OBJECT ASSOCIATION ACROSS MULTIPLE MOVING CAMERAS IN PLANAR SCENES

by

YASER SHEIKH
B.S. GIK Institute of Engineering Sciences and Technology, 2001
M.S. University of Central Florida, 2005

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2006

Major Professor:
Mubarak Shah

UMI Number: 3210379

UMI Microform 3210379

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

In this dissertation, we address the problem of object detection and object association across multiple cameras over large areas that are well modeled by planes. We present a unifying probabilistic framework that captures the underlying geometry of planar scenes, and present algorithms to estimate geometric relationships between different cameras, which are subsequently used for co-operative association of objects. We first present a local[1] object detection scheme that has three fundamental innovations over existing approaches. First, the model of the intensities of image pixels as independent random variables is challenged and it is asserted that useful correlation exists in intensities of spatially proximal pixels. This correlation is exploited to sustain high levels of detection accuracy in the presence of dynamic scene behavior, nominal misalignments and motion due to parallax. By using a non-parametric density estimation method over a joint domain-range representation of image pixels, complex dependencies between the domain (location) and range (color) are directly modeled. We present a model of the background as a *single* probability density. Second, temporal persistence is introduced as a detection criterion. Unlike previous approaches to object detection that detect objects by building adaptive models of the background, the *foreground* is modeled to augment the detection of objects (without explicit tracking), since objects detected in the preceding frame contain substantial evidence for detection in the current frame. Finally, the

[1] Local refers to processes occurring at each individual camera.

background and foreground models are used competitively in a MAP-MRF decision framework, stressing spatial context as a condition of detecting interesting objects and the posterior function is maximized efficiently by finding the minimum cut of a capacitated graph. Experimental validation of the method is performed and presented on a diverse set of data.

We then address the problem of associating objects across multiple cameras in planar scenes. Since cameras may be moving, there is a possibility of both spatial and temporal non-overlap in the fields of view of the camera. We first address the case where spatial and temporal overlap can be assumed. Since the cameras are moving and often widely separated, direct appearance-based or proximity-based constraints cannot be used. Instead, we exploit geometric constraints on the relationship between the motion of each object across cameras, to test multiple correspondence hypotheses, without assuming any prior calibration information. Here, there are three contributions. First, we present a statistically and geometrically meaningful means of evaluating a hypothesized correspondence between multiple objects in multiple cameras. Second, since multiple cameras exist, ensuring coherency in association, i.e. transitive closure is maintained between more than two cameras, is an essential requirement. To ensure such coherency we pose the problem of object associating across cameras as a k-dimensional matching and use an approximation to find the association. We show that, under appropriate conditions, re-entering objects can also be re-associated to their original labels. Third, we show that as a result of associating objects across the cameras, a concurrent visualization of multiple aerial video streams is possible. Results are shown on a number of real and controlled scenarios with multiple objects observed by multiple cameras, validating our qualitative models.

Finally, we present a unifying framework for object association across multiple cameras and for estimating inter-camera homographies between (spatially and temporally) overlapping and non-overlapping cameras, whether they are moving or non-moving. By making use of explicit polynomial models for the kinematics of objects, we present algorithms to estimate inter-frame homographies. Under an appropriate measurement noise model, an EM algorithm is applied for the maximum likelihood estimation of the inter-camera homographies and kinematic parameters. Rather than fit curves locally (in each camera) and match them across views, we present an approach that simultaneously refines the estimates of inter-camera homographies and curve coefficients *globally*. We demonstrate the efficacy of the approach on a number of real sequences taken from aerial cameras, and report quantitative performance during simulations.

To my parents,

Shahina and Muhammad Ajmal Sheikh,

for their inspiration,

for their love,

for their sacrifices.

~

To my wife,

Erum Arif Khan,

for her companionship,

and for humanizing my life.

ACKNOWLEDGMENTS

I would like to thank Dr. Shah for his guidance and direction. He taught me the worth of single-minded determination and hard-work and many, many lessons beyond those. I am grateful to the 'old-guard' of the Computer Vision Lab at UCF, Sohaib Khan, Omar Javed, Khurram Shafique, Zeeshan Rasheed and Cen Rao, for their company, advice, and for taking many falls so I didn't have to. I'm indebted to the newer members too, Alexei Gritai, Asaad Hakeem, Yun Zhai, Paul Smith and Saad Ali, for enduring my many late-night rants and for their (sometimes unwilling) service as soundboards! I gratefully acknowledge Dr. Xin Li for the invaluable advice he provided during our discussions. I would like to thank Niels Haering for the enjoyable experience I had during my internship. Finally, I thank Drs. Charles Hughes, Annie Wu and Huaxin You for serving on my committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

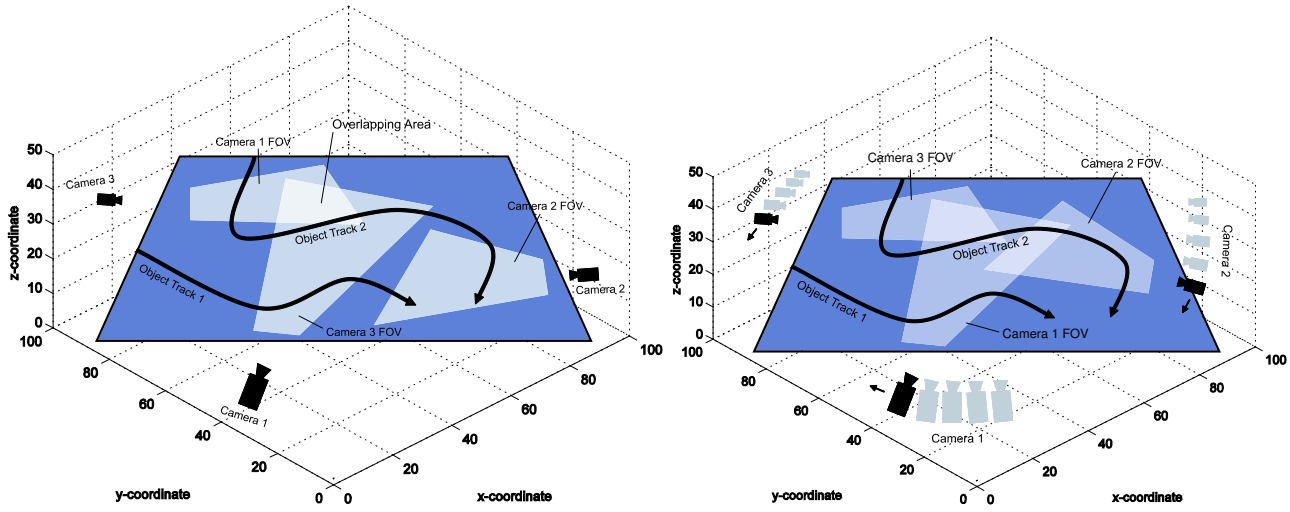
INTRODUCTION AND NOTATION

"*Birds do it, Bees do it, Even educated fleas do it.*" - Cole Porter

At its lowest abstraction, information about the world comes to us through the elementary senses of sight, smell, touch, taste and sound and from it we reconstruct a particular perception of the world. Of these different modalities, the sense that is thought to be most fundamental to the human experience is the sense of sight. Studies indicate that over 40% of the human brain is dedicated to processing visual information, and near 80% of a human child's first 12 years of learning is through vision, [Big06, Whi98, RKK02, VAF92]. Vision and hearing are also intrinsically cooperative, using stereoscopy and stereophony, for example, to infer depth. In fact, recent research has uncovered evidence that rats use the sense of smell in stereo to locate the source of a scent, [RCB06]. Cooperative sensing, in a more distributed sense, is exploited by a number of species, from schools of fish to flocks of birds, for achieving many diverse goals such as foraging of food, evading predators and transportation. Cooperative sensing is also widely used in human society for relatively localized tasks like guarding prisons or refereeing sports games to more sophisticated,

global data collection operations by intelligence agencies or pollsters. In itself, sensing collectively presents an interesting paradigm: solving a difficult global sensing problem, with an ensemble of efficient, but simpler local sensors. In this dissertation, we introduce this paradigm to the problem of scene understanding over wide planar dynamic scenes. Dynamic scenes (as opposed to static scenes) are scenes containing non-stationary objects such as moving vehicles and pedestrians and/or non-stationary backgrounds such as water rippling or grass swaying in the wind.

The concept of a cooperative multi-camera ensemble, informally a 'forest' of cameras [LRS00], has recently received increasing attention from the research community. The idea is of great practical relevance, since cameras typically have limited fields of view, but are now available at low costs. Thus, instead of having a single high-resolution camera with a wide field of view that surveys a large area, far greater flexibility and scalability can be achieved by observing a scene 'through many eyes', using a multitude of lower-resolution COTS (commercial off-the-shelf) cameras. It is difficult to survey wide areas using one sensor due to occlusions in the scene and the trade-off between resolution and field of view. Several approaches with varying constraints have been proposed, highlighting the wide applicability of co-operative sensing in practice. For instance, the problem of associating objects across multiple *stationary* cameras with overlapping fields of view has been addressed in a number of papers, e.g. [NKI98], [QA99], [CG01], [DT01], [MD03], [KHM00], [DEP95], [AP96], [LRS00] and [KS95]. Extending the problem to associating across cameras with non-overlapping fields of view, geometric and appearance based approaches have also been proposed recently, e.g. [HR97], [KZ99], [CLF01], [JRS03], and [SG00]. Motion too has been introduced to the 'forest', where correspondence is estimated across pan-tilt-zoom cameras,



(a) (b)

Figure 1.1: (a) Multiple stationary cameras observe a scene with moving objects, two of the camera FOVs are overlapping and one is not. (b) Multiple moving cameras observing a scene with moving objects. The camera FOVs can move in and out of overlap.

[MU02], [CAK02] and [KI96]. Allowing motion is particularly attractive since it allows much wider areas to be monitored by fewer cameras and introduces the possibility of active target tracking. In general, when using sensors in such a decentralized but cooperative fashion, knowledge of inter-camera relationships becomes of paramount importance in understanding what happens in the environment. Without such information it is difficult to tell, for instance, whether an object viewed in each of two cameras is the same object or a new object. These inter-camera relationships may be in the form of prior knowledge of relative positions and this information can be assumed known, through calibration, or can otherwise be learned over a training period. In the scenario under study in this dissertation (see Figure 1.1), obtaining calibration information usually requires sophisti-

cated equipment, such as a global positioning system (GPS) or an inertial navigation system (INS) (see [SKS03b] or [SKS03a]), perhaps with a geodetically aligned elevation map. Furthermore, the telemetry provided by such equipment is usually noisy, and for cameras mounted on aerial vehicles even nominal noise projects to errors of hundreds of meters due to the altitude of the aerial vehicle. As a result, approaches to recovering or refining inter-frame relationships based on video data are particularly useful.

Before objects can be associated *across* cameras, some degree of local sensing must occur at each camera. In this work, the local requirement at each camera is the detection of objects within each camera. The assumption that the sensor remains stationary (or that ego-motion can be compensated) between the incidence of each video frame allows the use of statistical background modeling techniques for the detection of moving objects such as [WAD97, SG00] and [EHD02]. Since 'interesting' objects in a scene are defined to be moving ones, such object detection provides a reliable foundation for other surveillance tasks like tracking ([HHD00, IB98, CRM00]) and is often also an important prerequisite for action or object recognition. However, the assumption of a stationary sensor does not necessarily imply a stationary *background*. Examples of 'nonstationary' background motion abound in the real world, including periodic motions, such as a ceiling fans, pendulums or escalators, and dynamic textures, such as fountains, swaying foliage or ocean ripples. The assumption that the sensor remains stationary is also often *nominally* violated by common phenomena such as wind or ground vibrations and to a larger degree by (stationary) hand-held cameras. Thus, if natural scenes are to be modeled it is essential that object detection algorithms operate reliably in such circumstances. Background modeling techniques have also been used for

foreground detection in pan-tilt-zoom cameras, [WM96]. Since the focal point does not change when a camera pans or tilts, planar-projective motion compensation can be performed to create a background mosaic model. Often, however, due to independently moving objects motion compensation may not be exact, and background modeling approaches that do not take such nominal misalignment into account usually perform poorly. Furthermore, for aerial video, which constitutes the primary source of test data in this dissertation, small misalignments and parallax can also violate a 'stationary' camera assumption. Thus, a principal proposition in this work is that modeling spatial uncertainties is important for real world deployment, and we describe an intuitive and novel representation of the scene background that consistently yields high detection accuracy.

1.1 Problem Stratification

In a planar scene with multiple cameras, there are several possible configurations that can arise. In literature, a distinction has been drawn between approaches that assume spatially overlapping and non-overlapping fields of view for stationary cameras[1]. Evidently, associating objects across stationary cameras can be treated as a special case of associating across *moving* cameras (where there is zero camera motion). If the cameras are moving independently, the fields of view of different cameras can alternatively move in and out of overlap and as a result the problem of correspondence becomes considerably more complicated than that of the stationary camera case. It is useful to think of the problem in terms of *spatio-temporal overlap*, analogous to spatial overlap

[1] A detailed literature review is provided in Chapter 2.

in the case of stationary cameras, i.e. for some duration of time, the FOV of each camera overlaps (spatially) with the FOV of another camera while observing the moving objects. In terms of spatio-temporal overlap, we identify four possible cases:

1. **Each object is simultaneously visible by all cameras, all the time:** In this instance, there is continuous spatial and temporal overlap between the fields of view. This is a reasonable assumption for stationary cameras configurations, but rarely occurs when cameras are continuously moving, especially over extended sequences.
2. **Each object is simultaneously visible by some cameras, all the time:** This is an instance of limited spatial overlap, where all objects are within the 'collective' field of view of all the cameras all the time (but not necessarily within *each* camera's field of view). This situation occurs most often when each camera is in pursuit of a separate target.
3. **Each object is simultaneously visible by some cameras for a limited duration of time:** This is the case where all objects are visible in some subset of cameras simultaneously. For stationary cameras that would mean (at least) pairwise overlap between fields of view.
4. **Each object is visible by some cameras, but not necessarily simultaneously:** This is the most general case where *spatiotemporal* overlap does not necessarily occur between any two cameras, while objects are visible in their field of view. Without making some strong assumptions about object or camera motion it is difficult to address this case. This case is the spatio-temporal analog of the problem of associating across stationary cameras with non-overlapping fields of view.

1.2 The Approach

In this dissertation, we present a unifying probabilistic framework that captures the underlying geometry of planar scenes, and present algorithms to estimate geometric relationship between different cameras. We subsequently use these relationships for scene understanding across a collection of cameras, gaining a global picture of the behavior of objects in the world. The thesis describes the local process of detecting objects in dynamic scenes, i.e. detecting moving objects of interest. To account for commonly encountered dynamic phenomena in video like temporal textures, nominal misalignments, and residual motion due to parallax we present a novel model of the entire background as a distribution in 5-space. We present a new constraint for object detection and demonstrate significant improvements in detection. The central criterion that is traditionally exploited for detecting moving objects is *background difference*, some examples being [JN79, WAD97, ORP00] and [SG00]. When an object enters the field of view it partially occludes the background and can be detected through background differencing approaches if its appearance differs from the portion of the background it occludes. Sometimes, however, during the course of an object's journey across the field of view, some colors may be similar to those of the background, and in such cases detection using background differencing approaches fails. To address this limitation and to improve detection in general, a new criterion called *temporal persistence* is presented here and exploited in conjunction with background difference for accurate detection. True foreground objects, as opposed to spurious noise, tend to maintain consistent colors and remain in the same spatial area (i.e. frame to frame color transformation and motion are small). Thus, fore-

ground information from the frame incident at time t contains substantial evidence for the detection of foreground objects at time $t + 1$. In this dissertation, this fact is exploited by maintaining both background and foreground models to be used competitively for object detection in stationary cameras, without explicit tracking. Finally, once pixel-wise probabilities are obtained for belonging to the background, decisions are usually made by direct thresholding. Instead, we assert that *spatial context* is an important constraint when making decisions about a pixel label, i.e. a pixel's label is not independent of the pixel's neighborhood labels (this can be justified on Bayesian grounds using Markov Random Fields, [GG84, Li95]). We introduce a MAP-MRF framework, that competitively uses both the background and the foreground models to make decisions based on spatial context. We demonstrate that the *maximum a posteriori* solution can be efficiently computed by finding the minimum cut of a capacitated graph, to make an optimal inference based on neighborhood information at each pixel.

Once detection and tracking (through any number of tracking algorithms) is performed locally, we turn our attention to global object association across cameras. There are three main objectives to be achieved: (1) Computing inter-camera associations, (2) Computing location parameters of the cameras, and (3) recovering best estimates of the true underlying trajectories that was viewed in the cameras. First we describe an approach that requires at least partial spatiotemporal overlap between the fields of view of the cameras (Case 3). This is the *minimal* assumption that is required by this approach to discern the relationship of observations in the uncalibrated moving cameras. We describe an extension to the re-projection error for the estimation of the set of homographies for multiple views, providing a geometrically and statistically sound means of evaluating the like-

lihood of a candidate association. We formulate the problem of maximizing this joint likelihood function as a k-dimensional matching problem and use an approximation that maintains transitive closure. The underlying concept of co-operative sensing is to use these relationships to give global context to 'locally' obtained information at each camera. It is desirable, therefore, that the data collected at each camera and the inter-camera relationship discerned by the system be presented in a coherent visualization. For moving cameras, particularly airborne ones where large swaths of areas may be traversed in a short period of time, coherent visualization is indispensable for applications like surveillance and reconnaissance. Thus, in addition to presenting an algorithm to track objects across multiple moving cameras with spatiotemporal overlap of fields of view, we provide a means to simultaneously visualize the collective field of view of all the airborne cameras and demonstrate that under special conditions, trajectories interrupted due to occlusion or missing detections can be repaired. For this approach, no constraints are placed on the object motion.

To include Case 4 configurations, explicit kinematic models of objects are used and this allows us to describe a general global object association algorithm. By including kinematic models, such as constant velocity, constant acceleration and higher order models we demonstrate that as long as the kinematic models are valid, global association of objects is possible. Under an appropriate measurement noise model, an EM algorithm is presented for the maximum likelihood estimation of the inter-camera homographies and the parameters of the kinematic model, where the associations are treated as hidden variables. Experiments are presented both qualitatively and quantitatively on videos collected from aerial cameras and on simulated data respectively.

1.3 Overview of the Thesis

The rest of the thesis is divided into five chapters, covering object detection, global association across multiple cameras with overlapping fields of view, general global association and a chapter on concluding remarks. In Chapter 2, we review different solutions proposed in current literature to the problem stratifications described earlier and place the ideas contained in this thesis in context of earlier work. In Chapter 3, we describe an algorithm to detect objects in dynamic scenes, in the presence of dynamic textures, misalignments and residual parallax motion. A general MAP-MRF decision framework is used and graph cuts are used to efficiently maximize the posterior term. In Chapter 4, algorithms that require at least limited spatio-temporal overlap between the fields of view of the cameras (Case 3) are described. The planarity assumption is exploited to posit the existence of a homography between corresponding trajectories, which is then used to associate object across frames and compute maximum likelihood estimates of inter-camera homographies and the 'true' underlying trajectories. Finally, in Chapter 5, we propose a unifying framework for learning inter-camera homographies between overlapping and non-overlapping cameras (both spatially and temporally), whether they are moving or non-moving (Case 4). Kinematic models of the objects are used to associate object across views and the Expectation Maximization algorithm is used to compute the maximum likelihood estimate of the inter-camera homographies and the kinematic model parameters.

CHAPTER 2

LITERATURE REVIEW

Since the seminal work of Sittler in [Sit64] on data association, multitarget-multisensor tracking has been extensively studied in the past three decades. In the data association community it is typically assumed that the sensors are calibrated and data is available in a common coordinate system (a good summary is available in [Ed90]). Optimal multi-target multi-sensor association is known to be NP-Hard [GJ79] and with n sensors and k objects there are $(k!)^n$ possible configurations which makes exhaustive evaluation computationally prohibitive. Sequential logic techniques include nearest neighbor filters, strongest neighbor filters, one-to-few assignments and one-to-many assignments. These methodologies are computationally very efficient, but since decisions are irreversible at each time step they are prone to error rates particularly when the number of objects is large. Deferred logic techniques typically use some form of multiple hypothesis testing and many variations have been proposed in literature. It was shown in [Poo94] and [PDB90] that this data association problem can be formulated as a multi-dimensional assignment problem. The analysis contained in this area is important and many ideas are relevant, however these approaches assume a registered setting with overlapping fields of view which cannot be used directly in the context of this work, where the coordinate systems differ up to a homography.

In this chapter we provide a context for our algorithms in the backdrop of previous work. Prior work can be broadly classified into three categories based on the assumptions they make on the camera setup: (1) Multiple stationary cameras with overlapping fields of view, (2) Multiple stationary cameras with non-overlapping fields of view and (3) Multiple pan-tilt-zoom cameras. In addition, we review related work on object detection in single cameras and discuss the limitations that we address.

2.1 Multiple Stationary Cameras with Overlapping Fields of View

By far, the largest body of work in associating objects across multiple camera make the assumptions that the cameras are stationary and have overlapping fields of view. The earliest work involving associating objects across cameras with overlapping fields of view stemmed from an interest in Multiple Perspective Interative Video in the early 90s, in which users observing a scene selected particular views from multiple perspectives. In [SMK94], Sato *et al.*, CAD based environment models were used to extract 3D locations of unknown moving objects; once objects entered overlapping views of two agents, stereopsis was used to recover exact 3D positions. Jain and Wakimoto, [JW95], also assumed calibrated cameras to obtain 3D locations of each object in an environment model for Multiple Perspective Interactive video. Although the problem of associating objects across cameras was not explicitly addressed, several innovative ideas were proposed, such as choosing the best view given a number of cameras and the concept of interactive television. In [KKK95], Kelly *et al.* constructed a 3D environment model using the voxel feature. Humans

were modelled as a collection of these voxels and they used this model to resolve the camera-handoff problem. These works were characterized by the use environment models, and calibrated cameras.

Tracking across multiple views was addressed in its own right, in a series of papers from the latter half of the 90s. In [NKI98], Nakazawa *et al.* constructed a state transition map that linked regions observed by one or more cameras, along with a number of action rules to consolidate information between cameras. Cai and Aggarwal, [QA99], proposed a method to track humans across a distributed system of cameras, employing geometric constraints between neighboring cameras for tracking. Spatial matching was based on the Euclidean distance of a point with its corresponding epipolar line. Bayesian Networks were used in several papers as well. In [CG01], Chang and Gong used Bayesian networks to combine geometry (epipolar geometry, homographies and landmarks) and recognition (height and appearance) based modalities to match objects across multiple sequences. Bayesian networks were also used by Dockstader and Tekalp in [DT01] to track objects and resolve occlusions across multiple calibrated cameras. Integration of stereo pairs was another popular approach, adopted by Mittal and Davis [MD03], Krumm *et al.* [KHM00] and Darrell *et al* [DEP95].

Several approaches were proposed that did not require prior calibration of cameras, but instead learned minimal relative camera information. Azarbayejani and Pentland, [AP96], developed an estimation technique for recovering 3D object tracks and the multi-view geometry from 2D blob features. Lee *et al.*, [LRS00], made an assumption of scene planarity and learned the homography related views by robust sampling methods. They then recovered 3D camera and plane configura-

tions to construct a common coordinate system, and used this coordinate to analyze object motion across cameras. In [KS95], Khan *et al.* proposed an approach that avoided explicit calibration of cameras and instead used constraints on the field of view lines between cameras, learned during a training phase, to track objects across the cameras.

2.2 Multiple Stationary Cameras with Non-Overlapping Fields of View

The assumption of overlapping fields of view restricts the area over which cameras can be dispersed. It was realized that meaningful constraints could be applied to tracking objects across cameras with non-overlapping fields of view as well. This allowed the collective field of view of the system of cameras to be dispersed over a far wider area. In the research community, this sub-field seems to initially have been an offshoot of object recognition, where it was viewed as a problem of recognizing objects previously viewed in other cameras. A representative work was [HR97], in which Huang and Russell proposed a probabilistic appearance based approach for tracking vehicles across consecutive cameras on a highway. Constraints on the motion of the objects across cameras were first proposed by Kettnaker and Zabih, [KZ99], where positions, object velocities and transition times across cameras were used in a setup of known path topology and transition probabilities. In [CLF01], Collins *et al.* used a system of calibrated cameras with an environment model to track objects across multiple views. The method proposed by Javed *et al.*, in [JRS03], did not assume a site model or explicit calibration of cameras, instead they learned the inter-camera illumination and transition properties during a training phase, which were then used to track ob-

jects across the cameras. Recently in [SG00], Stauffer and Tieu tracked across multiple cameras with both overlapping and non-overlapping fields of view, building a correspondence model for the entire set of cameras. They made an assumption of scene planarity and recovered the inter-camera homographies.

Some work has been published for recovering the pose and/or tracks between cameras with non-overlapping fields of view. Fisher, in [Fis02], showed that, given a set of randomly placed cameras, recovering pose was tractable using distant moving features and nearby linearly moving features. In [MEB04], Makris *et al.* also extracted the topology of a number of cameras based on the co-occurrence of entries and exits. Rahimi *et al.*, in [RDD04], presented an approach that reconstructed the trajectory of a target and the external calibration parameters of the cameras, given the location and velocity of each object.

2.3 Multiple Pan-Tilt-Zoom Cameras

So far, the discussion has addressed approaches that assumed the camera remained stationary, with overlapping and non-overlapping FOVs. Clearly, the collective field of view of the sensors can be further increased if motion is allowed in sensors. With the introduction of motion, the camera fields of view can be overlapping or non-overlapping at different times, and one of the challenges of tracking across moving cameras is that both situations need to be addressed. A limited type of camera motion has been examined in previous work: motion of the camera about the camera

center, i.e. pan-tilt-zoom (PTZ) motion. One such work is [MU02], where Matsuyama and Ukita present a system-based approach using active cameras, developing a fixed point PTZ camera for wide area imaging. In [KI96] Kang *et al.* proposed a method that involved multiple stationary and PTZ cameras. It was assumed that the scene was planar and that the homographies *between* cameras were known. Using these transformations, a common coordinate frame was established and objects were tracked across the cameras using color and motion characteristics. A related approach was also proposed in [CAK02], where Collins *et al.* presented an active multiple camera system that maintained a single moving object centered in each view, using PTZ cameras.

2.4 Object Detection

Since the late 70s, differencing of adjacent frames in a video sequence has been used for object detection in stationary cameras, [JN79]. However, it was realized that straightforward background subtraction was unsuited to surveillance of real-world situations and statistical techniques were introduced to model the uncertainties of background pixel colors. In the context of this work, these background modeling methods can be classified into two categories: (1) Methods that employ *local* (pixel-wise) models of intensity and (2) Methods that have *regional* models of intensity.

Most background modeling approaches tend to fall into the first category of pixel-wise models. Early approaches operated on the premise that the color of a pixel over time in a static scene could be modeled by a single Gaussian distribution, $N(\mu, \Sigma)$. In their seminal work, Wren *et*

al [WAD97] modeled the color of each pixel, $I(x, y)$, with a single 3 dimensional Gaussian, $I(x, y) \sim N(\mu(x, y), \Sigma(x, y))$. The mean $\mu(x, y)$ and the covariance $\Sigma(x, y)$, were learned from color observations in consecutive frames. Once the pixel-wise background model was derived, the likelihood of each incident pixel color could be computed and labeled as belonging to the background or not. Similar approaches that used Kalman Filtering for updating were proposed in [KBG90] and [KWH94]. A robust detection algorithm was also proposed in [HHD00]. While these methods were among the first to principally model the uncertainty of each pixel color, it was quickly found that the single Gaussian *pdf* was ill-suited to most outdoor situations, since repetitive object motion, shadows or reflectance often caused multiple pixel colors to belong to the background at each pixel. To address some of these issues, Friedman and Russell, and independently Stauffer and Grimson, [FR97], [SG00] proposed modeling each pixel intensity as a *mixture* of Gaussians, instead, to account for the multi-modality of the 'underlying' likelihood function of the background color. An incident pixel was compared to every Gaussian density in the pixel's model and if a match (defined by threshold) was found, the mean and variance of the matched Gaussian density was updated, or otherwise a new Gaussian density with the mean equal to the current pixel color and some initial variance was introduced into the mixture. Thus, each pixel was classified depending on whether the matched distribution represented the background process. While the use of Gaussian mixture models was tested extensively, it did not explicitly model the *spatial dependencies* of neighboring pixel colors that may be caused by a variety of real nominal motion. Since most of these phenomena are 'periodic', the presence of multiple models describing each pixel mitigates this effect somewhat by allowing a mode for each periodically observed pixel

intensity; however, performance notably deteriorates since dynamic textures usually do not repeat exactly (see experiments in Section 3.4). Another limitation of this approach is the need to specify the number of Gaussians (models), for the E-M algorithm or the *K*-means approximation. Still, the mixture of Gaussian approach has been widely adopted, becoming something of a standard in background subtraction, as well as a basis for other approaches ([JSS02],[Har02]).

Methods that address the uncertainty of spatial location using local models have also been proposed. In [EHD02], El Gammal *et al* proposed nonparametric estimation methods for per-pixel background modeling. Kernel density estimation (KDE) was used to establish membership, and since KDE is a data-driven process, multiple modes in the intensity of the background were also handled. They addressed the issue of nominally moving cameras with a local search for the best match for each incident pixel in neighboring models. Ren *et al* too explicitly addressed the issue of background subtraction in a nonstationary scene by introducing the concept of a spatial distribution of Gaussians (SDG), [RCH03]. After affine motion compensation, a MAP decision criteria is used to label a pixel based on its intensity and spatial membership probabilities (both modeled as Gaussian *pdf*s). There are two primary points of interest in [RCH03]. First, the authors modeled the spatial position as a *single* Gaussian, negating the possibility of bimodal or multi-modal *spatial* probabilities, i.e. that a certain background element model may be expected to occur in more than one position. Although, not within the scope of their problem definition, this is, in fact, a definitive feature of a temporal texture. Analogous to the need for a mixture model to describe intensity distributions, unimodal distributions are limited in their ability to model spatial uncertainty. 'Nonstationary' backgrounds have most recently been addressed by Pless *et*

al [PLS03] and Mittal *et al* [MP04]. Pless *et al* proposed several pixel-wise models based on the distributions of the image intensities and spatio-temporal derivatives. Mittal *et al* proposed an adaptive kernel density estimation scheme with a joint pixel-wise model of color (for a normalized color space), and optical flow at each pixel. Other notable pixel-wise detection schemes include [SRP00], where topology free HMMs are described and several state splitting criteria are compared in context of background modeling, and [RKJ00], where a (practically) non-adaptive three-state HMM is used to model the background.

The second category of methods use region models of the background. In [TKB99], Toyama *et al* proposed a three tiered algorithm that used region based (spatial) scene information in addition to per-pixel background model: region and frame level information served to verify pixel-level inferences. Another global method proposed by Oliver *et al* [ORP00] used eigenspace decomposition to detect objects. For k input frames of size $N \times M$ a matrix $\mathbf{B}$ of size $k \times (NM)$ was formed by row-major vectorization of each frame and eigenvalue decomposition was applied to $\mathbf{C} = (\mathbf{B} - \mu)^T(\mathbf{B} - \mu)$. The background was modeled by the eigenvectors corresponding to the η largest eigenvalues, $\mathbf{u}_i$, that encompass possible illuminations in the field of view (FOV). Thus, this approach is less sensitive to illumination. The foreground objects are detected by projecting the current image in the eigenspace and finding the difference between the reconstructed and actual images. The most recent region-based approaches are by Monnet *et al* [MMP03], Zhong *et al* [ZS03]. Monnet *et al* and Zhong *et al* simultaneously proposed models of image regions as an autoregressive moving average (ARMA) process, which is used to incrementally learn (using PCA) and then predict motion patterns in the scene.

The foremost assumption made in background modeling is the assumption of a stationary scene. However, this assumption is violated fairly regularly, through common real world phenomenon like swaying trees, water ripples, fountains, escalators etc. The local search proposed in [EHD02], the SDG of [RCH03], the time series models of [MMP03], [ZS03] and KDEs over color and optical flow in [MP04] are several formulations proposed for detection non-stationary backgrounds. While each method demonstrated degrees of success, the issue of spatial dependencies has not been addressed in a principled manner. In context of earlier work (in particular [MP04]), our approach falls under the category of methods that employ regional models of the background. We assert that useful correlation exists in the intensities of spatially proximal pixels and this correlation can be used to allow high levels of detection accuracy in the presence of general non-stationary phenomenon.

2.5 Formulation

In the presented work, objects are to be tracked across several cameras, each mounted on aerial vehicles, without any telemetry or calibration information (see Figure 1). Unlike earlier approaches involving PTZ cameras, we track objects across cameras while the camera center is allowed to move freely. Such a system finds obvious application in the monitoring of large areas where several aerial vehicles provide different views of the scene, with alternately overlapping and non-overlapping fields of view. Since the cameras are moving and are often distant, direct appearance-based or proximity-based constraints cannot be used. Instead, we exploit constraints

on the relationship between the motion of each observed object across cameras. The principal assumption that is made in this work is that the altitude of the camera allows the scene to be modeled closely by a plane. Scene planarity in turn allows geometric constraints to be used for evaluating the probability that trajectories observed in two different sequences originated from the same object.

In the context of multiple objects viewed by multiple cameras, global coherency is desired in object tracking, i.e. multiple assignments are not made to a single object and that transitive closure is maintained in correspondence across multiple views. We formulate the problem in probabilistic terms, obtaining the Maximum Likelihood assignment of objects using graph matching. We show that reassociation of re-entering objects is possible under certain conditions. In addition, while mosaics provide an excellent means of summarizing aerial video information from a *single* view, trying to simultaneously monitor information from several mosaics is awkward and inconvenient. Instead, we show that as a consequence of automatically tracking objects across multiple views, a *concurrent mosaic* can be computed summarizing the information from several aerial videos.

This detection approach has three novel contributions. First, the method presented here provides a principled means of modeling the spatial dependencies of observed intensities. The model of image pixels as independent random variables, an assumption almost ubiquitous in background subtraction methods, is challenged and it is further asserted that there exists useful structure in the spatial proximity of pixels. This structure is exploited to sustain high levels of detection accuracy in the presence of nominal camera motion and dynamic textures. By using nonparametric density estimation methods over a joint domain-range representation, the background data is modeled as

a single distribution and multi-modal spatial uncertainties can be directly handled. Second, unlike previous approaches, the foreground is explicitly modeled to augment the detection of objects without using tracking information. The criterion of temporal persistence is exploited for simultaneous use with the conventional criterion of background difference. Third, instead of directly applying a threshold to membership probabilities, which implicitly assumes independence of labels, we present a MAP-MRF framework that competitively uses the foreground and background models for object detection, while enforcing spatial context in the process.

CHAPTER 3

OBJECT DETECTION

Before associating objects across cameras, some degree of local processing much be performed within each camera. In this chapter we describe an approach to object detection that takes does not make the pixel-wise independence assumption, and as a result can provide high quality detection in the presence of many real world phenomena such as dynamic textures (water waves, foliage swaying in the wind etc) and nominal misalignments. The ability to handle nominal misalignments is critical because a primary scenario where the planarity assumption is valid is video taken from aerial vehicles. In these videos motion is compensated using frame-to-frame motion compensation methods such as [MP97][1]. However, due to the presence of outlier motions (from the independently moving objects) and parallax, nominal misalignments and residual parallax motion can be expected and object detection methods that do not account for spatial correlation perform poorly. To account for these issues we now describe a novel representation of the background, the use of temporal persistence to pose object detection as a binary classification problem, and the overall MAP-MRF decision framework. For an image of size $M \times N$, let $\mathcal{S}$ discretely and regularly index

[1] Since independently moving objects are expected robust estimation methods must be used when compensating motion.

the image lattice, $\mathcal{S} = \{(i,j)|\ 1 \leq i \leq N, 1 \leq j \leq M\}$. The objective is to assign a binary label from the set $\mathcal{L} = \{\text{background}, \text{foreground}\}$ to each of the sites in $\mathcal{S}$.

3.1 Joint Domain-Range Background Model

If the primary source of spatial uncertainty of a pixel is image misalignment, a Gaussian density would be an adequate model since the corresponding point in the subsequent frame is equally likely to lie in any direction. However, in the presence of dynamic textures, cyclic motion, and non-stationary backgrounds in general, the 'correct' model of spatial uncertainty often has an arbitrary shape and may be bi-modal or multi-modal, but structure exists because by definition, the motion follows a certain repetitive pattern. Such arbitrarily structured data can be best analyzed using nonparametric methods since these methods make no underlying assumptions on the shape of the density. Non-parametric estimation methods operate on the principle that dense regions in a given feature space, populated by feature points from a class, correspond to the modes of the 'true' *pdf*. In this work, analysis is performed on a feature space where the p pixels are represented by $\mathbf{x}_i \in \mathbb{R}^5$, $i = 1, 2, \ldots p$. The feature vector, $\mathbf{x}$, is a joint domain-range representation, where the space of the image lattice is the *domain*, (x, y) and some color space, for instance (r, g, b), is the *range*, [CM02]. Using this representation allows a *single* model of the entire background, $f_{R,G,B,X,Y}(r, g, b, x, y)$, rather than a collection of pixel-wise models. Pixel-wise models ignore the dependencies between proximal pixels and it is asserted here that these dependencies are important. The joint representation provides a direct means to model and exploit this dependency.

In order to build a background model, consider the situation at time t, before which all pixels, represented in 5-space, form the set $\psi_b = \{\mathbf{y}_1, \mathbf{y}_2 \dots \mathbf{y}_n\}$ of the background. Given this sample set, at the observation of the frame at time t, the probability of each pixel-vector belonging to the background can be computed using the kernel density estimator ([Par62], [Ros56]). The kernel density estimator is a nonparametric estimator and under appropriate conditions the estimate it produces is a valid probability itself. Thus, to find the probability that a candidate point, $\mathbf{x}$, belongs to the background, ψ_b, an estimate can be computed,

$$P(\mathbf{x}|\psi_b) = n^{-1} \sum_{i=1}^{n} \varphi_{\mathbf{H}}\Big(\mathbf{x} - \mathbf{y}_i\Big), \tag{3.1}$$

where $\mathbf{H}$ is a symmetric positive definite $d \times d$ bandwidth matrix, and

$$\varphi_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} \varphi(\mathbf{H}^{-1/2}\mathbf{x}), \tag{3.2}$$

where φ is a d-variate kernel function usually satisfying $\int \varphi(\mathbf{x})d\mathbf{x} = 1$, $\varphi(\mathbf{x}) = \varphi(-\mathbf{x})$, $\int \mathbf{x}\varphi(\mathbf{x})d\mathbf{x} = 0$, $\int \mathbf{x}\mathbf{x}^T\varphi(\mathbf{x})d\mathbf{x} = \mathbf{I}_d$ and is also usually compactly supported. The d-variate Gaussian density is a common choice as the kernel φ,

$$\varphi_{\mathbf{H}}^{(\mathcal{N})}(\mathbf{x}) = |\mathbf{H}|^{-1/2}(2\pi)^{-d/2} \exp\Big(-\frac{1}{2}\mathbf{x}^T\mathbf{H}^{-1}\mathbf{x}\Big). \tag{3.3}$$

It is stressed here, that using a Gaussian kernel does not make any assumption on the scatter of data in the feature space. The kernel function only defines the effective region of influence of each data point while computing the final probability estimate. Any function that satisfies the constraints specified after Equation 2, i.e. a valid pdf, symmetric, zero-mean, with identity covariance, can be used as a kernel. There are other functions that are commonly used, some popular alternatives to

the Gaussian kernel are the Epanechnikov kernel, the Triangular kernel, the Bi-weight kernel and the Uniform kernel, each with their merits and demerits (see [WJ95] for more details).

Within the joint domain-range feature space, the kernel density estimator explicitly models spatial dependencies, without running into difficulties of parametric modeling. Furthermore, since it is well known that the rgb axes are correlated, it is worth noting that kernel density estimation also accounts for this correlation. The result is a single model of the background.

Lastly, in order to ensure that the algorithm remains adaptive to slower changes (such as illumination change or relocation) a sliding window of length ρ_b frames is maintained. This parameter corresponds to the learning rate of the system.

3.1.1 Bandwidth Estimation

Asymptotically, the selected bandwidth $\mathbf{H}$ does not affect the kernel density estimate but in practice sample sizes are limited. Too small a choice of $\mathbf{H}$ and the estimate begins to show spurious features, too large a choice of $\mathbf{H}$ leads to an over-smoothed estimate, losing important structural features like multi-modality. In general, rules for choosing bandwidths are based on balancing bias and variance globally. Theoretically, the ideal or optimal $\mathbf{H}$ can be found by minimizing the mean-squared error,

$$MSE\{\hat{f}_{\mathbf{H}}(\mathbf{x})\} = E\{[\hat{f}_{\mathbf{H}}(\mathbf{x}) - f_{\mathbf{H}}(\mathbf{x})]^2\}, \tag{3.4}$$

where $\hat{f}$ is the estimated density and f is the true density. Evidently, the optimal value of $\mathbf{H}$ is data dependent since the MSE value depends on $\mathbf{x}$. However, in practice, one does not have access

to the true density function which is required to estimate the optimal bandwidth. Instead, a fairly large number of heuristic approaches have been proposed for finding $\mathbf{H}$. A survey is provided in [Tur93].

Adaptive estimators have been shown to considerably outperform (in terms of the mean squared error) the fixed bandwidth estimator, particularly in higher dimensional spaces, [Sai02]. In general two formulations of adaptive or variable bandwidth estimators have been considered [Jon90]. The first varies the bandwidth with the estimation point and is called the balloon estimator, given by,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}(\mathbf{x})}(\mathbf{x} - \mathbf{x_i})), \tag{3.5}$$

where $\mathbf{H}(\mathbf{x})$ is the bandwidth matrix at $\mathbf{x}$. The second approach, called the sample-point estimator, varies the bandwidth matrix depending on the sample point,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}(\mathbf{x_i})}(\mathbf{x} - \mathbf{x_i})). \tag{3.6}$$

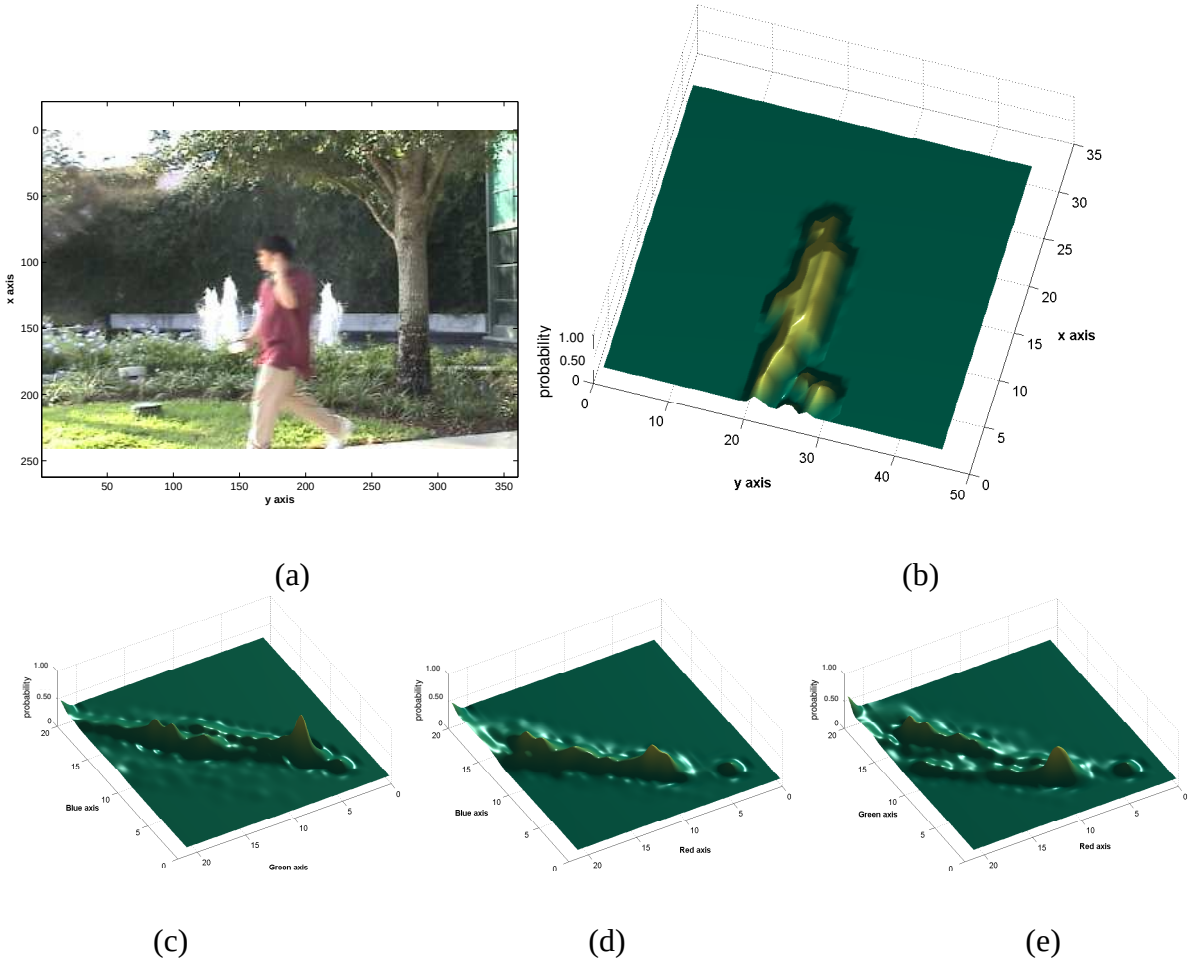
where $\mathbf{H}(\mathbf{x_i})$ is the bandwidth matrix at $\mathbf{x_i}$. However, developing variable bandwidth schemes for kernel density estimation is still research in progress, both in terms of theoretical understanding and in terms of practical algorithms, [Sai02].

In the given application, the sample size is large, and although it populates a 5 dimensional feature space, the estimate was found to be reasonably robust to the selection of bandwidth. Furthermore, choosing an optimal bandwidth in the MSE sense is usually highly computationally expensive. Thus, the balance between accuracy required (for matting, object recognition or action

recognition) and computational speed (for real-time surveillance systems) is application specific. To reduce the computational load, the Binned kernel density estimator provides a practical means of dramatically increasing computational speeds while closely approximating the kernel density estimate of Equation 3.1, ([WJ95], Appendix D). With appropriate binning rules and kernel functions the accuracy of the Binned KDE is shown to approximate the kernel density estimate in [HW95]. Binned versions of the adaptive kernel density estimate have also been provided in [Sai02]. To further reduce computation, the bandwidth matrix $\mathbf{H}$ is usually either assumed to be of the form $\mathbf{H} = h^2\mathbf{I}$ or $\mathbf{H} = diag(h_1^2, h_2^2, \ldots h_d^2)$. Thus, rather than selecting a fully parameterized bandwidth matrix, only two parameters need be defined, one for the variance in the spatial dimensions (x, y) and and one for the color channels, reducing computational load.

3.2 Modeling the Foreground

The intensity difference of interesting objects from the background has been, by far, the most widely used criterion for object detection. In this chapter, *temporal persistence* is presented as a property of real foreground objects, i.e. *interesting objects tend to remain in the same spatial vicinity and tend to maintain consistent colors from frame to frame*. The joint representation used here allows competitive classification between the foreground and background. To that end, models for both the background and the foreground are maintained. An appealing feature of this representation is that the foreground model can be constructed in a consistent fashion with the background model: a joint domain-range non-parametric density $\psi_f = \{\mathbf{z}_1, \mathbf{z}_2 \ldots \mathbf{z}_m\}$. Just as there was a



(a) (b)

(c) (d) (e)

Figure 3.1: Foreground Modeling. Using kernel density estimates on a model built from recent frames, the foreground can be detected in subsequent frames using the property of temporal persistence, (a) Current Frame (b) the X, Y-marginal, $f_{X,Y}(x, y)$. High membership probabilities are seen in regions where foreground in the current frame matches the recently detected foreground. The non-parametric nature of the model allows the arbitrary shape of the foreground to be captured accurately (c) the B, G-marginal, $f_{B,G}(b, g)$ (d) the B, R-marginal, $f_{B,R}(b, r)$ (e) the G, R-marginal, $f_{G,R}(g, r)$.

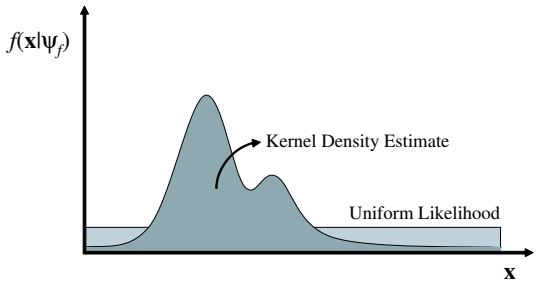


Figure 3.2: Foreground likelihood function. The foreground likelihood estimate is a mixture of the kernel density estimate and a uniform likelihood across the 5-space of features. This figure shows a conceptualization as a 1-D function.

learning rate parameter ρ_b for the background model, a parameter ρ_f is defined for the foreground frames. However, since the foreground changes far more rapidly than the background, the learning rate of the foreground is typically much higher than that of the background.

At any time instant the probability of observing a foreground pixel at any location (i, j) of any color is uniform. Then, once a foreground region is been detected at time t, there is an increased probability of observing a foreground region at time $t + 1$ in the same proximity with a similar color distribution. Thus, foreground probability is expressed as a mixture of a uniform function and the kernel density function,

$$P(\mathbf{x}|\psi_f) = \alpha\gamma + (1-\alpha)m^{-1}\sum_{i=1}^{m}\varphi_{\mathbf{H}}\Big(\mathbf{x} - \mathbf{z}_i\Big), \tag{3.7}$$

where $\alpha \ll 1$ is the mixture weight, and γ is a random variable with uniform probability, that is $\gamma_{R,G,B,X,Y}(r, g, b, x, y) = \frac{1}{R\times G\times B\times M\times N}$, where $0 \leq r \leq R,\ 0 \leq g \leq G,\ 0 \leq b \leq B,\ 0 \leq x \leq M,\ 0 \leq y \leq N$. This mixture is illustrated in Figure 3.2. If an object is detected in the preceding frame, the probability of observing the colors of that object in the same proximity increases

according to the second term in Equation 3.7. Therefore, as objects of interest are detected (the detection method will be explained presently), all pixels that are classified as 'interesting' are used to update the foreground model ψ_f. In this way, simultaneous models are maintained of both the background and the foreground, which are then used competitively to estimate interesting regions. Finally, to allow objects to become part of the background (e.g. a car having been parked or new construction in an environment), all pixels are used to update ψ_b. Figures 3.1 shows plots of some marginals of the foreground model.

At this point, whether a pixel vector $\mathbf{x}$ is 'interesting' or not can be competitively estimated using a simple *likelihood ratio classifier* (or a Parzen Classifier since likelihoods are computed using Parzen density estimates, [Fuk90]),

$$\tau = -\ln \frac{P(\mathbf{x}|\psi_b)}{P(\mathbf{x}|\psi_f)} = -\ln \frac{n^{-1}\sum_{i=1}^{n} \varphi_{\mathbf{H}}\Big(\mathbf{x}-\mathbf{y}_i\Big)}{\alpha\gamma + (1-\alpha)m^{-1}\sum_{i=1}^{m} \varphi_{\mathbf{H}}\Big(\mathbf{x}-\mathbf{z}_i\Big)} \tag{3.8}$$

Thus the classifier δ is,

$$\delta(\mathbf{x}) = \begin{cases} -1 & \text{if } -\ln \frac{P(\mathbf{x}|\psi_b)}{P(\mathbf{x}|\psi_f)} > \kappa \\ 1 & \text{otherwise} \end{cases}$$

where κ is a threshold which balances the trade-off between sensitivity to change and robustness to noise. The utility in using the foreground model for detection can be clearly seen in Figure 3.3. Figure 3.3(e) shows the likelihood values based only on the background model and Figure 3.3(f) shows the likelihood ratio based on both the foreground and the background models. In both histograms, two processes can be roughly discerned, a major one corresponding to the background pixels and a minor one corresponding to the foreground pixels. The variance *between* the clusters increases with the use of the foreground model. Visually, the areas corresponding to the tires of

the cars are positively affected, in particular. The final detection for this frame is shown in Figure 3.7(c). Evidently, the higher the likelihood of belonging to the foreground, the lower the overall likelihood ratio. However, as is described next, instead of using only likelihoods, prior information of neighborhood spatial context is enforced in a MAP-MRF framework. This removes the need to specify the arbitrary parameter κ.

3.3 Spatial Context: Estimation using a MAP-MRF Framework

The inherent spatial coherency of objects in the real world is often applied in a post-processing step, in the form of morphological operators like erosion and dilation, by using a median filter or by neglecting connected components containing only a few pixels, [SG00]. Furthermore, directly applying a threshold to membership probabilities implies conditional independence of labels, i.e. $P(\ell_i|\ell_j) = P(\ell_i), \text{ where } i \neq j$, and ℓ_i is the label of pixel i. We assert that such conditional independence rarely exists between proximal sites. Instead of applying such ad-hoc heuristics, Markov Random Fields provide a mathematical foundation to make a global inference using local information. While in some instances the morphological operators may do as well as the MRF for removing residual mis-detections at a reduced computational cost, there are two central reasons for using the MRF:

1. By selecting an edge-preserving MRF, the resulting smoothing will respect the object boundaries.

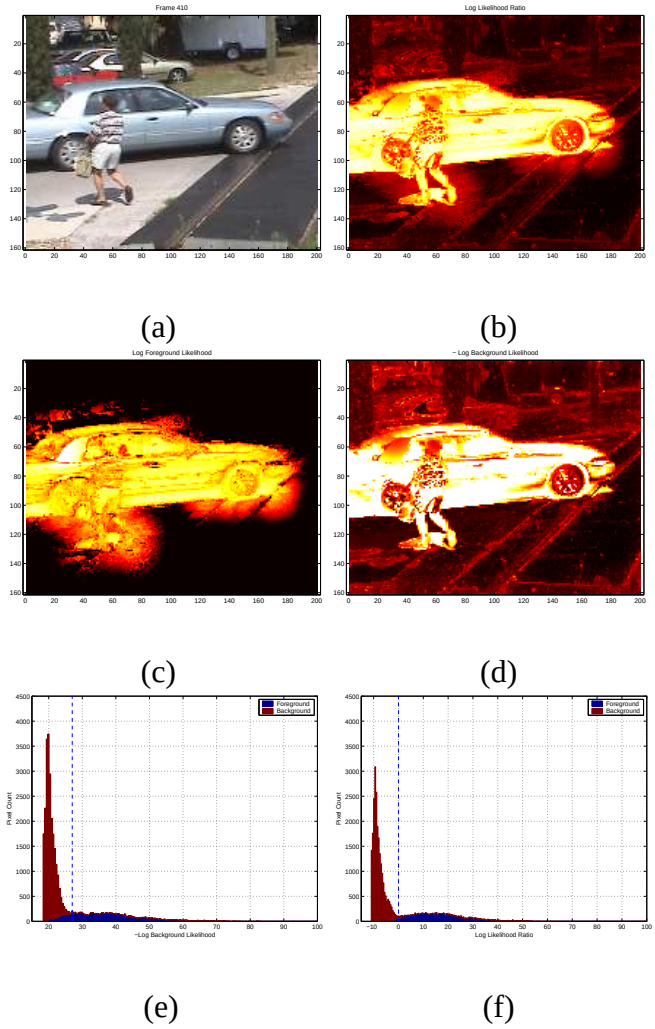
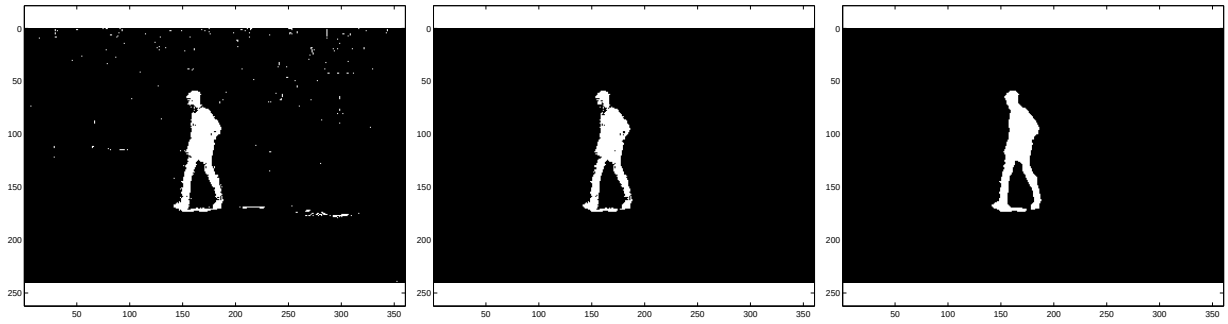


Figure 3.3: Improvement in discrimination using temporal persistence. Whiter values correspond to higher likelihoods of foreground membership. (a) Video Frame 410 of the Nominal Motion Sequence (b) Log-Likelihood Ratio values obtained using Equation 3.8. (c) Foreground likelihood map. (d) Background negative log-likelihood map. (e) Histogrammed negative log-likelihood values for background membership. The dotted line represents the 'natural' threshold for the background likelihood, i.e. $\log(\gamma)$. (f) Histogrammed log-likelihood ratio values. Clearly the variance *between* clusters is decidedly enhanced. The dotted line represents the 'natural' threshold for the log-likelihood ratio, i.e. zero.



(a) (b) (c)

Figure 3.4: Three possible detection strategies. (a) Detection by thresholding using only the background model of Equation 3.1. Noise can cause several spurious detections. (b) Detection by thresholding the Likelihood Ratio of Equation 3.8. Since some spurious detections do not persist in time, false positives are reduced using the foreground model. (c) Detection using MAP-MRF estimation, 3.13. All spurious detections are removed and false negative within the detected object are also removed as a result of their spatial context.

2. As will be seen, the formulation of the problem using the MRF introduces regularity into the final energy function that allows for the optimal partition of the frame (through computation of the minimum cut), without the need to pre-specify the parameter κ.

3. The MRF prior is precisely the constraint of spatial context we wish to impose on $\mathcal{L}$.

For the MRF, the set of neighbors, $\mathcal{N}$, is defined as the set of sites within a radius $r \in \mathbb{R}$ from site $\mathbf{i} = (i, j)$,

$$\mathcal{N}_{\mathbf{i}} = \{\mathbf{u} \in \mathcal{S} | \; \textit{distance}(\mathbf{i}, \mathbf{u}) \leq r, \mathbf{i} \neq \mathbf{u}\}, \tag{3.9}$$

where $distance(\mathbf{a}, \mathbf{b})$ denotes the Euclidean distance between the pixel locations $\mathbf{a}$ and $\mathbf{b}$. The 4-neighborhood (used in this chapter) and 8-neighborhood cliques are two commonly used neighborhoods. The pixels $\hat{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_p\}$ are conditionally independent given $\mathcal{L}$, with conditional density functions $f(\mathbf{x}_i|\ell_i)$. Thus, since each $\mathbf{x}_i$ is dependant on $\mathcal{L}$ only through ℓ_i, the likelihood function may be written as,

$$l(\hat{\mathbf{x}}|\mathcal{L}) = \prod_{i=1}^{p} f(\mathbf{x}_i|\ell_i) = \prod_{i=1}^{p} f(\mathbf{x}_i|\psi_f)^{\ell_i} f(\mathbf{x}_i|\psi_b)^{1-\ell_i}. \tag{3.10}$$

Spatial context is enforced in the decision through a pairwise interaction MRF prior. We use the Ising Model for its discontinuity preserving properties,

$$p(\mathcal{L}) \propto \exp\Big(\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\Big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\Big)\Big), \tag{3.11}$$

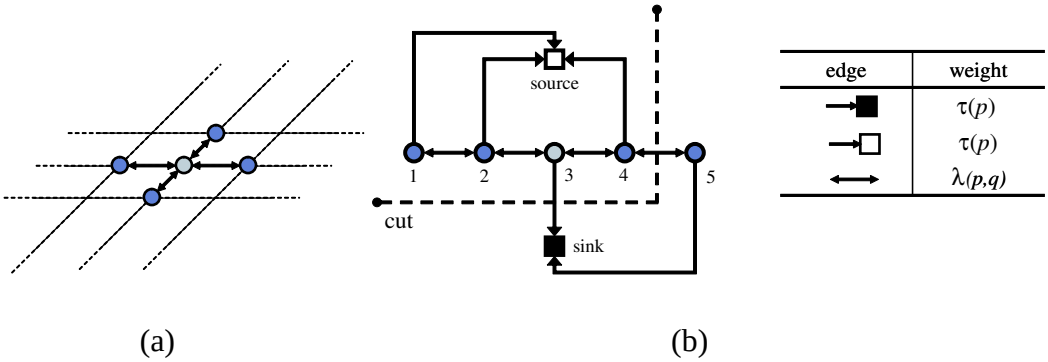
where λ is a positive constant and $i \neq j$ are neighbors. By Bayes Law, the posterior, $p(\mathcal{L}|\hat{\mathbf{x}})$, is then equivalent to

$$\begin{aligned} p(\mathcal{L}|\hat{\mathbf{x}}) &= \frac{p(\hat{\mathbf{x}}|\mathcal{L})p(\mathcal{L})}{p(\hat{\mathbf{x}})} \propto \\ &\frac{\Big(\prod_{i=1}^{p} f(\mathbf{x}_i|\psi_f)^{\ell_i} f(\mathbf{x}_i|\psi_b)^{1-\ell_i}\Big)p(\mathcal{L})}{p(\hat{\mathbf{x}}|\psi_f) + p(\hat{\mathbf{x}}|\psi_b)}. \end{aligned} \tag{3.12}$$

Ignoring constant terms, the log-posterior, $\ln p(\mathcal{L}|\hat{\mathbf{x}})$, is then equivalent to,

$$\begin{aligned} L(\mathcal{L}|\hat{\mathbf{x}}) = &\sum_{i=1}^{p} \ln\left(\frac{f(\mathbf{x}_i|\psi_f)}{f(\mathbf{x}_i|\psi_b)}\right)\ell_i + \\ &\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\Big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\Big). \end{aligned} \tag{3.13}$$

The MAP estimate is the binary image that maximizes L and since there are 2^{NM} possible configurations of $\mathcal{L}$ an exhaustive search is usually infeasible. In fact, it is known that minimizing



(a) (b)

Figure 3.5: A 4-neighborhood system. Each pixel location corresponds to a node in the graph, connected by a directed edge to the source and the sink, and by an undirected edge to it's four neighbors. For purposes of clarity the edges between node 3 and nodes 5 and 1 have been omitted in (b).

discontinuity-preserving energy functions in general is NP-Hard, [BVZ01]. Although, various strategies have been proposed to minimize such functions, e.g. Iterated Condition Modes [Bes86] or Simulated Annealing [GG84], the solutions are usually computationally expensive to obtain and of poor quality. Fortunately, since L belongs to the $\mathcal{F}^2$ class of energy functions, defined in [KZ04] as a sum of function of up to two binary variables at a time,

$$E(x_1, \ldots x_n) = \sum_i E^i(x_i) + \sum_{i,j} E^{(i,j)}(x_i, x_j), \tag{3.14}$$

and since it satisfies the regularity condition of the so-called $\mathcal{F}^2$ theorem, efficient algorithms exist for the optimization of L by finding the minimum cut of a capacitated graph, [GPS89, KZ04], described next.

To maximize the energy function (Equation 3.13), we construct a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with a 4-neighborhood system $\mathcal{N}$ as shown in Figure 3.5. In the graph, there are two distinct terminals s and t, the sink and the source, and n nodes corresponding to each image pixel location, thus

Algorithm

Initialize ψ_b using 1^{st} frame, $\psi_f = \emptyset$. At frame t, for each pixel,

Detection Step

1. Find $P(\mathbf{x}_i|\psi_f)$ (Eq. 3.7) and $P(\mathbf{x}_i|\psi_b)$ (Eq. 3.1) and compute the Likelihood Ratio τ (Eq. 3.8).
2. Construct the graph to minimize Equation 3.13.

Model Update Step

1. Append all pixels detected as foreground to the foreground model ψ_f.
2. Remove all pixels in ψ_f from ρ_f frames ago.
3. Append all pixels of the image to the background model ψ_b.
4. Remove all pixels in ψ_b from ρ_b frames ago.

Figure 3.6: Object Detection Algorithm

$\mathcal{V} = \{v_1, v_2, \cdots, v_n, s, t\}$. A solution is a two-set *partition*, $\mathcal{U} = \{s\} \cup \{i|\ell_i = 1\}$ and $\mathcal{W} = \{t\} \cup \{i|\ell_i = 0\}$. The graph construction is as described in [GPS89], with a directed edge (s, i) from s to node i with a weight $w_{(s,i)} = \tau_i$ (the log-likelihood ratio), if $\tau_i > 0$, otherwise a directed edge (i, t) is added between node i and the sink t with a weight $w_{(i,t)} = -\tau_i$. For the second term in Equation 3.13, undirected edges of weight $w_{(i,j)} = \lambda$ are added if the corresponding pixels are neighbors as defined in $\mathcal{N}$ (in our case if j is within the 4-neighborhood clique of i) . The capacity of the graph is $C(\mathcal{L}) = \sum_i \sum_j w_{(i,j)}$, and a cut defined as the set of edges with a vertex

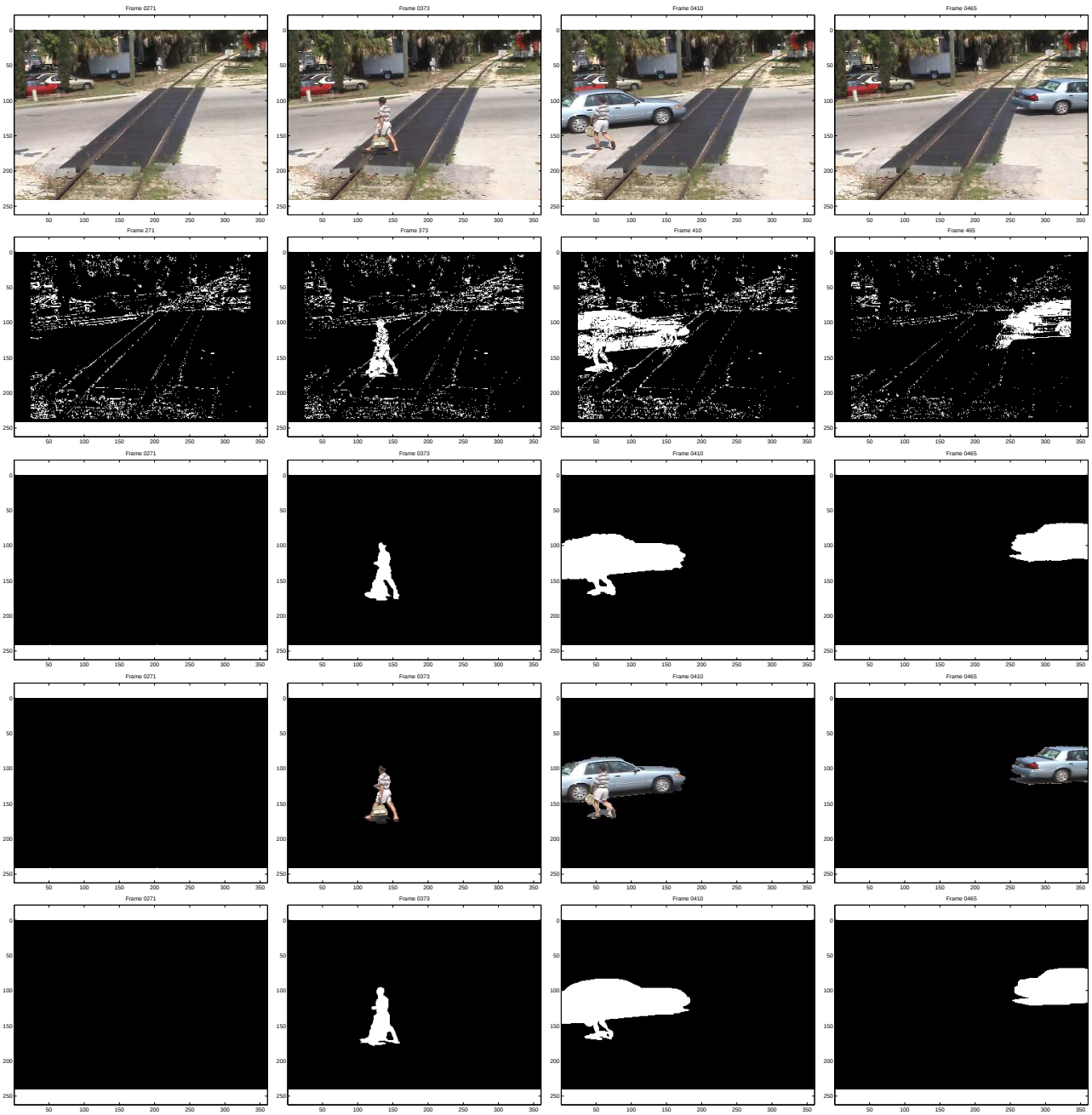
in $\mathcal{U}$ and a vertex in $\mathcal{W}$. As shown in [FF62], the minimum cut corresponds to the maximum flow, thus maximizing $L(\mathcal{L}|\hat{\mathbf{x}})$ is equivalent to finding the minimum cut. The minimum cut of the graph can be computed through a variety of approaches, the Ford-Fulkerson algorithm or a faster version proposed in [GPS89]. The configuration found thus corresponds to an optimal estimate of $\mathcal{L}$. The complete algorithm is described in Figure 4.4.

3.4 Results and Discussion

The algorithm was tested on a variety of sequences in the presence of nominal camera motion, dynamic textures, and cyclic motion. The sequences were all taken with a COTS camera (the Sony DCR-TRV 740). Comparative results for the mixture of Gaussians method have also been shown. For all the results the bandwidth matrix $\mathbf{H}$ was parameterized as a diagonal matrix with three equal variances pertaining to the range (color), represented by h_r and two equal variances pertaining to the domain, represented by h_d. The values used in all experiments were $(h_r, h_d) = (16, 25)$.

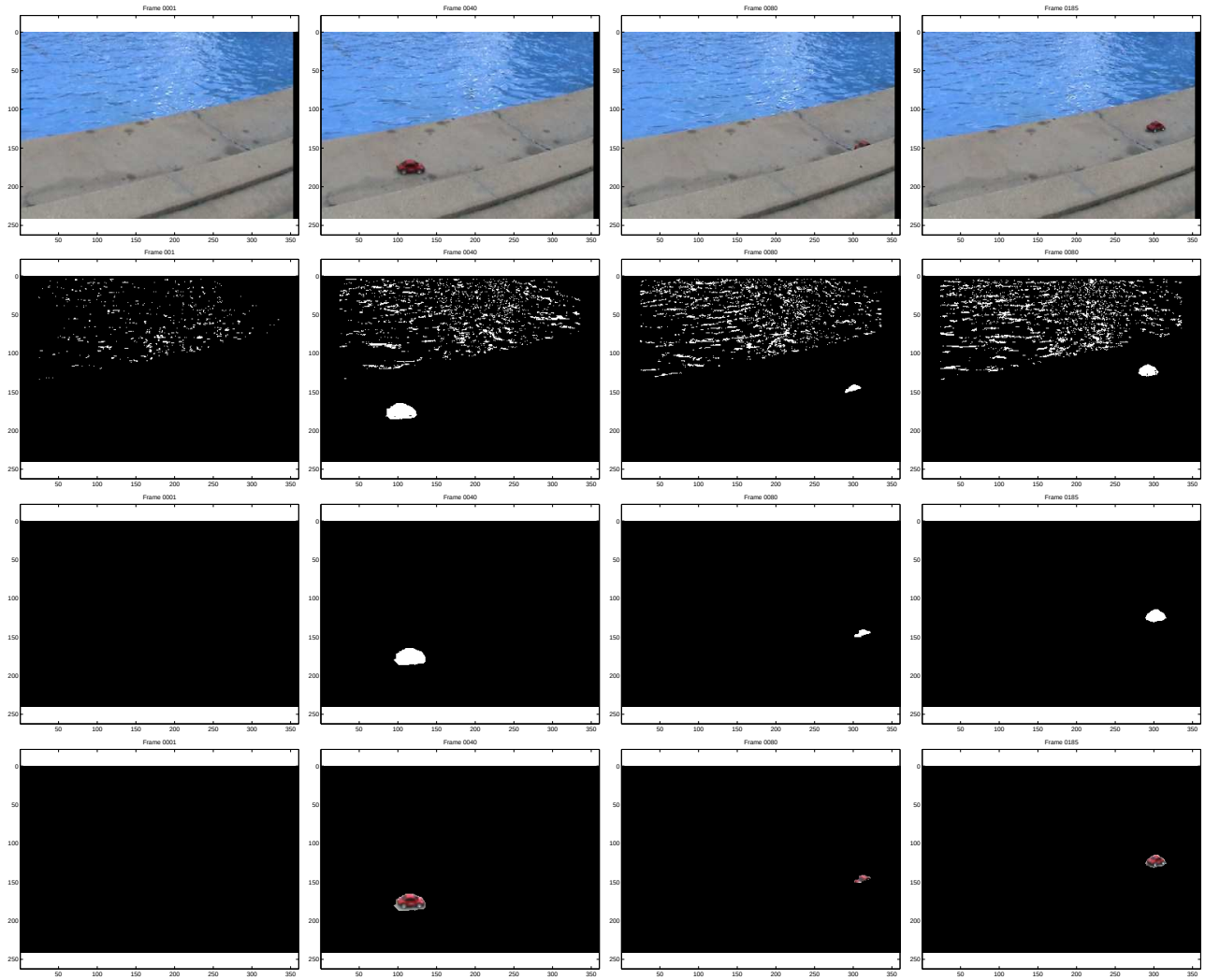
3.4.1 Qualitative Analysis

Qualitative results on seven sequences of dynamic scenes are presented in this section. The first sequence that was tested involved a camera mounted on a tall tripod. The wind caused the tripod to sway back and forth causing nominal motion of the camera. Figure 3.7 shows the results ob-



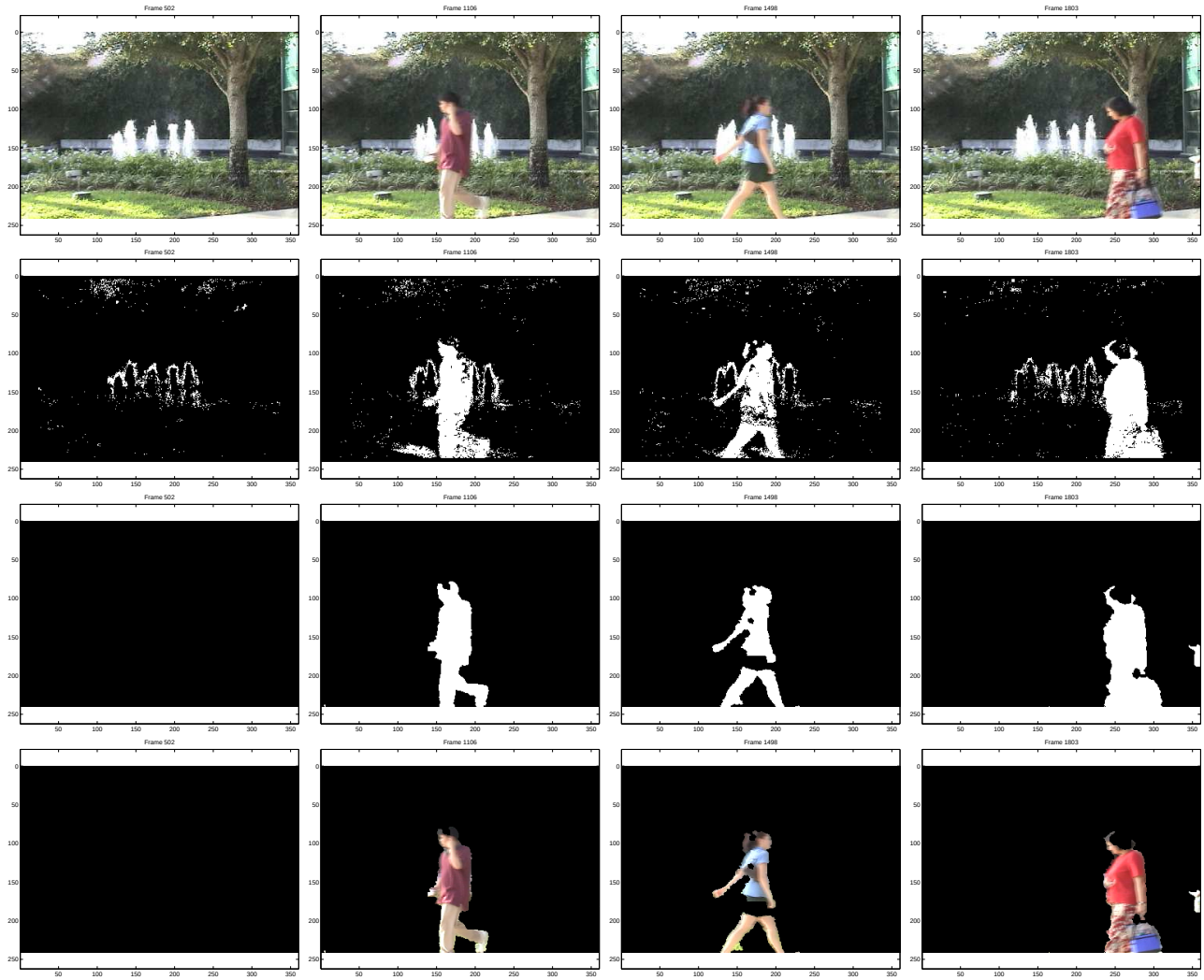
(a) (b) (c) (d)

Figure 3.7: Background Subtraction in a nominally moving camera (motion is an average of 12 pixels). The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row results obtained by our method. The fourth row is the masked original image. The fifth row is the manual segmentation. Morphological operators were not used in the results.



(a) (b) (c) (d)

Figure 3.8: Poolside sequence. The water in this sequence shimmers and ripples causing false positive in conventional detection algorithms, as a remote controlled car passes on the side. The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row are the results obtained by our method. The fourth row is the masked original image. Morphological operators were not used in the results.



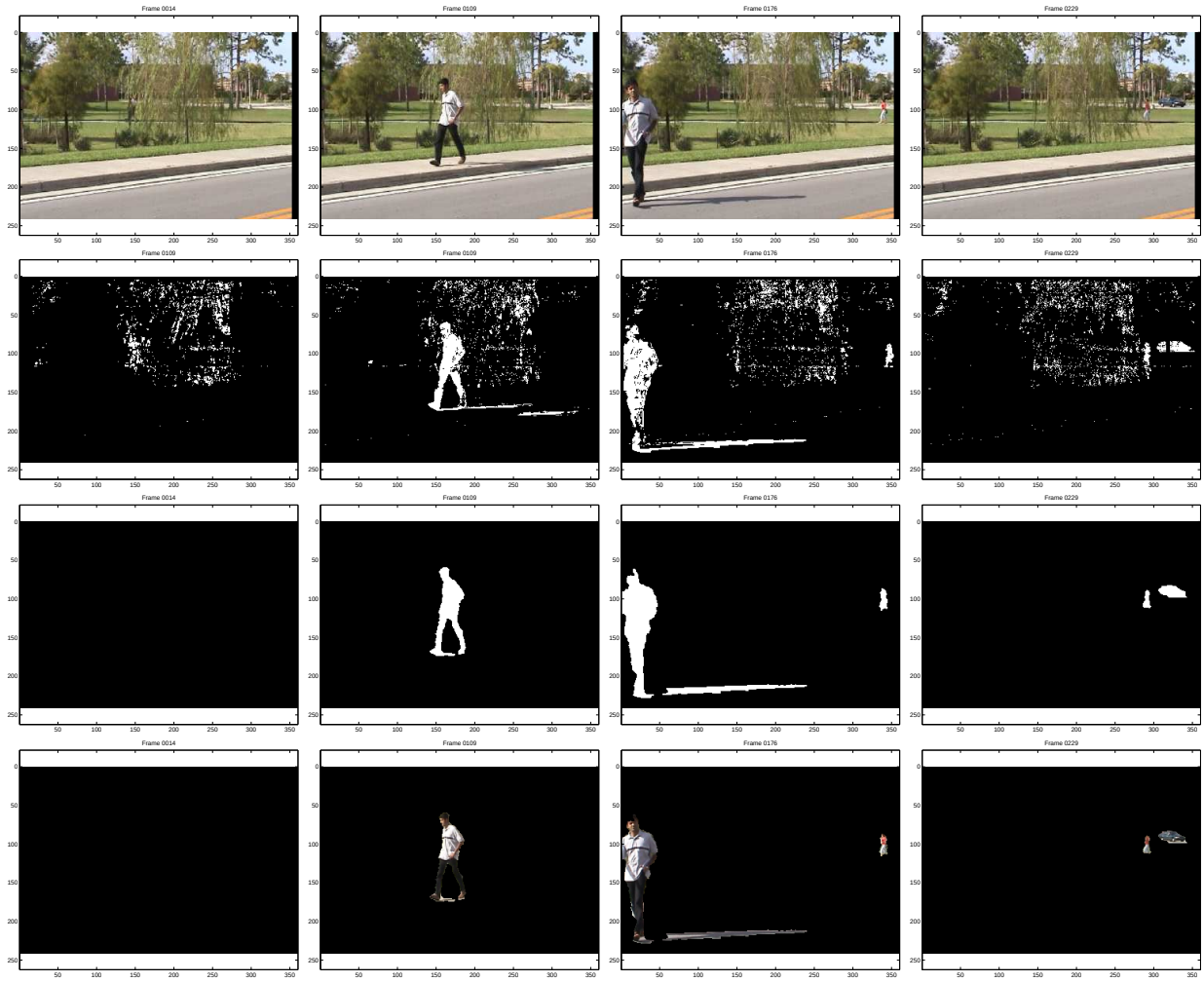
(a) (b) (c) (d)

Figure 3.9: Fountain Sequence. Background Subtraction in the presence of dynamic textures. There are three sources of nonstationarity: (1) The tree branches oscillate (2) The fountains (3) The shadow of the tree on the grass below. The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row results obtained by our method. The fourth row is the masked original image. Morphological operators were not used in the results.

(a) (b) (c)

Figure 3.10: Three more examples of detection in the presence of dynamic backgrounds. (a) The lake-side water is the source of dynamism in the background. The contour outlines the detected foreground region. (b) The periodic motion of the ceiling fans is ignored during detection. (c) A bottle floats on the oscillating sea, in the presence of rain.

tained by our algorithm. The first row contains the recorded images and the second row shows the detected foreground as proposed in [SG00]. It is evident that the nominal motion of the camera causes substantial degradation in performance, despite a 5-component mixture model and a relatively high learning rate of 0.05. The third row shows the foreground detected using our approach. It is stressed that *no* morphological operators like erosion / dilation or median filters were used in the presentation of these results. Manually segmented foreground regions are shown in the bottom row. This sequence exemplifies a set of phenomena, including global motion caused by vibrations, global motion in static hand-held cameras, and misalignment in the registration of mosaics. Quantitative experimentation has been performed on this sequence and is reported subsequently.

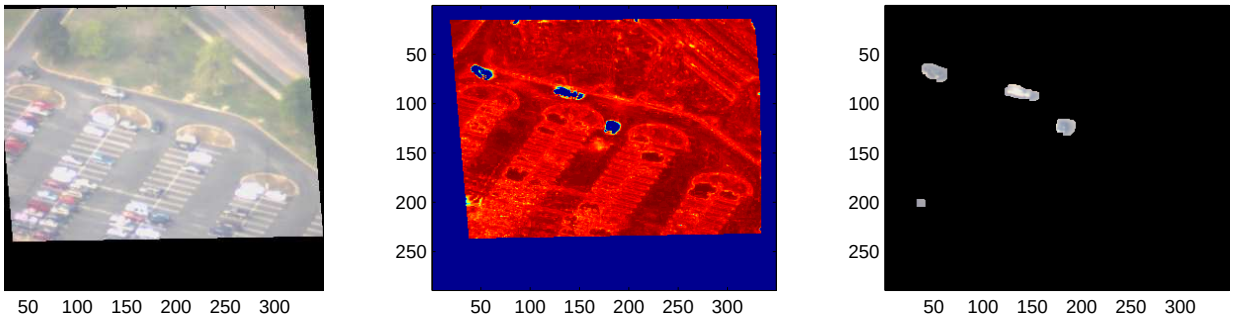


(a) (b) (c) (d)

Figure 3.11: Swaying trees sequence. A weeping willow sways in the presence of a strong breeze. The top row shows the original images, the second row are the results obtained by using the mixture of Gaussians method, and the third row are the results obtained by our method. The fourth row is the masked original image. Morphological operators were not used in the results.

Figures 3.8, 3.9, and 3.11 show results on scenes with dynamic textures. In Figure 3.8, a red remote controlled car moves in a scene with a backdrop of a shimmering and rippling pool. Since dynamic textures like the water do not repeat exactly, pixel-wise methods, like the mixture of Gaussians approach, handle the dynamic texture of the pool poorly, regularly producing false positives. On the other hand, our approach handled this dynamic texture immediately, while detecting the moving car accurately as well. Figure 3.9 shows results on a particularly challenging outdoor sequence, with three sources of dynamic motion: (1) The fountain, (2) the tree branches above, and (3) the shadow of the trees branches on the grass below. Our approach disregarded each of the dynamic phenomena and instead detected the objects of interest. In Figure 3.11, results are shown on sequence where a weeping willow is swaying in a strong breeze. There were two typical paths in this sequence, one closer to the camera, and another one farther back, behind the tree. Including invariance to the dynamic behavior of the background, both the larger objects closer by and the smaller foreground objects farther back were detected as shown in Figure 3.11(c) and (d).

Figure 3.10(a) shows detection in the presence of period motion, due to a number of ceiling fans. Despite a high degree of motion, the individual is detected accurately. Figure 3.10(b) shows detection with the backdrop of a lake, and and 3.10(c) shows detection in the presence of substantial wave motion and rain. In each of the results of 3.10, the contour outlines the detected region, demonstrating accurate detection. Finally, we applied the algorithm to motion stabilized video data collected from aerial videos, and despite nominal misalignment and residual parallax motion, objects were reliably detected. Figure 3.13 and Figure 3.12 show background membership



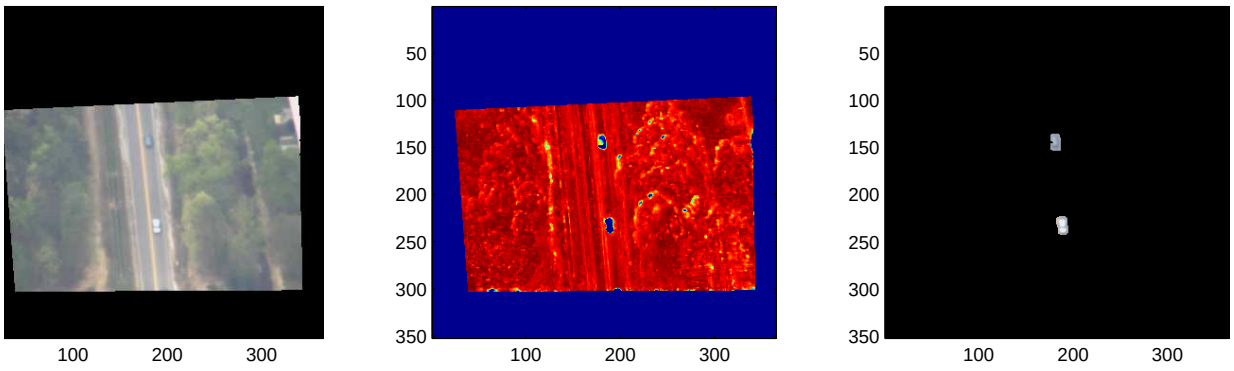
(a) (b) (c)

Figure 3.12: Aerial Video - Example 1. (a) Frame 1 of 80, (b) Background likelihood map, (c) Masked image frame based on foreground decision. Reliable object detection is obtained despite residual parallax motion of the tree and light poles. The small object detected in the bottom left of the frame is the shadow of an object entering the field of view.

likelihoods. There is significant residual parallax due to the trees and the light poles in the scene. Despite these, the end detection is highly accurate.

3.4.2 Quantitative Analysis

We performed quantitative analysis at both the pixel-level and object-level. For the first experiment, we manually segmented a 500-frame sequence (as seen in Figure 3.7) into foreground and background regions. In the sequence, the scene is empty for the first 276 frames, after which two objects (first a person and then a car) move across the field of view. The sequence contained an average nominal motion of approximately 14.66 pixels. Figure 3.14(a) shows the number of pixels



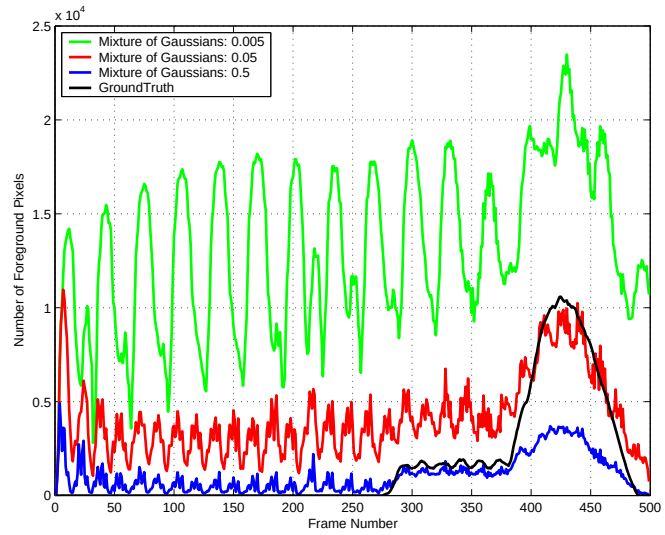
(a) (b) (c)

Figure 3.13: Aerial Video - Example 2. (a) Frame 1 of 80, (b) Background likelihood map, (c) Masked image frame based on foreground decision.

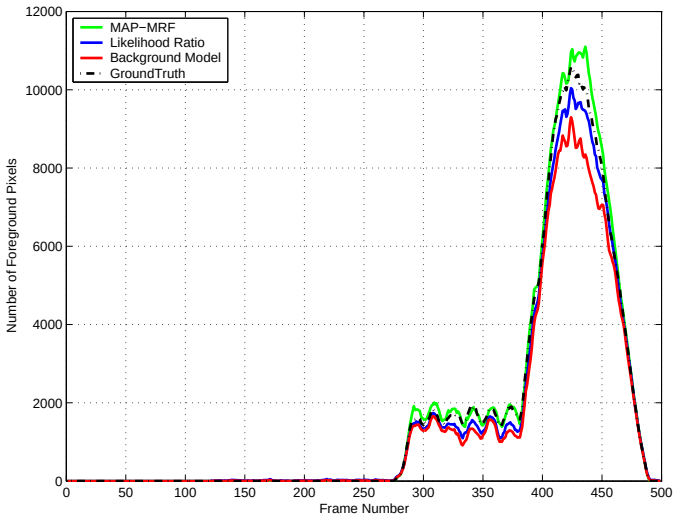
detected in selected frames by the mixture of Gaussians method at various values of the learning parameter and the ground truth. The periodicity apparent in the detection by the mixture of Gaussians method is caused by the periodicity of the camera motion. The initial periodicity in the ground truth is caused by the periodic self-occlusion of the walking person and the subsequent peak is caused by the later entry and then exit of the car. In Figure 3.14(b) the corresponding plot at each level of our approach is shown. The threshold for the detection using only the background model was chosen as $\log(\gamma)$ (see Equation 3.7), which was equal to -27.9905. In addition to illustrating the contribution of background model to the over-all result, the performance at this level is also relevant because, in the absence of any previously detected foreground, the system essentially uses only the background model for detection. For the log-likelihood ratio, the obvious value for κ (see Equation 3.8) is zero, since this means the background is less likely than the foreground. Clearly, the results reflect the invariance at each level of the approach to mis-detections caused by the nominal

camera motion. The per-frame detection rates are shown in Figure 3.15 and Figure 3.16 in terms of precision and recall, where Precision $= \frac{\#\text{ of true positives detected}}{\text{total }\#\text{ of positives detected}}$ and Recall $= \frac{\#\text{ of true positives detected}}{\text{total }\#\text{ of true positives}}$. The detection accuracy both in terms of recall and precision is consistently higher than the mixture of Gaussians approach. Several different parameter configurations were tested for the mixture of Gaussians approach and the results are shown for three different learning parameters. The few false positives and false negatives that were detected by the approach were invariably at the edges of true objects, where factors such as pixel sampling affected the results.

Next, to evaluate detection at the object level (detecting whether an object is present or not), we evaluated five sequences, each (approximately) an hour long. The sequences tested included an extended sequence of Figure 3.7, a sequence containing trees swaying in the wind, a sequence of ducks swimming on a pond, and two surveillance videos. If a contiguous region of pixels was consistently detected corresponding to an object during its period within the field of view, a correct 'object' detection was recorded. If two separate regions were assigned to an object, if an object was not detected or if a region was spuriously detected, a mis-detection was recorded. Results, shown in Table 1, demonstrate that our approach had an overall average detection rate of 99.708% and an overall mis-detection rate of 0.41%. The mis-detections were primarily caused by break-ups in regions, an example of which can be seen in Figure 3.9(c).

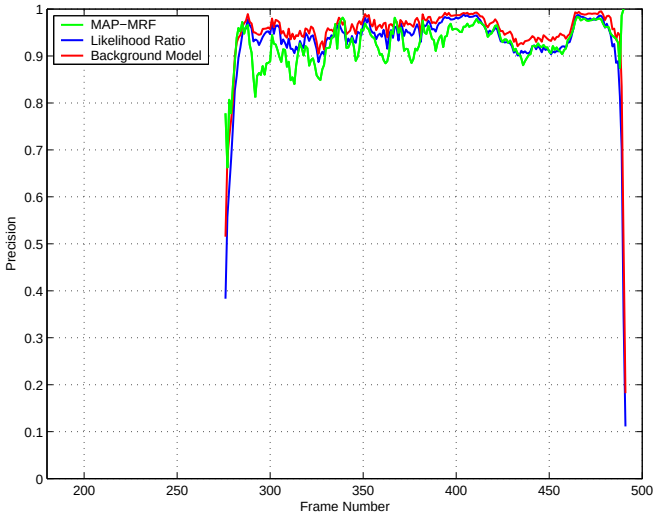


(a)

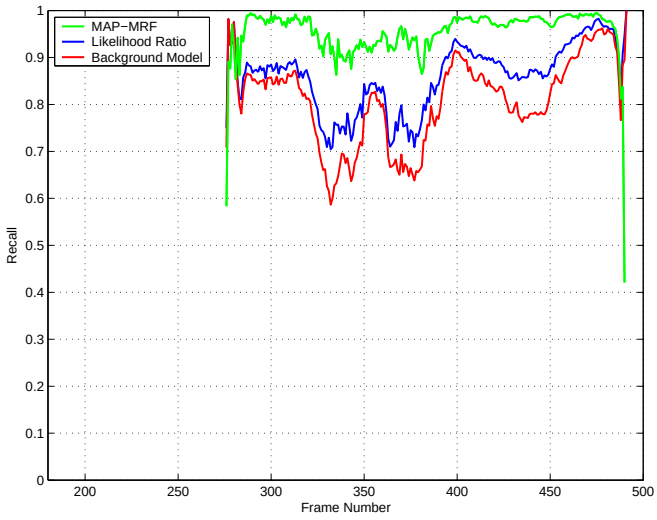


(b)

Figure 3.14: Numbers of detected pixels for the sequence with nominal motion (Figure 3.7). (a) This plot shows the number of pixels detected across each of 500 frames by the Mixture of Gaussians method at various learning rates. Because of the approximate periodicity of the nominal motion, the number of pixels detected by the Mixture of Gaussians method shows periodicity. (b) This plot shows the number of pixels detected at each stage of our approach, (1) using the background model, (2) using the likelihood ratio and (3) using the MAP-MRF estimate.

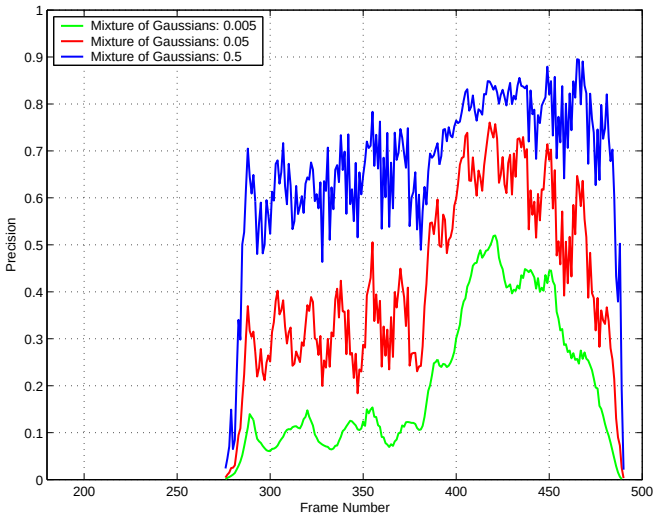


(a)

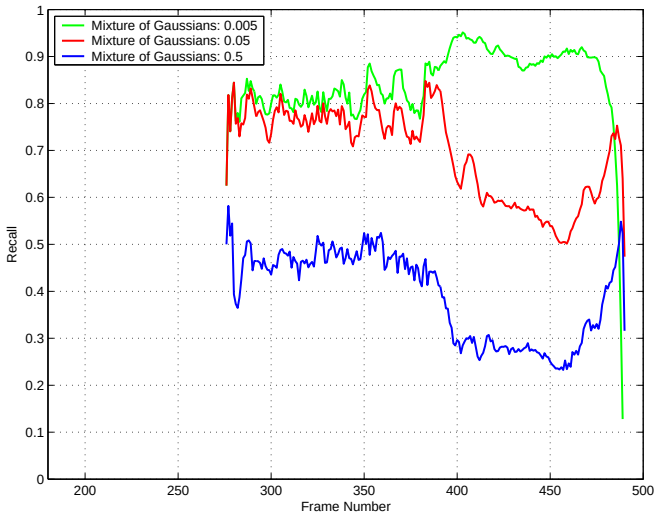


(b)

Figure 3.15: Pixel-level detection recall and precision at each level of our approach. (a) Precision and (b) Recall.



(a)



(b)

Figure 3.16: Pixel-level detection recall and precision using the Mixture of Gaussians approach at three different learning parameters: 0.005, 0.05 and 0.5. (a) Precision and (b) Recall.

Table 3.1: Object level detection rates. Object detection and mis-detection rates for 5 sequences (each 1 hour long).

	Objects	**Det.**	**Mis-Det.**	**Det. %**	**Mis-Det. %**
Seq. 1	84	84	0	100.00%	0.00%
Seq. 2	115	114	1	99.13%	0.87%
Seq. 3	161	161	0	100.00%	0.00%
Seq. 4	94	94	0	100.00%	0.00%
Seq. 5	170	169	2	99.41%	1.18%

3.5 Conclusion

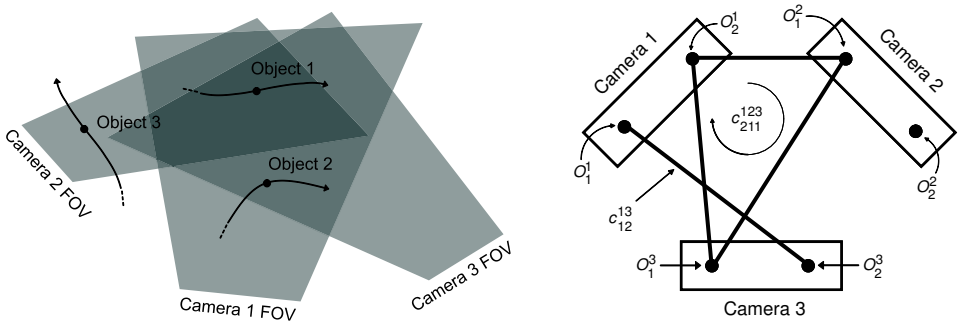
There are a number of innovations in this work. From an intuitive point of view, using the joint representation of image pixels allows local spatial structure of a sequence to be represented explicitly in the modeling process. The entire background is represented by a *single* distribution and a kernel density estimator is used to find membership probabilities. The joint feature space provides the ability to incorporate the spatial distribution of intensities into the decision process, and such feature spaces have been previously used for image segmentation, smoothing [CM02] and tracking [EDD03]. A second novel constraint in this work is temporal persistence as a criterion for detection without feedback from higher-level modules (as in [Har02]). The idea of using both background and foreground color models to compete for ownership of a pixel using the log likelihood ratio has been used before for improving tracking in [CL03]. However, in the context of object detection, making coherent models of both the background and the foreground, changes the

paradigm of object detection from identifying outliers with respect to a background model to explicitly classifying between the foreground and background models. The likelihoods obtained are utilized in a MAP-MRF framework that allows an optimal global inference of the solution based on local information. The resulting algorithm performed suitably in several challenging settings.

CHAPTER 4

OBJECT ASSOCIATION ACROSS MULTIPLE OVERLAPPING CAMERAS

In this chapter, we present an algorithm that requires at least limited spatiotemporal overlap between the fields of view of the cameras (Case 3 of the introduction). This is the *minimal* assumption that is required to discern the relationship of observations in the uncalibrated moving cameras. The underlying concept of cooperative sensing is to use these relationships to give global context to 'locally' obtained information at each camera. It is desirable, therefore, that the data collected at each camera and the inter-camera relationship discerned by the system be presented in a coherent visualization. For moving cameras, particularly airborne ones where large swaths of areas may be traversed in a short period of time, coherent visualization is indispensable for applications like surveillance and reconnaissance. Thus, in addition to presenting an algorithm to track objects across multiple moving cameras with spatiotemporal overlap of fields of view, we provide a means to simultaneously visualize the collective field of view of all the airborne cameras.



(a) (b)

Figure 4.1: Graphical representation. (a) Three trajectories observed in three cameras. (b) The graph associated with the scenario in (a).

Notation The scene is modeled as a plane in 3-space, $\mathbf{\Pi}$, with K moving objects, observed by N cameras. The k-th object[1] moves along a trajectory on $\mathbf{\Pi}$, represented by a time-ordered set of points. A particular object k, present in the field of view of camera n, is denoted as O_k^n and the imaged location of O_k^n at time t is $\mathcal{X}_k^n(t) = (x_{k,t}^n, y_{k,t}^n, \lambda_{k,t}^n)^\mathsf{T} \in \mathbb{P}^2$, the homogenous coordinates of the point in sequence n. The imaged trajectory of O_k^n is the sequence of points $\mathcal{X}_k^n = \{\mathcal{X}_k^n(i), \mathcal{X}_k^n(i+1), \ldots \mathcal{X}_k^n(j)\}$. When referring to inhomogeneous coordinates, we will refer to a point as $\mathbf{x}_k^n(t) = (x_{k,t}^n/\lambda_{k,t}^n, y_{k,t}^n/\lambda_{k,t}^n)^\mathsf{T} \in \mathbb{R}^2$. For two cameras, an association or correspondence $c_{k,l}^{n,m}$ is an ordered pair (O_k^n, O_l^m) that represents the hypothesis that O_k^n and O_l^m are images of the same object. Formally, it defines the event,

$$c_{k,l}^{n,m} \doteq \{O_k^n \text{and } O_l^m \text{arise from the same object in the world}\}, l = 1, \cdots, \mathbf{z}(m),$$

$$c_{k,0}^{n,m} \doteq \{O_k^n \text{was not viewed in camera } m\},$$

[1] The abstraction of each object is as a point.

where $\mathbf{z}(m)$ is the number objects observed in camera m. Since these events are mutually exclusive and exhaustive,

$$\sum_{l=0}^{\mathbf{z(m)}} p(c_{k,l}^{n,m} | \mathcal{X}_k^n, \mathcal{X}_l^m) = 1.$$

Similarly, for more than two cameras, a correspondence $c_{i,j,\ldots k}^{m,n,\ldots p}$ is a hypothesis defined by the tuple $(O_i^m, O_j^n, \ldots O_l^p)$. Note that O_1^1 does not necessarily correspond to O_1^2, the numbering of objects in each sequence is in the order of detection. Thus, the problem is to find the set of associations C such that $c_{i,j,\ldots l}^{m,n,\ldots p} \in C$ if and only if $O_i^m, O_j^n, \ldots O_l^p$ are images of the same object in the world. Graphical illustration allows us to more clearly represent these different relationships (Figure 4.1). We abstract the problem of tracking objects across cameras as follows. Each observed trajectory is modeled as a node and the graph is partitioned into N partitions, one for each of the N cameras. A hypothesized association, c, between two observed objects (nodes), is represented as an edge between the two nodes. This N-partite representation is illustrated in Figure 4.1. Clearly, in this instance, Object 1 is visible in all cameras, and the association across the cameras is represented by c_{211}^{123}. Object 2 is visible only in Camera 1 and Camera 3 and therefore an edge exists only between Camera 1 and 3. Object 3 is visible only in the field of view of Camera 2, therefore there is a unconnected node in the partition corresponding to Camera 2.

4.1 Estimating Inter-Camera Relationships

In this section, an unsupervised approach is presented to estimating the inter-camera relationships in terms of the inter-frame homography. We describe how the likelihood that trajectories, observed in different cameras, originating from the same world object, is estimated. The use of this, in turn, for multiple objects assignment across multiple cameras is then described next. Thus, at a certain instant of time, we have $\mathbf{z}(n)$ trajectories for the n-th camera corresponding to the objects visible in that camera. The measured image positions of objects, $\mathbf{x}_k^n = \{\mathbf{x}_k^n(i), \mathbf{x}_k^n(i+1), \ldots \mathbf{x}_k^n(j)\}$ are described in terms of the true image positions, $\bar{\mathbf{x}}_k^n = \{\bar{\mathbf{x}}_k^n(i), \bar{\mathbf{x}}_k^n(i+1), \ldots \bar{\mathbf{x}}_k^n(j)\}$, with independent normally distributed measurement noise, $\mu = 0$ and covariance matrix $\mathbf{R}_k^n(i)$, that is

$$\mathbf{x}_k^n(i) = \bar{\mathbf{x}}_k^n(i) + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k^n(i)). \tag{4.1}$$

It is assumed in this work that the trajectories are compensated for global egomotion of the camera, through the estimation of frame-to-frame homographies, and are therefore in a single coordinate system for each camera. The covariance matrix captures the uncertainty in detection at each frame, uncertainty that is propagated by the sequential estimation of inter-frame homographies, separately for each camera.

The principal assumption upon which the similarity between two trajectories is evaluated is that due to the altitude of the aerial camera, the scene can be well approximated by a plane in 3-space and as a result a homography exists between any two frames of any sequence ([HZ00]). This assumption of planarity dictates that a homography $\mathbf{H}_{k,l}^{n,m}$ must exist between any two trajectories that correspond, i.e. for any association hypothesis $c_{k,l}^{n,m}$. This constraint can be exploited

to compute the likelihood that 2D trajectories observed by two different cameras originate from the same 3D trajectory in the world - in other words, to estimate $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$ (which we describe presently). Furthermore, we show how this can be extended to multiple views to evaluate $p(c_{i,j,\ldots k}^{n,m,\ldots l}|\{\mathcal{X}_i^n, \mathcal{X}_j^m, \ldots \mathcal{X}_k^l\})$. By assuming conditional independence between each association c, the probability of a candidate solution C given the trajectories in multiple cameras is,

$$p(C|\{\mathcal{X}\}) = \prod_{c_{i,j,\ldots k}^{n,m,\ldots l} \in C} p(c_{i,j,\ldots k}^{n,m,\ldots l}|\{\mathcal{X}_i^n, \mathcal{X}_j^m, \ldots \mathcal{X}_k^l\}). \tag{4.2}$$

We are interested in the Maximum Likelihood solution,

$$C^* = \arg\max_{C \in \mathcal{C}} p(C|\{\mathcal{X}\}), \tag{4.3}$$

where $\mathcal{C}$ is the space of solutions. We now describe how to compute the likelihood that two trajectories observed in two or more cameras originated from the same real world object. Using these likelihoods, we describe how to maximize Equation 4.3 in Section 4.1.2.

4.1.1 Evaluating an Association Hypothesis

In this sub-section we discuss how to evaluate the likelihood of association between tracks in two cameras, i.e. we describe how to compute $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$, and its extension for multiple cameras. The evaluation of this likelihood is complicated by the imaging process, so despite the fact that trajectories in correspondence can be viewed as samples from a single trajectory on the plane $\mathbf{\Pi}$, the coordinates of the 'samples' are not registered. We can compute $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$ by computing the maximum likelihood estimate of the homography, $\mathbf{H}_{k,l}^{n,m}$ and two new trajectories

$\bar{\mathcal{X}}_k^n$ and $\bar{\mathcal{X}}_l^m$, related *exactly* by $\mathbf{H}_{k,l}^{n,m}$, as described in [HZ00], by minimizing the reprojection error. The re-projection error is a cost function that explicitly minimizes the *transfer* error between the trajectories and was proposed by Sturm in [Stu97], with further work with Chum and Pajdla in [CPS05]. Using this estimate of the homography and the 'true' trajectories,

$$p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m) \propto L(\mathcal{X}_k^n, \mathcal{X}_l^m|c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}) = L(\mathcal{X}_k^n|c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}) L(\mathcal{X}_l^m|c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}). \tag{4.4}$$

The proportionality follows from Bayes Theorem assuming a uniform prior on all associations and ignoring the constant evidence term. Since the errors at each point are assumed independent, the conditional probability of the association given the trajectories in the pair of sequences can be estimated,

$$L(\mathcal{X}_k^n, \mathcal{X}_l^m|\, c_{k,l}^{n,m}; \mathbf{H}_{k,l}^{n,m}, \bar{\mathcal{X}}_{k,l}^{n,m}) = \prod_i \frac{1}{2\pi \|\mathbf{R}_k^n(i)\|^{\frac{1}{2}} \|\mathbf{R}_l^m(i)\|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(d(\mathcal{X}_k^n(i), \bar{\mathcal{X}}_k^n(i))_{\mathbf{R}_k^n(i)} + d(\mathcal{X}_l^m(i), \bar{\mathcal{X}}_l^m(i))_{\mathbf{R}_l^m(i)}\right)}. \tag{4.5}$$

where $d(\cdot)_{\mathbf{R}(i)}$ is the Mahalanobis distance and $\mathbf{R}_k^n(i)$ is the error covariance matrix,

$$\begin{aligned} &d\big(\mathcal{X}_k^n(i), \bar{\mathcal{X}}_k^n(i)\big)_{\mathbf{R}_k^n(i)} + d\big(\mathcal{X}_l^m(i), \bar{\mathcal{X}}_l^m(i)\big)_{\mathbf{R}_l^m(i)} = \\ &\quad \big(\mathbf{x}_k^n(i) - \bar{\mathbf{x}}_k^n(i)\big)^T \mathbf{R}_k^n(i)^{-1} \big(\mathbf{x}_k^n(i) - \bar{\mathbf{x}}_k^n(i)\big) + \big(\mathbf{x}_l^m(i) - \bar{\mathbf{x}}_l^m(i)\big)^T \mathbf{R}_l^m(i)^{-1} \big(\mathbf{x}_l^m(i) - \bar{\mathbf{x}}_l^m(i)\big) \end{aligned} \tag{4.6}$$

Thus, to estimate the data likelihood, we compute the optimal estimates of the homography and exact trajectories and use them to evaluate Equation 4.7.

For situations where there are more than two cameras, this analysis extends directly. To evaluate, for instance, $p(c_{1,1,\ldots 1}^{1,2,\ldots,N}|\mathcal{X}_1^1, \mathcal{X}_1^2, \ldots, \mathcal{X}_1^N)$, we proceed by computing the maximum likelihood estimate of the set of $N-1$ homographies, and one 'canonical' trajectory related to each view by

the set of homographies. Using these estimates, we have,

$$p(c_{1,1,\ldots 1}^{1,2,\ldots,N}|\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N) \propto L(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\{\mathbf{H}_{1,1}^{1,2},\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{N-1,N}\},\bar{\mathcal{X}}_1^1), \quad (4.7)$$

where the *pdf* of $L(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\{\mathbf{H}_{1,1}^{1,2},\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{N-1,N}\},\bar{\mathcal{X}}_1^1)$, is[2] ,

$$L(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\{\mathbf{H}_{1,1}^{1,2},\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{N-1,N}\},\bar{\mathcal{X}}_i) = \prod_i \frac{1}{\left(2\pi\|\mathbf{R}\|\right)^{\frac{N}{2}}} e^{-d_r/2} \quad (4.8)$$

where

$$d_r = \sum_j \Big(d(\mathcal{X}_1^1(i),\bar{\mathcal{X}}_1(i))_{\mathbf{R}} + \sum_{j=2}^{N} d(\mathcal{X}_1^j(i),\mathbf{H}_{1,1}^{j-1,j}\bar{\mathcal{X}}_1(i))_{\mathbf{R}} \Big). \quad (4.9)$$

The Direct Linear Transform algorithm or RANSAC can be used as an initial estimate, followed by a Levenberg-Marquardt minimization over $9(N-1)+2\Delta t$ variables: $9(N-1)$ unknowns for the set of homographies and $2\Delta t$ unknowns for the canonical Δt 2D points. Equation 4.8 is used to compute the maximum likelihood estimates of the homography and the canonical trajectory and then used to evaluate the probability of the association hypothesis.

4.1.2 Maximum Likelihood Assignment of Global Correspondence

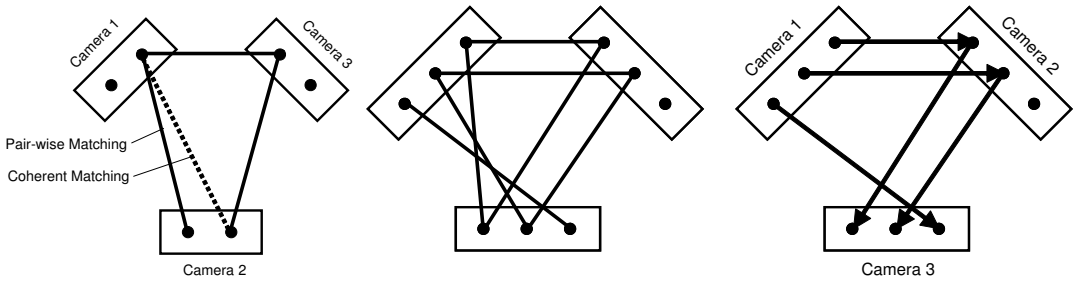
In the previous section, we developed a model to evaluate the probability of association between several imaged trajectories for a single object. Generally, however, when several objects are observed simultaneously by multiple cameras we require an optimal *global* assignment of object correspondences. We show that within this formulation, this global optimality too can

[2] For notational convenience we assume the covariance matrices are all equal.

be described in a maximum likelihood sense. As mentioned earlier, the problem of establishing association between trajectories can be posed within a graph theoretic framework. Consider first, the straightforward case of several objects observed by *two* airborne cameras. This can be modeled by constructing a complete bi-partite graph $G = (U, V, E)$ in which the vertices $U = \{u(\mathcal{X}_1^p), u(\mathcal{X}_2^p) \dots u(\mathcal{X}_k^p)\}$ represent the trajectories in Sequence p, and $V = \{v(\mathcal{X}_1^q), v(\mathcal{X}_2^q) \dots v(\mathcal{X}_k^q)\}$ represent the trajectories in Sequence q, and E represents the set of edges between any pair of trajectories from U and V. The bi-partite graph is complete because any two trajectories may match hypothetically. The weight of each edge is the probability of correspondence of Trajectory $\mathcal{X}_l^q$ and Trajectory $\mathcal{X}_k^p$, as defined in Equation 4.7. By finding the maximum matching of G, we find a unique set of correspondence C', according to the *maximum likelihood* solution,

$$C' = \arg\max_{C \in \mathcal{C}} \sum_{c_{k,l}^{p,q} \in C} \log p(c_{k,l}^{p,q} \mid \mathcal{X}_k^p, \mathcal{X}_l^q). \tag{4.10}$$

where $\mathcal{C}$ is the solution space. Several algorithms exist for the efficient maximum matching of a bi-partite graph, for instance [Kuh55] or [HK73] which are $O(n^3)$ and $O(n^{2.5})$ respectively. It should be noted that during the construction of the graph we need to ensure that 'left-over' objects are not assigned association. For instance, consider the case when all but one object in each of two cameras have been assigned association. Although the 'left-over' objects in each camera correspond to the two different objects in the real world (each that did not appear in one of the camera FOVs), they would be assigned association. In order to avoid this we prune all edges whose edge weights are below a certain likelihood. This is equivalent to ignoring measurements



(a) (b) (c)

Figure 4.2: Tracking across 3 moving cameras. (a) A possible association between objects in three cameras. (b) The digraph associated with correspondence in (a). Correspondence in 3 or more moving cameras. (a) An impossible matching. Transitive closure in matching is an issue for matching in three of more cameras. The dotted line shows the desirable edge whereas the solid line shows a possible solution from pairwise matching. (b) Missing observations. This matching shows the case of missing observations, with three objects in the scene, each visible in two cameras at a time. (c) The digraph associated with (b).

outside a 'validation' region, as described in [Ed90], ensuring that association hypotheses with low likelihoods are ignored.

This formulation generalizes to *multiple* airborne cameras by considering k-partite hypergraphs instead of the bipartite graphs considered previously, shown in Figure 4.2. Once again, we wish to find the set of associations C',

$$C' = \arg\max_{C \in \mathcal{C}} \sum_{c^{p,q\ldots r}_{k,l\ldots m} \in C} \log p(c^{p,q\ldots r}_{k,l\ldots m} \,|\mathcal{X}^p_k, \mathcal{X}^q_l \ldots, \mathcal{X}^r_m). \tag{4.11}$$

Each hyper-edge represents the hypothesized association $c^{p,q\ldots r}_{k,l\ldots m}$ between $(O^p_k, O^q_l \ldots O^r_m)$. However, it is known that the k-dimensional matching problem is NP-Hard for $k \geq 3$ ([Pap94]). A possible approximation that is sometimes used is pairwise, bipartite matching, however such an

approximation is unacceptable in the current context since it is vital that transitive closure is maintained while tracking. The requirements of consistency in the tracking of objects across cameras is illustrated in Figure 4.2. Instead, to address the complexity involved while accounting for consistent tracking, we construct a weighted digraph $D = (V, E)$ such that $\{V_1, V_2, \ldots V_k\}$ partitions V, where each partition corresponds to a moving camera. Direction is obtained by assigning an arbitrary order to the cameras (for instance by enumerating them), and directed edges exist between every node in partition V_i and every node in partition V_j where $i > j$ (due to the ordering). By forbidding the existence of edges against the ordering of the cameras, D is constructed as an acyclic digraph. This can be expressed as $E = \{v(\mathcal{X}_k^p)v(\mathcal{X}_l^q)|v(\mathcal{X}_k^p) \in V_p, v(\mathcal{X}_l^q) \in V_q\}$, where $e = v(\mathcal{X}_k^p)v(\mathcal{X}_l^q)$ represents an edge and $q > p$. The solution to the original association problem is then equivalent to finding the edges of maximum matching of the split G^* of the digraph D (for a proof see [SS05]). It should be noted that with this approach we need only define pairwise edge-weights. Figure 4.2 shows a possible solution and its corresponding digraph.

Once this solution, using an approximation, is provided, we evaluate $p(C|\mathcal{X})$ as follows. We observe that all homographies mapping pairs of corresponding tracks in Sequences p and q are equal (up to a scale factor), and are, in turn, the same homography that maps the reference coordinate of Sequence p to that of Sequence q. Since all the objects lie on the same plane, the homography relating the image of the trajectory of any object $\mathbf{H}_{k,l}^{p,q}$ in Sequence p to the image of the trajectory of that object in Sequence q is the same as the homography $\mathbf{H}_{i,j}^{p,q}$ relating any other object's trajectories in the two sequences (i.e. $i \neq p$ and $j \neq q$). Since these trajectories lie on the scene plane, these homography are equal to $\mathbf{H}^{p,q}$, the homography that related the images of

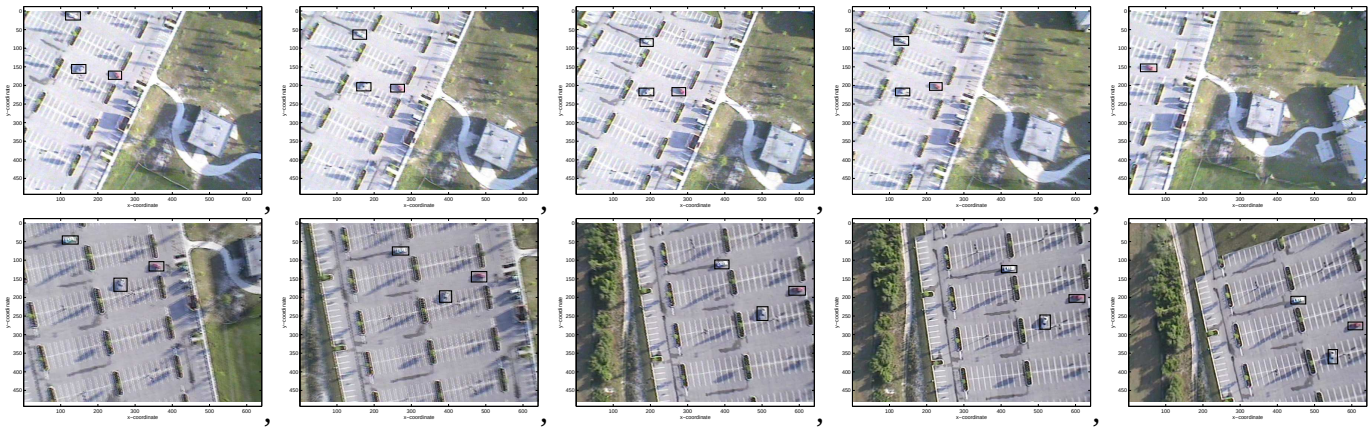


Figure 4.3: Corresponding frames from two sequences. Both rows show frames recorded from different cameras.

Sequence p to the images of Sequence q. This allows us to express $p(C|\mathcal{X})$ as,

$$p(C|\mathcal{X}) = \prod_i \frac{1}{\left(2\pi\|\mathbf{R}\|\right)^{\frac{N}{2}}} e^{-d_r/2}, \tag{4.12}$$

where

$$d_r = \sum_j \sum_k \left(d(\mathcal{X}_k^1(i), \bar{\mathcal{X}}_k^1(i))_{\mathbf{R}} + \sum_{j=2}^{N} d(\mathcal{X}_k^j(i), \mathbf{H}^{j-1,j} \bar{\mathcal{X}}_k(i))_{\mathbf{R}} \right). \tag{4.13}$$

By using all trajectories between cameras simultaneously to estimate the inter-camera homography, the spatial separation of different trajectories enforces a strong non-collinear constraint on association despite the near collinear motion of individual objects. In this way, even with relatively small durations of observation the correct correspondence of objects can be discerned. Once again the optimal value of the set of homographies and the canonical trajectories are estimated using Levenberg-Marquardt minimization, and measure the 'goodness of fit'. The final algorithm is summarized in Figure 4.4.

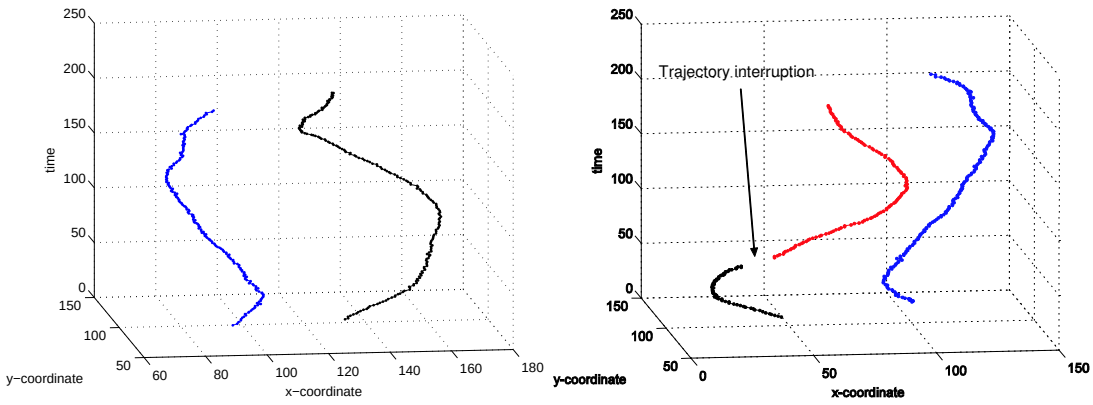
Objective

Given object trajectories from all cameras for $\Delta t \geq 5$, estimate the inter-camera spatial transformations.

Algorithm

1. **Number cameras arbitrarily**
2. **For all pairwise** c**, compute** $p(c_{k,l}^{n,m} | \mathcal{X}_k^n, \mathcal{X}_l^m)$
3. **Construct Split Graph** G^*: Find the maximum matching of the split of the acyclic directed graph described in Section 4.1.2.
4. **Evaluate Global Likelihood Function**: Using the estimated maximum matching, compute the canonical trajectories and the maximum likelihood estimate of the inter-frame homographies.
5. **Repair Trajectories** For each unassociated trajectory, evaluate association likelihood with respect to all canonical trajectories, and re-associate broken trajectories.
6. **Concurrent Visualization**: Use the inter-frame homographies to construct a concurrent mosaic using all videos simultaneously.

Figure 4.4: Algorithm for object association across moving cameras



(a) (b)

Figure 4.5: Trajectory Interruption. (a) Complete trajectories observed in Camera 1. (b) The second trajectory (black) is interrupted as the object exits and then re-enters the field of view. The re-entering trajectory is recorded as a new trajectory (red).

4.1.3 Repairing Trajectories

During single camera tracking, object trajectories can sometimes be interrupted because of missing detections, noise, specularities, or feature similarity to the background. Trajectory interruption can also occur due to scene events like occlusion of the object by some other object, such as clouds, bridges or tree cover, or due to the exiting and re-entering of an object from the field of view. This causes the object's motion to be recorded by two different trajectories. Figure 4.5 shows trajectories in two cameras, plotted in space and time. In the second camera the second trajectory is interrupted as the object exited and re-entered the scene. Several methods have been proposed to account for this problem at the single camera level using predictive methods. However, we show that the canonical tracks and the estimated inter-camera homographies can be used to repair

broken trajectories in a straightforward way. Since matching ensures a one-to-one correspondence, all such broken trajectories should be unassociated after matching. For each free trajectory $\mathcal{X}_i^n$, we evaluate with respect to each canonical trajectory $\bar{\mathcal{X}}_j$,

$$j^* = \arg\max_{j \in 1 \ldots N} p(\mathcal{X}_i^n | \bar{\mathcal{X}}_j; \mathbf{H}^n). \tag{4.14}$$

$p(\mathcal{X}_i^n | \bar{\mathcal{X}}_{j^*}, \mathbf{H}^n)$ is evaluated asymmetrically,

$$p(\mathcal{X}_i^n | \bar{\mathcal{X}}_{j^*}, \mathbf{H}^n) \propto \prod_k \frac{1}{\sqrt{2\pi} \|\mathbf{R}_i^n(k)\|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(d(\mathcal{X}_i^n(k), \bar{\mathcal{X}}_i(k))_{\mathbf{R}_i^n(k)}\right)}. \tag{4.15}$$

If this is greater than an empirical threshold $\gamma(k)$ and if there is no temporal overlap between $\mathcal{X}_i^n$ and $\mathcal{X}_j^n$ (the trajectory in Camera n currently associated with $\mathcal{X}_j$) then $\mathcal{X}_i^n$ and $\mathcal{X}_j^n$ are re-connected and both associated to $\mathcal{X}_j$ - the trajectory is repaired. It is noteworthy, here, that unlike single camera methods, the duration of occlusion is irrelevant as long as the object is continuously viewed in any other camera.

4.2 Concurrent Mosaic

The purpose of aerial surveillance is to obtain an understanding of what occurs in an area of interest. While it is well known that video mosaics can be used to compactly represent a single aerial video sequence, they cannot compactly represent several such sequences *simultaneously*. If, on the other hand, the homographies between each of the mosaics (corresponding to each aerial sequence) are known, a *concurrent* mosaic can be created of all the sequences simultaneously. Since each sequence is aligned to a single coordinate frame during the construction of individual

mosaics, we can register mosaics from multiple sequences onto one concurrent mosaic. To this end, the known point-wise correspondences (see Figure 4.14(c)) from object tracking can be used to compute the inter-camera homography. The alignment is then refined using direct registration.

Although the moving cameras observe the same scene, the color values of corresponding points in the scene differ across the cameras. Figure 4.6 (a) shows a concurrent mosaic generated from two different sequences. Clearly, directly using these mosaics to create a concurrent mosaic causes noticeable artifacts, as shown in Figure 4.6. Assuming a Lambertian scene with a distant light source (the sun), the scene radiance, $L(\mathbf{X})$ depends only on material properties and the surface normal, i.e $L(\mathbf{X}) = \rho(\mathbf{X})\ \mathbf{I} \cdot \mathbf{n}$, where $\rho(\mathbf{X})$ is the surface albedo at the world point $\mathbf{X}$, $\mathbf{I}$ is the scene irradiance, and $\mathbf{n}$ is the surface normal. Clearly, under the Lambertian model, the scene radiance does not vary with respect to the viewing direction. The image irradiance $I(\mathbf{x})$, in turn is linear in the scene radiance,

$$I_i(\mathbf{x}) = P_i e_i L(\mathbf{X}), \tag{4.16}$$

where P_i is an optical factor of camera i, and e_i is the exposure. Finally, let $M_i(\mathbf{x})$ be the intensity measurements at the world point $\mathbf{X}$ available from the images obtained from camera i. These values are related to the image irradiance by the radiometric response function f_i, $M_i(\mathbf{x}) = f_i(I_i(\mathbf{x}))$. Since f is a monotonically increasing function it is also invertible and we can define $g_i = f_i^{-1}$. We have

$$I_i(\mathbf{x}) = g_i(M_i(\mathbf{x})). \tag{4.17}$$

The source of intensity variation across images captured by different cameras can then only arise from the different radiometric response functions of the cameras. In [GN02], the measured

intensities in image captured by two cameras, $M_p(\mathbf{x})$ and $M_q(\mathbf{x})$ are related by a *intensity mapping function*, $M_p(\mathbf{x}) = G(M_q(\mathbf{x}))$. Since the scene radiance, L, does not vary with the viewpoint, from Equation 4.16,

$$\frac{I_p(\mathbf{x})}{P_p e_p} = \frac{I_q(\mathbf{x})}{P_q e_q}.$$

Using Equation 4.17 we then have,

$$M_p(\mathbf{x}) = g_p^{-1}\left(\frac{P_p e_p}{P_q e_q} g_q(M_q(\mathbf{x}))\right) = g_p^{-1}(k g_q(M_q(\mathbf{x}))) = G(M_q(\mathbf{x})). \tag{4.18}$$

This discussion can be extended to color (spectral reflectance) as well. The response of the ith sensor (red, green or blue) of camera p is expressed as,

$$M_p^{(i)}(\mathbf{x}) = f_p^{(i)}\left(\int_\Lambda \sigma_p^{(i)}(\lambda) I(\mathbf{X}, \lambda) d\lambda\right), \tag{4.19}$$

where λ is the wavelength, Λ is the range of visible wavelengths, σ is the spectral sensitivity of the ith sensor, I is the image irradiance. This approach of using three response functions for each color channel separately has been used in [GN04], [SS04] and [CR96]. The three spectral response functions produce three corresponding *color transference functions*, G_r, G_b and G_g. color transference functions for color images are the analogue of intensity mapping function for grayscale images. However, the analogue is not direct since it was shown in [BG83] that for any spectrally broadband color signal, each channel sensor produces outputs that are correlated[3]. This correlation stems mainly (but not exclusively) from the spectral overlap of $\sigma^{(r)}, \sigma^{(g)}$ and $\sigma^{(b)}$. Thus, in order to model the color transference functions between the two views, it is important that the correlations between the channels be considered. Principally, a color transference functions matrix ought to be

[3]Similar conclusions based instead on analysis of natural images directly, have also been reported in [RCC98].

defined, however, by ignoring dispersion effects, we model the system as a multiple input single output system using multiple regression. We approximate each color transference functions by a cubic trivariate polynomial,

$$r' = G_r(r,g,b) = \sum_{i+j+k\leq 3} a^{(r)}_{i,j,k} r^i g^j b^k - a^{(r)}_{0,0,0} + \epsilon, \tag{4.20}$$

$$b' = G_b(r,g,b) = \sum_{i+j+k\leq 3} a^{(g)}_{i,j,k} r^i g^j b^k - a^{(g)}_{0,0,0} + \epsilon, \tag{4.21}$$

$$g' = G_g(r,g,b) = \sum_{i+j+k\leq 3} a^{(b)}_{i,j,k} r^i g^j b^k - a^{(b)}_{0,0,0} + \epsilon. \tag{4.22}$$

where $a_{i,j,k}$ are the coefficients of the polynomial, $\{r,g,b\}$ and $\{r',g',b'\}$ are the color values the two images, and $\epsilon \sim \mathcal{N}(0,\sigma)$ is the i.i.d. random error. The property that $G_i(0,0,0) = 0$ is ensured by ignoring $a^{(i)}_{0,0,0}$. The error term exists because the available measurements are expected to contain noise. Notably, the cross-product terms in this formulation take the correlation of the rgb axes into account. The coefficient of the polynomial can be solved through an over-constrained linear system of equations, since the homography is known between each pair of mosaics, where every point-to-point correspondence provides an $[r\ g\ b]^\top \leftrightarrow [r'\ g'\ b']^\top$ constraint.

Figure 4.6(b) shows a concurrent mosaic blended using the computed color transference functions. It should be noted that although there are a number of phenomenon that are not modeled in this model of the color transference functions, e.g. specular objects, saturation, or quantization in space and color, nominal misalignment, this approach provides satisfactory approximations. In addition, while accurate results have been obtained using the ordinary least squares (OLS) approach, since outliers can be expected due to the phenomenon mentioned, a robust approach can be used to solve the linear system, such as the least median square approach or iteratively reweighted least

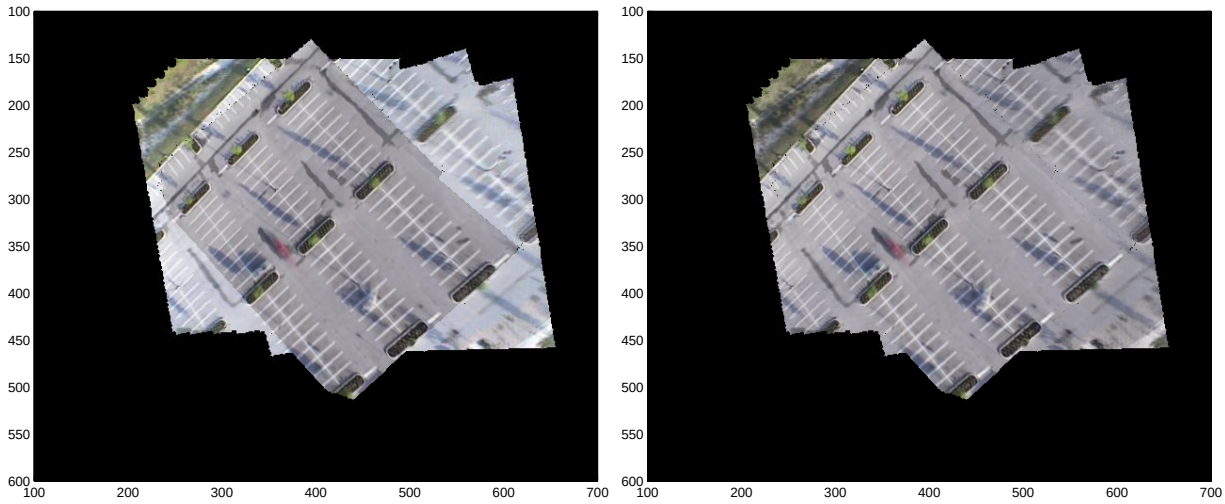


Figure 4.6: Concurrent visualization of two sequences. (a) Concurrent mosaic before blending, (b) Concurrent mosaic after blending.

squares (used for the results in this chapter). Of course, various further simplifications can be made to the multiple regression, such as assuming lower order models, or reducing the degree of the polynomial.

4.3 Results

In this section, we report the experimental performance of our object association approach qualitatively for data from airborne cameras and also in a more controlled setting. We also perform simulations and report the quantitative performance of the algorithm. In each experiment, we demonstrate that the approach is able to accurately associate moving objects across multiple moving cameras despite short durations of observations, nearly linear motion and noisy detections.

The detection and tracking *within* each sequence were done automatically for some sequences and manually for others. For automatic detection and tracking, we used the COCOA system [AS06]. It is recommended that the subsequent results be viewed in color. Additional results and videos associated with these results have been included in the supplementary folder.

4.3.1 Data Generation

In order to run simulations, a generator was designed to randomly synthesize data for quantitative experimentation. The camera parameters included the number of cameras and the number of frames of observation, and the object parameters included the number of objects and the mean and variance of the object motion $(\hat{\rho}, \sigma_\rho)$. For each object an initial position, $X(0)$ and $Y(0)$, was determined by sampling from a uniform distribution over a spatial support region, assuming the world plane Π was the plane $Z = 0$. To closely imitate the smooth motion of real world objects, The object motion $(\rho, \Delta\theta)$ was sampled from the Normal distributions $\mathcal{N}(\hat{\rho}, \sigma_\rho)$ and $\mathcal{N}(0, \sigma_\theta)$, and initial θ was a (single) sample from a uniform distribution over the interval $[-\pi, \pi]$. Thus,

$$X(t) = X(t-1) + \rho\cos(\theta + \Delta\theta), \tag{4.23}$$

$$Y(t) = Y(t-1) + \rho\sin(\theta + \Delta\theta). \tag{4.24}$$

For each camera, a reference to frame homography P was randomly generated, by sampling from a uniform distribution over the support of the camera extrinsic and intrinsic parameters, and the

imaged trajectories of each object in each camera are generated as,

$$\mathcal{X}(t) = \mathsf{P}[X(t) \;\; Y(t) \;\; 1]^T + \epsilon, \tag{4.25}$$

where ϵ is the zero-mean measurement noise, that is specified by a noise variance parameter σ_ϵ. The ratio ρ/σ_ϵ is referred to as the motion-to-noise ratio, measuring the expected strength of noise. Randomly generated trajectories of five objects observed by 5 cameras are shown in Figure 4.7.

4.3.2 Simulations

We conducted simulations on thousands of problem instances, to test the approach. In order to analyze the accuracy of the estimated inter-camera homography as the ratio of mean motion to noise variance we recorded the mean squared error of difference between the maximum likelihood estimate of the homography and the true homography over 100 runs. At each run a new set of trajectories and homographies were generated. As expected, the estimation error decreased as the number of frames increase and the objects began to show more non-collinear motion, shown in Figure 4.8. We then analyzed the quality of the estimate of the canonical tracks with respect to the ground truth, by computing the average log-likelihood of the canonical frame given the ground truth. Here too, the average of 100 runs were taken.

We then analyzed the association accuracy with respect to larger increase in noise as the number of cameras and objects increased. In Figure 4.9(a) reports the association accuracy 10 objects viewed across 10 cameras as the number of frames increase. The motion-to-noise ratio was varied

from infinity (divide-by-zero) to 5×10^{-5} while the number of frames were tested for 5, 50, 100 and 200 frames. Clearly, as the number of frames increased the accuracy increased too. A hundred runs were executed (with randomly generated trajectories) per noise strength and the average accuracy was reported. The accuracy is shown in Figure 4.9(b) as it varies with respect to the number of cameras/objects. As expected, as the number of cameras and objects decrease the accuracy of the approach reduces too. The trajectory length was 60 frames (2 seconds at 30fps). Please note that the motion-to-noise ratio in both experiments is *n*ot linearly increasing.

4.3.3 Experiments on Controlled Sequences

Two controlled experiments were carried out, where remote controlled cars were observed by moving camcorders (Sony DCR-TRV 740). In the first experiment, two cameras were used, along with two remote controlled cars. The cars were operated on a tiled (planar) floor with the two moving cameras viewing their motion from the height of about 12 feet. The mosaic is shown in Figure 4.10 along with the trajectories of the car on the registered coordinate frame of Sequence 1. The variation of the first two hypotheses with respect to time is shown in Figure 4.13(a).

The second controlled experiment was carried out to test the performance of the system for more than two cameras. Three moving cameras at various zooms observed a scene with two remote controlled cars. Figure 4.11 (c) shows the final, correct assignment of correspondence established by our approach. Figure 4.11 (d) shows the associated directed graph. The inter-

sequence homographies were estimated and all three mosaics were registered together to create the concurrent mosaic, as shown in Figure 4.11 (a). Figure 4.11 (b) shows the tracks of both objects, overlaid after blending each mosaic. Figure 4.12 shows the correspondence of the three sequence trajectories.

4.3.4 Experiments on UAV Sequences

In these experiments, two unmanned aerial vehicles (UAVs) mounted with cameras viewed real scenes with moving cars, typically with a smaller duration of overlap than the controlled sequence. Two sequences were recorded with three objects in the scene. Since the motion of aerial vehicles is far less controlled than that of controlled sequences, the duration of time in which a certain object is seen in both cameras is smaller. We show that despite the challenge of smaller overlap, object can be successfully tracked across the moving cameras.

In first experiment, a small number of frames were used, observing the motion of three moving cars. All three objects were simultaneously visible in the field of view for the entire duration of observation. The individual tracks of each sequence, on a single registered coordinate are shown in Figure 4.14 (a) and (b). The result of correspondence is shown in Figure 4.14 (c). The correct correspondence, (Hypothesis: 1 2 3), is clearly higher as the process reaches an equilibrium. Using this correspondence, the concurrent mosaic of the scene was generated, shown in Figure 4.6.

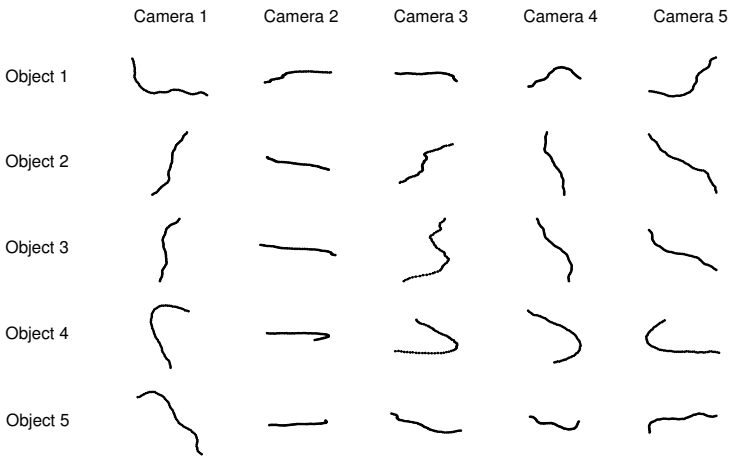
In the second experiment, a longer sequence was used. This time, object exited and entered the field of view, and all three objects were only briefly visible together in the field of view. The individual tracks of each sequence, on a single registered coordinate are shown in Figure 4.16 (a) and (b). The result of correspondence is shown in Figure 4.16 (c). The variation of the 'goodness' of each hypothesis is shown in Figure 4.13(b). The correct correspondence, (Hypothesis: 1 2 3), is clearly higher as the process reach an equilibrium. Using this correspondence, the concurrent mosaic of the scene was generated, shown in Figure 4.16(c).

The final experiment involved association across IR and EO cameras. Since only motion information is used in discerning association, the modality of the cameras do not affect the viability of the algorithm. In the first set, six objects were recorded by one EO and one IR camera. Although the relative positions of the cameras were fixed in this sequence, no additional constraints were used during experimentation. The vehicles in the field of view moved in a line, and one after another performed a u-turn and the durations of observation of each object varied in both cameras. Since only motion information is used, the different modalities did not pose a problem to this algorithm. Figure 4.15 shows all six trajectories color coded in their correspondence. Despite the fact that the sixth trajectory (color coded yellow in Figure 4.15) was viewed only briefly in both sequences and underwent mainly collinear motion in this duration, due to the matching correct global correspondence was obtained. In the second set, two objects were observed by an EO and IR camera as shown in Figure 4.17. Both objects were continuously viewed in the EO camera, but one object repeatedly exited and re-entered the field of view of the IR camera. Using the trajectory repairing algorithm the object was successfully re-associated.

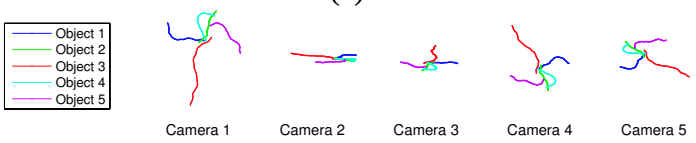
4.4 Discussion and Conclusion

Using multiple UAVs for aerial reconnaissance is an idea of wide applicability. While several algorithms have been proposed for rearranging the positions of the UAVs based on some sensors like GPS or INS for optimal coverage, object association across multiple UAVs presents an interesting option once the 'control loop' is closed, namely that of rearranging multiple sensors using image information and knowledge of object associations. Instead of a cost function of maximum coverage, or maximum overlap between UAVs, more intelligent cost functions based on object positions, proximity or object importance can be autonomously used. In this chapter, a method to associate objects across multiple airborne cameras was presented. We make two fundamental assumptions about the data: (1) That the altitude of the aerial vehicle upon which the camera is mounted is significantly high with respect to the ground, that a planar assumption is viable, and (2) that at least one object is seen simultaneously between every pair of cameras for at least 5 frames (1/6th of a second). Given these assumptions, and taking as input the time-stamped trajectories of objects observed in each camera, we estimate the inter-camera transformations, the association of each object across the views, and 'canonical' trajectories, which are the best estimate (in a maximum likelihood sense) of the original object trajectories up to a 2D projective transformation. To that end, we describe an extension to the re-projection error for multiple views, providing a geometrically and statistically sound means of evaluating the likelihood of a candidate correspondence set. We formulate the problem of maximizing this joint likelihood function as a k-dimensional matching problem and use an approximation that maintains transitive closure. The estimated so-

lution is verified using a strong global constraint for the complete set of correspondences across all cameras. In addition, we show that all the available data can be conveniently viewed in a concurrent mosaic. We evaluated our approach with both simulated and real data. In the simulations we tested the sensitivity of the approach to noise strength (in terms of the motion-to-noise ratio), the number of cameras, the number of frames viewed, and the 'collinearity' of the trajectories. We demonstrated qualitative results on several real sequences, including the standard VIVID data set and the ARDA VACE data, for multiple cameras and between IR and EO video. There are several future directions and applications that can be explored, such as resolving occlusions and re-entries, relaxing the planar constraint and relaxing the constraint of spatiotemporal overlap.

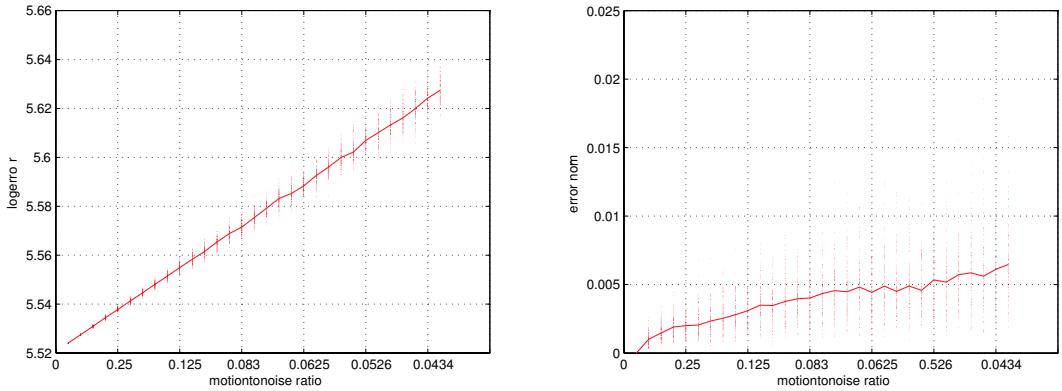


(a)



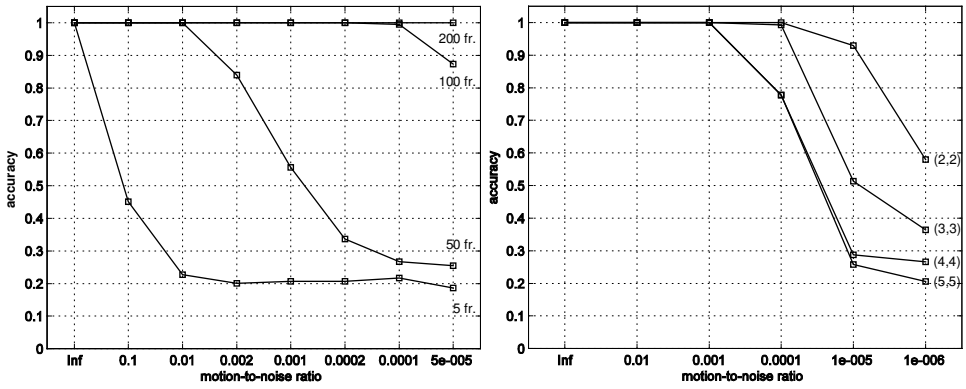
(b)

Figure 4.7: Data generation. The randomly generated data captures the smoothness of real trajectories. There are 5 cameras observing 5 objects, for 100 frames. The mean of motion magnitude, $\bar{\rho}$ was set to 50, the noise variance, σ_ϵ was 2. (a) 5 objects viewed in 5 cameras. Each row corresponds to the image of the trajectory in that camera. (b) The image of all objects in each camera, with trajectories color-coded for association.



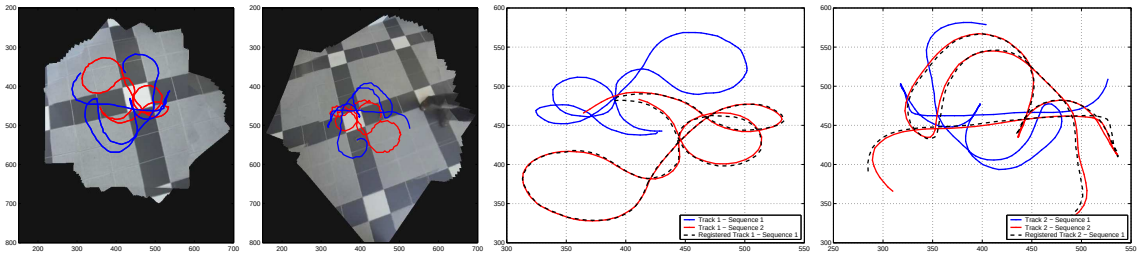
(a) (b)

Figure 4.8: Accuracy of the Estimated Parameters. (a) The log-likelihood of the canonical tracks, as the motion-to-noise ratio was increased, across 3 cameras observing 3 objects. (b) The error norm of the estimated to the true homography. A hundred iterations were run for each noise level which are plotted (dots) along with the median value (line).



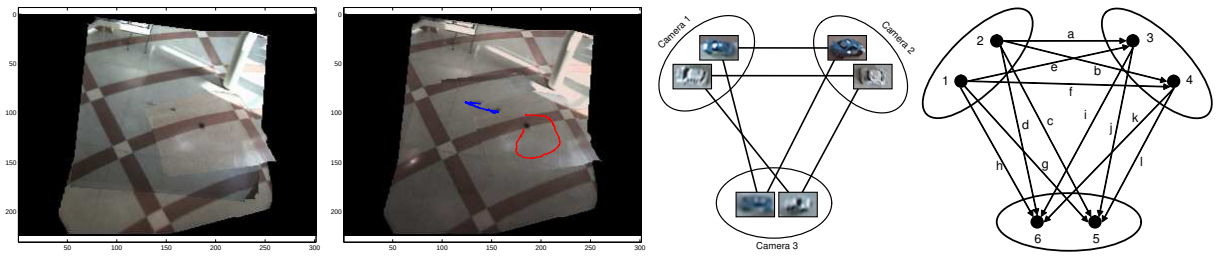
(a) (b)

Figure 4.9: Association accuracy w.r.t number of cameras, number of objects, number of frames and motion-to-noise ratio. Note: the horizontal axis is not progressing linearly. (a) For ten cameras with ten objects the percentage of correct associations to the total number of associations. (b) As the number of cameras and objects increase linearly, for a fixed 60 frames, the association accuracy decreases. The results are the average of 100 runs.



(a) (b) (c) (d)

Figure 4.10: Controlled Experiment 1. Two remote controlled cars move around on a tiled floor. The trajectory of the first car is shown by the red curve and the trajectory of the second car is shown by the blue curve for the first camera in (a) and for the second camera in (b). Registered Tracks. The trajectories of each object in Sequence 1 (red) and Sequence 2 (blue) are shown, along with the trajectory of Sequence 2 registered to Sequence 1 (dashed black) using the inter-camera homography for the first and second camera in (c) and (d) respectively.



(a) (b) (c) (d)

Figure 4.11: Concurrent visualization of three sequences. (a) Concurrent mosaic before blending, (b) Blended concurrent mosaic with the track overlayed. Matching in three sequences. (c) Matching of the tripartite graph, (d) The corresponding directed graph.

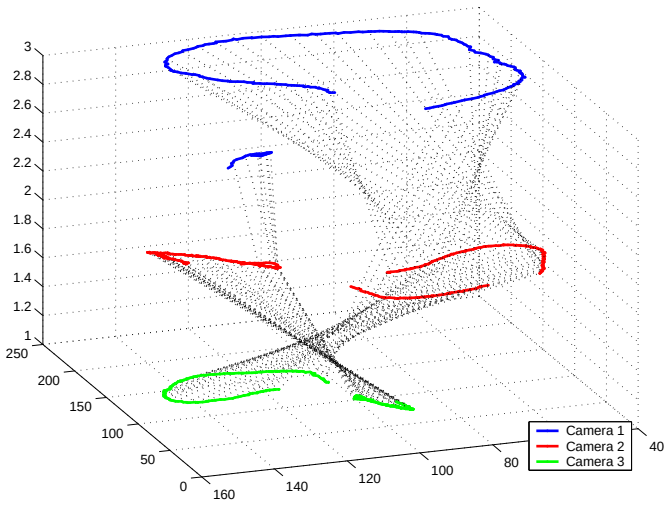
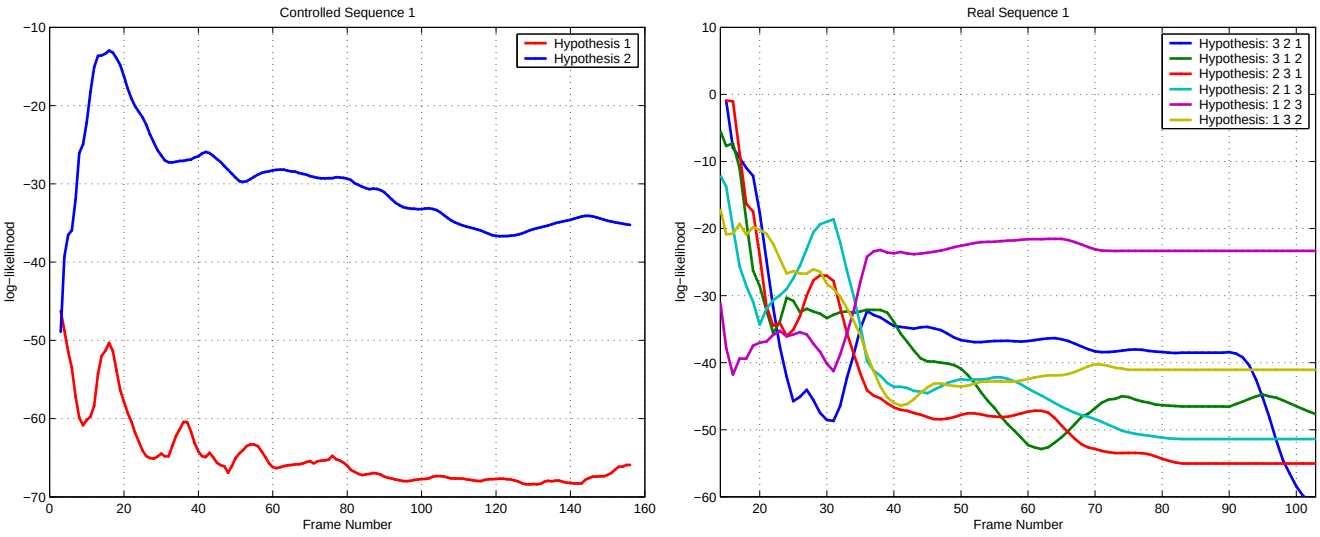
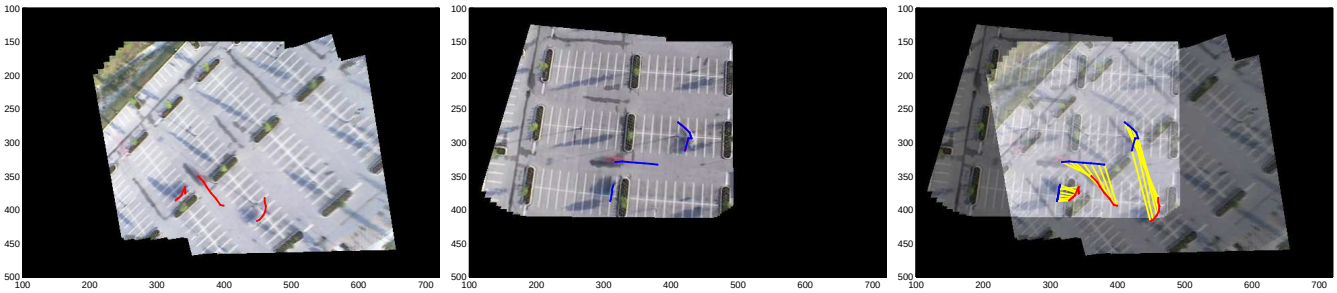


Figure 4.12: Correspondence across the 3 moving cameras. For each sequence, each pair of tracks is plotted at a level. The point-wise correspondence is show by the dotted black line.



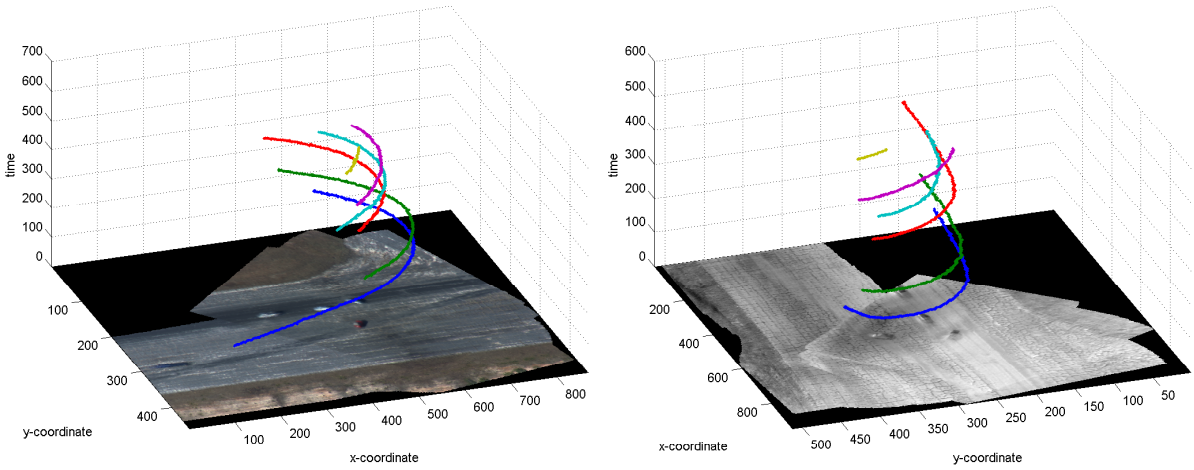
(a) (b)

Figure 4.13: Variation of some global correspondence hypotheses. (a) Variation for Controlled Experiment 1. (b) Variation for UAV Experiment 2. Due to colinear motion of the object, ambiguity in correspondence exists initially which is quickly resolved as the object begin to show more non-colinear behavior.



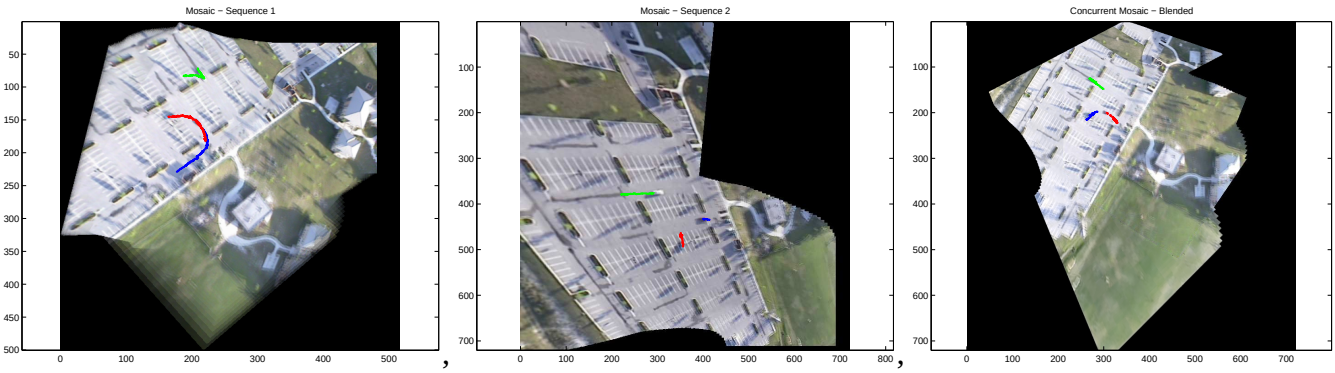
(a) (b) (c)

Figure 4.14: Object association across two real sequences. (a) The red points show tracks of three objects detected and tracked in the first sequence (b) The blue points show the tracks of the same three objects detected and tracked in the second sequence and (c) Correspondences between the points are shown in a single plot by the yellow lines.



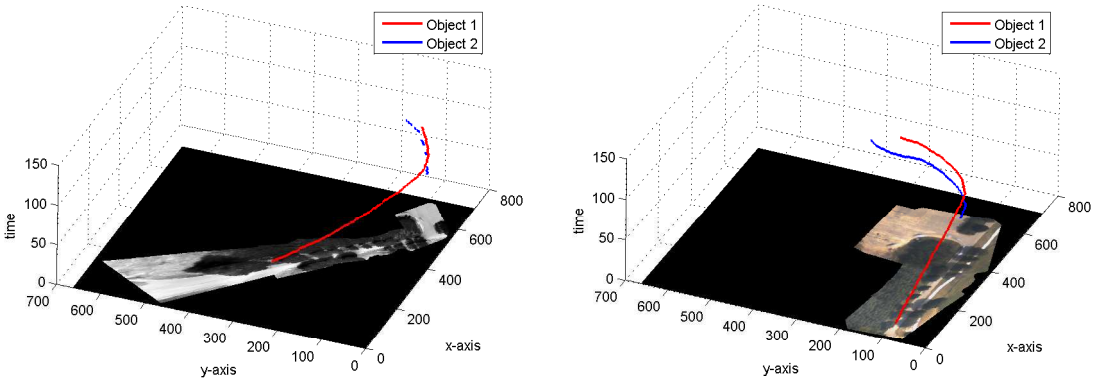
(a) (b)

Figure 4.15: First UAV Experiment - two cameras, six objects. (a) The EO video, (b) The IR video. Since we are using only motion information, association can be performed across different modalities.

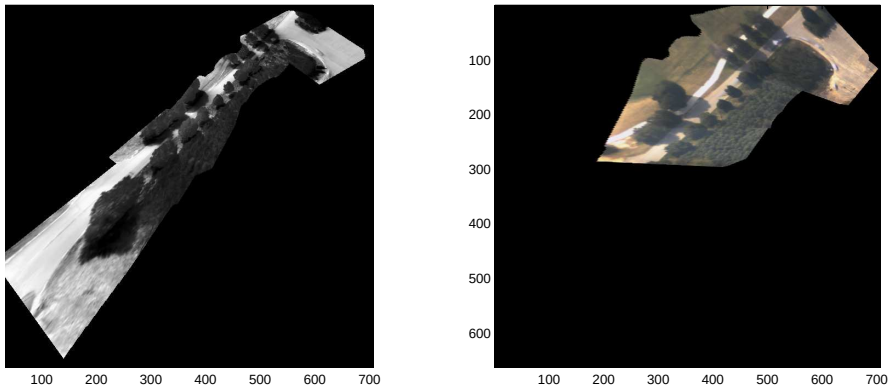


(a) (b) (c)

Figure 4.16: Second UAV experiment - Short temporal overlap. Despite a very short duration of overlap, correct correspondence was estimated. (a) Mosaic of Sequence 1 (b) Mosaic of Sequence 2 (c) Concurrent visualization of two sequences. The two mosaics were blended using a quadratic color transfer function. Information about the objects and their motion is compactly summarized in the concurrent mosaic.



(a)



(b)

Figure 4.17: Repairing broken trajectories. (a) Due to rapid motion of the camera, the object corresponding to the blue trajectory exited and re-entered the field of view of the IR camera several times. On the other hand the same object in the EO camera remained continuously visible. The trajectories were successfully re-associated. (b) The aligned mosaics.

CHAPTER 5

OBJECT ASSOCIATION ACROSS MULTIPLE CAMERAS

In this chapter we present a unified framework for the association of multiple objects across multiple cameras in planar scenes. This approach makes additional assumptions on the object kinematics but is able to recover object associations, inter-camera transformations and canonical trajectories across cameras irrespective of whether the cameras are stationary or moving, or whether the fields of view are overlapping or not as long as the kinematic model is valid. The intuition used to solve this problem is that association across cameras with spatiotemporally non-overlapping fields of view can be achieved by explicitly modeling the motion of objects, thus providing constraints for the estimation of inter-camera homographies, as shown in Figure 5.1. We use polynomial kinematic models for the motion of objects and under this model an Expectation Maximization algorithm is formulated to estimate the inter-camera homographies and motion parameters.

There are two principal applications where the algorithms in this chapter can be used. First, where multiple aerial cameras at high altitudes, observing objects such as vehicles or people move along the ground, and the problem is to recover the association of the objects across cameras and estimate the inter-camera transformations. Second, for a single camera in this setting if, due to

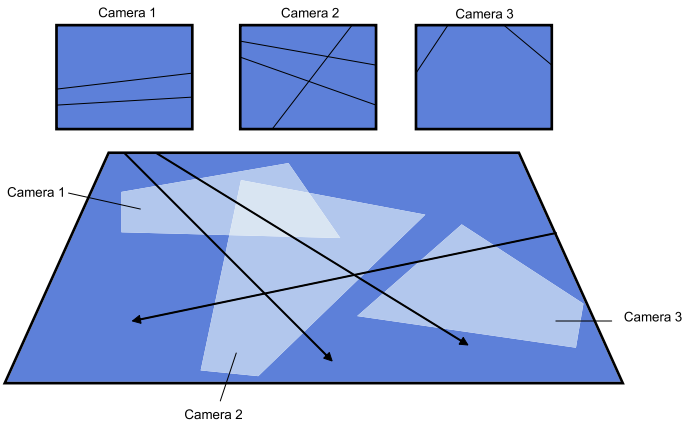


Figure 5.1: A unified framework for estimating inter-camera transformations. Overlapping and non-overlapping views are handled identically, since we look at global object motion models rather than pairwise correspondences.

the motion of the camera, an object exits and then re-enters the field of view of one camera, the problem of reassociation can also be solved in this context.

Most existing approaches to estimating inter-camera homographies from curves, such as conics, perform the matching given the parameters of the curves. The general theory is covered in [KS04]. A separate portion of literature cover the problem of fitting curves to points - a survey for conics can be found in [FF95]. In this chapter, we fuse the two problems, of estimating curve parameters and the homographies simultaneously. The benefit of this approach is two-fold. First, it is difficult to characterize an error model for curve coefficients, since they are not usually directly measurable. On the other hand, it is reasonable to assume an error model for point detection, and then develop statistically meaningful estimation algorithms for estimating homographies between views. Second, since only a portion of the curve is observed in each view, it is likely that the curve

may be erroneously fit. This is due to the fact that samples from the curve are localized in small intervals for each view (partial occlusion). By estimating curve parameters and homographies simultaneously, recovery is possible from local (in each camera) over-fitting.

5.1 Data Model

The scene is modeled as a plane in 3-space, $\mathbf{\Pi}$, with K objects moving at constant velocity. The k-th object[1], O_k, moves along a trajectory on $\mathbf{\Pi}$, represented by a time-ordered set of points, $\mathbf{x}_k(t) = \big(x_k(t), y_k(t)\big) \in \mathbb{R}^2$, where $x_k(t)$ and $y_k(t)$ evolve according to some spatial algebraic curve such as a line, a quadratic or a cubic. The finite temporal support is denoted by Δt. The scene is observed by N perspective cameras, each observing some subset of the entire scene motion, due to a spatially limited field of view and temporally limited window of observation (due to camera motion). The imaged trajectory observed by the n-th camera for O_k is $\mathbf{x}_k^n(t)$. As we did in the last chapter, we assume that within each sequence frame-to-frame motion within camera has been compensated so $\mathbf{x}_k^n(t)$ is in a single reference coordinate. The measured image positions of objects, $\bar{\mathbf{x}}_k^n$ are described in terms of the canonical image positions, $\mathbf{x}_k^n$, with independent normally distributed measurement noise, $\mu = \mathbf{0}$ and covariance matrix $\mathbf{R}_k^n$, that is

$$\bar{\mathbf{x}}_k^n(i) = \mathbf{x}_k^n(i) + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k^n). \tag{5.1}$$

[1]The abstraction of each object is as a point, such as the centroid. It should be noted, however, that since the centroid is not preserved under general perspective transformations using the centroid will introduce bias.

The imaged trajectory is related to $\mathbf{x}_k(t)$ by a projective transformation denoted by an invertible 3×3 matrix, $\mathbf{H}^n \in PGL(3)$. The homogeneous representation of a point $\mathbf{x}_k^n(t)$ is $\mathcal{X}_k^n(t) = (\lambda x_k^n(t), \lambda y_k^n(t), \lambda) \in \mathbb{P}^2$. Thus, we have,

$$\mathcal{X}_k^n(t) = \mathbf{H}^n \mathcal{X}_k(t).$$

Finally, we introduce the association or correspondence variables $\mathbf{C} = \{c_k^n\}_K^N$, where $c_j^i = m$ that represents the hypothesis that O_i^j is the image of O_m, where $p(c)$ is the probability of association c. Since the association of an imaged trajectory with different scene trajectories are mutually exclusive and exhaustive,

$$\sum_{l=1}^{K} p(c_k^n = l) = 1. \tag{5.2}$$

A term $p(c_k^n = 0)$ may be included to model the probability of spurious trajectories but we do not consider this in the remainder of this work (i.e. we assume $p(c_k^n = 0) = 0$).

5.1.1 Kinematic Polynomial Models

The position $\mathbf{x}_j(t)$ of an Object O_j is modeled as an $d - th$ order polynomial in time,

$$\mathbf{x}_j(t) = \sum_{i=0}^{d} \mathbf{p}_i t^i, \tag{5.3}$$

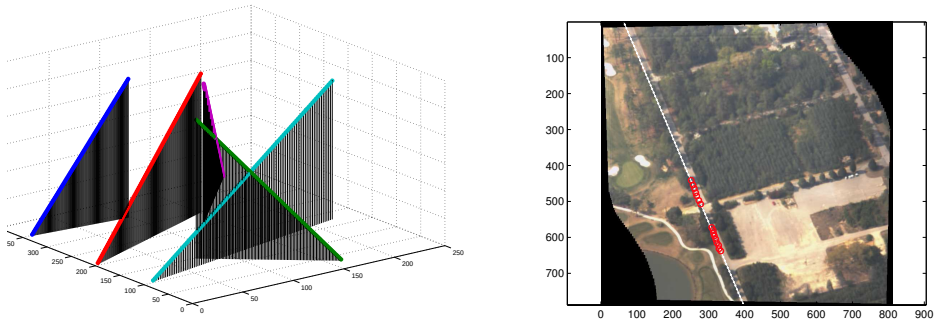
where $\mathbf{p}_i$ are the coefficients of the polynomial. In matrix form,

$$\mathbf{x}_j = \mathbf{P}_j \mathbf{t}^{(d)} = \begin{bmatrix} p_{x,0} & p_{x,1} & \dots & p_{x,d} \\ p_{y,0} & p_{y,1} & \dots & p_{y,d} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \vdots \\ t^d \end{bmatrix}.$$

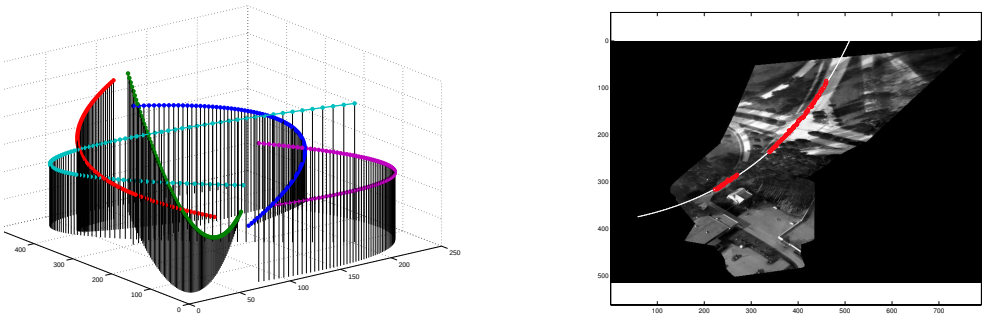
Selecting the appropriate order of polynomials is an important consideration. If the order is too low, the polynomial may not correctly reflect the kinematics of the object. On the other hand, if the order is too high, some of the estimated coefficients may not be statistically significant, [Ed90]. This problem is even more important in the situation under study since oftentimes only a segment of the polynomial is observed and over or under-fitting is likely. Thus, numerical considerations while estimating the coefficients of the curve are of paramount important, especially during the optimization routine. Readers are advised to refer to [HZ00] for information on numerical conditioning during estimation. The monograph by Fitzgibbon and Fischer [FF95] on conic fitting is also informative.

5.1.1.1 Linear (Constant Velocity) Model

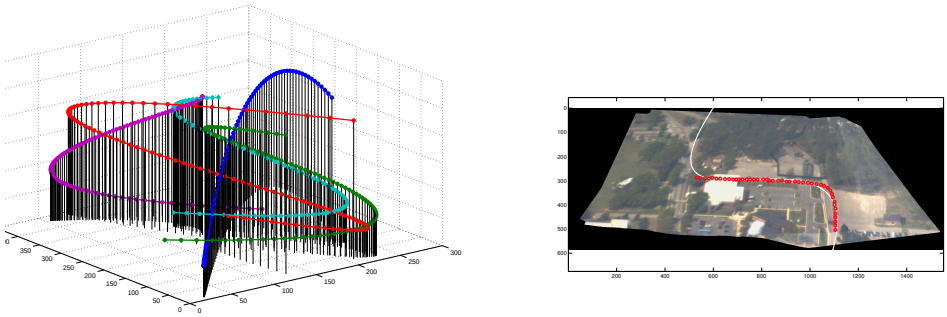
For first order polynomials, the number of parameters reduces to $3 \times K + 9 \times N$. The number of parameters that need to be estimated when a parametric quadratic curve is to be fit to the trajectories is at most $6 \times K + 9 \times N$, since there are K curves which are described by 8 parameters each, with N homographies, each with 9 unknowns. At least two points per object must be observed and four



(a)



(b)



(c)

Figure 5.2: Space-time plots of different models. Synthetic (left) and real (right) trajectories following (a) a Linear Model (b) a Quadratic Model and (c) a Cubic Model.

lines must be observed between a pair of views. We can then use the parametric representation of a line as,

$$\mathbf{x}(t) = \mathbf{p}_1 t + \mathbf{p}_0, \tag{5.4}$$

where $\mathbf{p}_0 = [p_{x,0}\ p_{y,0}]^T$ and $\mathbf{p}_2 = [p_{x,1}\ p_{y,1}]^T$ and therefore in this case

$$\mathcal{P} = \begin{bmatrix} p_{x,0} & p_{x,1} \\ p_{y,0} & p_{y,1} \\ 1 & 1 \end{bmatrix}.$$

5.1.1.2 Quadratic (Constant Acceleration) Model

The number of parameters that need to be estimated when a parametric quadratic curve is to be fit to the trajectories is at most $6 \times K + 9 \times N$, since there are K curves which are described by 8 parameters each, with N homographies, each with 9 unknowns. At least three points per object must be observed. The parametrization for a quadratic curve is,

$$\mathbf{x}(t) = \mathbf{p}_2 t^2 + \mathbf{p}_1 t + \mathbf{p}_0, \tag{5.5}$$

In this case

$$\mathcal{P} = \begin{bmatrix} p_{x,0} & p_{x,1} & p_{x,2} \\ p_{y,0} & p_{y,1} & p_{y,2} \\ 1 & 1 & 1 \end{bmatrix}.$$

5.1.1.3 Cubic Model

The number of parameters that need to be estimated when a parametric cubic curve is to be fit to the trajectories is at most $8 \times K + 9 \times N$, since there are K curves which are described by 8 parameters each, with N homographies, each with 9 unknowns. At least four points per object must be observed and just one curve must be observed between a pair of views. The parametrization for a cubic curve is,

$$\mathbf{x}(t) = \mathbf{p}_0 t^3 + \mathbf{p}_1 t^2 + \mathbf{p}_2 t + \mathbf{p}_3. \tag{5.6}$$

In this case

$$\mathcal{P} = \begin{bmatrix} p_{x,0} & p_{x,1} & p_{x,2} & p_{x,3} \\ p_{y,0} & p_{y,1} & p_{y,2} & p_{y,3} \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

5.1.2 Imaging and the Error Model

Since the scene is modeled as a plane, a point on $\mathbf{\Pi}$ is related to its image in the n-th camera by $\mathbf{H}$. Thus a measured point $\mathcal{X}_j^i$ at time t under the model $\mathbf{P}_m$ is,

$$\bar{\mathcal{X}}_j^i = \mathbf{H}^i \mathcal{P}_j \mathbf{t}^{(d)} + \tilde{\epsilon}, \tag{5.7}$$

or more explicitly,

$$\begin{bmatrix} \lambda\bar{x}(t) \\ \lambda\bar{y}(t) \\ \lambda \end{bmatrix} = \begin{bmatrix} h_1^{(i)} & h_2^{(i)} & h_3^{(i)} \\ h_4^{(i)} & h_5^{(i)} & h_6^{(i)} \\ h_7^{(i)} & h_8^{(i)} & h_9^{(i)} \end{bmatrix} \begin{bmatrix} p_{x,0}^{(j)} & p_{x,1}^{(j)} & \dots & p_{x,d}^{(j)} \\ p_{y,0}^{(j)} & p_{y,1}^{(j)} & \dots & p_{y,d}^{(j)} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \vdots \\ t^d \end{bmatrix} + \begin{bmatrix} \lambda\epsilon \\ 0 \end{bmatrix}. \tag{5.8}$$

5.1.3 Problem Statement

Given the trajectory measurements for each camera $\{\bar{\mathbf{x}}_k^n\}_K^N$, find associations $\mathbf{C}$ of each object across cameras and the Maximum Likelihood Estimate of $\mathbf{\Theta} = (\{\mathbf{P}_k\}_K, \{\mathbf{H}_n\}^N)$, where $\{\mathbf{P}_k\}_K$ are the motion parameters of the K objects, and $\{\mathbf{H}_n\}_N$ are the set of homographies to $\mathbf{\Pi}$. For the remainder of this chapter, θ_k^n represents $(\mathbf{P}_k, \mathbf{H}^n)$.

5.2 Maximum Likelihood Estimation

We wish to find the Maximum Likelihood Estimate of the scene parameters, $\mathbf{\Theta}$, and recover the correct associations of objects, $\mathbf{C}$ from the observed trajectories $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_k^n\}_K^N$. For each individual observed trajectory $\bar{\mathbf{x}}_k^n$ we have,

$$p(\bar{\mathbf{x}}_j^i | c_j^i, \mathbf{\Theta}) = p(\bar{\mathbf{x}}_j^i | \theta_{c_j^i}^i) = \prod_{t=\delta_{\alpha(i,j)}}^{\delta_\omega(i,j)} p(\bar{\mathbf{x}}_j^i(t) | \mathbf{x}_j^i(t)), \tag{5.9}$$

where $\delta_\alpha(i,j)$ and $\delta_\omega(i,j)$ are the start-time and end-time of O_j^i respectively[2]. Computing $\mathbf{x}_m^n(t)$ requires description of the object kinematic model, which we described in Section 5.1.1. Applying Bayes Theorem to Equation 5.11 and assuming conditional independence between trajectories we then have,

$$p(\bar{\mathbf{X}}, \mathbf{C}|\mathbf{\Theta}) = \prod_{i=1}^{N}\prod_{j=1}^{\mathbf{z}(i)} p(\bar{\mathbf{x}}_j^i|c_j^i, \mathbf{\Theta})p(c_j^i) = \prod_{i=1}^{N}\prod_{j=1}^{\mathbf{z}(i)} \frac{1}{K} p(\bar{\mathbf{x}}_j^i|\theta_{c_j^i}^i). \tag{5.10}$$

Thus, the complete data log-likelihood,$p(\bar{\mathbf{X}}, \mathbf{C}|\mathbf{\Theta})$ is,

$$\log p(\bar{\mathbf{X}}, \mathbf{C}|\mathbf{\Theta}) = \sum_{i=1}^{N}\sum_{j=1}^{z(i)} \log \frac{1}{K} p(\bar{\mathbf{x}}_j^i|\theta_{c_j^i}^i). \tag{5.11}$$

The problem, of course, is that we do not have measurements of $\mathbf{C}$. Therefore, the best we can do is to find the Maximum Likelihood Estimate of $\mathbf{\Theta}$ given $\bar{\mathbf{X}}$, i.e.

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} p(\bar{\mathbf{X}}|\mathbf{\Theta}). \tag{5.12}$$

To evaluate the MLE we need to (i) describe how to evaluate $p(\bar{\mathbf{X}}|\mathbf{\Theta})$ and (ii) describe a maximization routine. By marginalizing out the association in Equation 5.9, $p(\bar{\mathbf{x}}_k^n|\mathbf{\Theta})$ can be expressed as a mixture model,

$$p(\bar{\mathbf{x}}_j^i|\mathbf{\Theta}) = \frac{1}{K}\sum_{m=1}^{K} p(\bar{\mathbf{x}}_j^i|\theta_m^i). \tag{5.13}$$

Then, the incomplete data log-likelihood from the data is given by,

$$\log \mathcal{L}(\mathbf{\Theta}|\bar{\mathbf{X}}) = \log \prod_{i=1}^{N}\prod_{j=1}^{z(i)} p(\bar{\mathbf{x}}_j^i|\mathbf{\Theta})) = \sum_{i=1}^{N}\sum_{j=1}^{z(i)} \log \frac{1}{K}\sum_{m=1}^{K} p(\bar{\mathbf{x}}_j^i|\theta_m^i). \tag{5.14}$$

[2]Evaluating $p(\bar{\mathbf{x}}_j^i(t)|\mathbf{x}_j^i(t))$ requires a measurement error model to be defined, e.g. normally distributed in which case $p(\bar{\mathbf{x}}_j^i(t)|\mathbf{x}_j^i(t)) = \mathcal{N}(\bar{\mathbf{x}}_j^i(t)|\mathbf{x}_j^i(t), \mathbf{R}_j^i)$.

This function is difficult to maximize since it involves the logarithm of a large summation. The Expectation-Maximization Algorithm provides a means to maximize $p(\bar{\mathbf{X}}|\mathbf{\Theta})$, by iteratively maximizing a lower bound,

$$\mathbf{\Theta}^{+} = \arg\max_{\mathbf{\Theta}} \mathcal{Q}(\mathbf{\Theta}, \mathbf{\Theta}^{-})) = \arg\max_{\mathbf{\Theta}} \sum_{\mathbf{C}\in\mathcal{C}} p(\mathbf{C}|\bar{\mathbf{X}}, \mathbf{\Theta}^{-}) \log p(\bar{\mathbf{X}}, \mathbf{C}|\mathbf{\Theta}), \tag{5.15}$$

where $\mathbf{\Theta}^{-}$ and $\mathbf{\Theta}^{+}$ are the current and the new estimates of $\mathbf{\Theta}$, respectively, and $\mathcal{C}$ is the space of configurations that $\mathbf{C}$ can assume. To evaluate this expression we have,

$$p(\mathbf{C}|\bar{\mathbf{X}}, \mathbf{\Theta}^{-}) = \prod_{i=1}^{N} \prod_{j=1}^{z(i)} p(c_j^i|\bar{\mathbf{x}}_j^i, \mathbf{\Theta}^{-}), \tag{5.16}$$

where

$$p(c_j^i|\bar{\mathbf{x}}_j^i, \mathbf{\Theta}^{-}) = \frac{p(\bar{\mathbf{x}}_j^i|c_j^i, \mathbf{\Theta}^{-})p(c_j^i)}{p(\bar{\mathbf{x}}_j^i|\mathbf{\Theta}^{-})} = \frac{\frac{1}{K}p(\bar{\mathbf{x}}_j^i|\theta_{c_j^i}^{i-})}{\sum_{j=1}^{K}\frac{1}{K}p(\bar{\mathbf{x}}_j^i|\theta_{c_j^i}^{i-})}. \tag{5.17}$$

After manipulation (see [Bil97]), we get an expression for $\mathbf{\Theta}$,

$$\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Theta}^{-}) = \sum_{\mathbf{C}\in\mathcal{C}} p(\mathbf{C}|\bar{\mathbf{X}}, \mathbf{\Theta}^{-}) \log p(\bar{\mathbf{X}}, \mathbf{C}|\mathbf{\Theta}) = \sum_{m=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{z(i)} p(c_j^i = m|\bar{\mathbf{x}}_j^i, \theta_m^{i-}) \log \frac{1}{K} p(\bar{\mathbf{x}}_j^i|\theta_m^i). \tag{5.18}$$

In order to derive the update terms for $\mathbf{H}$ and $\mathbf{P}$, we need to make explicit the algebraic curve we are using to model the object trajectory and the measurement noise model.

If noise is normally distributed,

$$p(\bar{\mathbf{x}}_k^n|\theta_m^n) = \prod_{t=\delta_\alpha(n,k)}^{\delta_\omega(n,k)} \frac{1}{(2\pi\|\mathbf{R}\|)^{\frac{1}{2}}} e^{-\frac{1}{2}d(\bar{\mathbf{x}}_k^n(t), \mathbf{x}_m^n(t))}, \tag{5.19}$$

where $\delta_\alpha(i,j)$ and $\delta_\omega(i,j)$ are the start-time and end-time of O_j^i respectively. The probability $p(\bar{\mathbf{X}}|\mathbf{C}, \mathbf{\Theta})$ can be evaluated as follows,

$$p(\bar{\mathbf{X}}|\mathbf{C}, \mathbf{\Theta}) = \prod_{n=1}^{N}\prod_{k=1}^{z(n)}\prod_{t=\delta_\alpha(n,k)}^{\delta_\omega(n,k)} \frac{1}{(2\pi\|\mathbf{R}\|)^{\frac{1}{2}}} e^{-\frac{1}{2}d(\bar{\mathbf{x}}_k^n(t), \mathbf{x}_{c_k^n}^n(t))}, \tag{5.20}$$

where

$$d(\bar{\mathbf{x}}_k^n(t), \mathbf{x}_{c_k^n}^n(t)) = (\bar{\mathbf{x}}_k^n(t) - \mathbf{x}_{c_k^n}^n(t))^T \mathbf{R}^{-1} (\bar{\mathbf{x}}_k^n(t) - \mathbf{x}_{c_k^n}^n(t)),$$

and $\mathbf{x}_{c_k^n}(t)$ is the corresponding point that lies *exactly* on the curve described by $\mathbf{P}_k$, and is computed using t. It is transformed to the coordinate system of Camera n using $\mathbf{H}_n$. Explicitly,

$$[\lambda x_{c_k^n}^n(t)\ \lambda y_{c_k^n}^n(t) \lambda]^T = \mathbf{H}_n [x_{c_k^n}(t)\ y_{c_k^n}(t)\ 1]^T. \tag{5.21}$$

Taking the logarithm,

$$\log p(\bar{\mathbf{X}}|\mathbf{C}, \mathbf{\Theta}) = \sum_{n=1}^{N} \sum_{k=1}^{z(n)} \sum_{t=\delta_\alpha(n,k)}^{\delta_\omega(n,k)} -\frac{1}{2} d(\bar{\mathbf{x}}_k^n(t), \bar{\mathbf{x}}_{c_k^n}^n(t)) + \text{constant}. \tag{5.22}$$

It is instructive to note that unlike the Maximum Likelihood term for independent point detections defined in terms of the reprojection error in [Stu97], where the parameters of re-projection error function include 'error free' data points, the curve model fit on the points allows the error function to be written compactly in terms of the parameters of the curve and a scalar value denoting the position along the curve (taken here to be the time index t). This drastically reduces the number of parameters that need to be estimated.

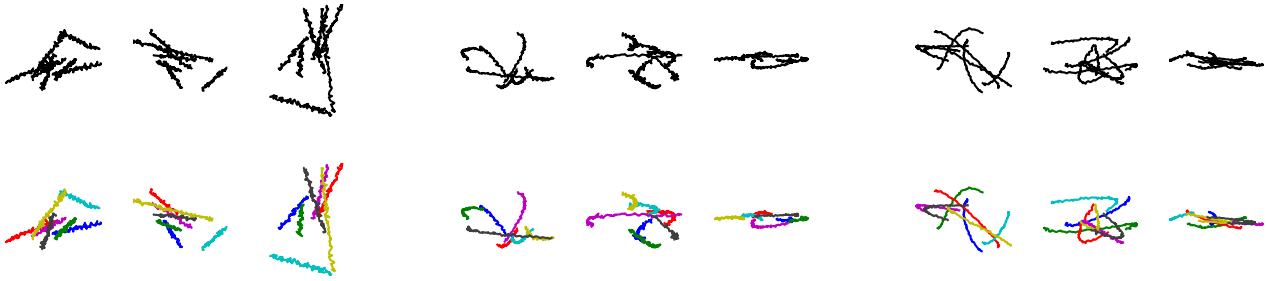
We need an analytical expression for $\log \frac{1}{K} p(\bar{\mathbf{x}}_j^i | \theta_m)$, which will then be maximized the so-called 'M-step'. We then need to evaluate $\{\frac{df}{dh_1^i}, \ldots, \frac{df}{dh_9^i}, \frac{df}{dp_1^i}, \ldots, \frac{df}{dp_4^i}\}$ for each of the cameras (except the reference camera) and all the world objects, which is straightforward. The Jacobian can then be created to guide nonlinear minimization algorithms (such as the Levenberg-Marquardt algorithm).

5.3 Initialization

Good initialization of Θ is an important requirement of the EM algorithm. There are several initialization methods that can be used. Ideally, for the inter-frame homographies, telemetry information, which is usually noisy, can be used for initialization. Alternatively, a rough correspondence can be computed using appearance values and initial estimates of homographies and curve coefficients can be estimated using robust methods. For the second application, i.e. reacquisition of objects in single views, the initialization is simpler: estimate of the initial homography can be computed using the frame-to-frame homography estimation, and the curve coefficients can be initialized by estimating them w.r.t to the original trajectories (before exit).

5.4 Experimentation and Results

We performed quantitative analysis through simulations to test the behavior of the proposed approach to noise. In addition, we show qualitative results on a number of real sequences, recovering the true underlying scene geometry and object kinematics. For the real sequences the video was collected by cameras mounted on aerial vehicles. Frame to frame registration was performed using robust direct registration methods. Object detection and tracking were performed partially using the COCOA system and partly through manual tracking.



(a) (b) (c)

Figure 5.3: Randomly generated imaged trajectories. Seven object seen from three cameras. (a) Linear Model (b) Quadratic Model (c) Cubic Model. The top row show the trajectories unlabeled, bottom row shows them labeled.

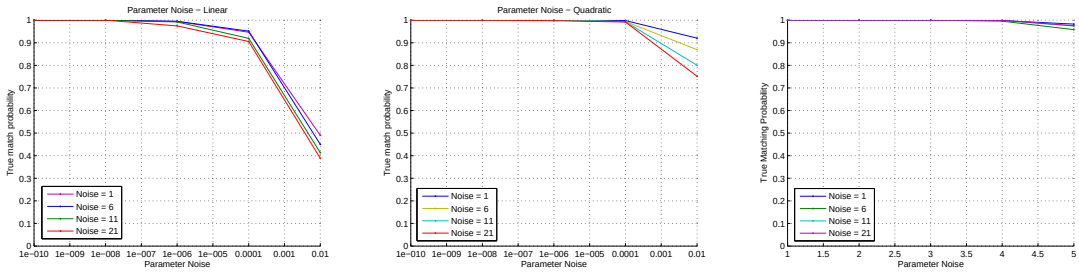


Figure 5.4: Simulations

5.4.1 Simulations

In this set of experiments we generated random trajectories fitting the prescribed model. The variable scene descriptors included number of objects, number of cameras, number of frames (observations). For each camera there was a separate probability of observation of an object, and for each object a duration of observation was randomly selected. In this way, spatio-temporal overlap

was not guaranteed during data generation. A noise parameter was set for introducing errors into the true parameter values (camera parameters and curve coefficients), which were then treated as initial estimates. The homographies subtended by the camera parameters were calculated and used to project each curve onto the image, depending on its probability of observation and its duration of observation. Zero mean noise was then added to the projected points.

We tested the sensitivity of the algorithm with respect to corruption of the curve coefficients by white noise and with respect to measurement error. For these experiments 5 object trajectories were randomly generated according to linear, quadratic and cubic models, and two homographies (two cameras) were generated. The probability of observation was set to 1 so that both cameras were guaranteed to see both object (just not at the same time). Only 10 frames were observed, and 10 iterations of the EM algorithm were run. Four measurement noise levels were tested: 1, 6, 11 and 21, against five coefficient noise levels of $1 \times 10^{-10}, 1 \times 10^{-8}, 1 \times 10^{-6}, 1 \times 10^{-4}$ and 1×10^{-2} and each configuration was repeated 25 times (to generate meaningful statistics). We performed this for both linear and quadratic curves. This experiment shows that quadratic curves are less susceptible to noise, which follows intuition since more information on the underlying homography is placed by a quadratic curve than a line.

5.4.2 Real Sequences

In this set of experiments, we study the association of objects across multiple sequences in real videos. We tested the proposed approach on three sequences. In the first sequence, several cars were moving in succession along a road, shown in Figure 5.5. From the space-time plot it is clear that one of the objects is moving quicker than the rest of the objects (indicated by the angle with the horizontal plane). The linear (constant velocity) model was used for this experiment. Figure 5.6 shows the association probabilities arranged in an adjacency matrix between the model lines and the observed trajectories. In just six iterations the correct associations are discerned, and as shown in Figure 5.7 the trajectories are correctly aligned. It should be noted that in this case the lines were parallel they did not strictly constrain the homography. However, the correct association was still found, and the homography estimate was also reasonable.

In the second experiment, both humans and vehicles were moving for short durations, with the two views shown in Figure 5.8. The initial misalignment is over 100 pixels but our approach successfully recovers the correct alignment (shown in Figure 5.9). Tables 5.4.2 and 5.4.2 show the Adjacency matrix before and after the application of the approach respectively. The correct associations have been made, despite the close parallelism and proximity of Objects 4 and 5 (see legend in Figure 5.9). A linear kinematic model was used in this experiment.

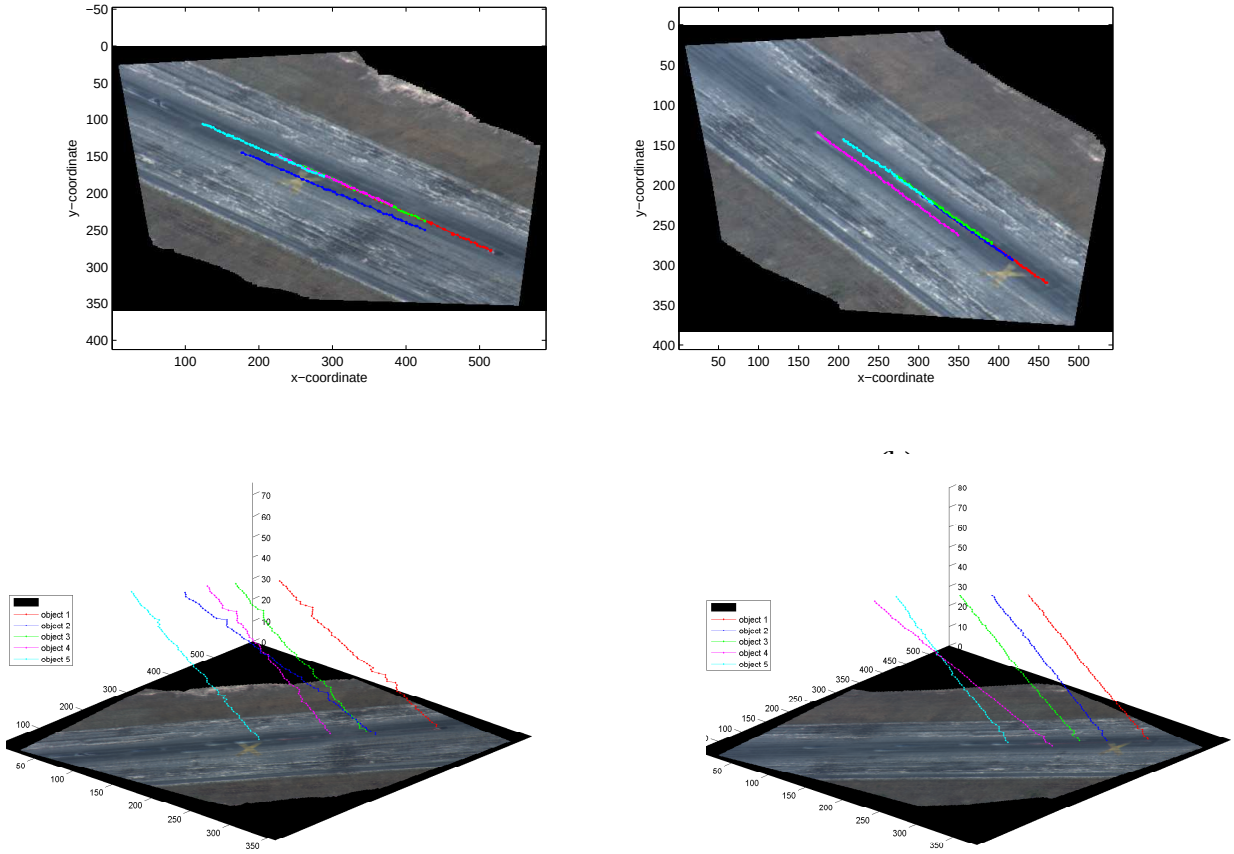
In the second experiment a quadratic kinematic model was used during experimentation in two sequences. Figure 5.10 shows the relative positions of the first set of sequences before (a) and after (b) running the proposed approach. It can be observed that the initial misalignment was almost

Table 5.1: Initial association table for objects in the disconnected segment. The values represent the probability of the i-th object matching the j-th model. $i = j$ are the ground truth associations.

	Object 1	**Object 2**	**Object 3**	**Object 4**	**Object 5**
Model 1	0.99999312	0.00948986	0.50000×10^{-8}	0.99813573	0.99181073
Model 2	0.6870×10^{-5}	0.99028289	0.615×10^{-5}	0.00024566	0.00008245
Model 3	0	0	0.97620477	0	0
Model 4	0.680×10^{-9}	0.00009831	0.01515987	0.00139656	0.00667508
Model 5	0.255×10^{-10}	0.00012892	0.00862918	0.00022202	0.00143172

Table 5.2: Final association table for objects in the disconnected segment. The values represent the probability of the i-th object matching the j-th model. $i = j$ are the ground truth associations. Despite correct resolution of association, the ambiguity between Object 4 and Object 5 is due to their spatial proximity (see Figure 5.8).

	Object 1	**Object 2**	**Object 3**	**Object 4**	**Object 5**
Model 1	0.99997424	0.1304×10^{-6}	0	0.00001085	0.7834×10^{-6}
Model 2	0.1445×10^{-6}	0.99999986	0	0.2×10^{-13}	0
Model 3	0	0	0.99999997	0.1158×10^{-10}	0.16064×10^{-9}
Model 4	0.00002465	0.1000×10^{-13}	0.1244×10^{-8}	0.58989874	0.47724456
Model 5	0.9585×10^{-6}	0	0.2099×10^{-7}	0.41009040	0.52275465



(c) (d)

Figure 5.5: Experiment 1 - Reacquisition of objects. (a) Trajectories overlayed on the first segment mosaic, (b) Trajectories overlayed on the second segment mosaic (c) Space time plot of trajectories show that object 2 is moving faster than the rest of the objects, (d) Space time plot of trajectories of segment 2.

400-500 pixels. It took 27 iterations of the algorithm to converge. For the second set of videos, Figure 5.11 shows the objects (a) before and (b) after running the proposed algorithm. In this case the initial estimate of the homography was good (within 50 pixels), but the initial estimate of the curve parameters was poor. The final alignment of the sequences is shown in Figure 5.12. The algorithm took only 6 iterations to converge. Finally, in Figure 5.13 we show association on video

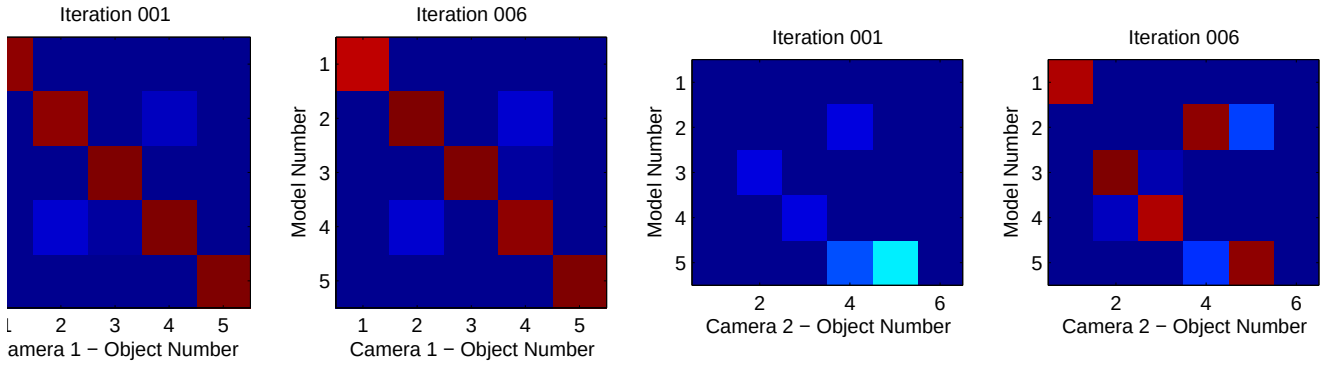
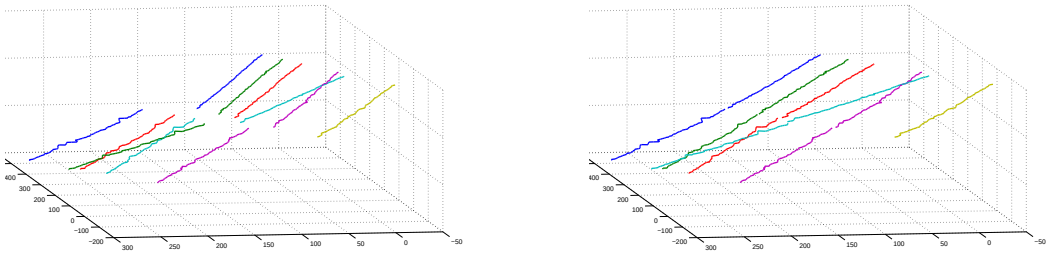


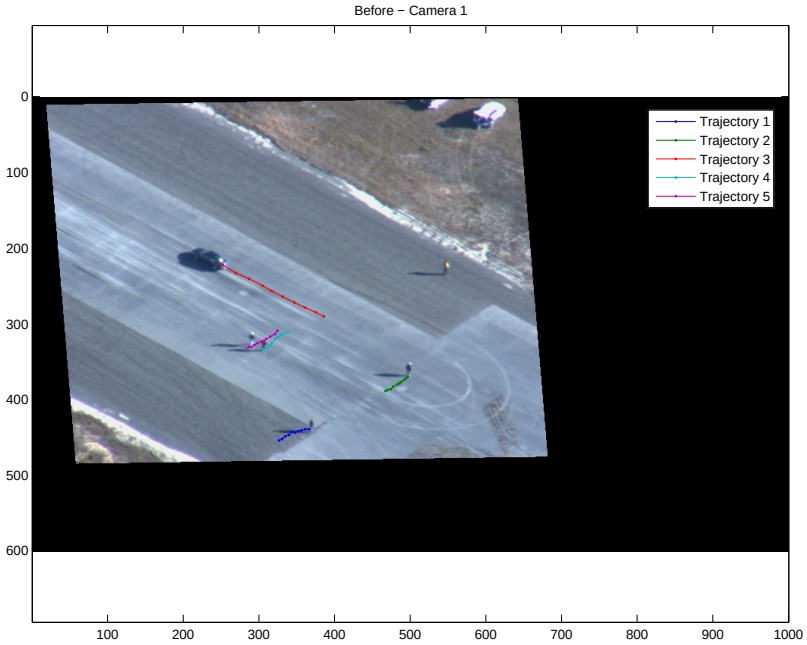
Figure 5.6: Adjacency matrix across EM Iterations containing the probabilities of association. (a) Adjacency Matrix at the first iteration between Camera 1 and the model lines, (b) Adjacency Matrix at the after convergence (6 iterations) (c) Adjacency Matrix at the first iteration between Camera 2 and the model lines, (d) Adjacency Matrix at the after convergence (6 iterations).

taken from two overhead cameras looking at people walking. The color-code of each trajectory shows the association across views recovered by the algorithm. Due to the large rotation present between the views the algorithm took a large number of iterations were executed (39 iterations).

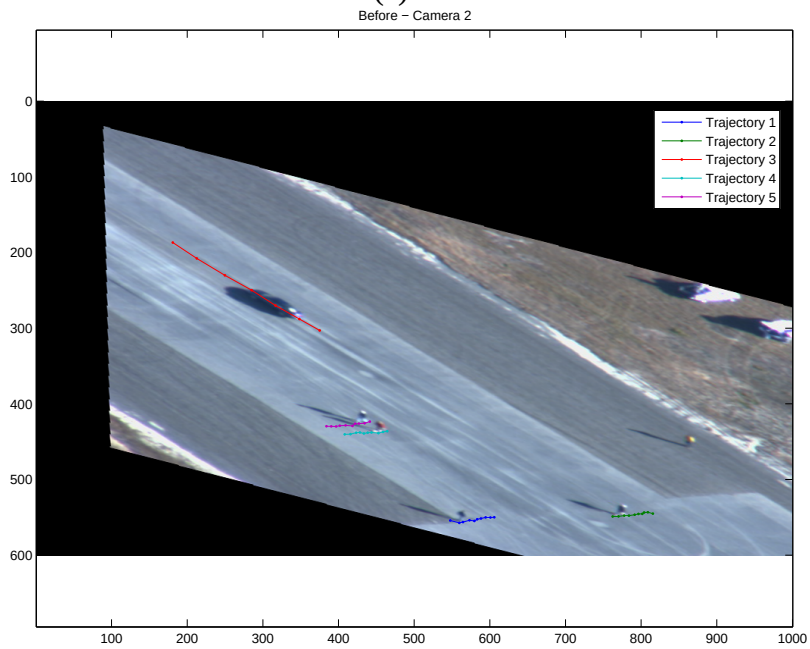


(a) (b)

Figure 5.7: Adjacency matrix across EM Iterations. (a) Adjacency Matrix at the first iteration between Camera 1 and the model lines, (b) Adjacency Matrix at the after convergence (6 iterations) (c) Adjacency Matrix at the first iteration between Camera 2 and the model lines, (d) Adjacency Matrix at the after convergence (6 iterations).

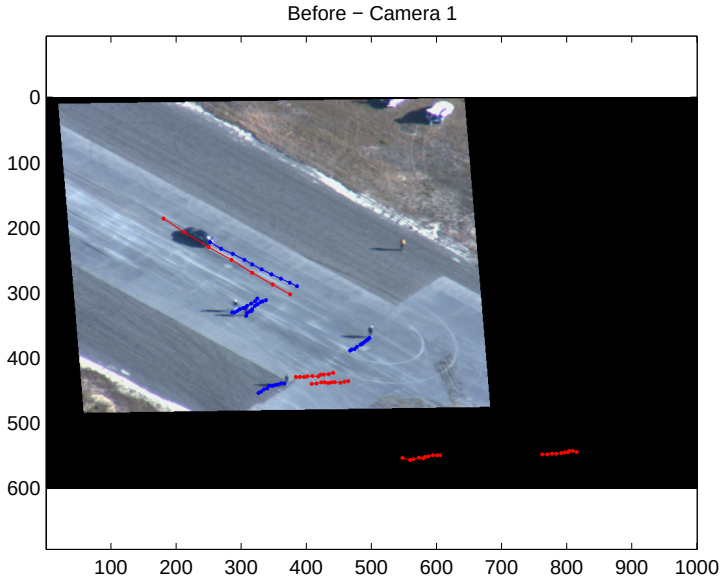


(a)

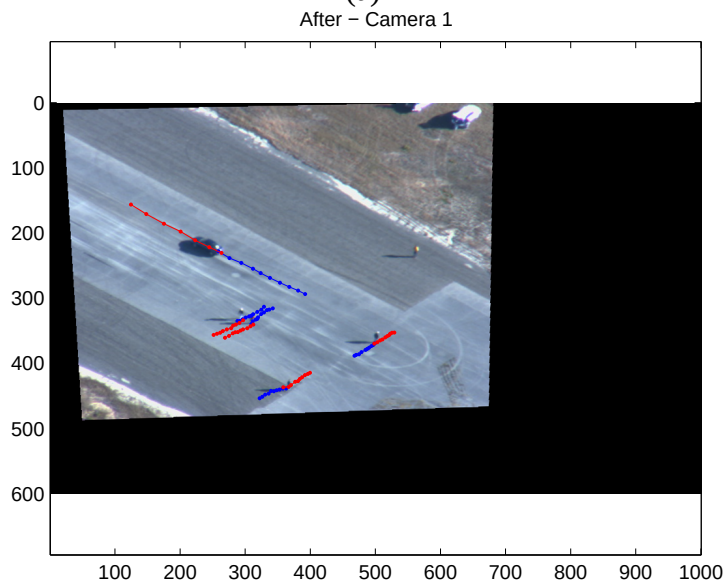


(b)

Figure 5.8: Experiment 1b. (a) Trajectories observed in Camera 1. (b) Trajectories observed in Camera 2 warped to coordinate system of Camera 1.

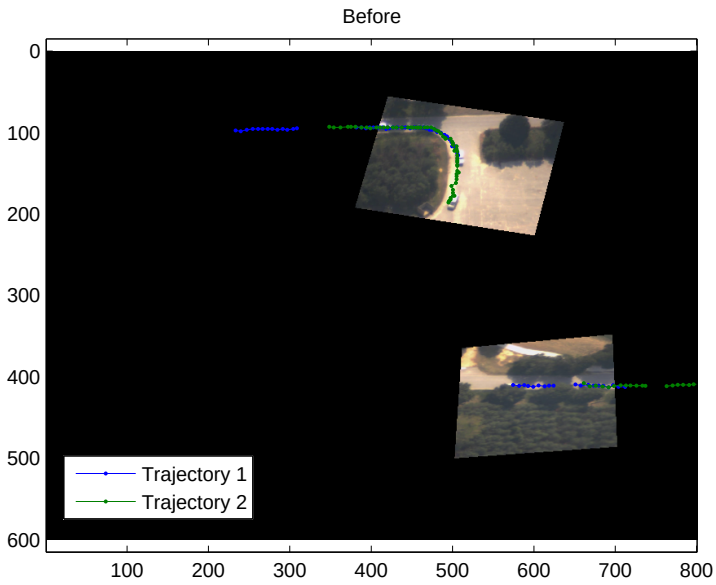


(a)

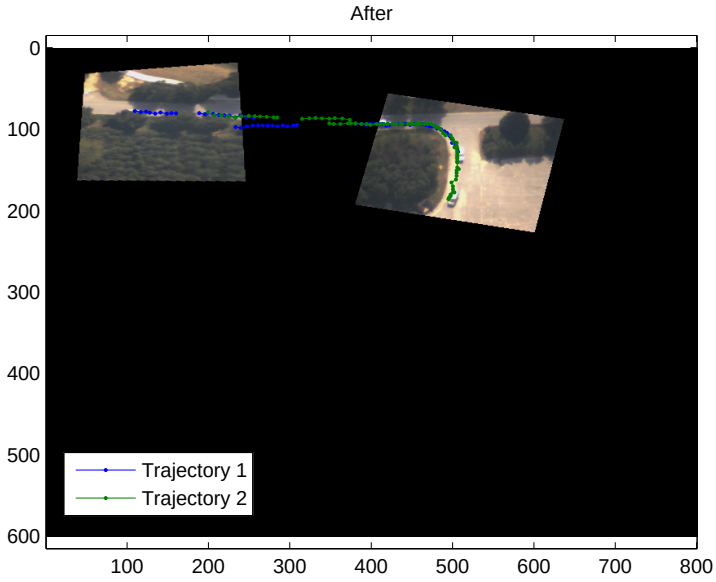


(b)

Figure 5.9: Object reacquisition. (a) Before running the proposed approach. The blue trajectories are the trajectories observed in the first camera, and the red trajectories are the trajectories observed in the second camera warped to the coordinate of the first camera. The initial misalignment can be observed to be over 300 pixels. (b) After running the proposed algorithm. The trajectories are now aligned.

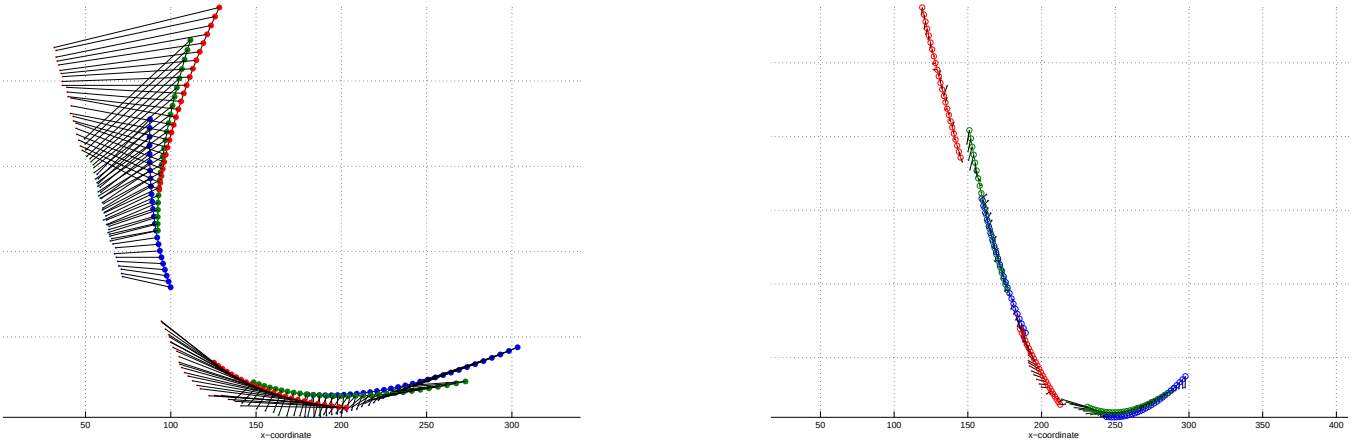


(a)



(b)

Figure 5.10: Object Association across multiple non-overlapping cameras - Quadratic curve. (a) Initialization, (b) Converged Solution.



(a) (b)

Figure 5.11: Object Association across multiple non-overlapping cameras - Quadratic curve. (a) Initialization, (b) Converged Solution.

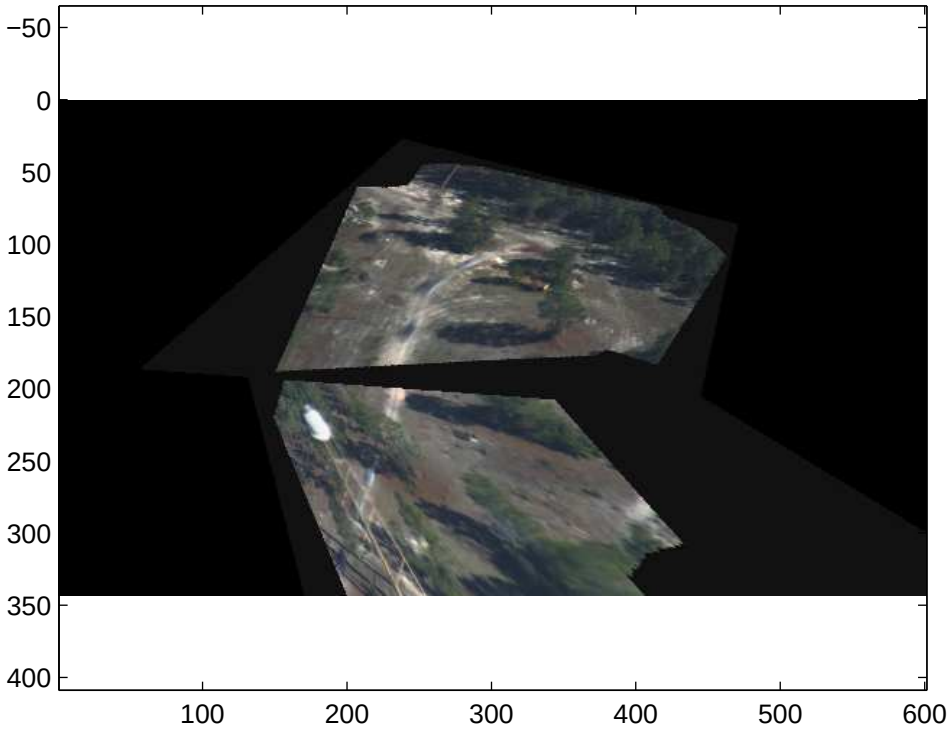
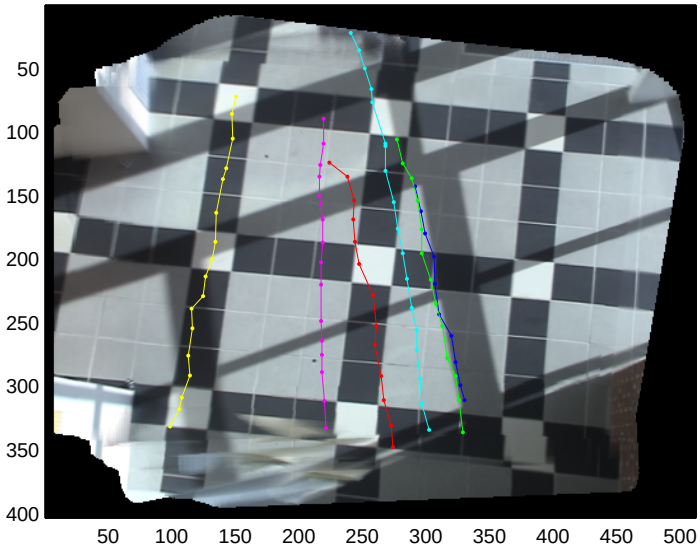
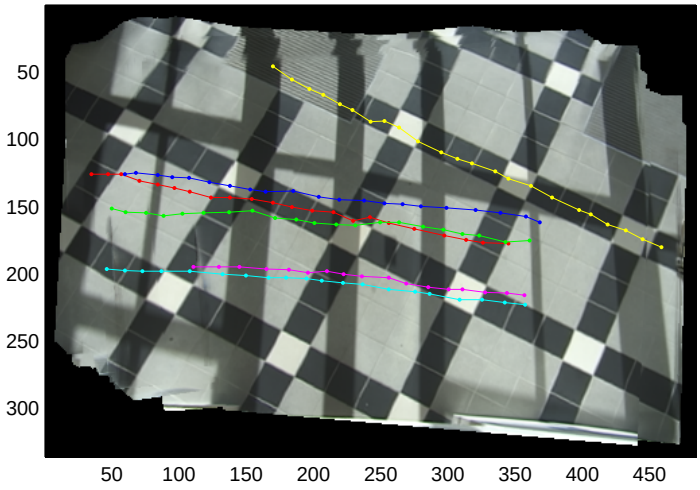


Figure 5.12: Object Association across multiple non-overlapping cameras - Quadratic curve. (a) Initialization, (b) Converged Solution.



(a)



(b)

Figure 5.13: Overhead view of people walking. (a) Shows the color-coded trajectories viewed from the first camera, (b) shows the same trajectories from the second camera.

CHAPTER 6

CONCLUSION

The main theme in this dissertation has been the recovery of a coherent reconstruction of the world (homographies of cameras and canonical trajectories) given imaged data (imaged data at each camera). To that end, we investigated better models for the scene, both for object detection in single cameras and for association across multiple cameras. This theme has led us to pose models that reflect the geometry of the scene and the imaging process, while capturing the uncertainty and incompleteness of data at each camera.

6.1 Summary of Contributions

1. Object detection

   (a) Representation of background as a single 5D distribution for object detection

   (b) Proposal of temporal persistence as a criterion for detection.

   (c) Formulation of object detection in a MAP-MRF framework - find minimum cut of a capacitated graph to minimize the functional.

2. Object Association across multiple spatiotemporally overlapping cameras.

   (a) Definition of a joint re-projection error term for multiple cameras.

   (b) Formulation to ensure that transitive closure is maintained between cameras.

   (c) Algorithm to repair interrupted trajectories.

   (d) Description of the concurrent mosaic for visualization of multiple aerial video streams.

3. Unified Framework for the association of objects across multiple cameras

   (a) Description of novel scene model to explicitly include a polynomial kinematic model for object motion.

   (b) Definition of a likelihood functional for the scene model

   (c) Use of the Expectation Maximization algorithm for parameter and association estimation.

6.2 Future Directions

In this dissertation, in order to render the problem tractable we imposed several assumptions on the scene. Investigating ideas towards relaxing these assumptions is fertile ground for future research. We describe some open problems and discuss future directions,

6.2.1 Global refinement of association and tracking

In this work, we assume tracking within each camera has been performed using any one of the methods proposed in the vast literature on tracking. One interesting direction to take would be to use the tracks as initializations to a final optimization where detections are simultaneously refined. This falls neatly within the proposed approach, where each point can be taken to be the unique observation in each trajectory, and finding the best association *per detection*. This would allow occlusion resolution and the repairing of broken trajectories, as well as an opportunity to correct inaccuracies in tracking.

6.2.2 Non-planar Scenes

The assumption of planarity is reasonable when the altitude of the sensor is much greater than the change in depth in the scene. The aerial video data used in this dissertation are examples, but in general, the question of object association across cameras in non-planar scenes is largely unanswered, though some work has been reported in [YS05]. In addition to looking at full 3D scenes, intermediate relaxations such as the use of layers (multiple planes) to model the scene can also be investigated.

6.2.3 General Kinematic Models

While it is necessary to use a kinematic model to recover the inter-camera homographies across non-overlapping cameras, the assumption of polynomial kinematics is reasonable over limited areas of motion. Instead of using a single polynomial, a spline or piece-wise polynomial could used to parameterize the trajectory. Another potentially interesting direction would be to learn the dynamics of objects in a scene. There is good reason to believe that trajectories in a scene are going to show a lot of redundancy, because of roads, pathways etc. A learning algorithm can be used to fit likely polynomials or better to act as priors during polynomial coefficients estimation.

6.2.4 Spatiotemporal Alignment

In this work we assume that each sequence is time-stamped according to some global time coordinate. An additional parameter over which to minimize could be a temporal displacements for each camera, and further a scaling parameter could also be incorporated for varying frame-rates.

6.3 Discussion

As computers become faster, and the interface between cameras and computers improves, most low level vision tasks have started to show significant maturity in terms of their 'readiness to be

deployed'. However, since the data received by computer through cameras is always noisy and often incomplete, there is an expected threshold of reconstructibility with a single sensor. The larger objective of this line of work is to make cameras conscious of other sensors in a scene and to accumulate evidence synergistically to perceive a reconstruction of the world. In this dissertation, we have set foundations for such co-operative sensing through the use of principled scene and data modeling.

Finally, automated surveillance cannot be discussed without some mention of the Orwellian overtones of this sort of work. A straightforward argument for justification that is often made is that it is not technology, ultimately, that is dangerous but how it is used. Unfortunately, the precedence of misuse of technology makes it important for scientists and researchers to consider fully the implications of their work. Despite this, it is my view that it is not the place of a scientist to stop investigating or thinking about honestly interesting problems, no matter what the possible application and implication. But, at the same time, each scientist is an informed individual with a voice and sometimes, geography permitting, means to influence policy makers. It becomes the responsibility of scientists and researchers to make their concerns known as members of society, vociferously, if the situation demands it.

LIST OF REFERENCES

[AP96] A. Azarbayejani and A. Pentland. "Real-Time Self-Calibrating Stereo Person Tracking Using 3D Shape Estimation from Blob Features." In *Proceedings on International Conference on Pattern Recognition*, 1996.

[AS06] S. Ali and S. Shah. "COCOA: Tracking in Aerial Imagery." In *SPIE Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications*, 2006.

[Bes86] J. Besag. "On the Statistical Analysis of Dirty Pictures." In *Journal of the Royal Statistical Society*, volume 48 of *B*, 1986.

[BG83] G. Buchsbaum and A. Gottschalk. "Trichromacy, Opponent Colours Coding and Optimum Colour Information Transmission in the Retina." In *Proceedings of the Royal Society of London*, 1983.

[Big06] J. Bigun. "Vision with Direction." In *Springer-Verlag*, 2006.

[Bil97] J. Bilmes. "A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models." In *Technical Report, University of Berkeley*, 1997.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. "Fast Approximate Energy Minimization via Graph Cuts." In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2001.

[CAK02] R. Collins, O. Amidi, and T. Kanade. "An active camera system for acquiring multi-view video." In *IEEE International Conference on Image Processing*, 2002.

[CG01] T.-H. Chang and S. Gong. "Tracking Multiple People with a Multi-Camera System." In *IEEE Workshop on Multi-Object Tracking*, 2001.

[CL03] R. Collins and Y. Liu. "On-Line Selection of Discriminative Tracking Features." In *IEEE International Conference on Computer Vision*, 2003.

[CLF01] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. "Algorithms for Cooperative Multisensor Surveillance." In *Proceedings of the IEEE*, 2001.

[CM02] D. Comaniciu and P. Meer. "Mean shift: A Robust Approach Toward Feature Space Analysis." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[CPS05] O. Chum, T. Pajdla, and P. Sturm. "The Geometric Error for Homographies." In *Computer Vision and Image Understanding*, 2005.

[CR96] Y.-C. Chang and J. Reid. "RGB Calibration for Color Image Analysis in Machine Vision." In *IEEE Transactions on Image Processing*, 1996.

[CRM00] D. Comaniciu, V. Ramesh, and P. Meer. "Real-time Tracking of Non-Rigid Objects using Mean Shift." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[DEP95] T.J. Darrell, I.A. Essa, and A.P. Pentland. "Task-specific Gesture Analysis in Real-Time using Interpolated Views." In *IEEE Trans. PAMI*, 1995.

[DT01] S. Dockstader and A. Tekalp. "Multiple Camera Fusion for Multi-Object Tracking." In *IEEE International Workshop on Multi-Object Tracking*, 2001.

[Ed90] Y. Bar-Shalom (Editor). *Multitarget-Multisensor Tracking: Advanced Applications*. Artech House, 1990.

[EDD03] A. Elgammal, R. Duraiswami, and L. Davis. "Probabilistic Tracking in Joint Feature-Spatial Spaces." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[EHD02] A. Elgammal, D. Harwood, and L. Davis. "Background and Foreground Modeling Using Non-parametric Kernel Density Estimation for Visual Surveillance." In *Proceedings of the IEEE*, 2002.

[FF62] L. Ford and D. Fulkerson. "Flows in Networks." In *Princeton University Press*, 1962.

[FF95] A. Fitzgibbon and R. Fischer. "A Buyer's Guide to Conic Fitting." In *British Machine Vision Conference*, 1995.

[Fis02] R. Fisher. "Self-Organization of Randomly Placed Sensors." In *Proceedings of the European Conference on Computer Vision*, 2002.

[FR97] N. Friedman and S. Russell. "Image Segmentation in Video Sequences: A Probabilistic Approach." In *Proceedings of the Thirteenth Conference on Uncertainity in Artificial Intelligence*, 1997.

[Fuk90] K. Fukunaga. "Introduction to Statistical Pattern Recognition." In *Academic Press*, 1990.

[GG84] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[GJ79] M. Garey and D. Johnson. *Computers and Intractibility: A Guide to Theory of NP-Hardness*. Freeman, 1979.

[GN02] M. Grossberg and S. Nayar. "What can be Known about the Radiometric Response from Images?" In *Proceedings of the European Conference on Computer Vision*, 2002.

[GN04] M. Grossberg and S. Nayar. "Modeling the Space of Camera Response Functions." In *IEEE Transactions on Pattern Analysis and Machine Vision*, 2004.

[GPS89] D. Greig, B. Porteous, and A. Seheult. "Exact Maximum A Posteriori Estimation for Binary Images." In *Journal of the Royal Statistical Society*, volume 51 of *B*, 1989.

[Har02] M. Harville. "A framework of high-level feedback to adaptive, per-pixel, mixture of Gaussian background models." In *Proceedings of the European Conference on Computer Vision*, 2002.

[HHD00] I. Haritaoglu, D. Harwood, and L. Davis. "W4: Real-time of people and their activities." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[HK73] J. Hopcroft and R. Karp. "A $n^{2.5}$ Algorithm for Maximum Matching in Bi-Partite Graph." In *SIAM Journal of Computing*, 1973.

[HR97] T. Huang and S. Russell. "Object Identification in a Bayesian Context." In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 1997.

[HW95] P. Hall and M. Wand. "On the Accuracy of Binned Kernel Estimators." In *Journal of Multivariate Analysis*, 1995.

[HZ00] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, September 2000.

[IB98] Michael Isard and Andrew Blake. "CONDENSATION – conditional density propagation for visual tracking." In *Int. J. Computer Vision*, volume 29, pp. 5–28, 1998.

[JN79] R. Jain and H. Nagel. "On the analysis of accumulative difference pictures from image sequences of real world scenes." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[Jon90] M. Jones. "Variable Kernel Density Estimates." In *Austrailian Journal of Statistics*, 1990.

[JRS03] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. "Tracking in Multiple Cameras with Disjoint Views." In *IEEE International Conference on Computer Vision*, 2003.

[JSS02] O. Javed, K. Shafique, and M. Shah. "A Hierarchical Appraoch to Robust Background Subtraction Using Color and Gradient Information." In *IEEE Workshop on Motion and Video Computing*, 2002.

[JW95] R. Jain and K. Wakimoto. "Multiple Perspective Interactive Video." In *IEEE International Conference on Multimedia Computing and Systems*, 1995.

[KBG90] K.-P. Karmann, A. Brandt, and R. Gerl. "Using adaptive tracking to classify and monitor activities in a site." In *Time Varying Image Processing and Moving Object Recognition*. Elsevier Science Publishers, 1990.

[KHM00] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. "Multi-camera Multi-person tracking for Easy Living." In *IEEE Workshop on Visual Surveillance*, 2000.

[KI96] S.B. Kang and K. Ikeuchi. "Toward automatic robot instruction from perception – Mapping human grasps to manipulator grasps." In *IEEE Trans. on Robotics and Automation*, volume 12, Dec. 1996.

[KKK95] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain. "An Architecture for Multiple Perspective Interactive Video." In *ACM Proceedings of the Conference on Multimedia*, 1995.

[KS95] S. Khan and M. Shah. "Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View." In *ACM Proceedings of the Conference on Multimedia*, 1995.

[KS04] J. Kaminski and A. Shashua. "Multiple View Geometry of General Algebraic Curves." In *International Journal of Computer Vision*, 2004.

[Kuh55] H. Kuhn. "The Hungarian Method for Solving the Assignment Problem." In *Naval Reserach Logistics Quarterly*, 1955.

[KWH94] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell. "Towards robust automatic traffic scene analysis in real-time." In *International Conference of Pattern Recognition*, 1994.

[KZ99] V. Kettnaker and R. Zabih. "Bayesian Multi-Camera Surveillance." In *IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[KZ04] V. Kolmogorov and R. Zabih. "What Energy Functions can be Minimized via Graph Cuts?" In *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 2004.

[Li95] S. Li. "Markov Random Field Modeling in Computer Vision." In *Springer-Verlag*, 1995.

[LRS00] L. Lee, R. Romano, and G. Stein. "Learning Patterns of Activity Using Real-Time Tracking." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[MD03] A. Mittal and L. Davis. "M_2 Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene." In *International Journal of Computer Vision*, 2003.

[MEB04] D. Makris, T. Ellis, and J. Black. "Bridging the Gaps between Cameras." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[MMP03] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. "Background Modeling and Subtraction of Dynamic Scenes." In *IEEE Proceedings of the International Conference on Computer Vision*, 2003.

[MP97] S. Mann and R. Picard. "Video Orbits of the Projective Group: A Simple Approach to Featureless Estimation of Parameters." In *IEEE Transactions on Image Processing*, 1997.

[MP04] A. Mittal and N. Paragios. "Motion-Based Background Subtraction using Adaptive Kernel Density Estimation." In *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2004.

[MU02] T. Matsuyama and N. Ukita. "Real-Time Multitarget Tracking by a Cooperative Distributed Vision System." In *Proceedings of the IEEE*, 2002.

[NKI98] A. Nakazawa, H. Kato, and S. Inokuchi. "Human Tracking Using Distributed Vision Systems." In *Proceedings of the International Conference on Pattern Recognition*, 1998.

[ORP00] N. Oliver, B. Rosario, and A. Pentland. "A Bayesian Computer Vision System for Modeling Human Interactions." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[Pap94] C. Papadimitriou. *Computational Complexity*. 1994.

[Par62] E. Parzen. "On Estimation of a Probability Density and Mode." In *Annals of Mathematical Statistics*, 1962.

[PDB90] K. Pattiapati, S. Deb, and Y. Bar-Shalom. "Passive Multisensor Data Association using a New Relaxation Algorithm. In: Y. Bar-Shalom (Ed.), Multisensor-multitarget Tracking: Advanced Applications." Artech House, 1990.

[PLS03] R. Pless, J. Larson, S. Siebers, and B. Westover. "Evaluation of Local models of Dynamic Backgrounds." In *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2003.

[Poo94] A. Poore. "Multidimensional assignment fomulation of daa association problem arising from multitarget and multisensor tracking." Computational Optimization and Applications, 1994.

[QA99] Q.Cai and J.K. Aggarwal. "Tracking Human Motion in Structured Environments using a Distributed Camera System." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[RCB06] R. Rajan, J. Clement, and U. Bhalla. "Rats smell in stereo." In *Science*, 2006.

[RCC98] D. Ruderman, T. Cronin, and C.-C. Chiao. "Statistics of Cone Responses to Natural Images: Implications for Visual Coding." In *Journal of the Optical Society of America A*, 1998.

[RCH03] Y. Ren, C-S. Chua, and Y-K. Ho. "Motion Detection with Nonstationary Background." In *Machine Vision and Application*. Springer-Verlag, 2003.

[RDD04] A. Rahimi, B. Dunagan, and T. Darrell. "Simultaneous Calibration and Tracking with a Network of Non-Overlapping Sensors." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[RKJ00] J. Rittscher, J. Kato, S. Joga, and A Blake. "A Probabilistic Background Model for Tracking." In *Proceedings of the European Conference on Computer Vision*, 2000.

[RKK02] G. Rees, G. Kreiman, and C. Koch. "Neural correlates of consciousness in humans." In *Nature Reviews, Neuroscience*, 2002.

[Ros56] M. Rosenblatt. "Remarks on some nonparametric estimates of a density functions." In *Annals of Mathematical Statistics*, 1956.

[Sai02] S. Sain. "Multivariate Locally Adaptive Density Estimates." In *Computational Statistics and Data Analysis*, 2002.

[SG00] C. Stauffer and W. Grimson. "Learning Patterns of Activity using Real-time Tracking." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[Sit64] R. Sitler. "An Optimal Data Association Problem in Surveillance Theory." In *IEEE Transactions on Military Electronics*, 1964.

[SKS03a] Y. Sheikh, S. Khan, and M. Shah. *Feature-based Georegistration of Aerial Images*. 2003.

[SKS03b] Y. Sheikh, S. Khan, M. Shah, and R. Cannata. *Geodetic Alignment of Aerial Video Frame*. KLUWER Academic Publisher, 2003.

[SMK94] K. Sato, T. Maeda, H. Kato, and S. Inokuchi. "CAD-Based Object Tracking With Distributed Monocular Camera For Security Monitoring." In *Proceedings of IEEE Workshop on CAD-Based Vision*, 1994.

[SRP00] B. Stenger, V. Ramesh, N. Paragios, F Coetzee, and J. Buhmann. "Topology Free Hidden Markov Models: Application to Background Modeling." In *Proceedings of the European Conference on Computer Vision*, 2000.

[SS04] K. Shafique and M. Shah. "A Estimation of the Radiometric Response Functions of a Color Camera from Differently Illuminated Images." In *IEEE International Conference on Image Processing*, 2004.

[SS05] K. Shafique and M. Shah. "A Noniterative Greedy Algorithm for Multiframe Point Correspondence." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[Stu97] P. Sturm. "Vision 3D non calibrée - contributions à la reconstruction projective et étude des mouvements critiques pour l'auto-calibrage." In *PhD Thesis*, 1997.

[TKB99] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. "Wallflower: Principles and Practice of Background Maintenance." In *IEEE Proceedings of the International Conference on Computer Vision*, 1999.

[Tur93] B. Turlach. "Bandwidth Selection in Kernel Density Estimation: A Review." In *Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin*, 1993.

[VAF92] D. Van Essen, C. Anderson, and D. Felleman. "Information processing in the primate visual system: an intergrated systems perspective." In *Science*, 1992.

[WAD97] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. "Pfinder: Real time Tracking of the Human Body." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[Whi98] K. Whitlock. "The mind's eye." In *Perspectives, Ohio University*, 1998.

[WJ95] M. Wand and M. Jones. "Kernel Smoothing." In *Monographs on Statistics and Applied Probability*. Chapman & Hill, 1995.

[WM96] T. Wada and T. Matsuyama. "Appearance Sphere: Background Model for pan-tilt-zoom camera." *Proceedings of the International Conference on Pattern Recognition*, 1996.

[YS05] A. Yilmaz and M. Shah. "Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras." In *Proceedings of the IEEE International Conference on Computer Vision*, 2005.

[ZS03] J. Zhong and S. Sclaroff. "Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter." In *IEEE Proceedings of the International Conference on Computer Vision*, 2003.