surface in 3D is given by the locus of the point $\mathbf{x}$ that satisfies

$$(\mathbf{x} - \mathbf{x}_c)^T M(\mathbf{x} - \mathbf{x}_c) = 1 \tag{4.9}$$

where $\mathbf{x}_c$ is the center of the ellipsoid and M is a positive definite symmetric matrix. For any point $\mathbf{x}$ inside the ellipsoid, $(\mathbf{x} - \mathbf{x}_c)^T M(\mathbf{x} - \mathbf{x}_c) < 1$ and for every $\mathbf{x}$ outside it, $(\mathbf{x} - \mathbf{x}_c)^T M(\mathbf{x} - \mathbf{x}_c) > 1$.

Now, $(\mathbf{x} - \mathbf{x}_c)^T M(\mathbf{x} - \mathbf{x}_c)$ is, in fact, the Mahalanobis distance of the point $\mathbf{x}$ from the center $\mathbf{x}_c$ of the ellipsoid. Therefore, if a point $\mathbf{y}$ is equidistant for the centers of two ellipsoids $(\mathbf{x}_{c_1}, M_1)$ and $(\mathbf{x}_{c_2}, M_2)$ in the Mahalanobis sense, then,

$$(\mathbf{y} - \mathbf{x}_{c_1})^T M_1(\mathbf{y} - \mathbf{x}_{c_1}) = (\mathbf{y} - \mathbf{x}_{c_2})^T M_2(\mathbf{y} - \mathbf{x}_{c_2}) \tag{4.10}$$

Now, Eqn. 4.10 gives the locus of the points $\mathbf{y}$, which is exactly the same as that of the points on the bisector two neighbouring quadratic Voronoi regions parameterized by $(\mathbf{x}_{c_1}, M_1)$ and $(\mathbf{x}_{c_2}, M_2)$. We already have a set of points (P_{sparse})which are sparsely but uniformly z-sampled from the boundaries between neighbouring cells. The tracking

algorithm provides us with the exact cell pairs on which every point from this set belongs to. Therefore we can approximately estimate the distance parameters associated with every quadratic Voronoi region individually by fitting an ellipsoid to the sparse data-points belonging to the boundaries of that region. In this work, we choose to fit Minimum Volume Enclosing Ellipsoids (MVEE) to each of $\{P^{(1)}_{sparse}, P^{(2)}_{sparse}, \cdots P^{(C)}_{sparse}\}$ for C cells individually and obtain approximate estimates of $\{(\mathbf{s}_1, \Sigma_1), (\mathbf{s}_2, \Sigma_2), \cdots (\mathbf{s}_C, \Sigma_C)\}$. The estimation strategy is described later in this section.

As we are estimating the parameters of quadratic distance metric associated with every individual Voronoi cell separately and then using the distance metrics, thus obtained, to tessellate a dense point cloud, we choose to call the resulting Voronoi tessellation as the 'Adaptive Quadratic Voronoi Tessellation' (AQVT).

After registration, segmentation and identification of a cell in multiple slices in the 3-D stack, we can obtain (x, y, z) co-ordinates of the set of points on the perimeter of the segmented cell slices. Let this set of points on the c^{th} cell be $P^{(c)}_{sparse} = \{p_1, p_2, \cdots p_k\} \in \mathbb{R}^3$. We have to estimate the minimum volume ellipsoid which encloses all these k points in $\mathbb{R}^3$ and we denote that with $\mathcal{E}$. An ellipsoid in its center form is represented by

$$\mathcal{E}(s, \Sigma) = \{p \in \mathbb{R}^3 \mid (p - s)^T \Sigma (p - s) \leq 1\} \tag{4.11}$$

where $s \in \mathbb{R}^3$ is the center of the ellipsoid $\mathcal{E}$ and $\Sigma \in \mathbb{R}^{3\times 3}$. Since all the points in $P^{(c)}_{sparse}$ must reside inside $\mathcal{E}$, we have

$$(p_i - s)^T \Sigma (p_i - s) \leq 1 \text{ for } i \ = \ 1, 2, \cdots k \tag{4.12}$$

and the volume of this ellipsoid is

$$Vol(\mathcal{E}) \ = \ \frac{4}{3}\pi \{det(\Sigma)\}^{-\frac{1}{2}} \tag{4.13}$$

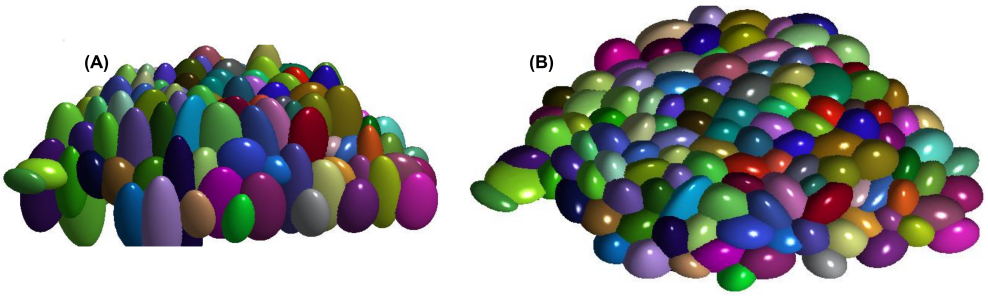


Figure 4.3: Ellipsoidal representation of the AQVT parameters estimated from the sparse data-points. (A) The Minimum Volume Enclosing Ellipsoids representing the (c, Σ) parameter pairs for individual cells are shown in different colors. (B) The same representation viewed from top.

Therefore, the problem of finding the Minimum Volume Enclosing Ellipsoid (MVEE) for the set of points $P^{(c)}_{sparse}$ can be posed as

$$\begin{aligned} \min_{\Sigma, s} \quad & - log\ det(\Sigma) \\ \text{s.t.} \quad & (p_i - s)^T \Sigma (p_i - s) \leq 1 \quad \text{for } i = 1, 2, \cdots k \\ & \Sigma \succ 0 \end{aligned} \tag{4.14}$$

To efficiently solve Problem 4.14 we convert the primal problem into its dual problem since the dual is easier to solve. A detailed analysis on the problem formulation and its solution can be found in [48, 53]. Solving this problem individually for each sparse point set $P^{(1)}_{sparse}, P^{(2)}_{sparse}, \cdots P^{(C)}_{sparse}$, the parameters of the quadratic distance metrics are estimated as $\{(\hat{\mathbf{s}}_1, \hat{\Sigma}_1), (\hat{\mathbf{s}}_2, \hat{\Sigma}_2), \cdots (\hat{\mathbf{s}}_C, \hat{\Sigma}_C)\}$. To visually represent these parameters, we have constructed the ellipsoids with each of these parameter pairs and color coded them to represent individual cells in a SAM tissue (Fig. 4.3).

4.3.2.2 3D Tessellation Based on the Estimated Parameters of AQVT: the Final Cell Shapes

As soon as the parameters of the quadratic distance metrics are estimated from the previous step, the dense point cloud P_{dense} obtained in Sec. 4.3.1.3 can be partitioned into different Voronoi regions based on Eqn. 4.8, i.e. the dense point cloud belonging to cell c is given as

$$\hat{P}^{(c)}_{dense} = \left\{ \mathbf{x} \in P_{dense} \mid (\mathbf{x} - \hat{\mathbf{s}}_c)^T \hat{\Sigma}_c (\mathbf{x} - \hat{\mathbf{s}}_c) \leq (\mathbf{x} - \hat{\mathbf{s}}_j)^T \hat{\Sigma}_j (\mathbf{x} - \hat{\mathbf{s}}_j) \ \forall j \in \{1, 2, \cdots C\} \right\} \tag{4.15}$$

For visualization purpose of the cell resolution 3D reconstruction results, we fit convex polyhedrons to $\hat{P}^{(1)}_{dense}, \hat{P}^{(2)}_{dense}, \cdots \hat{P}^{(C)}_{dense}$ to represent each cell.

4.4 Results and Discussion

We have tested the proposed 3D Reconstruction framework on a cluster of around two hundred and twenty cells spanning L1 and L2 layers of an Arabidopsis Shoot Apical Meristem. Once the 2D segmented cell slices are clustered together using the cell tracker, the sparse set of data points on each cell are extracted. These data points are then used to estimate the approximate location of the sites for Euclidean distance based Voronoi tessellation or in case of the proposed AQVT based method, the site and weight parameters for the quadratic distance metric for each cell.

4.4.1 3D Reconstruction Results

Fig. 4.4(A) shows a cell resolution reconstruction of the cell cluster in SAM using AQVT. Note that for 3D visualization purpose of the 3D structure only, we have represented each cell as a convex polyhedron fitted to the dense point cloud clustered to

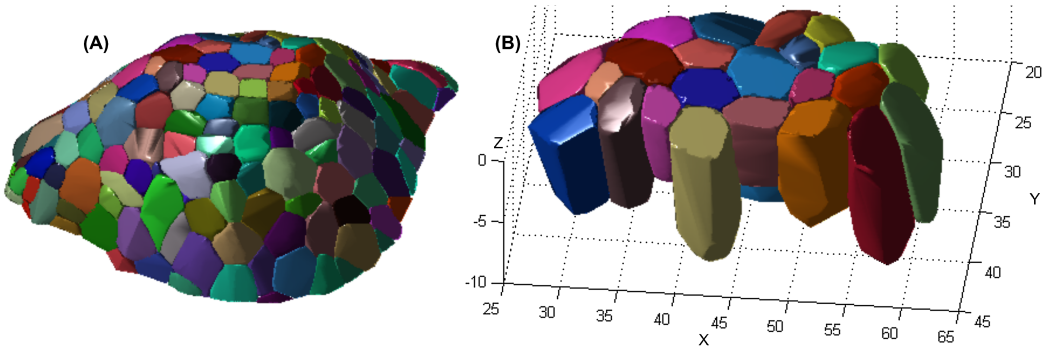


Figure 4.4: Visualization of the AQVT based 3D reconstruction of SAM cell cluster. (A) Visualization of the 3D reconstructed structure of a cluster of around 220 closely packed cells using convex polyhedron approximations of the densely clustered data-points for each cell, as obtained from the proposed 3D reconstruction scheme, (B) A subset of cells from the same tissue.

the cells, as obtained from our 3D reconstruction/ 3D segmentation scheme. For better understanding of the 3D structures of individual cells, we have shown the reconstructed shapes of a smaller cluster of cells in Fig. 4.4(B).

4.4.2 Validation of the Proposed Method

4.4.2.1 Validation on 3D SAM Data

There is hardly any biological experiment which can directly validate the estimated growth statistics for individual cells in a sparsely sampled multi layered cluster. In fact, the absence of a method to estimate growth statistics directly using non-computational methods in a live-imaging developmental biology framework is the motivation for the proposed work and we needed to design a method for computationally validating our 3D reconstruction technique. Once the 3D reconstruction is achieved, we can computationally re-slice the reconstructed shape along any arbitrary viewing plane by simply collecting the subset of reconstructed 3D point cloud that lies on the plane.

To show the validation of our proposed method, we have chosen a single time

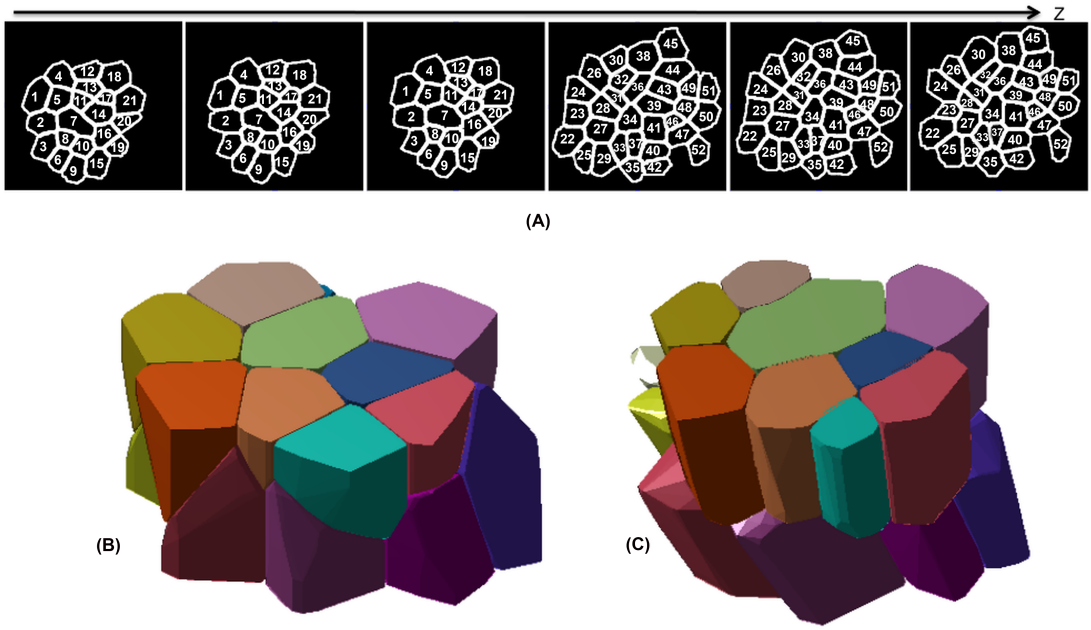


Figure 4.5: Reconstruction of a cluster of cells using Euclidean distance based Voronoi tessellation and the proposed AQVT for comparison of the 3D reconstruction accuracy. (A) Segmented and tracked cell slices for a cluster of fifty two cells from the L1 and L2 layers of SAM. A dense confocal image stack is subsampled at a z-resolution of 1.35 μm to mimic the 'z-sparsity' observed in a typical Live-Imaging scenario. The slices belonging to the same cell are marked with the same number to show the tracking results. (B) 3D reconstructed structure for a subset of these cells when reconstructed using the Euclidean distance based Voronoi Tessellation. (C) The AQVT based reconstruction result for the same subset of the cell cluster.

point dataset that is relatively densely sampled along Z (0.225 μm between successive slices). Then, we resampled this dense stack at a resolution of 1.35 μm to generate a sparser subset of slices that mimic the sparsity generally encountered in a live-imaging scenario. After 2D segmentation and tracking, different slices of the same cells imaged at different depths in Z are shown using the same number in Fig. 4.5(A). Next, we reconstructed the cell cluster first by the standard Voronoi tessellation using the Euclidean distance metric and then using our proposed method (AQVT) with a quadratic distance metric adapted for each of these cells. The reconstruction results for a subset of the cells for each of these methods are shown in Figs. 4.5(B) and 4.5(C) respectively for a direct comparison. It can be observed that not only our proposed method very accurately reconstructed the cell shapes but also it has captured the multi-layer architecture of these SAM cells more closely in comparison to its Voronoi counterpart with the Euclidean distance metric.

To better investigate the accuracy of the reconstruction and to show the clear advantage of using the proposed method of reconstruction quantitatively, we measure distances between the 2D cross sections of the 3D reconstruction results for each cell to the Watershed segmented cell boundaries. By choosing the reslicing plane as $z = 1.125, 2.475, 3.825\mu$m (different slices than those used for reconstruction) from top of the stack for the L1 layer cells and $z = 5.175, 6.525, 7.875\mu$m for the L2 layer cells, we can computationally regenerate the cell walls for these imaging planes along Z. The shapes of cells in the reconstructed slices can be compared against their counterparts in the 2D segmented images and the distance between the shapes would represent the reconstruction error. There are several different distance metrics that can be used to compute the dissimilarity between two shapes such as the Procrustes distance, Hausdorff

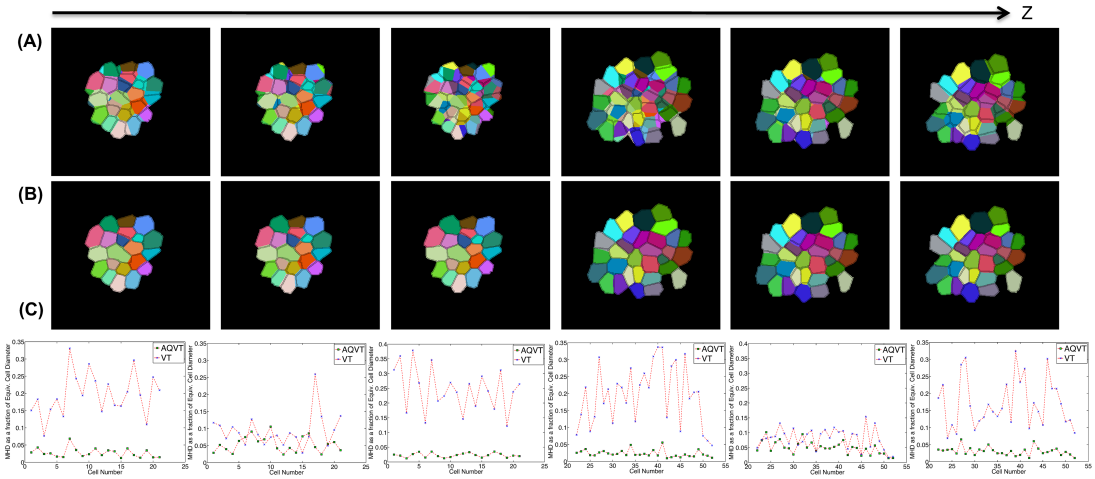


Figure 4.6: Comparison of the 3D reconstruction accuracy for the proposed AQVT based reconstruction against Euclidean distance based Voronoi tessellation. (A) The cells shown in Fig. 4.5(A) are reconstructed using the Euclidean distance based Voronoi tessellation and the computationally re-sliced cells are compared against the ground truth. (B) The same cells are reconstructed using the adaptive quadratic distance based Voronoi tessellation and then computationally re-sliced along various depths in z at which we also have the ground truth (in terms of the 2D segmentation results of the cell slices), but were not used in generating the reconstruction results. The computationally obtained cell slices are shown in different colors for different cells and they are superimposed by the ground truth segmentation results. (C) The error in reconstruction (similar to the reprojection error) is computed as the Modified Hausdorff Distance (MHD) between the computationally generated cell slices and the segmentation results on the ground truth images of the same cells. The MHD, computed for each of the 52 cells at different depths in the Z-stack are plotted for both the methods to compare the methods against each other. It can be clearly observed from the plots that the reconstruction error is much larger for the Euclidean distance based Voronoi tessellation (VT) than for AQVT, especially at the terminal ($3^{rd}, 4^{th}$and 6^{th}) slices, between consecutive layers of cells.

distance etc., each one having its own advantages. We have chosen to use the Modified Hausdorff Distance (MHD), one of the more popular distance measures in the Hausdorff distance family, to evaluate our reconstruction method. The advantages of MHD over other distance measures for object shape matching is described in details in [26]. It can be noted that the error, thus computed, is analogous to the reprojection error that is widely used in the 3D reconstruction community to quantify the accuracy of reconstruction.

In Fig. 4.6(A), we have shown the computationally resliced cell slices (color coded to represent the same cells at multiple slices) at various depths for Euclidean distance based Voronoi tessellation and Fig. 4.6(B) shows the 2D cross sections for the same cells as obtained by reslicing the 3D cell shapes in the proposed AQVT based reconstruction method. For both the image sets, the computationally obtained cross sections of each cell are superimposed by the Watershed segmentation results for the same cell slices (the ground truth). The MHD between the original and computationally generated cell slices are computed and for each of the cells in 6 different slices, this error is shown in Fig. 4.6(C).

It can be observed that each of the graphs in Fig. 4.6(C) comprises of the reconstruction errors for individual cell slices for both the methods of the reconstruction. For the slices closer to the center of the cells, both the methods show similar and acceptably small reprojection error. However, in case of the terminal slices for each cell along Z, the proposed adaptive quadratic distance based reconstruction method shows a far better reconstruction accuracy (as evident in the $1^{st}, 3^{rd}, 4^{th}, 6^{th}$ graphs in Fig. 4.6(C)). This improvement of result is more prominent for cells that have non-uniform shapes and growth such as elongation along any one of the axis. We further validate this observation by performing a similar experiment on a cluster of cells of

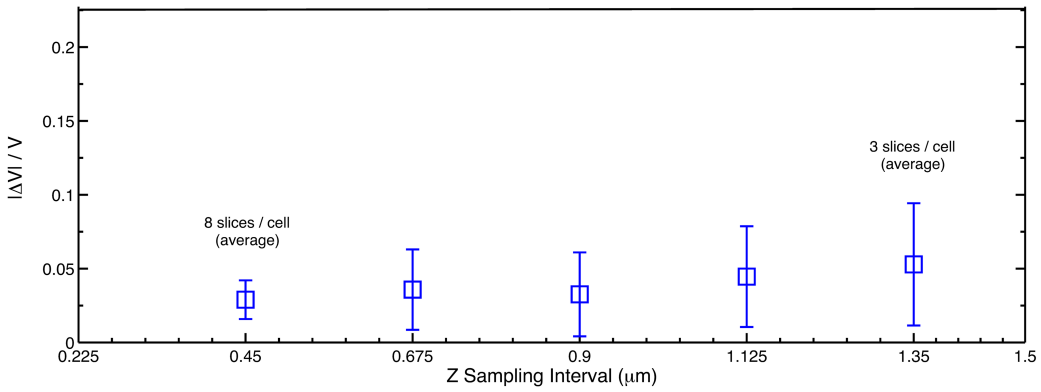


Figure 4.7: Errors in AQVT estimated cell volumes from their respective ground truth volumes at various levels of sparsity. A cluster of cells from a 3D confocal image stack with z resolution of 0.225μm is resampled to generate stacks of 5 different levels of sparsity. Each of these resampled stacks is 3D reconstructed using the proposed AQVT and volumes of each of the cells in the cluster are computed. The means and standard deviations of absolute errors in volumes (expressed as a ratio to the ground truth volumes) of all the cells for each sparser stacks are plotted. The average error slowly increases with increased sparsity but is less than 5.3% with a standard deviation of 4% even at 1.35 μm/slice (i.e. 3 slices/cell on an average).

various sizes and dimensions in a sample root meristem longitudinal cross sectional slice image. This experiment and the results are elaborated in the next subsection.

We have also evaluated the accuracy of our method by studying how much do the estimated volumes of the cells differ from ground truth for various levels of sparsity in the z-sampling. For the dense data described before (0.225 μm between slices along z), we first estimate the ground truth cell volumes of a cluster of cells by counting the total number of superpixels per cell and multiplying that with the superpixel size (0.2x0.2x0.225 μm^3). Then, we gradually resample the dense stack at 5 successive z resolutions (viz. 0.45, 0.675, 0.9, 1.125, 1.35 μm) and for each of these resampled stacks, we reconstruct the same cells using the proposed AQVT and estimate the cell volumes. It can be noted that with each of these respective resampling, the resultant 3D stack becomes more and more sparse. For example, at 0.45 μm z resolution of sampling, the average number of slices per cell is 7 or 8, whereas, for 1.35 μm, the same cells are

captured in an average of 3 slices. The errors in estimation of individual cell volumes at various sparsity levels are shown in Fig. 4.7. We have plotted the means and standard deviations of the absolute errors expressed as a ratio to the ground truth cell volumes. It can be observed that at the densest resampling (0.45 μm), the average estimation error is as low as 3% with a standard deviation of 1.3%. The estimation error slowly increases with increased sparsity in the stack and at a sparsity of about 3 slices/cell, the average estimation error is 5.3%, with standard deviation of around 4%.

4.4.2.2 Validation on 2D Root Meristem Data

For further evaluating the performance of the proposed AQVT on tightly packed cells of heterogeneous sizes and shapes, we perform similar experiments on a 2D image of root meristem longitudinal cross section with 226 cells. The ground truth segmentation of the tissue is shown in Fig. 4.8(A), the raw images for which can be seen in Fig. 5 of [94]. From Fig. 4.8(A), it can be seen that a large number of cells in the tissue slice (shown as x-z plane in Fig. 4.8(B)) is more elongated along the longitudinal direction (shown as z-axis in Fig. 4.8(B)), whereas the other cells have similar x and z diameters. Again, the cells towards the the periphery are much larger than the cells in the lower central region of the tissue. To mimic the scenario of sparse sampling along one dimension, we artificially sample the segmented cells along z (longitudinal) axis (see Fig. 4.8(B)) to generate point clouds for each cell at various sparsity ($\Delta z = d_z/9, d_z/6, d_z/4$ and $d_z/3$, where d_z is the average cell diameter along z axis). We assume sparse clustering of sampled points per cell (analogous to spatial tracking in 3D) is given to us as the contribution of the present work is dense reconstruction of cells from sparse point clouds, which lies in the post segmentation and tracking stage

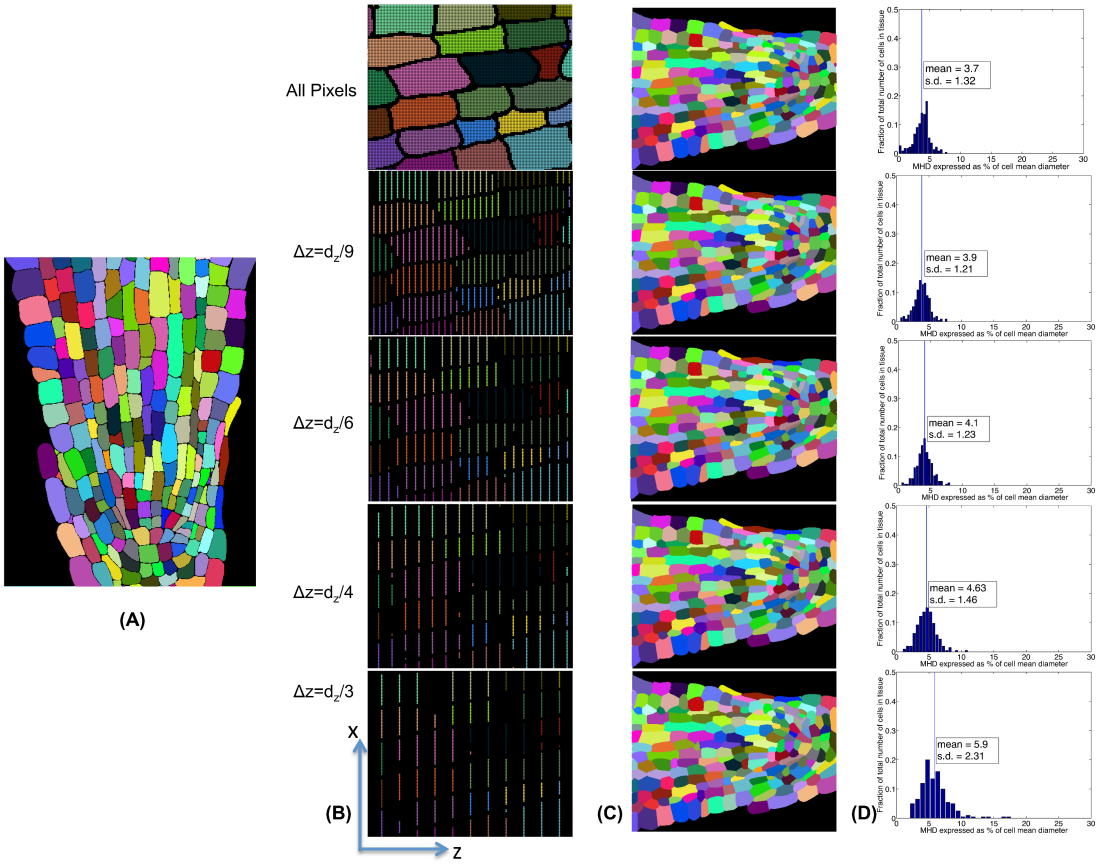


Figure 4.8: Validation of AQVT on 2D root apex longitudinal cross section data. (A) Ground truth segmentation of a sample cross sectional slice of root apical meristem tissue. The source images for this tissue can be found in [94] (Copyright (2003) National Academy of Sciences, U.S.A.). (B) The zoomed in tissue after segmentation (B-top) and sparser point clouds per cell (in the x-z plane) after resampling the tissue at various z-resolutions (zoomed in for clarity). (C) The cells (color-coded) are reconstructed using the proposed AQVT with the resampled point clouds as shown in B to present the change in reconstruction quality with increased sparsity in the sampled point clouds for both the larger and elongated cells towards the outer and upper part and the smaller cells towards the lower central part of the tissue. (D) Quantitative measure of reconstruction errors: the difference between actual and reconstructed cell shapes are computed using modified Hausdorff distance (MHD) and the histograms of MHDs for all the cells at every level of sparsity is plotted.

of the reconstruction pipeline. For each of the resampled point clouds, we use AQVT (in $\mathbb{R}^2$) to reconstruct the cells (Fig. 4.8(C)). We compute modified Hausdorff distances from every reconstructed cell shape to the ground truth segmentation for quantitative evaluation of the reconstruction accuracy (Fig. 4.8(D)). We can observe that the mean error in the reconstructed cell shapes is less than 6% of average cell longitudinal diameter and the maximum reconstruction error for most of the cells in the tissue is less that 10% for all sparsities upto $\Delta z = d_z/3$ (average 3 slices or lines/cell).

4.5 Conclusion

In this chapter, we have presented a method of reconstructing densely packed cluster of cells using a very sparsely z-sampled confocal live imaging dataset. We have provided a mathematically rigorous framework built on top of basic geometric tessellation concepts. We have first shown how the cell shapes can be approximated by the Voronoi tessellation based on Euclidean distance measure. Then, we proposed a quadratic distance metric based Voronoi tessellation framework to capture the asymmetry of the cell sizes and growth along their different axes. We described how the proposed tessellation can take care of the asymmetry by providing weights on the distance metric along each axis for each cell and how these weights as well as the location of the sites can be approximately estimated from the sparse image data for individual cells by fitting enclosing ellipsoids to the segmented sparse image slices. We have validated our method by showing that the reconstruction error (both in reconstructed cell shapes and estimated cell volumes) for the cells is sufficiently low and have provided a direct comparison of the reconstruction error for the proposed method against the popular Euclidean distance based Voronoi tessellation approach.

Chapter 5

Context-aware Activity Forecasting

5.1 Introduction

In computer vision literature, one major topic of interest is to automatically detect and recognize human activities in a video. The methods developed in the literature on activity recognition range from analyzing simple individual actions such as those discussed in [86,25] to more natural and complex human activities involving one or more actors in the scene [84, 73, 18]. However, these methods provide 'after-the-fact' recognition once the activity of interest is complete. Forecasting activities into the future much before they are observed is an important problem for many application scenarios and can be useful in designing anomalous event detection schemes. However, it hasn't yet received much attention in the computer vision community.

We have seen some recent developments in the field of activity prediction or forecasting and two classes of such problems have been introduced in the literature.

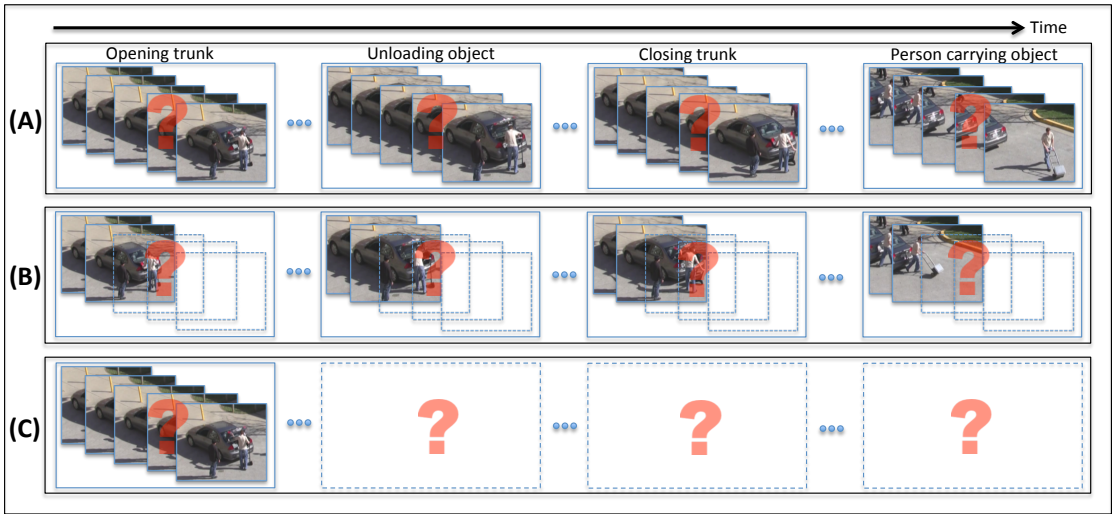


Figure 5.1: Different types of problems in human activity analysis. The figure shows four consecutive activity sequences for an actor - opening the trunk of a vehicle, unloading an object from the vehicle, closing the trunk, and the actor carrying the unloaded object, performed in that order. Three categories of activity analysis problems are presented on these sequences. (A) The classic activity recognition problem: Each of the activity sequences is fully observed before the activity labels are predicted. (B) Early prediction of ongoing activity: Only a few initial frames per activity sequence is observed and the goal is an early prediction of the activity classes from these incomplete observation sets. (C) Forecasting of future activities in absence of observation: At any point of time in a continuous video all activities occurring upto that time point are observed and the goal is to forecast the labels for future activities without the availability of observation for any of them.

The first class of problems looks into early recognition of ongoing activities [83, 98, 42] and is defined in the literature as an inference of the ongoing activity given temporally incomplete observations. In this problem, the first few frames of the video sequence containing an activity is observed and an early classification of the ongoing activity needs to be achieved. The second class of the problems seeks to forecast future activities in continuous videos [51] well before they are observed. This problem can be generally stated as an anticipation about future activity classes in a continuous video, where no observation of any future activity is available and all past activities are observed. The differences between these two problems and how each of them are principally different from a standard activity recognition problem are described through Fig. 5.1.

In this chapter, we present a method that not only attempts to solve the problem of forecasting unseen future activities (the second class of problem) but also jointly recognizes the activities that have already taken place and were observed. In most cases, it can be observed that activities performed by an actor occur following fixed temporal sequences. For example, if a person carries a bag and walks towards the trunk of a parked car, s/he is most likely to open the trunk, load the bag into it and then close the trunk. Also, for collective activities it can often be seen that actions of the actors involved are strongly synchronized with each other within a spatio-temporal window. All of these are collectively termed in this chapter as 'activity and scene context' and we leverage upon these contextual information for successful recognition of observed activities and forecasting of unobserved future activities.

We formulate the joint recognition and forecasting problem probabilistically. The past, present and future activities in a video are modeled as the nodes of a graph and the activity and scene context are modeled as a Markov Random Field on the proposed graph structure. Finally, a suitable inference strategy is adopted for recognition and

forecasting of the activity classes. We show experiments on a challenging and realistic activity dataset - the VIRAT ground dataset release 2 [74]. This dataset comprises of long duration video clips, each containing multiple activities that take place either simultaneously or sequentially, thereby making these datasets both challenging and suitable for testing the proposed spatio-temporal context based activity forecasting method.

5.2 Overview of The Proposed Method

In this chapter, we propose a strategy to jointly recognize and forecast activities in long duration continuous videos. The method attempts to recognize activities that have already been observed in a video while forecasting the most probable categories of future activities, yet to be observed in that video sequence.

A typical surveillance video contains multiple activities occurring simultaneously or in succession at different portions of the scene. In such videos, it can be observed that a specific activity by an actor is often followed by another activity by the same actor and this pattern repeats itself through and across the videos given the similarity in scenes. Therefore, an actor's future activities can often be inferred from one or more of its previous activities. Moreover, in group activities where multiple actors are involved, one actor's observed activity pattern can help us forecast another's future actions. We call this 'activity context' and it can be modeled on the edges of an *activity graph* to aid recognition and prediction. As the number of observed activities can increase with time, the graph formation strategy is dynamic with an aim to keeping the size of the graph constant. This is discussed in Sec. 5.3.1.

After graph formation, a 'Markov Random Field' (MRF) (see Sec. 5.3.2) is defined on the graph. The edge potentials defined on each of the edges of the graph

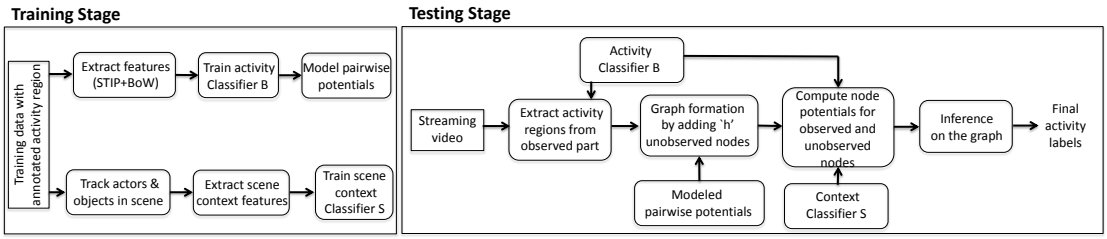


Figure 5.2: The overall activity forecasting pipeline: training and testing.

are modeled using the frequencies of occurrence of pairs of activities in a tight spatio-temporal proximity (Sec. 5.3.3) and are directly learned from a set of annotated training videos. The node potentials for the observed nodes (observed activities) are obtained using the likelihood of the activities, given by a set of activity classifiers when applied on the features (STIP+BoW) extracted from the observed activity regions (Fig. 5.2). The node potentials for all the unobserved nodes (unobserved future activities) are initially set as uniform distributions in absence of any other specific information. However, as in most cases, the activity can be characterized by the proximity and motion of the actor relative to a number of key points and detected secondary objects in the scene. These scene specific information, termed as the *scene context*, help modifying the observation/node potential of the first unobserved activity node in immediate future for every actor (Fig. 5.3). Please note that these scene contexts and the previously introduced spatio-temporal activity context are collectively termed as 'activity and scene context'.

An object detector and a person tracker is employed to extract and estimate various scene context features (see Sec. 5.3.4.3) for each actor in the scene at each time point. A trained classifier, when applied on the scene context features extracted from the observed video, provides us with the observation potentials of the aforementioned nodes. The edge potentials remain fixed for the graph across all time points. Finally, the joint

activity recognition and forecasting problem can be posed as an inference problem on the MRF just described, which is solved using an iterative 'message passing' algorithm.

5.3 Activity Forecasting Framework

Let a complete continuous video clip be V, v_t being the portion of V that is observed upto time t and let $v_t^{'}$ be the portion that is yet to be observed. Therefore, $v_t \cup v_t^{'} = V$. v_t contains a number of activity regions and the set of K most recent observations from these activity regions is given as $Y = \{y_1, y_2, \cdots y_K\}$. More clearly, the observation y_k denotes the image observation of an activity, i.e., the features computed from the k^{th} activity region amongst the most recent K activity observations. A subset of these observations is the set of observed activities by one individual actor. If there are n^o actors $O = \{o_1, o_2, \cdots o_{n^o}\}$ in the scene at time t, the set of activities by actor $o_i \in O$, observed so far would be $Y^i = \{y_1^i, y_2^i, \cdots y_N^i\}$. Further, we define a forecasting horizon h over which we intend to do activity forecasting. Note that h is not a time window, rather it denotes the number of future activities per actor we would be predicting ahead of the current time instant. Therefore, we can define a total of $(K + n^o.h)$ variables representing the hidden activity labels, which we estimate. Let the set of these labels be $X_t = \{x_1, x_2, \cdots x_K, x_{K+1}, \cdots x_{K+n^o.h}\}$. The two subsets of this label set are the one containing the labels with associated observations, $X_t^{obs} = \{x_1, x_2, \cdots x_K\}$ and another containing the labels for which no observation is available, $X_t^{unobs} = \{x_{K+1}, x_{K+2}, \cdots x_{K+n^o.h}\}$. Let the hidden variable/label for k^{th} activity by actor o_i be represented as $x_k^i \in X_t$.

In the next subsections we introduce the structure of an 'activity graph', the potential functions associated with an MRF defined on it and how to do recognition and

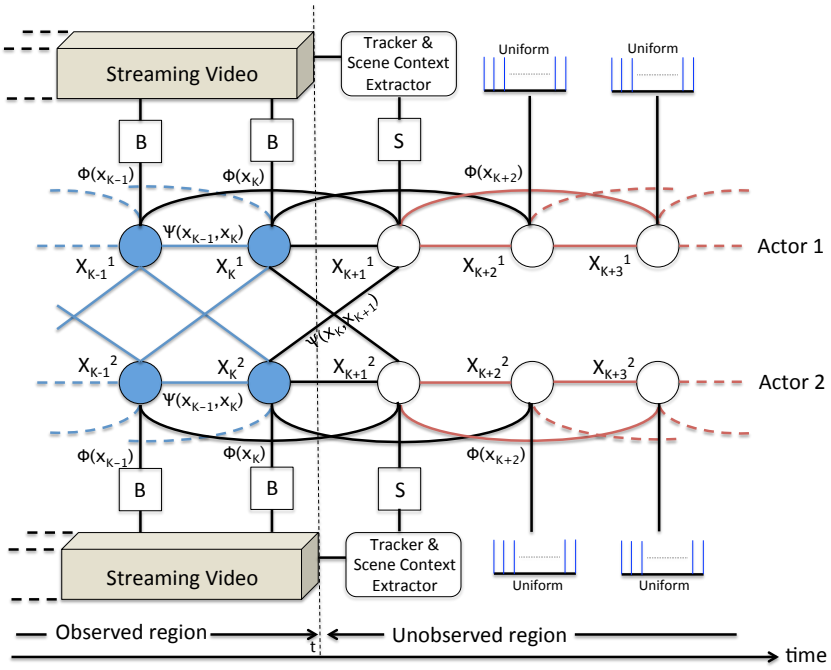


Figure 5.3: A snapshot of the graph structure for activity forecasting for two actors in the scene at any time instant 't'. 'B' denotes a trained activity classifier for observed activity recognition and 'S' denotes a scene-context classifier.

prediction as inference on this MRF to obtain the labels of the hidden states X_t.

5.3.1 Activity Graph Formation

A graph is built with the atomic activities (both observed and unobserved) as nodes and the activity contexts are modeled on the edges of the graph. The characteristics and definitions of various components of the graph are, as follows,

- Each node in the graph is an atomic activity. Let the set of all the nodes in the graph at any given time instant t be $\mathcal{N}_t$. Let a node corresponding to an activity by actor o_i be n_k^i. Then, $n_k^i \in \mathcal{N}_t$. The hidden variable corresponding to the node n_k^i is x_k^i, the value of which is to be estimated.

- An edge between two activity nodes represents the spatio-temporal context between them. Let the set of all the edges in the graph be $\mathcal{E}_t$.

- The nodes corresponding to the already observed activities in the video are called observed nodes ($\mathcal{N}_t^{obs}$, blue nodes in Fig. 5.3).
- The unobserved activities are represented by the observed nodes ($\mathcal{N}_t^{unobs}$, white nodes in Fig. 5.3).
- Observed edges are those which connect two observed nodes ($\mathcal{E}_t^{obs}$, blue lines in Fig. 5.3).
- If both the terminal nodes of an edge are unobserved, it is called an unobserved edge ($\mathcal{E}_t^{unobs}$, red lines in Fig. 5.3).
- If an edge connects two nodes one of which is observed and the other unobserved, it is called a semi-observed edge ($\mathcal{E}_t^{semi-obs}$, black lines in Fig. 5.3).

For a better pictorial understanding of the complete graph structure, please refer to Fig. 5.3.

Two observed nodes are connected by an edge if the corresponding activities occur in a predefined spatio-temporal proximity. But for the unobserved nodes and edges, this strategy cannot be adopted as we are unaware of both the spatial location and time of any future activity. Even the exact number of future activities in a video clip at any observational time point is also unknown. Therefore, whenever an activity is observed, we add one more unobserved node (corresponding to the actor of that activity) in the graph and drop the node corresponding to the oldest observed activity. Thus the total number of observed (K) and unobserved ($n^o.h$) nodes remains constant through the video. Please note that an actor might exit the scene, or the video sequence might end before all the future activity nodes are observed. The unobserved nodes are time ordered and two consecutive unobserved nodes pertaining to the activities to be

performed by the same actor are connected using an *unobserved edge*. Second order connections are also made for two unobserved nodes. Finally, the *semi-observed edges* are used to connect the last two observed nodes per actor and the two unobserved nodes in immediate future.

5.3.2 Markov Random Field Modeling

The set of random variables associated with nodes $\mathcal{N}_t$ is $X_t = \{x_1, x_2, \cdots x_K, x_{K+1}, \cdots x_{K+n^o.h}\}$, which are to be estimated given all observations Y_t. These random variables correspond to the state of each node in the graph and the support for each of these variables is the candidate set of activities (C).

Then the overall MRF is expressed as

$$P(X_t; Y_t) = \frac{1}{Z} \prod_{k=1}^{K+n^o.h} \phi(x_k, y_k) \prod_{\substack{(k,l) \\ : (n_k,n_l)\in\mathcal{E}_t}} \psi(x_k, x_l) \tag{5.1}$$

Here $\phi(x_k, y_k)$ represents the node potential of any node $n_k \in \mathcal{N}_t$, and $\psi(x_k, x_l)$ is the edge potential from node n_k to node n_l. To estimate the optimal states for every node, we have to maximize $P(X_t; Y_t)$. Towards that objective, we first estimate the approximate marginal distributions $P(x_k; y_k)$ at each node using belief-propagation scheme as described later. The optimal states that maximize the posterior distribution could be then estimated by maximizing the marginals independently.

5.3.3 Edge/Activity Context Potential

The activity context potential is defined on the edges of the graph, in each of $\mathcal{E}_t^{obs}$, $\mathcal{E}_t^{unobs}$, $\mathcal{E}_t^{semi-obs}$. This potential function models the association between any two activities occurring immediately one after the other or in close spatio-temporal succession. For any two nodes n_k^i and n_l^j (the corresponding labels being x_k^i and x_l^j

respectively), the inter-activity potential is given as, if $\left(n_k^i, n_l^j\right) \in \mathcal{E}_t$,

$$\begin{aligned}\psi\left(x_k^i = c_m, x_l^j = c_n\right) &= f_{mn,1}^s \text{ if } i = j, |l-k| = 1 \\ &= f_{mn,2}^s \text{ if } i = j, |l-k| = 2 \\ &= f_{mn}^d \text{ if } i \neq j \end{aligned} \tag{5.2}$$

All these values $f_{mn,1}^s$, $f_{mn,2}^s$ and f_{mn}^d are computed from the annotated training data. $f_{mn,1}^s$ is computed as the ratio of the number of times the same actor performs the activities c_m and c_n immediately one after another to the total number of times the activity c_m is performed in the training data. $f_{mn,2}^s$ is computed as the number of times the same actor performs activities c_m and c_n with the gap of exactly one activity in between them, and it is expressed as a ratio to the total number of times the activity c_m is performed. Finally, f_{mn}^d is obtained as the ratio of the number of times activities c_m and c_n are performed in a close spatio-temporal vicinity by two different actors to the number of times c_m is observed in the video. The same spatio-temporal proximity thresholds are also used in forming the graph, as discussed in Sec. 5.3.1.

For computing the activity context, only close spatio-temporal neighbors (1^{st} and 2^{nd} order connections) are considered, as we have observed that sub-sequences of relatively smaller length show stronger trends in repeating themselves than the longer activity sequences. Thus in the training videos, we examine all such 2 and 3-tuples of activity sub-sequences and model their pairwise relationships. This also helps us in correcting for any false positives and missing activities.

5.3.4 Node Potentials

The node potential is the likelihood of occurrence of a particular type of activity as observed in the video data. As there are specifically two types of nodes in our graph

(observed and unobserved), we devise separate strategies for computing node potentials for these two categories of nodes.

5.3.4.1 Observed Nodes:

From the annotated training data, we identify the activity regions and we train one activity classifier, the output of which is the probability of a given activity belonging to a particular category. Features at these activity regions are the observation variables and if any of the observation variables is associated with the k^{th} observed activity by actor o_i, it is denoted as y_k^i. A classifier can be employed to estimate the probability of an observation y_k^i resulting from an activity belonging to a particular category $c_p \in C$. Thus, if the set of trained baseline classifiers is B, then the observation/node potentials of the node n_k^i is given as

$$\phi\left(x_k^i, y_k^i\right) \;=\; p\left(x_k^i | y_k^i, B\right), \text{ if } n_k^i \in \mathcal{N}_t^{obs} \tag{5.3}$$

Although we have mentioned a particular feature and type of baseline classifier, any other discriminative classifier and low level motion features could be used for this purpose.

5.3.4.2 Unobserved Nodes:

The node potentials, thus obtained above, are potentials for the observed nodes. However, for an unobserved nodes y_k^i is yet to be obtained and hence a future activity is equally likely to belong to any category, i.e.,

$$\phi\left(x_k^i, y_k^i = \emptyset\right) \;=\; (1/M)\mathbf{1}^T, \;\; \text{if } n_k^i \in \mathcal{N}_t^{unobs} \tag{5.4}$$

Although no low level motion feature is available for a future activity, its likelihood of being categorized as a specific activity can sometimes be substantially improved over

$1/M$ with the help of some secondary observations from the scene, termed as the 'scene context' through this chapter.

5.3.4.3 Scene Context Classifier:

Often times, an activity is characterized by its interaction with other objects in the scene. For example, in [74], 'opening trunk'/ 'closing trunk', 'loading a vehicle'/'unloading a vehicle', 'getting into a vehicle'/'getting out of a vehicle' - all these activities have at least one thing in common, i.e. the actor interacts with a parked car in all of them. Similarly, 'entering a facility'/'exiting a facility' are both associated with a detectable entry-exit point of a facility in the scene, probably the doorway of a building. Therefore, knowledge about the locations of these objects, key scene elements and whether an actor is going to interact with either of them in near future could help us ascertain that the future activity belongs to a much smaller subset of all possible activities. This set of information, as a whole, is termed as the 'scene context'. It is represented by a set of variables comprising of the locations/bounding boxes of all the secondary objects, and key points in the scene that are related to one or more types of activities, location and motion information of the actor relative to these objects/key points. Please note that the scene context is computed individually for every actor in the scene and the values naturally change with time. Details on such context features in relation to experiments on VIRAT data is given in Sec. 5.4.

The computed scene context features are averaged over a predefined time window to generate a smoothed scene context feature vector per actor at each time point. Let, at time point t, the scene context feature computed for actor o_i be $f_t^{o_i} = \left\langle f_{t,1}^{o_i}, f_{t,2}^{o_i}, \cdots f_{t,Nf}^{o_i} \right\rangle$. As these features are computed in between two successive activities, the pair $(f_t^{o_i}, a_k)$ completes the representation of the scene context, where $f_t^{o_i}$ is

computed at a time t, after which the next activity o_i is going to perform is a_k. Such features for all the actors over the entire training dataset are combined and a scene-context classifier S is trained. Given a test video, at each time point, whenever we want to run the recognition and prediction, we compute the scene context features. If an actor o_i has already performed k activities and its computed scene context features at time t is $f_t^{o_i}$, then the classifier S provides us with the likelihood of the next activity (x_{k+1}^i) that o_i is going to perform, which is also the node potential for the first unobserved activity node for o_i at time t. Therefore,

$$\phi\left(x_{k+1}^i, y_{k+1}^i = \emptyset\right) \;=\; p\left(x_{k+1}^i \mid f_t^{o_i},\; S\right), \tag{5.5}$$

where $n_k^i \in \mathcal{N}_t^{obs}$, $n_{k+1}^i \in \mathcal{N}_t^{unobs}$. For all other future unobserved nodes for actor o_i, the node potentials remain uniform (see Eqn. 5.4) until the next observation is obtained. It can be noted that as $f_t^{o_i}$ is time varying, the estimated node potential also changes from frame to frame and needs to be re-estimated.

5.3.5 Inference: Loopy Belief Propagation

The next step is to do the inference on the MRF, which involves the computation of the marginal probability distributions for the states x_k of each node $n_k \in \mathcal{N}_t$, given the observations Y. For computation of the marginals at each node, we choose to use *Loopy Belief Propagation* (LBP) based on the *Sum-Product* algorithm [52]. The iterative message passing scheme associated with the LBP algorithm is given in Sec. 3.3.7. If LBP converges at iteration L, the estimated marginals at each node would be $P^{(L)}(x_k; y_k)$ and the MAP estimates for the most likely states is computed as $\hat{x_k} \;=\; \arg_{x_k} \max P^{(L)}(x_k; y_k)$. This optimum state corresponds either to the recognized or the predicted label of activity node n_k depending on the type of the node.

5.4 Experimental Results

To assess the effectiveness of our proposed method in activity forecasting, we perform experiments on the publicly available state-of-the-art VIRAT ground dataset [74]. We perform two similar sets of experiments corresponding to two recognition schemes used for labeling observed activities, viz., 1. An automated classifier (BOW + SVM), 2. Ground truth activity labels.

5.4.1 Dataset

VIRAT Ground dataset is a state-of-the-art activity dataset with many challenging characteristics, such as wide variation in the activities and clutter in the scene. The dataset consists of surveillance videos of realistic scenes with different scales and resolution. The videos are 2 to 15 minutes long and can contain upto 30 events. The dataset contains scenes captured on a single camera though the viewpoints can differ across scenes. The activities are: 1 - person loading an object to a vehicle; 2 - person unloading an object from a vehicle; 3 -person opening a vehicle trunk; 4 - person closing a vehicle trunk; 5 - person getting into a vehicle; 6 - person getting out of a vehicle; 7 - person gesturing; 8 - person carrying an object; 9 - person running; 10 - person entering a facility; 11 - person exiting a facility. We work on the all the scenes except scene 0002 and 0102. For experiment set 1, we use half of the data for training our model and the rest is used for testing. For experiment set 2, however, training is only needed for the scene context based future activity classifier and only a fifth of the entire dataset is used for training and we test our method on the rest of the data.

5.4.2 Preprocessing

Given a test video sequence, the first task is to obtain the observed activity regions. As activity regions overlap with the motion regions in a video, a background subtraction method [103] can be used to locate the motion regions. Moving persons and vehicles are identified using an available software [32]. Doors, bags, boxes etc. are detected using a detector similar to [20]. A tracking method [89], when applied on the detected actors' bounding boxes, provides us with the trajectories of the actors.

5.4.3 Extraction of Scene Context Features

For our experiments on VIRAT dataset, the set of scene context features computed are - 1. Are cars parked in the scene? (1-Y, 0-N), 2. Distance from the closest parked vehicle normalized by length of diagonal of the car bounding box, 3. Heading towards the closest parked vehicle, 4. Largest overlap of the actor bounding box with the bounding box of a parked vehicle normalized by area of the actor bounding box, 5. Is there one or more entry/exit points to facilities in the scene? (1-Y, 0-N), 6. Distance from the closest entry/exit point normalized by the length of the diagonal of the actor bounding box, 7. Heading towards the closest entry/exit point, 8. Is an object seen on the actor? (1-Y, 0-N), 9. Average velocity of the actor, 10. Time elapsed since last observed activity. For other datasets, the objects of interest will be recognized from the segmented training videos and the generalized scene context features can be estimated by keeping the same relationship between actor and objects. The features are estimated at each frame for every actor using the actor track and locations of detected objects in the scene. The features extracted from the training videos are further used to train a *bag of decision trees* containing 200 fully grown trees. At every frame, the next activity class

is used as label. In training videos, given a scene context feature vector extracted at any frame, the trees individually vote and the normalized votes are used as the likelihood for probable future activity class labels.

5.4.4 Motion feature extraction for observed activities

In experimental setup 1 (automated classifier based labels for observed activities), we have used a 'Bag-of-Features' approach over 'Space Time Interest Points' (STIP) [57] due to its popularity in the literature for recognition of atomic activities. The STIPs based on Harris and Forstner operators are computed for every activity region in the training data. Feature vectors computed at each point are clustered and quantized to generate a codebook during the training phase and each activity category is modeled as a distribution over this codebook. A multiclass SVM classifier is trained with these features and the corresponding activity labels obtained from the annotated training data. Similarly, for test video inputs, the STIPs are computed and probable activity regions are identified where a significant number of points from the trained vocabulary is observed.

5.4.5 Experiment Set 1

In this section, we present the experimental results when a classifier (BOW + SVM) is used to generate node potentials corresponding to already observed activities. At every fifth frame between two activities in a continuous video, we forecast the next activities that an actor is going to perform using the previously observed activities in the video as well as estimated scene context at that frame. At any given time before an activity is performed, the proposed method estimates the probabilities of various

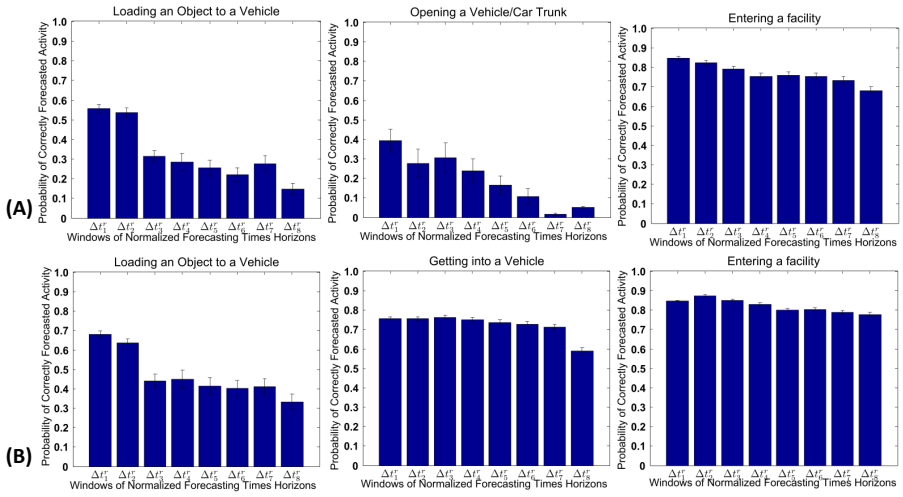


Figure 5.4: Increasing trend of forecasting probabilities for different classes of activity (observed in the test set) with time. The positive direction of the time axis indicates increasing time gap from the instant at which the activity to be forecasted happens. (A) Probability with which the ground truth activity is forecasted as the next activity in exp. setup 1, (B) Similar increasing trend as observed in exp. setup 2.

candidate future activity labels and these forecasting probabilities can vary with time as the actor moves and the scene context changes. In Fig. 5.4(A) and Fig. 5.5(A), we examine this variation in forecasting probabilities with time for the next activity that an actor may perform.

Let us assume that an actor has already performed an activity A_{last} upto time t_{last} (or just entered the scene) and is going to perform A_{next} at time t_{next}. At every time point t between t_{last} and t_{next}, we estimate the probability with which A_{next} is forecasted as the next unobserved activity label, and thus $t^r = (t - t_{next})$ is the forecasting horizon. The average probability of forecasting the ground truth activity (A_{next}) over all instances of A_{last} in the dataset is computed and its time evolution is observed. Please note that for the same future activity performed, the time gap $[t_{last}, t_{next}]$ varies for different instances and hence is normalized between $[-1, 0]$ (-1 denotes the end time of last observed activity or the time of actor's first appearance and 0 is the time when the next activity is going to occur in future). This time gap is

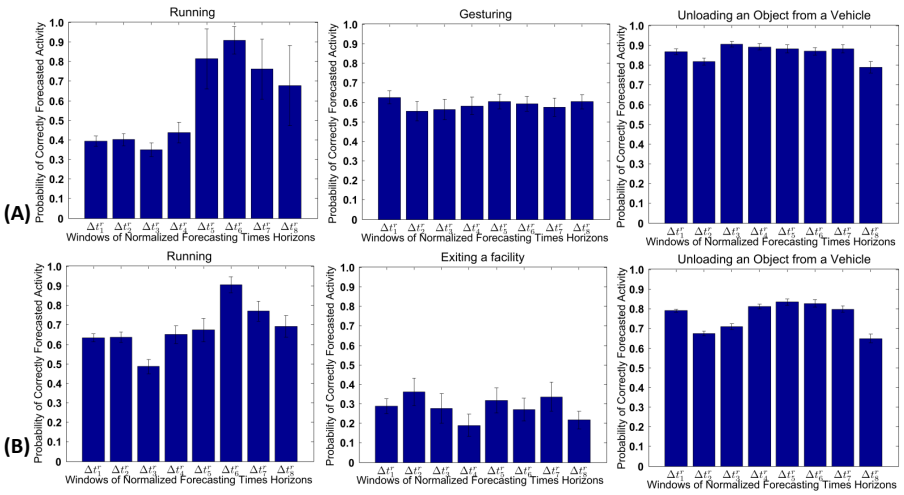


Figure 5.5: Time evolution of forecasting probabilities for different classes of activity (observed in the test set), where no apparent trend is observed. The positive direction of the time axis indicates increasing time gap from the instant at which the activity to be forecasted happens. (A) Probability with which the ground truth activity is forecasted as the next activity in exp. setup 1, (B) Similar absence of trend, as observed in exp. setup 2.

split into 8 equal ranges ($\Delta t_i^r, i = 1, \cdots 8$) and the average probabilities (with standard errors) of the next ground truth activities are plotted.

Fig. 5.4(A) shows the time evolution of probabilities of three activity types (loading obj. to vehicle, opening trunk, entering facility) as the next activity in experimental setup 1. It can be observed that for these activities, the probabilities rapidly increase as the forecasting horizon closes to zero (t closes to t_{next}), especially for the first two car related activities. This is because, during this time range an actor typically walks upto the vehicle and as the actor gets closer to the vehicle (t closes to t_{next}), the model gets more and more confident the person is going to interact with the car and hence one of these activities is going to be performed. The last observed activity label and the spatio-temporal context further refines the forecasting probabilities to put preference to a particular activity label.

However, this increasing trend in forecasting probability is largely activity spe-

cific as for some of the activities in the dataset, there may not be any tightly associated scene context variable. Thus, even large changes in computed scene context variables minimally affect the forecasting probabilities when these activities would occur in immediate future. For VIRAT, some examples of such activities are Running, Gesturing etc. As seen in Fig. 5.5(A), there is no visible trend in the time evolution of forecasting probabilities for these activities. Again, for activities such as 'unloading object from a vehicle', the relevant scene context variables (e.g. distance from a car, overlap with a car bounding box etc.) remain largely constant, thereby resulting in uniform average probabilities through the forecasting time range (Fig. 5.5(A)).

5.4.6 Experiment Set 2

The probabilities and accuracies for forecasting future activities are affected by the accuracy of the recognition module used for already observed activities. Therefore, to factor out the effect of the errors in the observed activity recognition module on the forecasting results and we repeat the same experiments as in set 1 with only the classifier replaced by a perfect recognition scheme. As we observe an activity, we retrieve its ground truth label and set the activity recognition probability for that particular activity at a very high value, and close to zero for the rest.

The evolution of forecasting probabilities with normalized horizon is shown in Fig. 5.4(B) for the activities that show an increasing trend in forecasting probabilities and Fig. 5.5(B) for the activities without any apparent trend in the time evolution of probabilities. The figures are visually similar to those for the same activities in experimental setup 1. However, the average forecasting probabilities for most of the activities is typically higher than that in the classifier based recognition case (set 1).

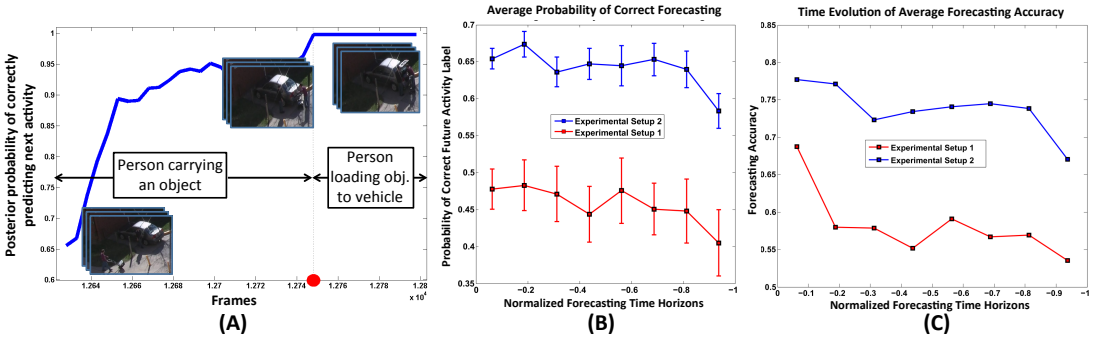


Figure 5.6: (A) An example showing how the posterior probability of forecasting increases with time and stabilizes once the next observation is obtained. (B) Comparison of time evolution of probability averaged over all activity classes with which any correct activity class is forecasted as the next activity in exp. setup 1 and 2. (C) Comparison of forecasting accuracy for immediately next activity in the two experimental setups.

An example showing the increasing forecasting probability for the next unobserved activity is presented in Fig. 5.6(A) (in exp. setup 2). In a video segment, an actor in observed to 'carry an object' and the *unobserved* future activity would be 'person loading object to a vehicle'. A parked car is detected in the scene and the posterior probability of the next activity being labeled as 'person loading obj.' rapidly increases as the person walks straight towards the car and gets closer to it. Fluctuation in the probability is seen due to occlusion of the detected object on person. The posterior probability gets close to 1 just before the start of the next activity (shown by red circle). Once the next observation is obtained, the posterior represents recognition probability and remains constant for the rest of the video.

The time evolutions of forecasting probability averaged over all activity classes for both exp. setup 1 and 2 are shown in Fig. 5.6(B) and in both the cases they show an overall increasing trend. As expected, the average probabilities in exp. setup 2 are higher than that in exp. setup 1. Similar trends are also observed in Fig. 5.6(C), which shows the time evolution of forecasting accuracies combined for all activity classes. An increasing trend very similar to that of forecasting probabilities is observed.

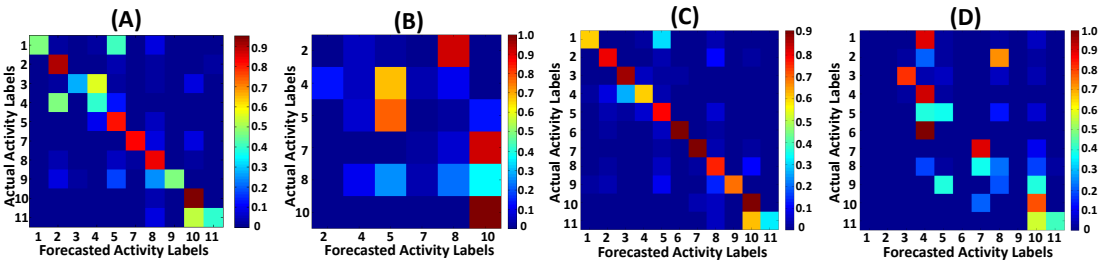


Figure 5.7: Confusion matrices showing the overall forecasting accuracies obtained for each class of activity. (A-B) Accuracies for forecasting activities in immediate future and one step ahead (next-to-next) in experimental setup 1, (C-D) Accuracies for next and next-to-next activities respectively in experimental setup 2. For activity types corresponding to the label numbers, refer to Sec. 5.4.1.

As the proposed method is capable of forecasting activities deeper into the future beyond the immediately next activity, we can also compute the time evolution of forecasting probabilities and accuracies for activities one step ahead (the 'next-to-next' activities). The overall forecasting accuracies of all next and next-to-next activities in exp. setup 1 are shown in Fig. 5.7(A-B). The accuracies for most activities happening in immediate future is high. As we predict activities deeper into future, the accuracies tend to go down, evidenced by Fig. 5.7(B). Similar trends are seen for activities in exp. setup 2 (Fig. 5.7(C-D)). Please note that, as there is no baseline activity classifier to train in exp. setup 2 and we need only 20% of the data for training the scene context classifier, we have the entire remaining dataset for testing and that is why results for all 11 activities could be investigated. The expected improvement in forecasting accuracy in setup 2 can immediately be evidenced by the confusion matrices. Most of the classes of immediate future activities were predicted near perfectly. Thus, as expected, the forecasting accuracy of activities will majorly improve with the improvement in activity recognition scheme.

5.5 Conclusion

In this chapter, we have presented a novel approach towards the problem of forecasting future activities in long duration continuous videos. We have shown that the forecasting problem can be posed as a graph inference problem on a MRF where individual activities in a sequence are nodes on the graph. The method combines the spatio-temporal inter activity context and inter-relationship between actors' tracks and detected key points and objects in the scene with a standard activity recognition classifier to forecast activities that are yet to be observed. We show detailed experimental results on the challenging VIRAT [74] dataset and achieve meaningful and encouraging results.

Chapter 6

Conclusion

6.1 Thesis Summary

In this thesis, we studied spatial and temporal context models and explored their usefulness in a number of computer vision applications, viz., multi-camera data association, spatio-temporal cell tracking and human activity forecasting. In Chapter 2, we presented the network consistent data association (NCDA) - an optimization based framework that not only maintains consistency when pairwise data association results are combined over a network of cameras or amongst observations across spatio-temporal locations, but also improves the data-association accuracy. The proposed method was applied to the person re-identification problem and results indicate a significant improvement in re-identification accuracy while establishing consistency.

In Chapter 3, we applied the proposed data association method in another challenging application area - the spatio-temporal cell tracking problem. To generate robust similarity measures between two dimensional projections of cells on various confocal image planes, we proposed a CRF based framework, where we leveraged upon the local spatial geometric structure and topology of the relative positions of the neighbor-

ing cells in presence of imaging noise. The pairwise similarity measures, thus obtained, are further combined together via the NCDA method proposed in Chapter 2 to generate consistent and highly accurate spatio-temporal cell lineages. Experimental results are shown on four dimensional confocal image stacks of Arabidopsis shoot apical meristem. In Chapter 4, we utilized the cell lineages obtained using the cell tracking framework in a cell resolution 3D reconstruction method to generate cell growth statistics for a large number of cells in a high throughput manner. The tight packing of the cells in a tissue enabled us to estimate the 3D structures of individual cells using the slice information of the cell as well as that of its nearest neighbors through an adaptive quadratic Voronoi tessellation model, which showed capability of handling the extreme z-sparsity of a typical live imaging dataset.

We explored the relatively new and exciting problem of activity forecasting in Chapter 5. We observed that in continuous videos, activities occurring in a scene are rarely independent and are often simultaneous/sequential in nature. We modeled the same on an activity graph as contextual information in a Markov Random Field (MRF) and combined that with the interrelationship between static scene cues and dynamic target trajectories to forecast into future. Labels of unobserved future activities were obtained through inference on the MRF and experiments on the challenging VIRAT ground dataset yielded promising results.

6.2 Future Work

Besides person re-identification and cell tracking, the idea of network consistency is relevant to many other data association problems. Feature tracking in video and association of feature points in a multi-robot system are some of such challenging

domains where consistent association is of utmost importance. We would like to explore each of these application areas in our future work. We would also investigate how the proposed NCDA can be formulated and solved for distributed network data association problems. This would be extremely useful in networks such as distributed multi-agent systems with limited communication, i.e., when every agent in the network can share information only with its neighbors. Besides the aforementioned problems, the future directions of our research in NCDA based multi-camera data association will be to apply our approach to bigger networks with large numbers of cameras and to cope with wider space-time horizons.

The spatio-temporal cell tracking and the cell resolution 3D reconstruction modules constitute two important components of an image analysis pipeline for live tissues. We would explore the applicability of this pipeline to various densely clustered plant and animal tissues other than Arabidopsis shoot meristems. We would use this image analysis pipeline to generate various cell division and cell growth statistics in a fully automated, high-throughput manner. The collected statistics could be extremely useful in building a dynamical model to quantitatively analyse the spatio-temporal correlation in cell division and cell growth in a complex multi-layered tissue, which is another goal of our future work.

The proposed activity forecasting framework was built using a set of predefined scene context variables, which are highly relevant for a typical human activity dataset. In future, we would like to investigate how scene context can be automatically defined and learned for any set of activities and scene in an online fashion. Specifically, we would explore the applicability of a deep neural network for this purpose. Also, if an activity is observed that was forecast with a very low probability, it might be anomalous with respect to the already observed pattern of activities in the scene. Thus the pro-

posed activity forecasting method, when combined with an online scene context learning framework, could be extremely useful for anomalous activity detection.

Bibliography

[1] A. Alavi, Y. Yang, M. Harandi, and C. Sanderson. Multi-shot person re-identification via relational stein divergence. In *International Conference on Image Processing*, 2013.

[2] M. Ankerst, G. Kastenmüller, H. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In *International Symposium on Advances in Spatial Databases*, pages 207–226, 1999.

[3] F. Aurenhammer and R. Klein. *Handbook of Computational Geometry.* Elsevier Publishing House, 1999.

[4] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 381–390, 2012.

[5] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. In *International Conference on Computer Vision*, pages 137–144, 2011.

[6] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013.

[7] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *ArXiv e-prints*, 2013.

[8] R. Benmokhtar and I. Laptev. Inria-willow at trecvid2010: Surveillance event detection. In *TRECVID*, 2010.

[9] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819, 2011.

[10] R. Bise, Z. Yin, and T. Kanade. Reliable cell tracking by global data association. In *International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1004–1010, 2011.

[11] J. D. Boissonnat, C. Wormser, and M. Yvinec. Curved Voronoi diagrams. In *Effective Computational Geometry for Curves and Surfaces*, pages 67–116. Springer-Verlag, Mathematics and Visualization, 2006.

[12] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] A. Chakraborty, M. Liu, K. Mkrtchyan, G. V. Reddy, and A. K. Roy Chowdhury. Cell volume estimation from a sparse collection of noisy confocal image slices. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 183–189, 2010.

[14] A. Chakraborty, M. M. Perales, G. V. Reddy, and A. K. Roy-Chowdhury. Adaptive geometric tessellation for 3D reconstruction of anisotropically developing cells in multilayer tissues from sparse volumetric microscopy images. *PLoS ONE*, 8(8):e67202, 08 2013.

[15] A. Chakraborty and A. Roy-Chowdhury. A conditional random field model for tracking in densely packed cell structures. In *International Conference on Image Processing*, 2014.

[16] A. Chakraborty, R. K. Yadav, G. V. Reddy, and A. Roy-Chowdhury. Cell resolution 3D reconstruction of developing multilayer tissues from sparsely sampled volumetric microscopy images. *International Conference on Bioinformatics And Biomedicine*, pages 378–383, 2011.

[17] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, pages 68.1–68.11, 2011.

[18] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition*, 2011.

[19] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Computer Vision and Pattern Recognition*, pages 44–51, 2000.

[20] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[21] A. Das, A. Chakraborty, and A. Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer vision*, 2014.

[22] D. Delibaltov, S. Karthikeyan, V. Jagadeesh, and B. S. Manjunath. Robust biological image sequence analysis using graph based approaches. In *Asilomar Conference On Signals, Systems and Computers*, 2012.

[23] J. D'Errico. Surface fitting using gridfit. In *MATLAB Central File Exchange*, 2005.

[24] M. Dikmen, E. Akbas, T. S Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Asian Conference on Computer vision*, pages 501–512, 2010.

[25] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 0:65–72, 2005.

[26] M.-P. Dubuisson and A. K. Jain. A modified hausdorff distance for object matching. *International Conference on Pattern Recognition*, pages 566–568, 1994.

[27] A. Dufour, V. Shinin, S. Tajbakhsh, N. Guillen-Aghion, J. C. Olivo-Marin, and C. Zimmer. Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. *IEEE Transactions on Image Processing*, 14(9):1396–1410, 2005.

[28] O. Dzyubachyk, W.A. Van Cappellen, J. Essers, W.J. Niessen, and E. Meijering. Advanced level-set-based cell tracking in time-lapse fluorescence microscopy. *IEEE Transactions on Medical Imaging*, 29(3):852–867, 2010.

[29] R. J. Errington, M. D. Fricker, J. L. Wood, A. C. Hall, and N. S. White. Four-dimensional imaging of living chondrocytes in cartilage using confocal microscopy: a pragmatic approach. *American Journal of Physiology*, 272(3):C1040–1051, 1997.

[30] D. Evans and S. Jones. Detecting voronoi (area-of-influence) polygons. *Mathematical Geology*, 19:523–537, 1987.

[31] J. Farinas, M. Kneen, M. Moore, and A. S. Verkman. Plasma membrane water permeability of cultured cells and epithelia measured by light microscopy with spatial filtering. *Journal of General Physiology*, 110(3):283–296, 1997.

[32] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[33] R. Fernandez, P. Das, V. Mirabet, E. Moscardi, J. Traas, J.-L. Verdeil, G. Malandain, and C. Godin. Imaging plant growth in 4d: robust tissue reconstruction and lineaging at cell resolution. *Nature Methods*, 7(7):547–553, 07 2010.

[34] M. M. A. Ferrero. Voronoi diagram: The generator recognition problem. *Computing Research Repository*, abs/1105.4246, 2011.

[35] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114:712–722, 2010.

[36] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *European Conference Computer Vision*, 2006.

[37] V. Gor, M. Elowitz, T. Bacarian, and E. Mjolsness. Tracking cell signals in fluorescent images. *Workshop on Computer Vision Methods for Bioinformatics*, pages 142–150, 2005.

[38] V. Gor, B. E. Shapiro, H. Jönsson, M. Heisler, G. V. Reddy, E. M. Meyerowitz, and E. Mjolsness. A software architecture for developmental modelling in plants: the computable plant project. *Bioinformatics of Genome Regulation and Structure*, 2005.

[39] H. Grabner, J. Matas, L. J. Van Gool, and P. C. Cattin. Tracking the invisible: Learning where the object might be. In *Computer Vision and Pattern Recognition*, pages 1285–1292, 2010.

[40] U. Guar, Y. Zhu, B. Song, and A. K. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *International Conference on Computer Vision*, 2011.

[41] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *Computer Vision and Pattern Recognition*, 2009.

[42] M. Hoai and F. De la Torre. Max-Margin early event detectors. In *Computer Vision and Pattern Recognition*, 2012.

[43] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008.

[44] N. N. Kachouie, P. Fieguth, J. Ramunas, and E. Jervis. Probabilistic model-based cell tracking. *International Journal of Biomedical Imaging*, 2006.

[45] Y. Kalaidzidis. Intracellular objects tracking. *European Journal of Cell Biology*, 86:569–578, 2007.

[46] S. Karthikeyan, D. Delibaltov, U. Gaur, M. Jiang, D. Williams, and B.S. Manjunath. Unified probabilistic framework for simultaneous detection and tracking of multiple objects with application to bio-image sequences. In *International Conference on Image Processing*, 2012.

[47] K. Kawahara, M. Onodera, and Y. Fukuda. A simple method for continuous measurement of cell height during a volume change in a single A6 cell. *Jpn. J. Physiol*, pages 411–419, 1994.

[48] L. G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, pages 307–320, 1996.

[49] M. F. Kircher, S. S. Gambhir, and J. Grimm. Noninvasive cell-tracking methods. *Nature Reviews Clinical Oncology*, 8:677–688, 2011.

[50] T. Kirubarajan, Y. Bar-Shalom, and K. R. Pattipati. Multiassignment for tracking a large number of overlapping objects. *IEEE Transactions on Aerospace and Electronic Systems*, 37(1):2–21, 2001.

[51] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214, 2012.

[52] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 1998.

[53] P. Kumar and E. A. Yildirim. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and Applications*, 126:1–21, 2005.

[54] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[55] D. Kwiatkowska and A. Routier-Kierzkowska. Morphogenesis at the inflorescence shoot apex of anagallis arvensis: surface geometry and growth in comparison with the vegetative shoot. *Journal of Experimental Botany*, 2009.

[56] T. Lan, Y. Wang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[57] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, pages 432–439, 2003.

[58] K. Li, M. Chen, T. Kanade, E. Miller, L. Weiss, and P. Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical Image Analysis*, 12(5):546 – 566, 2008.

[59] K. Li and T. Kanade. Cell population tracking and lineage construction using multiple-model dynamics filters and spatiotemporal optimization. In *Workshop on Microscopic Image Analysis with Applications in Biology*, 2007.

[60] W. Li and X. Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition*, 2013.

[61] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44, 2012.

[62] L. Liang, H. Shen, P. Rompolas, V. Greco, P. De Camilli, and J. S. Duncan. A multiple hypothesis based method for particle tracking and its extension for cell segmentation. In *Information Processing in Medical Imaging*, pages 98–109, 2013.

[63] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification : What features are important? In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 391–401, 2012.

[64] M. Liu, A. Chakraborty, D. Singh, R. K. Yadav, M. Gopi, G. V. Reddy, and A. Roy-Chowdhury. Adaptive cell segmentation and tracking for volumetric confocal microscopy images of a developing plant meristem. *Molecular Plant*, 4(5):922–31, 2011.

[65] M. Liu, R. K. Yadav, A. Roy-Chowdhury, and G. V. Reddy. Automated tracking of stem cell lineages of arabidopsis shoot apex using local graph matching. *Plant Journal*, 62:135–147, 2010.

[66] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *Computer Vision and Pattern Recognition Workshops*, pages 31–36, 2012.

[67] E. Mjolsness. The growth and development of some recent plant models: A viewpoint. *Journal of Plant Growth Regulation*, 25:270–277, 2006.

[68] K. Mkrtchyan, A. Chakraborty, and A. K. Roy-Chowdhury. Automated registration of live imaging stacks of Arabidopsis. In *International Symposium on Biomedical Imaging*, pages 672–675, 2013.

[69] K. Mkrtchyan, D. Singh, M. Liu, G. V. Reddy, A. K. Roy Chowdhury, and M. Gopi. Efficient cell segmentation and tracking of developing plant meristem. In *International Conference on Image Processing*, pages 2165–2168, 2011.

[70] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *Computer Vision and Pattern Recognition*, 2011.

[71] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in AI*, pages 467–475, 1999.

[72] T. Nakahari, M. Murakami, H. Yoshida, M. Miyamoto, Y. Sohma, and Y. Imai. Decrease in rat submandibular acinar cell volume during ACh stimulation. *American Journal of Physiology*, 258(6):G878–886, 1990.

[73] N. M. Nayak, Y. Zhu, and A. K. Roy-Chowdhury. Exploiting spatio-temporal scene structure for wide-area activity analysis in unconstrained environments. *IEEE Transactions on Information Forensics and Security*, 8(10):1610–1619, 2013.

[74] S. Oh, A. Hoogs, A. G. A. Perera, N. P. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. S. Davis, E. Swears, X. Wang, Q. Ji, K. K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. K. Roy Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Computer Vision and Pattern Recognition*, pages 3153–3160, 2011.

[75] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley and Sons, Inc., 2nd edition, 2000.

[76] D. R. Padfield, J. Rittscher, N. Thomas, and B. Roysam. Spatio-temporal cell cycle phase analysis using level sets and fast marching methods. *Medical Image Analysis*, 13(1):143–155, 2009.

[77] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[78] S. Pedagadi, J. Orwell, and S. Velastin. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[79] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[80] F. Porikli and M. Hill. Inter-camera color calibration using cross-correlation model function. In *International Conference on Image Processing*, pages 133–136, 2003.

[81] B. Prosser, S. Gong, and T. Xiang. Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions. In *British Machine Vision Conference*, 2008.

[82] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *International Conference on Computer Vision*, 2007.

[83] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *International Conference on Computer Vision*, pages 1036–1043, 2011.

[84] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *International Conference on Computer Vision*, 2009.

[85] A. Schrijver. *Theory of linear and integer programming.* John Wiley and Sons, 1998.

[86] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *International Conference on Pattern Recognition*, pages 32–36, 2004.

[87] K. Shafique and M. Shah. A noniterative greedy algorithm for multiframe point correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):51–65, 2005.

[88] Z. Si, M. Pei, B. Yao, and S. C. Zhu. Unsupervised learning of event AND-OR grammar and semantics from video. In *International Conference on Computer Vision*, 2011.

[89] B. Song, T.-Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury. A Stochastic Graph Evolution Framework for Robust Multi-target Tracking. In *European Conference on Computer Vision*, pages 605–619, 2010.

[90] C. Sutton and A. McCallum. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[91] M. Taj, E. Maggio, and A. Cavallaro. Multi-feature graph-based object tracking. In *Multimodal Technologies for Perception of Humans*, volume 4122, pages 190–199. Springer Berlin Heidelberg, 2007.

[92] K. Tang, L. Fei-Fei, and D. Koller. Learning Latent Temporal Structure for Complex Event Detection. In *Computer Vision and Pattern Recognition*, 2012.

[93] O. Tataw, M. Liu, R. Yadav, G. V. Reddy, and A. Roy-Chowdhury. Pattern Analysis of Stem Cell Growth Dynamics in the Shoot Apex of Arabidopsis. In *International Conference on Image Processing*, 2010.

[94] J. Wyrzykowska and A. Fleming. Cell division pattern influences gene expression in the shoot apical meristem. *Proceedings of the National Academy of Sciences*, 100(9):5561–5566, 2003.

[95] G. Yang, A. Matov, and G. Danuser. Reliable Tracking of Large Scale Dense Antiparallel Particle Motion for Fluorescence Live Cell Imaging. In *Computer Vision and Pattern Recognition - Workshops*, page 138, 2005.

[96] L. Yang and R. Jin. Distance metric learning : A comprehensive survey, 2006.

[97] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human object interaction activities. In *Computer Vision and Pattern Recognition*, 2010.

[98] G. Yu, J. Yuan, and Z. Liu. Predicting human activities using spatio-temporal structure of interest points. In *ACM Multimedia*, pages 1049–1052, 2012.

[99] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. In *Computer Vision and Pattern Recognition*, 2013.

[100] Q. Zhu, P. Tekola, J. P. Baak, and J. A. Belin. Measurement by confocal laser scanning microscopy of the volume of epidermal nuclei in thick skin sections. *Analytical and Quantitative Cytology and Histology*, 16(2):145–52, 1994.

[101] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury. Context-Aware Activity Recognition and Anomaly Detection in Video. *Journal of Selected Topics in Signal Processing*, 7(1):91–101, 2013.

[102] Y. Zhu, N. M. Nayak, and A. K. Roy Chowdhury. Context-Aware Modeling and Recognition of Activities in Video. In *Computer Vision and Pattern Recognition*, pages 2491–2498, 2013.

[103] Z. Zivkovic. Improved Adaptive Gaussian Mixture Model for Background Subtraction. In *International Conference on Pattern Recognition*, pages 28–31, 2004.