2009

# Joint Visual and Wireless Tracking System

Viswajith Karapoondi Nott
*University of Kentucky*, viswajithkn@uky.edu

ABSTRACT OF THESIS

Joint Visual and Wireless Tracking System

Object tracking is an important component in many applications including surveillance, manufacturing, inventory tracking, etc. The most common approach is to combine a surveillance camera with an appearance-based visual tracking algorithm. While this approach can provide high tracking accuracy, the tracker can easily diverge in environments where there are much occlusions. In recent years, wireless tracking systems based on different frequency ranges are becoming more popular. While systems using ultra-wideband frequencies suffer similar problems as visual systems, there are systems that use frequencies as low as in those in the AM band to circumvent the problems of obstacles, and exploit the near-field properties between the electric and magnetic waves to achieve tracking accuracy down to about one meter. In this dissertation, I study the combination of a visual tracker and a low-frequency wireless tracker to improve visual tracking in highly occluded area. The proposed system utilizes two homographies formed between the world coordinates with the image coordinates of the head and the foot of the target person. Using the world coordinate system, the proposed system combines a visual tracker and a wireless tracker in an Extended Kalman Filter framework for joint tracking. Extensive experiments have been conducted using both simulations and real videos to demonstrate the validity of our proposed scheme.

KEYWORDS: Tracking, wireless, visual, kalman filter, homography

(Viswajith Karapoondi Nott)

(27th March, 2009)

Joint Visual and Wireless Tracking System

By

Viswajith Karapoondi Nott

Dr. Sen-ching, Samson, Cheung

(Director of Thesis)

Dr. YuMing Zhang

(Director of Graduate Studies)

27th March, 2009

(Date)

RULES FOR THE USE OF THESES

Unpublished theses submitted for the Master's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgements.

Extensive copying or publication of the dissertation in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this thesis for use by its patrons is expected to secure the signature of each user.

| Name | Date |
| --- | --- |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |

THESIS

Viswajith Karapoondi Nott

The Graduate School

University of Kentucky

2009

Joint Visual and Wireless Tracking System

---

THESIS

---

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Electrical Engineering in the
College of Engineering
at the University of Kentucky
By

Viswajith Karapoondi Nott

Lexington, Kentucky

Director: Dr. Sen-ching, Samson, Cheung , Department of Electrical and Computer

Engineering

Lexington, Kentucky

2009

I would like to dedicate this thesis to Dr. Samson Cheung and my family without whom none of this would have been possible.

# ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere gratitude towards my advisor Prof. Sen-ching Samson Cheung for his valuable guidance, encouragement and continuous support throughout this thesis work. It has been a wonderful experience in working in his research group not only in the field of research but also in terms of personal growth. Besides the technical guidance, I am very thankful for the financial support as well that I received throughout this work. Next, I like to thank other members of my thesis advisory committee, Dr. Parker and Dr. Hassebrook for taking time to read my thesis and providing valuable comments.

I would like to thank my friends, especially the group mates for being very helpful and supportive to me. I would like to thank people at the Center of Visualization and Virtual Environment for providing excellent educational, conducive and healthy environment for students and also for participating in some of the experiments that I conducted to complete this work.

Finally, I am particularly grateful to my parents, and friends for their unremitting support and patience throughout this work. Without their love and encouragement this work would probably not have been completed.

Table of Contents

List of Tables

List of Figures

# Chapter 1

## Introduction

In recent years, there has been a flurry of research and development in the use of smart surveillance systems. The combination of inexpensive cameras, intelligent object identification and tracking algorithms allow such systems to be used in diverse applications from infrastructure protection to smart home environments. On a separate front, the myriad of radio frequency identification and tracking systems have also enjoyed an enormous growth mostly in the area of inventory tracking. The combination of these different tracking systems, can potentially increase the tracking accuracy significantly and unlock interesting applications. In this dissertation, I investigate one such combination by using a visual tracker with a low frequency wireless tracker to improve visual objecting in highly occluded area. Before discussing the contributions of this work lets first review a number of applications that can benefit from combining visual tracking and wireless tracking.

- Smart Rooms act as invisible butlers. They have microphones, cameras, and other sensors to interpret what people are trying to do in this room. In such rooms to track individuals instead of just using a camera and a visual tracker a wireless tracker can be deployed and used to track the people. Such smart rooms can be used for crisis management and mobile command posts to deal with emergencies.

- Wireless sensors that monitor the traffic speed and lane occupancy like the

Digital Traffic Pulse sensor network. These sensor networks are used to identify the speed limits of the vehicles and if such wireless sensors are used along with the cameras and visual sensors it might be possible to identify the license numbers too of the vehicles that crossed the speed limits.

- While video surveillance can provide the most direct visual information, it can be obstructed by clutters and other environmental factors. A wireless tracking system used in this proposal can solve this problem as it supports accurate tracking through walls and other obstructions. On the other hand, tags carried by the human subjects (which emit radio frequency signals) are prone to a host of security attacks, especially in a hostile environment like a correctional facility. Tags may be replaced, destroyed or even under sophisticated attacks like RF jamming, cloning or replay attacks. By relating the tag locations to the visual biometrics from a video surveillance system, the identity of the person with the wireless tag can be easily validated.

- Privacy protection requires the subject whose privacy to be protected to be tracked and identified over each frame. Along with a simple visual object tracker, the use of a wireless tracker, with the subject wearing a wireless tag with a specific frequency helps in identifying the subject over each frame.

## 1.1 Motivation

In this research work we propose a joint wireless and visual tracking scheme based on graphical models, to track individuals in videos especially during occlusion. Tracking using a camera as a sensor and tracking using wireless sensors are complimentary.

Visual tracking highly depends on the environmental conditions such as lighting, occlusion, etc. Low frequency wireless tracker usually does not depend much on such environmental conditions and can track people even through occlusion or when the lighting is very poor. The disadvantages of using a low frequency wireless tracker is that it is highly sensitive and noisy. It does not convey any other visual information other than just the location of the subject being tracked. For example in existing surveillance environments like an office, occlusion would be naturally present when the subject being tracked is standing in a line at the counter or the teller at the counter is sitting behind a computer or walking behind a cupboard. In such an environment the joint visual and wireless tracker would be advantageous due to the complimentary nature of the visual tracker and wireless tracker.

## 1.2 Contribution of thesis

In this thesis we present a joint visual and low frequency wireless tracker to track subjects even when the subjects are occluded for a long duration. Such a scenario can occur in regular office environments, airports, shopping malls or even it can be tried to track sports players in soccer fields. We propose to use just the ground plane homography and perform tracking in the world coordinate system. The fusion of the visual tracker with the wireless tracker is achieved using a probabilistic framework based on graphical models. The major contribution of this thesis is the data fusion scheme based on graphical models. Secondly we evaluate the proposed scheme using both simulated data and real videos to test its robustness and efficiency.

## 1.3 Thesis Outline

The thesis is organized as follows: In Chapter 1 the applications of the proposed joint visual and wireless tracking scheme are discussed and a brief motivation for this research work is proposed. Chapter 2 analyzes the existing literature on visual tracking, wireless tracking and data fusion schemes for tracking. In Chapter 3 we propose the joint visual and wireless tracking system. We explain graphical models, define the state vector and derive the equations for prediction and update for a kalman filter. In Chapter 4 we discuss about synchronizing the wireless tracker and the visual tracker using the network time protocol. We also discuss the background subtraction algorithm used in the tracker, propose a method to detect occlusion in the videos, and analyze on how the model parameters are estimated. The proposed scheme is evaluated in Chapter 5 using simulations on synthetic data and experiments on real videos too. The thesis is concluded in the Chapter 6 where the scope for future work is also discussed.

# Chapter 2

# Related Works

This section analyzes the existing research in visual tracking, wireless tracking and data fusion. The first section discusses previous work in visual tracking.

## 2.1 Visual Tracking

In this section we discuss the existing literature on visual tracking and how they deal with occlusion. Broida et al make use of the kalman filter to track points in the noisy images in [1]. They propose a recursive solution to estimate the motion parameters of an object over a sustained period of time by considering a large sequence of frames or images. The tracking scheme is based on the extended kalman filter using correspondences between object points in a sequence of images. The dynamics of the system is linear while the measurement model is non linear. The scheme is efficient with the data being considered one frame at a time and the estimates of the motion parameters are improved and bettered upon as additional data are used. The authors do not discuss how they deal with occlusion in this research work. Since the extended kalman filter is used for tracking, during occlusion the absence of new measurements would result in the prediction gradually diverging from the actual position of the subject being tracked.

Rosales et al use extended kalman filter to estimate the 3D trajectory of an object from 2D motion in [2]. The extended kalman filter formulation that the authors propose is very interesting. The authors model each object as a 3D box and assume

that the 2D bounding box in each frame is a projection of the 3D bounding box. The tracking is performed with a state vector of the real world coordinates. Hence the measurement model in this case is non linear. They deal with occlusion using higher level mechanisms which provide a feedback mechanism based on the prediction and error from the extended kalman filter. They also discuss a reasonable solution for tracking during occlusion by suggesting a temporal analysis and trajectory prediction. The authors suggest a scheme that maintains a map of the previously segmented and processed frame. The authors then use the map as an approximation of the connected elements in the current frame and compare the connectivity in the current frame and the previous frame. Through the trajectory prediction the authors make use of the extended kalman filter to estimate the 3D motion trajectories based on a 3D linear trajectory model. The issue with the proposed scheme is that during occlusion the extended kalman filter does not have any new measurements from the image and if the duration of occlusion is large, the prediction from the extended kalman filter gradually diverges from the actual position.

Needham and Boyle in track multiple sports players through occlusion using a multi target tracking scheme based on particle filters. The tracker in this proposed scheme has been developed for the sports science industry (to analyze players movements within the soccer field) and also in general the interaction between the players and teams. The authors perform tracking using the ground plane coordinates in this scheme as in a playing field the players might occlude each other even if they are a meter away from each other on the ground plane. The authors do mention that the use of predicted estimates and improving them using the kalman filter helps tracking

during occlusion but do not explain how it does [3].

One of the issues with existing visual tracking schemes alone is that when multiple objects are being tracked using kalman filter or particle filters one needs to establish the correspondence between the measurement for a particular object to the that of the objects state. This issue can be resolved with the combined visual and wireless tracking scheme as the subjects being tracked wear a RFID tag of unique frequency and even under conditions of proximity the two subjects can be distinguished on the basis of the RFID tag frequency. On the other hand when visual tracking alone is used to track multiple objects, one can use techniques such as Multiple Hypothesis Tracking or Joint Probability Data Association Filtering [4].

Sidenbladh et al provide a Bayesian formulation for tracking. The authors define a generative model for 3D human figures. The authors model almost all the parts of the human body as a cylinder except for the torso which is modeled as an elliptical cross section. All the cylinders are right circular. The authors propose an approach using particle filtering which is used to propagate the posterior distribution over time. The authors do account for self occlusion from the limbs and body parts movements, but not much has been discussed about occlusion between one subject and another. Also the results provided in the article involve tracking of individual subjects in the environment. The case of tracking multiple individuals is not being discussed by the authors in this paper [5].

Konstantinos Moustakas et al present a framework to synthesize stereoscopic video using as input the monochromatic image alone [6]. Extended kalman filters are used to recover the 3D structure and motion. A new bayesian framework is also proposed

to deal with occlusion and misclassification of pixels.

Otsuka et al model the spatial structure of the occlusion process between objects and uncertainity based on 2D silhouette based visual angles [7] . Occlusion structure is defined as the tangency between the objects and the edges of the visual angles. The authors then formulate the problem of occlusion as one of recursive bayesian estimation for the hypothesis generation of occlusion structure and estimation of the posterior probability of the object posture and position.

While these probabilistic techniques excel in resolving multiple object matching, it is the visual feature used which creates the most problem during occlusion. One of the most commonly used feature for visual tracking is the silhouette of the object. The silhouette matching trackers match either the shape or the contour of the object in the current frame with those from previous frames. The idea behind citing these trackers are that I believe that Shape Matching trackers and contour matching trackers should fail under occlusion.

Huttenlocher et al propose a shape-model based tracker in [8]. The author performs shape matching using edge based representation and Hausdorff distance is a metric to compare two sets of points in terms of the least similar members. So in the case of matching the edges the hausdorff distance is used as a measure to obtain mismatched edges. The author uses the edge map of the head and the torso of the body as these are the parts that are least susceptible to changes when the subject being tracked is walking. The authors again do not even discuss how occlusion affects their tracking scheme. Comaniciu et al propose a new framework for tracking of non-rigid objects in [9]. The authors define a similarity function which is spatially smooth by

masking the target spatially using an isotropic kernel. The authors propose that after the above step, the target localization problem now changes to identifying the basin of attraction of this similarity function. The authors make use of a metric derived from the Bhattacharya coefficient to measure the similarity between the target model and the target candidates. The authors represent the target model as an ellipsoidal region in the image. To find the location of the target corresponding to the current frame the authors propose to minimize the derived metric from the Bhattacharya coefficient as a function of the locations of the target candidates. The proposed scheme works for only partial occlusions and complete occlusions are not discussed.

Shafique et al establish a correspondence of points between successive frames in a graph theoretic approach in [10]. The authors propose a multi frame approach to obtain the speed and position of the object and maintain coherency of the same. The authors define the problem of tracking as one which involves identification of track points which corresponds to only one set of real world points. In other words the authors define the problem with conditions which state that each real world point has one and only one track point in the image. The authors proposed framework also deals with the false positives and missed detections also. The framework optimizes a gain function over multiple frames. The authors deal with the problem of occlusion in the research work by using the greedy optimization scheme which is non iterative.

Tracking using multiple cameras to deal with occlusion has been proposed by Javed et al [11]. The authors combine multiple cues such as object velocity, inter camera intervals and the location of the exit and entrance in the environment are combined within a bayesian framework. The tracking system has a prior training

phase when the kernel density estimators are used to estimate the probability of an object entering a certain camera at a certain time given the location, time and velocity of its exit from the other cameras.

Maccormick et al propose an exclusion principle permitting the observation model used to interpret the image measurements to let two objects to occupy the same point in configuration space thereby working under occlusion too. The authors mode the target objects by their outlines as B splines. The authors call such an outline as a contour. Tracking is performed by making use of the particle filter. The authors though do propose a probabilistic method for tracking in occluded scenarios the duration of occlusion is not discussed again [12]. Sudderth et al make use of non parametric belief propagation to track a three dimensional model of the human hand. The authors represent the different human hand model constraints as undirected graphs and build a tracking algorithm using Non-parametric Belief Propagation. The authors state that the hands of a human body never form mutually occluding configurations. The authors do not discuss about track recovery, when the tracker fails and tracks arbitrarily and its ability to recover and track the hand model again [13].

Senior et al propose a tracking scheme making use of appearance models to track objects through complex real world interactions. The authors make use of a two tier architecture in which the higher lever architecture associates foreground regions in the adjacent frames to construct the tracks. This is achieved by constructing a distance matrix by computing the bounding box distance between the bounding boxes. The authors then propose to use an appearance model to resolve and improve the tracks during occlusions. The appearance model is built by associating the foreground pixels

to it in each track [14]. Object tracking by template matching and further smoothening of the track results by making use of a kalman filter for each pixel is performed in by Nguyen et al [15]. The authors propose that the tracker is also resistant to changes in the lighting conditions and severe occlusions.

In the above feature based tracking techniques, the tracker depends on features extracted from each frame like color or hue. The external environmental conditions change with changing lighting conditions affecting the tracking.

Gabriel et al summarize the techniques and systems to deal with occlusion based on single cameras and multiple cameras in [16]. The authors classify the techniques as merge-split approach and straight through approach. In the merge-split approach the authors mention that the if there are multiple blobs, the attributes of these blobs are updated till the point of occlusion and at the point of occlusion when a combined blob is obtained, it is considered as an entity and its attributes are updated. Once the combined blob splits again, the problem to be solved remains that of associating and re-identifying the attributes of each blob. Mackenna et al make use of color cues to disambiguate occlusions and color and gradient information to cope with shadows during background subtraction [17]. For the purposes of tracking the authors in this paper describe *regions* as a set of connected components that are tracked over a set of consecutive frames. *People* is described as a person comprising of one or more regions grouped together and *Groups* are defined as one or more people grouped together. Bremond et al propose a tracking scheme to track multiple non rigid objects in video sequences [18]. The authors make use of the merge-split approach during occlusions. When the occlusion mentioned as ambiguous correspondence is detected the authors

call it a Compound Target. The compound target is tracked as a new temporary target and then when more information is available the ambiguous and temporary targets are associated. The authors state that in the straight through approach the individual objects are tracked even through occlusion. According to the authors most of the tracking schemes using this approach have been based on appearance features of the object to associate each pixel with a label.

## 2.2 Wireless Radio Frequency Tracking

In recent years we have witnessed an explosive growth in the use of radio frequency equipment for tracking and identification of assets. National Scientific Corporation have developed Wi-Fi tags to enable tracking. The advantage of the Wi-Fi technologies are one does not need to deploy new sensors, but integrate it with the existing Wi-Fi networks within the office complex. It also has a range of a few hundreds of meters. Another manufacturer *Airetrak* claim to have developed an asset tracking system that works across standard 802.11 wireless LAN. The manufacturer also claims that since the Wi-Fi tracking system operates at a frequency of 2.4 GHz there is very little interference from other wireless equipment.

Ultra Wideband (UWB) technologies are being developed in NASA-Johnson Space Center. The advantages of a UWB tracking system is that UWB has low spectral density enabling it to be used with other communication systems. It is also resistant to multipath interference and has a time resolution up to the picoseconds. And due to this fine time resolution the UWB technologies can be used for precise position tracking. To estimate the location of the radio source different approaches like angle

of arrival, time of arrival, time difference of arrival, relative signal strength are being used. The authors in [19] make use of a UWB tracking system with a time difference of arrival algorithm to build a prototype of a tracking system targeted for space applications.

Foursa et al propose a wireless infrared motion tracking system to track a stylus and the head of a user. The authors make use of infra red monochrome cameras and an RGB frame grabber. The authors then detect the infrared beacons on the image. Then a 2-D transformation is performed to transform the distorted image coordinates to the undistorted camera sensor coordinates. Since three different monochrome cameras are made use of, by using an epipolar constraints are used to obtain the corresponding image points and finally obtain the 3-D coordinates [20].

Most of existing wireless technologies are not capable of accurately tracking a subject in indoor environment due to complicated signal propagation characteristics and challenging radio frequency (RF) propagation environments. Traditional high frequency wireless tracking technologies like ultra-high frequency (UHF), 2.4 GHz, and ultra-wideband (UWB) systems do not work well at significant ranges in the highly reflective indoor environments. Conversely, more accurate short-range tracking technologies, like infrared (IR) or ultrasonics, require an uneconomically dense network of sensors to provide tracking in a correctional environment. In this thesis, we use a new low-frequency (LF) tracking system.

The LF wireless tracking system we used is called QT-400 Starter Kit. It is based on Near-Field Electromagnetic Ranging (NFER®) developed by Q-track Inc [21]. Each user wears an active RFID tag that broadcasts a RF signal of unique

frequency within the AM broadcast band (530-1710 kHz) that is detected by three antennas for triangulation. After a careful calibration to establish the correspondence between the RF signals and the ground coordinates of many pre-selected calibration points, the active tag can then be continuously tracked in real-time. Unlike other RFID systems, NFER® exploits the properties of medium and low-frequency signals within about a half wavelength of a transmitter. The low frequencies used by NFER® are more penetrating and less prone to multi-path than the typically-used microwave frequencies. The manufacturer claims to provide real-time tracking performance of uncertainty of 60 centimeters when the antennas are 55 meters apart [22].

### 2.3 Data fusion for object tracking

In this section we discuss existing literature on data integration schemes used to track assets and subjects.

Siebel et al combine a head detector, a shape tracker and a region tracker to provide a multiple cue tracking system. The authors use a motion detector which detects the moving objects in the background [23]. The region tracker tracks the moving regions detected by the motion detector. The authors perform background subtraction to detect people moving in the video. The authors achieve the same by maintaining a background model of the scene. It splits regions if these regions contain more than one person. The region detector checks whether a significant part of the region covered by the region tracker was not covered by the shapes in the region. If the region detector establishes that there are more than one subject in the given region, then the given region is further divided in to sub regions and each individual

sub region is further tracked. In the subsequent frames again the region detector performs a region splitting if there are multiple persons in the video and each sub region are individually processed by the active shape tracker and the head detector. The authors do provide a reasonable solution to deal with occlusion by making use of both the region tracker and the active shape tracker. The region tracker makes use of region splitting when an occlusion is detected and a single whole region which is the output of the motion detector is split in to multiple regions. These multiple regions form the input to the active shape tracker and the head detector. The authors do not discuss on the duration of the occlusion and the effect of sustained occlusion in this paper.

Spengler et al implement two integration schemes in their research work [24]. The authors implement a democratic integration scheme for combining multiple cues and also a particle filter scheme. Democratic integration scheme was proposed by [25] in which the multiple modalities used for data integration agree upon a result. This result serves as a basis when the environment changes ensuring that the modalities are adaptive. The democratic integration scheme provides single hypothesis tracking. In this scheme five visual cues are used to agree upon a common position. They propose that the democratic integration can be used for single object tracking. On the other hand the authors test the particle filter cue integration scheme with a two person sequence to prove that the particle filter scheme works for tracking multiple people. They propose an effective scheme of dealing with occlusion by making use of the particle filter, visual cue integration scheme. The authors do not discuss about the case when perspective projection is involved. In the results provided by the authors

the two subjects being tracked their height does not change at all. The authors do not discuss the case of how the tracking is affected in case the subjects height keeps changing.

Perez et al combine audio cues and visual cues to provide an effective multi modal tracking scheme for the purposes of teleconferencing [26]. The authors also demonstrate combining visual cue with motion for video surveillance. Particle filters are proposed for fusing the multiple modalities as it can be used for non gaussian distributions too. The authors provide a data fusion scheme for both highly localized environments like a teleconference setting and also a generic environmental setting like office where video surveillance is essential. The disadvantage of the audio cue is it usually does not give the vertical height factor for the subject being tracked as only the two dimensional location of the person speaking can be obtained using the audio cue. So it can be used in only specific environments for tracking.

We discuss probabilistic and graphical model based schemes for data integration in this section. Xue et al propose a graphical model based cue integration scheme for head tracking [27]. The authors also propose a new inference procedure based on a non parametric belief propagation. The proposed scheme considers that an object is represented by $M$ modalities and corresponding to a state space. The state spaces of all the cues are dependent on each other and are connected to be a Gibbs Field. The authors proceed to describe the new inference procedure using non parametric belief propagation by representing each message and belief as a weighted sample set. Head tracking is demonstrated by selecting three cues-Color, Shape and Intensity. The authors approximate the projection of the head in the image as an

ellipse. The authors demonstrate head tracking under conditions of partial occlusion using the proposed graphical model cue integration scheme. Leichter et al propose a probabilistic framework to combine multiple tracking algorithms [28]. The framework is built on the assumption that each tracking algorithm is conditionally independent of the other algorithms. The advantage of the proposed framework is that there is no necessity to switch between the different tracking algorithms explicitly. The weighting of the cues from the individual trackers are done implicitly. Also trackers having different state spaces can be combined using the proposed framework. The above framework fails from occlusion even from a tree. The proposed framework fails to work for occlusions of short durations. Wu et al integrate rough models of multiple cues based on a probabilistic factorized graphical model [29]. The authors use importance sampling and a sequential monte carlo algorithm to for the co-inferencing of multiple cues. The authors do experiment cases of occlusion in this paper, but the duration of occlusion upon which the algorithm is tested is relatively small. The authors test for object to object occlusion where one subject's face occludes the second subject's face. The duration of object to object occlusion is relatively small when both the objects are human subjects. The authors do not explain the effect of sustained occlusion in this paper.

Oruc et al test the combination of cues for the task of slant estimation [30]. Though this paper does not entirely relate to object tracking it tests out different cue combination strategies. The authors test out basic cue combination like weighted linear cue combination for combining correlated cues. The authors test out the experiment of adjusting the slant of a plane on eight different observers. The two cues

used where linear perspective cue and a texture cue. The authors notice that the observers make use of the lesser reliable cue when the cue with higher reliability becomes noisier.

Branson et al propose a scheme that makes use of a multiple blob tracker and a contour tracker to deal with severe occlusions [31]. The authors use the proposed scheme to track the mice using a video of the side view of the cage. The area over which the tracking has been done seems to be substantially small compared to tracking for the applications proposed in this thesis. Mice on the other hand are small tracking objects which when tracked over a video might create occlusions over small durations only.

# Chapter 3

# Model Design

In this chapter, we discuss the sensor and object models, as well as the Kalman Filter framework used in the proposed combined tracker.

## 3.1 Sensor Model

There are two types of sensors used in our system: a camera and an active low-frequency wireless tracking system. A camera captures the 3-D world by projecting it onto a 2-D image plane. The wireless tracker provides the two dimensional coordinates of the subject carrying an radio frequency transmitter. In order to develop a system that can simultaneously make use of both sensors for tracking, the geometrical information they provide must be combined in a uniform coordinate system. Assuming a pinhole camera model, the 3D world coordinates of a point and its 2D image coordinates are related by a projective transform:

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{pmatrix} \tag{3.1}$$

$$X_{image} = \frac{(p_{11}X_{world} + p_{12}Y_{world} + p_{13}Z_{world} + p_{14})}{(p_{31}X_{world} + p_{32}Y_{world} + p_{33}Z_{world} + p_{34})} \tag{3.2}$$

$$Y_{image} = \frac{(p_{21}X_{world} + p_{22}Y_{world} + p_{23}Z_{world} + p_{24})}{(p_{31}X_{world} + p_{32}Y_{world} + p_{33}Z_{world} + p_{34})} \tag{3.3}$$

The ground plane in the world coordinate system can be considered as a two dimensional plane as the third dimension $Z_{world}$ happens to be 0 for the ground plane. If the $Z_{world}$ is set as zero then the perspective projection matrix can be simplified as follows to represent a mapping between the two dimensional ground plane and the image plane.

$$H = homography = \begin{pmatrix} p11 & p12 & p14 \\ p21 & p22 & p24 \\ p31 & p32 & p34 \end{pmatrix} \tag{3.4}$$

$$X_{image} = \frac{(p_{11}X_{world} + p_{12}Y_{world} + p_{14})}{(p_{31}X_{world} + p_{32}Y_{world} + p_{34})} \tag{3.5}$$

$$Y_{image} = \frac{(p_{21}X_{world} + p_{22}Y_{world} + p_{24})}{(p_{31}X_{world} + p_{32}Y_{world} + p_{34})} \tag{3.6}$$

The use of homography simplifies the calibration process as we are concerned only with calibration points on a plane. This process is not limited only to the ground plane but can be applied to any plane parallel to the ground plane, using of course a different homography. In fact, two homographies are used in our human model which captures the characteristics of our tracking object.

The human model is assumed to be that of a planar rectangle in the 3-D world. The height of the rectangle is assumed to remain constant but the width of the rectangle is assumed to be changing with that of the pose and posture. In the image plane the subject being tracked has a Top Point (denoted by TP - image coordinates of the head) and a Bottom Point (denoted by BP - image coordinates of the foot of

the subject being tracked). In the world coordinate system, all the TP's of the same subject at different time instants are coplanar to each other due to the constant height assumption. Similarly all the BP's are coplanar to each other as they all lie on the ground plane. Hence the two dimensional coordinates of the subject being tracked in plane of the head map to image coordinates of the head by a homography. Similarly the two dimensional coordinates of the subject being tracked with respect to the ground plane or the foot, map to the image coordinates of the foot by a homography.

### 3.1.1 State Vector

In this section we discuss the state vector for the joint wireless and visual tracker. The state vector captures all the essential information of the tracked subject and is used to combine information from various sensors at different time instances. The internal state of the tracking is defined as follows:

$$s_t = \begin{pmatrix} x_f(t) \\ y_f(t) \\ h(t) \\ w(t) \\ x_{f1}(t) \\ y_{f1}(t) \end{pmatrix} \tag{3.7}$$

where $h$ represents the height of the subject being tracked, $w$ represents the width of the person being tracked and $x_f$ and $y_f$ represents the two dimensional world coordinates of the foot. The above state would be updated based on the sensor measurements and the assumed dynamics of the subject.

We assume that the dynamics of the subject is linear, for example a constant velocity or constant acceleration model. The sensor measurement model for the wireless tracker is also linear as the measured 2D coordinates relate to the actual ground

plane by a similarity transform. For simplicity, we absorb this similarity transform in our calibration of the wireless tracker and use an identity when comes to propagating the information to the state vector. On the other hand the image coordinates map to the ground plane coordinates based on the homography between the ground plane and the image plane. This is a non linear measurement model for the image coordinates. Thus the dynamics, the measurement on the wireless tracker, and the visual measurement can be represented as follows:

$$s_{t+1} = As_t + Gw_t \tag{3.8}$$

$$m_{t+1}^{Image} = f(s_{t+1}) + Image_{measnoise} \tag{3.9}$$

$$m_{t+1}^{RFID} = Cs_{t+1} + RFID_{measnoise} \tag{3.10}$$

Equation 3.12 represents the dynamics in which the state vector $s_t$ at time instant $t$ is moved to $s_{t+1}$ at time $t+1$ according to the motion matrix $A$ and process zero-mean white noise $w_t$ with covariance. The motion matrix applies mainly to the foot coordinates. It has no effect on the height as we assume that to be constant. We also ignore the complex dynamics of the change of width due to posture and models the variability with a high noise variance. Equation 3.9 represent the measurement of the camera sensor, while equation 3.13 represents the measurement on the wireless tracker where $m_t^{RFID}$ is the 2D world coordinate measurement from the wireless tracker and C defined as follows:

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.11}$$

The function $f$ in 3.9 represents the homography $H$. We assume that the camera can provide us, after some low-level vision processing, the four corners of the bounding box of the subject in the image plane. These corners are the images of the four corners of our world human model. The world coordinates of the four corners are derived in Equation . The non-linear function f is composed of two homographies: the first homography maps the world coordinates of the two corners on the ground plane to the image coordinates and the second one transform the remaining two corners.

Next, we discuss how measurements at different instances are combined and used to predict future states. The primary tool used is a Kalman Filter and we will use a probabilistic graphical model to represent each state and measurement [32]. A probabilistic graphical model represents each random variable as a node in a graph and conditional dependency as edges between nodes. The advantage of using a graphical model is that the notation is easy to understand and interpret. Also the inference procedure is well studied for almost all types of graphs regardless of their complexity. Inference is used to find the unknown probabilities of unobserved random variables based on the available evidence in other random variables. There are many algorithms for performing inference in a graphical model. Some of these algorithms are Elimination, Belief Propagation, Message Passing Algorithms and the Junction Tree algorithm [32]. The inference in the Kalman filter is similar to inference in the Hidden markov models. The Hidden Markov models and Kalman filter are special cases of the junction tree algorithm which is used to perform inference in graphical models. Inference in kalman filter comprises the computation of the posterior probability of the states given an output sequence. The output sequence is the measurement that

is made using the wireless sensors and the image coordinates that are obtained from the visual tracker. The posterior probability of the states in the kalman filter are computed recursively.

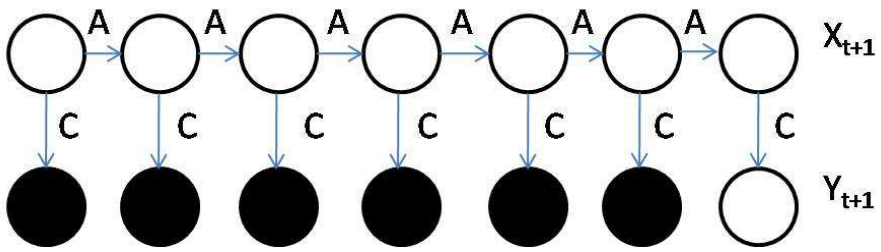
Calculating the posterior probability can be a tedious procedure if arbitrary density function is allowed. In Kalman filter, all the process and measurement noise processes are assumed to be multivariate Gaussian. Assuming that the initial state vector is also Gaussian, linear dynamics and measurement equations will imply that all the subsequent state vectors and measurements are Gaussian. As Gaussian distribution is parametrized by mean and covariance, the recursive procedure only involves linking the conditional means and covariances at neighboring moments in time.

Before discussing this recursive procedure, we note that our image measure procedure is non-linear due to the use of homographies. Rather than abandoning the entire framework of Kalman filter, the non-linear measurement can be approximated by a linear version via Taylor series expansion. Such an approximation is called Extended Kalman Filter (EKF) and its formulation will be discussed in Section 3.3.

## 3.2 Kalman filter with single measurement

In this section we derive the recursions for mean and covariance of the internal state for a Kalman Filter with a single output sequence.

In Figure 3.1 the state space model or graphical model for kalman filter is shown. The shaded nodes are known measurements while the hollow nodes are unknown random variables. If we assume $x_t$ represents the state of the system and $y_t$ represents the measurement of the system then the dynamics and the measurement are shown




(a) State Space Model

Figure 3.1: Graphical Model for Kalman Filter

below:

$$x_{t+1} = Ax_t + Gw_t \tag{3.12}$$

$$y_{t+1} = Cx_{t+1} + v_t \tag{3.13}$$

where $v_t$ is a gaussian random variable with zero mean and covariance matrix $R$, and $G$ is the input matrix. We wish to estimate the state $x_{t+1}$ based on the partial output sequence $y_0$,....$y_{t+1}$. At the beginning of the time instant $y_{t+1}$ is not available. Hence the $y_{t+1}$ is unshaded in Figure 3.1. That is we wish to calculate $P(x_{t+1}|y_0, ...., y_t, y_{t+1})$. Let us have a notation which uses $\hat{x}_{t+1|t+1}$ to represent the mean of $x_{t+1}$ conditioned on the partial sequence $y_0$,....$y_{t+1}$. The covariance matrix of $x_{t+1}$ conditioned on $y_0$,....$y_{t+1}$ is denoted $P_{t+1|t+1}$. We assume that we have already calculated $P(x_t|y_0, ...., y_t)$, that is we have calculated $\hat{x}_{t|t}$ and $P_{t|t}$ as the mean and the covariance matrix can be used to define a gaussian distribution. We now have a time update and a measurement update.

Let us consider a time update step. The dynamic equation is as follows:

$$x_{t+1} = Ax_t + Gw_t \tag{3.14}$$

We take conditional expectation on both sides of this equation. Since $w_t$ is noise and is independent of the conditioning variables $y_0$,....$y_t$, the second term vanishes and we have the following:

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t} \tag{3.15}$$

Similarly taking the conditional covariance on both the sides of the dynamic equation we have:

$$P_{t+1|t} = AP_{t|t}A' + Q \tag{3.16}$$

where we have $Q$ as the covariance of the noise in the dynamic equation. Now we proceed further in the graphical model fragment and calculate the conditional mean and covariance of $y_{t+1}$ as well as the conditional covariance of $x_{t+1}$ and $y_{t+1}$.

$$E[y_{t+1}|y_0, ....y_t] = E[Cx_{t+1} + v_{t+1}|y_0, ....y_t] = C\hat{x}_{t+1|t} \tag{3.17}$$

$$E[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})'|y_0, ....y_t] = CP_{t+1|t}C' + R \tag{3.18}$$

$$E[(y_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})'|y_0, ....y_t] = CP_{t+1|t} \tag{3.19}$$

Now the conditional distribution of $x_{t+1}$ given $y_{t+1}$ can be obtained as follows:

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}) \tag{3.20}$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}CP_{t+1|t} \tag{3.21}$$

The above recursions constitute the Kalman filter. From the above equations Kalman gain is defined as follows:

$$K_{t+1} = P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1} \tag{3.22}$$

Using the above notation we have:

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C\hat{x}_{t+1|t}) \tag{3.23}$$

### 3.2.1 Kalman Filter using two measurements

In this section we introduce the joint wireless and visual tracking graphical model and derive the equations for mean and variance of the internal state of the graphical model.

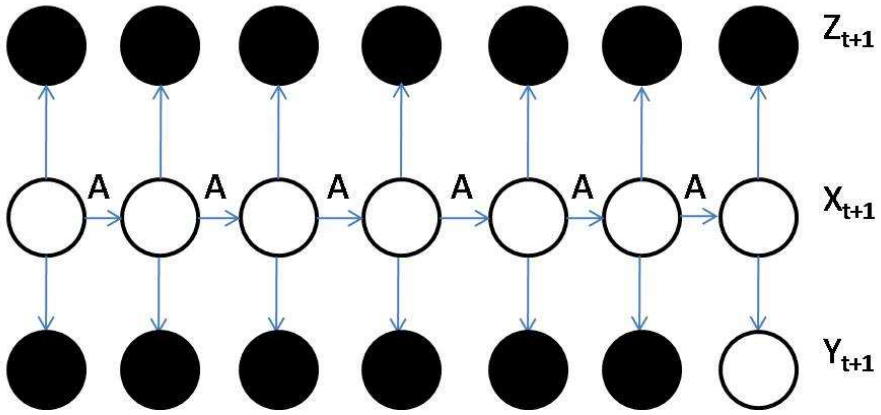



Figure 3.2: Graphical Model for Joint Wireless and Visual Tracker

Based on the joint wireless and visual tracker and graphical model in 3.2, the state space models for the kalman filter can be represented as follows:

$$x_{t+1} = Ax_t + w_t \tag{3.24}$$

$$y_{t+1} = f(x_{t+1}) + measnoise_{t+1} \tag{3.25}$$

$$z_{t+1} = Cx_{t+1} + n_{t+1} \tag{3.26}$$

$z$ represents the wireless sensor state measurement and $y$ represents the visual measurement. Before deriving the mean and the variance of the state, the measurement for the visual tracker has to be linearized. The non linear measurement of the visual tracker can be linearized by using the taylor's series which can be used to represent a function as an approximate sum of its derivatives evaluated in the neighborhood of a real or complex number.

$$y_{t+1} \approx f(\hat{x}_{t+1}) + \nabla f(\hat{x}_{t+1})(x_{t+1} - \hat{x}_{t+1}) + measnoise_{t+1} \tag{3.27}$$

After linearizing and then rearranging 3.27 we redefine a new measurement variable as follows:

$$\tilde{y}_{t+1} = y_{t+1} - f(\hat{x}_{t+1}) + f \cdot \hat{x}_{t+1} \tag{3.28}$$

and result in the following linearized measurement sequence:

$$y_{t+1} - f(\hat{x}_{t+1}) + f'(\hat{x}_{t+1})\hat{x}_{t+1} = f'(\hat{x}_{t+1})(x_{t+1}) \tag{3.29}$$

$$y_{visual} = f'(\hat{x}_{t+1})(x_{t+1}) \tag{3.30}$$

$$D = f'(\hat{x}_{t+1}) \tag{3.31}$$

We assume that we already have the distribution of the state estimate $x$ at instant $t$ conditioned on the measurements from the first instant upto $t$, that is we have $P(x_t|y_0, ...., y_t, z_0, ...., z_t)$. We wish to compute $P(x_{t+1}|y_0, ...., y_t, z_0, ...., z_t)$ in the time update step. This can be then updated to either $P(x_{t+1}|y_0, ...., y_t, z_0, ...., z_t, z_{t+1})$ followed by computing $P(x_{t+1}|y_0, ...., y_t, y_{t+1}, z_0, ...., z_t, z_{t+1})$ in the measurement update step. We denote mean of $x_t$ conditioned on $y_0, ...y_t, z_0, ...z_t$ as $\hat{x}_{t|y_0...y_t,z_0,...z_t}$.

The mean and covariance for the state variables for the graphical model stated

above can be written as follows:

$$\hat{x}_{t+1|y_0 \ldots y_t, z_0, \ldots z_{t+1}} = \hat{x}_{t+1|y_0 \ldots y_t, z_0, \ldots z_t} + K^z_{t+1}(z_{t+1} - C\hat{x}_{t+1|y_0 \ldots y_t, z_0, \ldots z_t}) \tag{3.32}$$

$$K^z_{t+1} = P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} C'(C P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} C' + R)^{-1} \tag{3.33}$$

where $R$ is the covariance of the noise term $n_{t+1}$ in 3.26. $K_{t+1}$ is the kalman gain in 3.32 and in 3.33. To compute the covariance $P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t, z_{t+1}}$

$$P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t, z_{t+1}} = P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} - K^z_{t+1} C P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} \tag{3.34}$$

The next step would then be to compute $\hat{x}_{t+1|y_0 \ldots y_t, y_{t+1}, z_0, \ldots z_{t+1}}$. This can be written as follows:

$$\hat{x}_{t+1|y_0 \ldots y_t, y_{t+1}, z_0, \ldots z_{t+1}} = \hat{x}_{t+1|y_0 \ldots y_t, z_0, \ldots z_{t+1}} + K^y_{t+1}(y_{t+1} - D\hat{x}_{t+1|y_0 \ldots y_t, z_0, \ldots z_{t+1}}) \tag{3.35}$$

The kalman gain can be written as follows:

$$K_{y(t+1)} = P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} D'(D P_{t+1|y_0, \ldots y_t, z_0, \ldots z_t} D' + R)^{-1} \tag{3.36}$$

### 3.2.2 Kalman Filter having irregular measurements

In the previous sections we discussed about how multiple measurements can be fused using a graphical model. But in the proposed graphical model in the section 3.2.1 the measurements from the multiple trackers are at an equal rate and hence the measurements are available for the joint tracker at the same time instants. The graphical model would be different when the measurements from the multiple trackers are available for the joint tracker at different time instants. Under such a scenario at certain specific instants either one of the two trackers would not have any new mea-

surements while the other tracker would have a new measurement available. This can be dealt in a manner similar to the way in which occlusion is dealt in the joint tracker. During occlusion in the joint tracker there might not be any new measurements from the visual tracker.

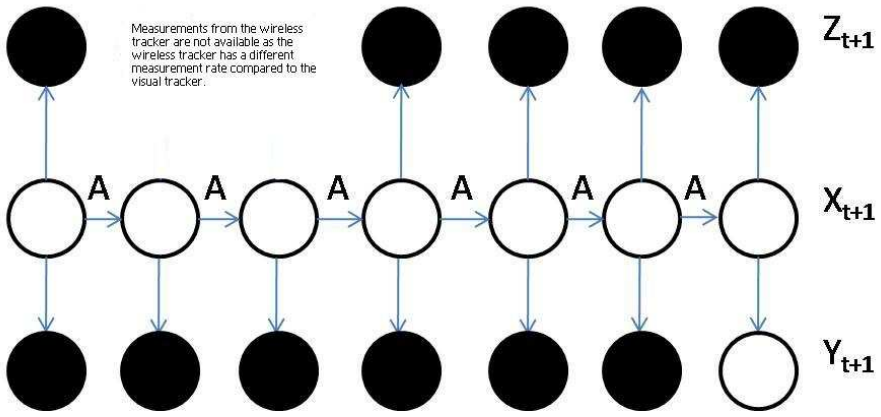



Figure 3.3: Joint Tracker with irregular measurements

In Figure 3.3 the wireless measurements at time instants 1 and 2 are not available as the wireless tracker has a different measurement rate when compared to the visual tracker. Hence the prediction at time instant 2 and 3 will be weighted more by the visual tracker and less by the wireless tracker.

# Chapter 4

# Estimating Model Parameters

In this chapter we discuss about the background subtraction algorithm used to identify the objects in motion in videos. We also discuss about the calibration procedure used to calibrate the wireless tracker system. We further discuss about how occlusion is detected when multiple subjects are walking in the environment before concluding the chapter explaining on how we obtain the model parameters such as initial velocity along both the horizontal and vertical direction, the covariances of the process noise and the measurement noise, how the measurement noise is obtained and the initial parameters of the combined kalman filter.

## 4.1 Foreground Extraction

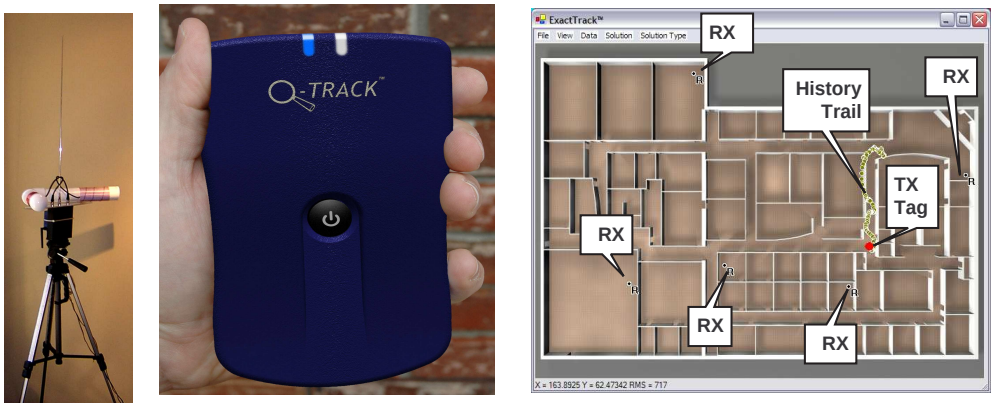
The moving blobs and the bounding boxes around them are extracted using a background subtraction algorithm proposed by Horpraset et al [33] and [34]. The background subtraction algorithm needs to be trained using a static background. The background is modeled as a 4-tuple using the expected color value, the standard deviation of the color value, the variation of the brightness distortion and the variation of the chromaticity distortion. During classifying phase, each pixel of the incoming frame is classified as a foreground if the chromaticity exceeds a color threshold. On the other hand the pixel in the incoming frame is classified as a shadow if they have similar color chromaticity but lower brightness than those of the same pixel in the

background image. The classification threshold is set based on confidence level so that the static background does not get classified as foreground.

## 4.2 Calibration of the wireless tracker system

The wireless tracker system is calibrated manually by standing on a set of predetermined points on the ground plane and establishing the corresponding point on the floor map used in the wireless tracker system. So a number of calibration points are marked in the environment. Then to calibrate the system one needs to traverse through all the calibration points wearing an active tag and at the same time alerting the system to establish the correspondences between the received RF signals and the known 2D coordinates of the calibration points.

In our system, we have chosen to use a wireless tracking system based on Near-Field Electromagnetic Ranging (NFER®) developed by Q-track Inc. [21]. Each user wear an active RFID tag that broadcasts a RF signal of unique frequency within the AM broadcast band (530-1710 kHz) that is detected by three antennas for triangulation. After a careful calibration to establish the correspondence between the RF signals and the ground coordinates of many pre-selected calibration points, the active tag can then be continuously tracked in real-time. Unlike other RFID systems, NFER® exploits the properties of medium and low-frequency signals within about a half wavelength of a transmitter. The low frequencies used by NFER® are more penetrating and less prone to multi-path than the typically-used microwave frequencies [22].




(a) Antenna (b) Active Tag (c) Tracking Software

Figure 4.1: The QT$^{TM}$-400 antenna, tag and tracking software

## 4.3 Synchronization of the visual tracker with the wireless tracker

To establish the correspondence between the RFID coordinates and the video frames the corresponding time stamps (times at which these frames are captured and the time instants at which the RFID coordinates are obtained) are also obtained. But for these time stamps to make sense the two different computers on which the wireless tracker and the video capturing mechanism are running are synchronized. This is achieved by making use of Network Time Protocol (NTP) which synchronizes multiple computers to within 10 ms, which is below the capturing period of both the RFID and the camera systems.

## 4.4 Obtaining the homographies

A training sequence is captured to obtain the homography of the ground plane with the image plane. After background subtraction, by computing the extremum of the blobs the image coordinates of the head and foot of the subject being tracked is

obtained as shown in Figure 4.2. At each instant we obtain the corresponding RFID 2D coordinates from the wireless tracker. The homography can be obtained by using singular value decomposition to obtain a transformation between the ground plane and the image plane.

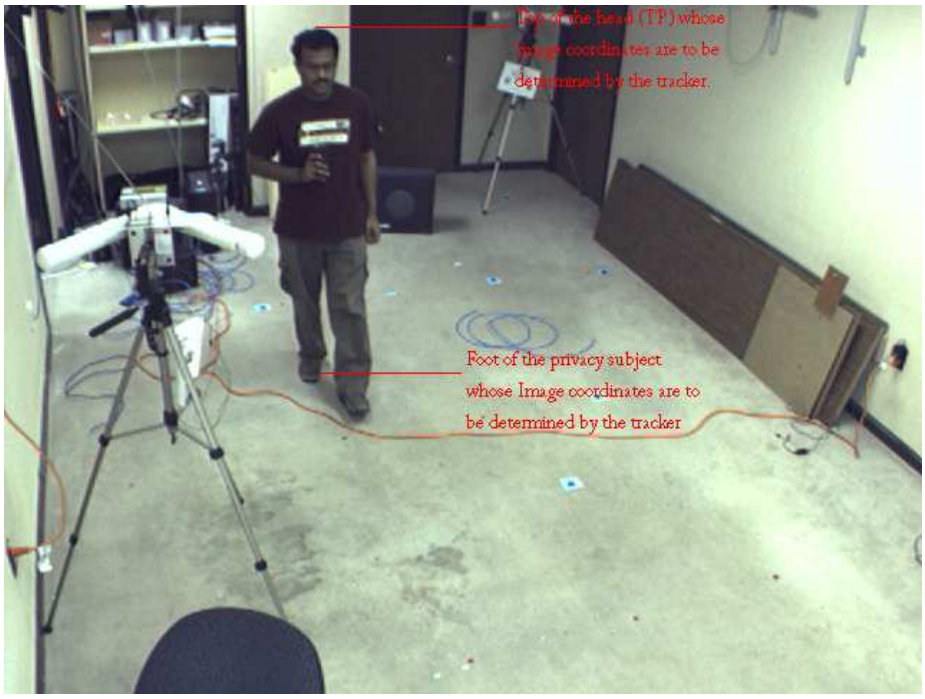



Figure 4.2: Top and foot of the privacy subject

## 4.5 Occlusion Detection

When there are multiple dynamic objects in a video, the background subtraction yields multiple motion blobs. During occlusion the multiple motion blobs combine together to form a huge single blob. The instant at which this happens is considered to be the instant when the occlusion starts. So at any time instant we can detect

occlusion if the number of blobs are less than the number of objects as identified by the wireless tracker. Under conditions when the background subtraction yields multiple blobs but of the same object, that is the background subtraction does not give a single blob for a single moving object, the blob with the largest area is considered as the blob associated with the subject being tracked. This is necessary because if the background subtraction yields multiple blobs for a single object and the total number of blobs in a given frame might exceed the number of objects detected by the wireless tracker. So if the background subtraction yields poor results the occlusion detector might tag each and every individual frame as one in which the subject being tracked is occluded.

## 4.6 Estimation of Model Parameters

The wireless sensor system is calibrated and the measurement noise is estimated by carrying the RFID tag and standing on all the calibration points itself. The two dimensional ground plane coordinates are obtained from the wireless tracker system as a measurement. The corresponding point on the floor map of the wireless tracker is known. The difference between the actual point on the floor plan and the measured point is considered to be the measurement noise. This difference is averaged over all the calibration points to obtain an approximate estimate of the measurement noise from the wireless sensor. The motion model consists of the constant velocity model. Similarly for a given ground plane coordinate we can obtain the corresponding image plane coordinate using the ground plane homography obtained from Section 4.4. The extremum of blob discussed in Section 4.4 yields the foot coordinate of the

subject in the image plane. By finding the difference in the image plane coordinate obtained using the homography and the image plane coordinate obtained using the blob extremum, the measurement noise is obtained for the visual tracker.

# Chapter 5

## Evaluation of the proposed scheme

In this section the joint RFID and visual tracking scheme is evaluated by means of simulations and real video experiments under different test cases.cases. These cases include no occlusion, occlusion with multiple moving objects, and sustained occlusion by a stationary object.

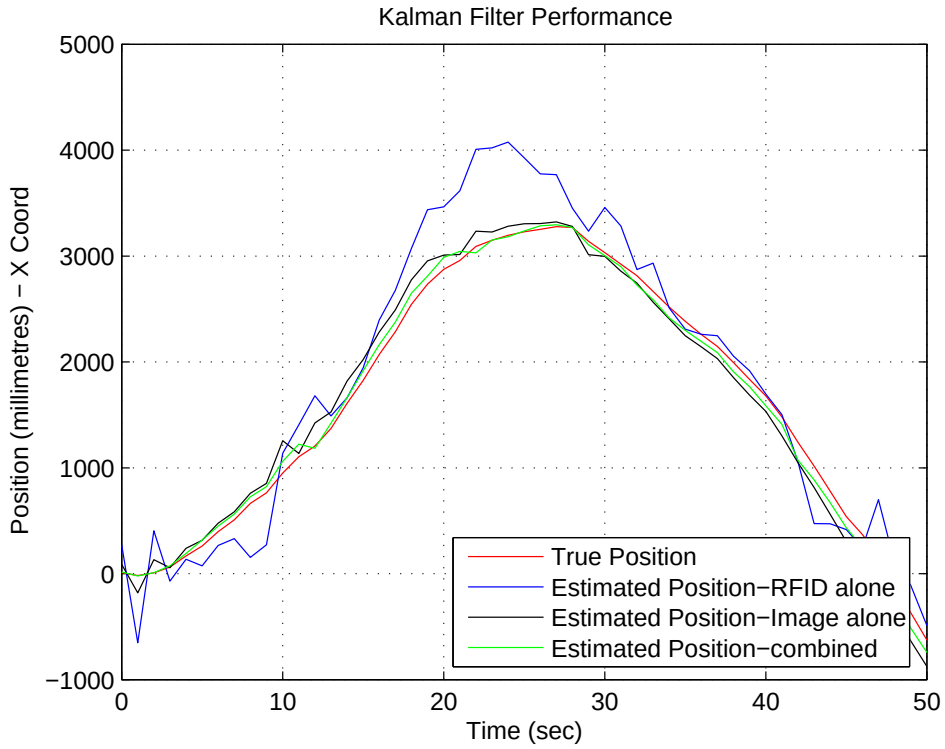
### 5.1 Simulation Results

The motivation behind performing the simulations is that we can test the extended kalman filter for different camera centers, different focal lengths, assume different arbitrary heights of the subject under tracking, etc. The performance of the joint visual and wireless tracker under different varying conditions can be tested to ensure a robust performance. While testing the joint visual and wireless tracker with synthetic data, one does not need to calibrate the camera as we assume a certain homography between the ground plane and the image plane. The simulations are performed on synthetic data that is generated using different motion models like the constant velocity model and constant acceleration model. The process noise and measurement noises of known covariances are added.The measurement noise for the RFID system was set to .6 meters and .4 meters for the horizontal directions and the vertical directions respectively. The noise in the measurement of the visual tracker has been modeled as 1 pixel each along the rows and columns respectively. For the simulations based on the constant acceleration model, the velocity for the system dynamics

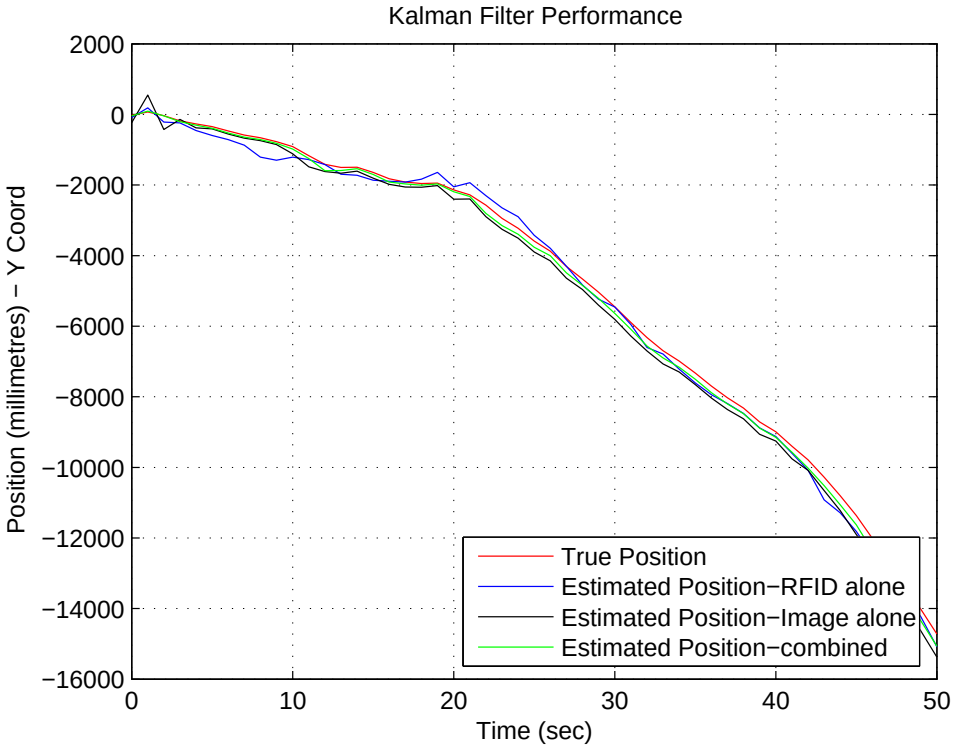
was measured from the wireless tracker. The wireless sensor maps the tracking coordinates to a floor map. In the floor map the time taken to traverse between two sets of points was used as the distance traversed over a duration of time. From this simple experiment the velocity used for the constant acceleration model was set as 3 feet/second along the horizontal direction and 2 feet per second along the vertical direction. The homography used for the camera model in our simulations are obtained using a chequered board and the Camera Calibration toolbox.

Figure 5.1 shows tracking when there is no occlusion. In this case the dynamics of the system are constant acceleration. It compares the individual trackers with the joint visual and wireless tracker, and the true position of the subject being tracked. The $x$ axis of the plots show the time instants and the $y$ axis show the position of the subject being tracked in the horizontal direction and the vertical direction. From the two plots it can be interpreted that the joint wireless and visual tracking scheme works better than the RFID tracking scheme alone or a visual tracking scheme alone. The wireless tracker is more sensitive to noise. On the other hand when there is no occlusion the visual tracker seems to work as good as the joint tracker.

Figure 5.2shows the joint RFID and Visual tracking scheme along with the wireless tracker and the visual tracking scheme alone under conditions of no occlusion. The dynamics used for this simulation is a constant velocity model. The $x$ axis of the plots show the time instants and the $y$ axis show the position of the subject being tracked in the horizontal direction and the vertical direction. At any time instant the difference between the position predicted by the joint visual and wireless tracker from the actual position of the subject being tracked is less. The wireless tracker and also




(a) X coordinate vs Time



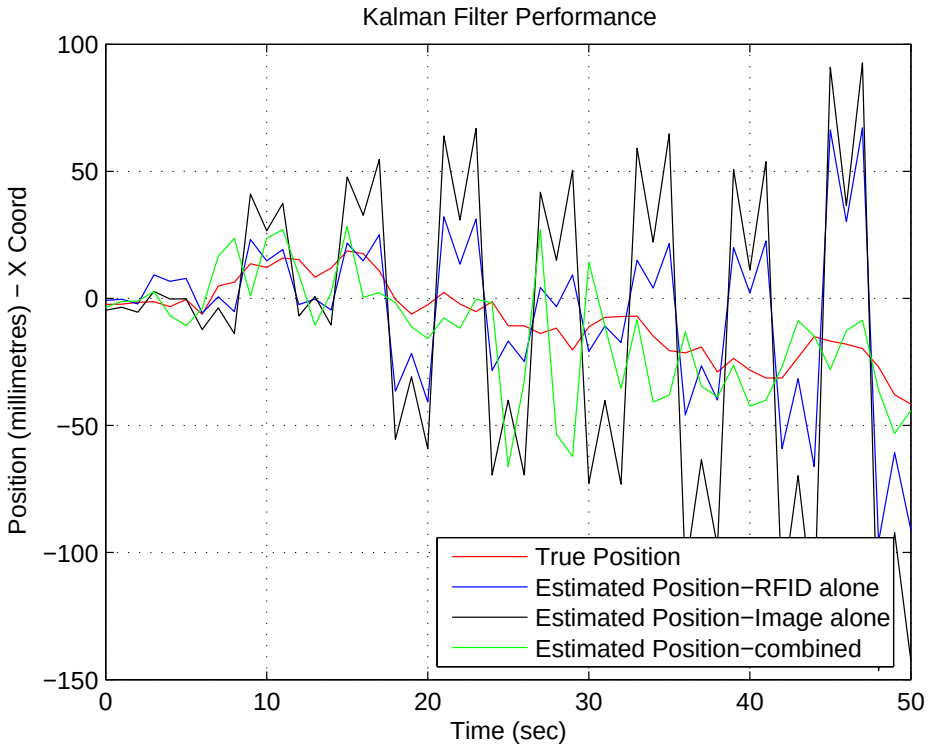

(b) Y coordinate vs Time

Figure 5.1: Performance of tracking schemes:no occlusion - Constant Acceleration Model

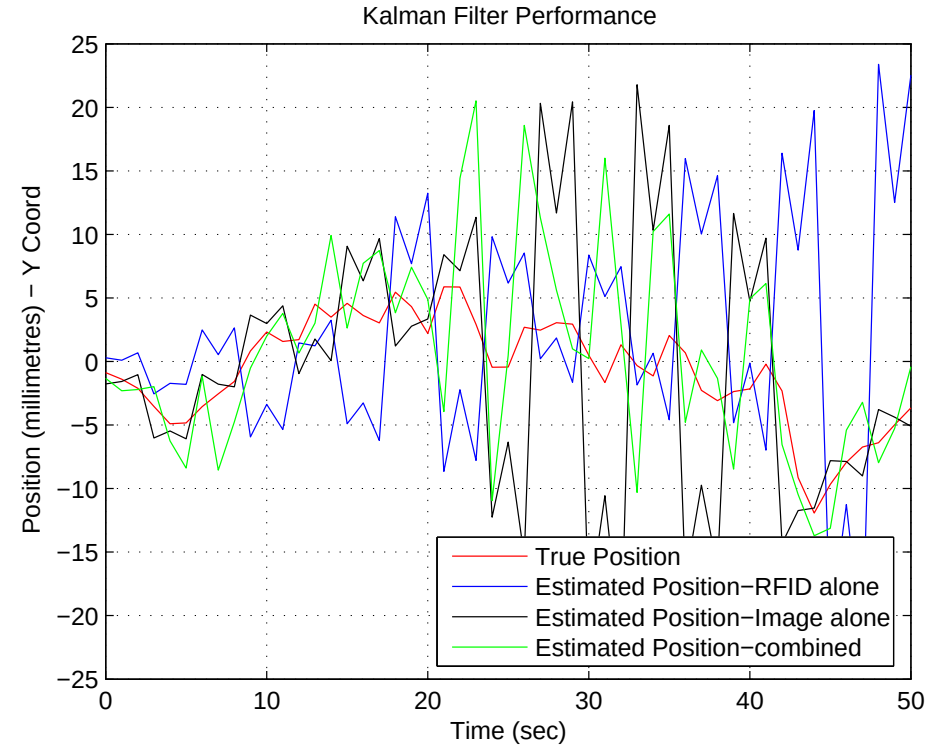
the visual tracker are much more noisy than the joint tracker.

The proposed joint wireless and visual tracking scheme has also been evaluated with simulations of occlusion. We simulated occlusion between specific time intervals during which there are no new measurements from the visual tracker. The combined tracker works under occlusions also with the predictions being weighted by the inverse of the covariance matrix of the noise from the wireless tracker and the visual tracker. The occlusions were simulated for short durations of 2 - 3 time instants and also for longer durations of 15 time instants.

In Figure Figure 5.3 the effect of the three different tracking schemes given that the duration of occlusion is relatively small has been shown. The $x$ axis of the plots show the time instants and the $y$ axis show the position of the subject being tracked in the horizontal direction and the vertical direction. Under such conditions when the occlusion is relatively low the combined tracker ensures that the prediction of the location of the subject is more accurate compared to switching to the wireless tracker during occlusion and tracking subject. This is because even though there are no new measurements from the visual tracker, the variance in the measurements from the visual tracker is much lower compared to the variance of the wireless sensor measurements. When the occlusion is for a longer period as in Figure 5.4 the visual tracker diverges gradually from the actual position. Under such a circumstance the joint wireless and visual tracker performs better.

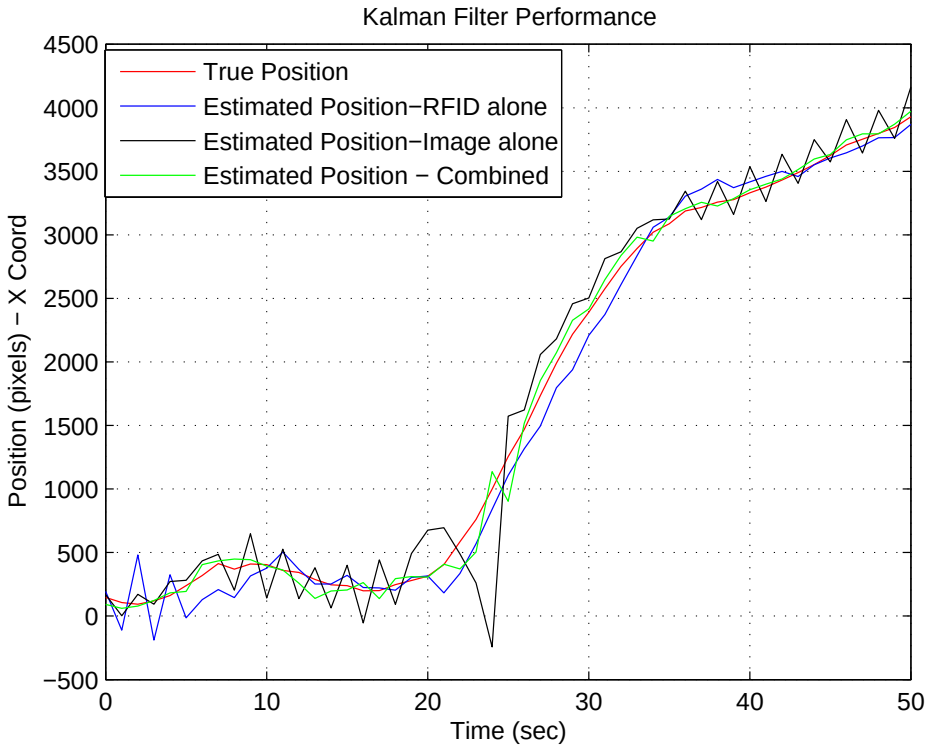



(a) X coordinate vs Time




(b) Y coordinate vs Time

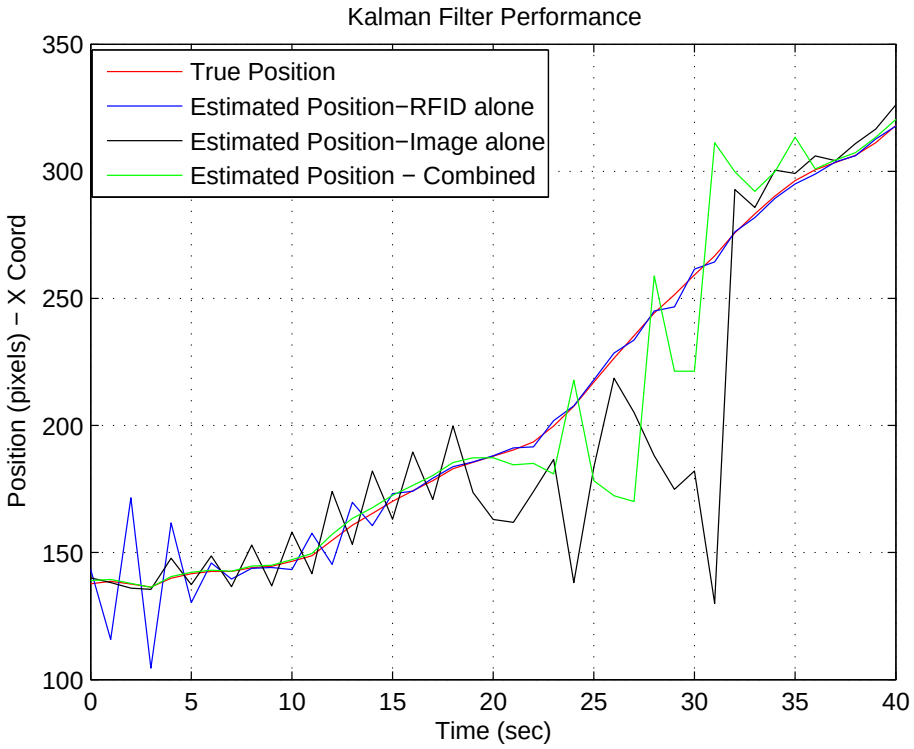
Figure 5.2: Performance of tracking schemes-no occlusion-Constant Velocity Model



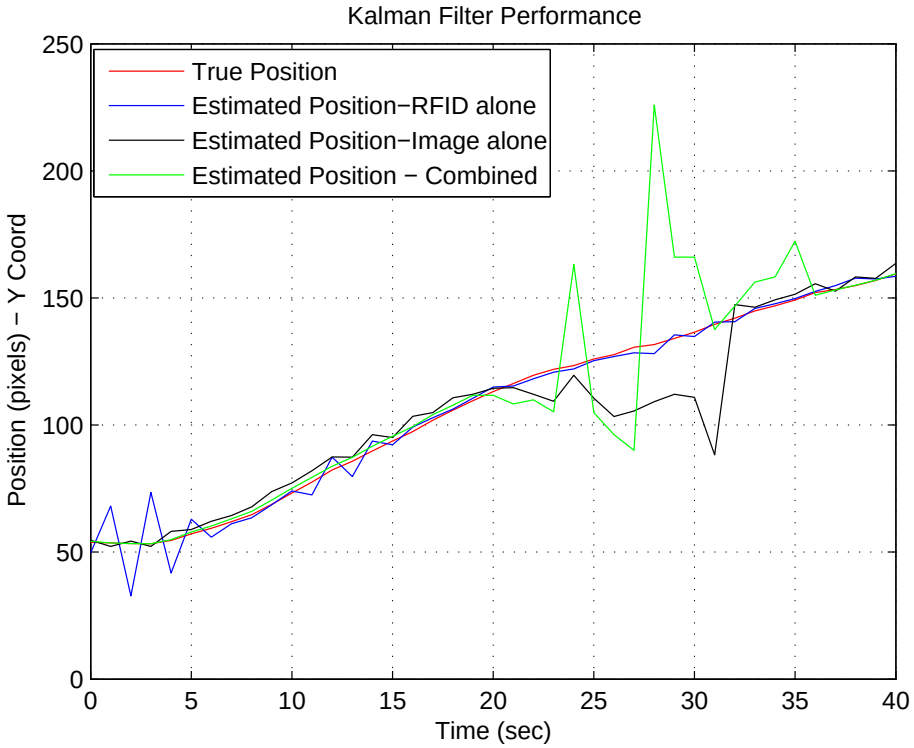

(a) X coordinate vs Time


(b) Y coordinate vs Time

Figure 5.3: Performance of tracker: when the duration of occlusion is low




(a) X coordinate vs Time




(b) Y coordinate vs Time

Figure 5.4: Occlusion for a longer duration

## 5.2 Experimental Results from Real Videos

In this section, we focus on the algorithmic components of our system design and measure their performance with physical data captured at a realistic indoor office environment.We first measure the performance of the RFID system. As such, the calibration of the RFID system is done through the interface provided by the manufacturer – by putting the RFID tag at various calibration points of known world coordinates established using a floor plan image. A subset of the calibration points are also used to establish the homography between the floor plan image and the camera plane so that the RFID coordinates after calibration can be re-projected on the camera plane. Figure 5.5 shows a 16-meter hallway where we have put 32 calibration points. After the calibration, we measure the tracking accuracy at six different locations within the area as shown in Table 5.1. The average error is 0.458 meters with standard deviation equal to 0.273 meters. The error is smaller at the open area (E and F) but larger in the hallway where there are power-lines behind the walls and computer servers inside the offices.

Figure 5.5 shows a 16-meter hallway where we have put 32 calibration points. After the calibration, we measure the tracking accuracy at six different locations within the area as shown in Table 5.1. The average error is 0.458 meters with standard deviation equal to 0.273 meters. The error is smaller at the open area (E and F) but larger in the hallway where there are power-lines behind the walls and computer servers inside the offices.

We perform three different experiments on the real videos that we captured. In the first experiment we estimate the homography of the floor plan with the image


Figure 5.5: RFID Floor Plan

Table 5.1: RFID Tracking Error

| Testing Points | Ground-truth (meters) | RFID (meters) | Error (meters) |
|---|---|---|---|
| A | (6.456, 2.653) | (7.013, 2.840) | 0.588 |
| B | (7.144, 5.150) | (7.013, 4.837) | 0.339 |
| C | (6.587, 6.897) | (7.013, 6.710) | 0.465 |
| D | (8.029, 9.550) | (7.111, 9.488) | 0.920 |
| E | (2.982, 11.953) | (3.277, 11.891) | 0.301 |
| F | (7.013, 11.797) | (7.111, 11.891) | 0.136 |

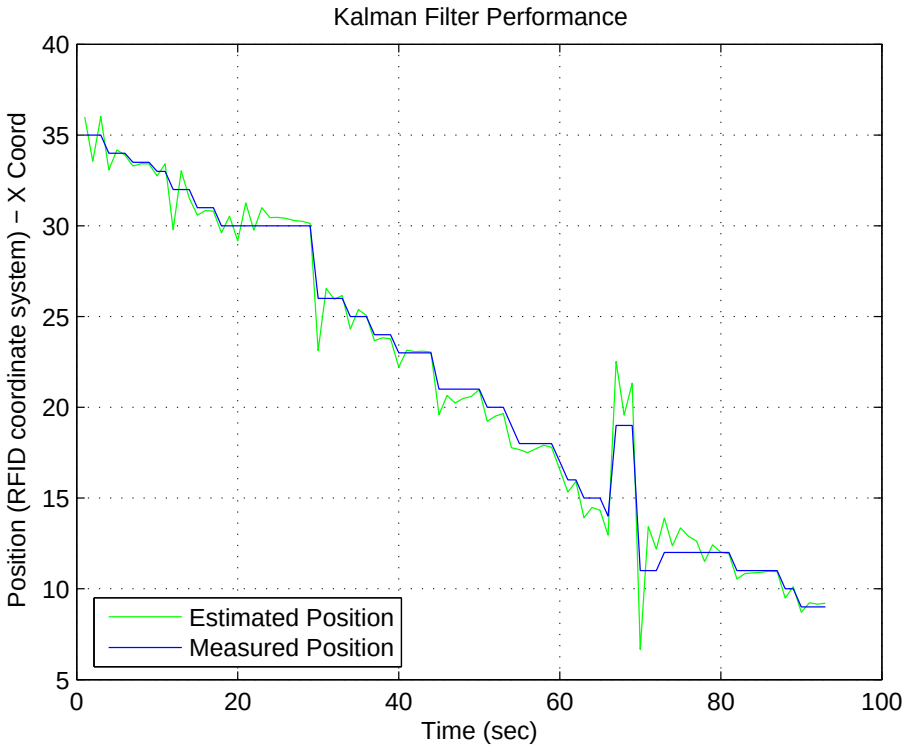
plane by using a training sequence. We compare this with a manually computed homography. In the second experiment we use the combined wireless and visual tracker to track a single individual walking completely in the view of the camera and in an unoccluded scenario. This experiment is conducted to ensure that the combined tracker works on real videos and real data captured from the wireless tracker. In the third experiment we evaluate the combined tracker under for occlusions of short durations and occlusions of longer durations. In the first scenario the occlusion is generated by another individual, that is an object to object occlusion. In the second scenario the occlusion is generated by an artificial wall and the subject being tracked walks behind the wall. We provide the results for the individual trackers - kalman

filter for the wireless tracker and the extended kalman filter for the visual tracker-below.

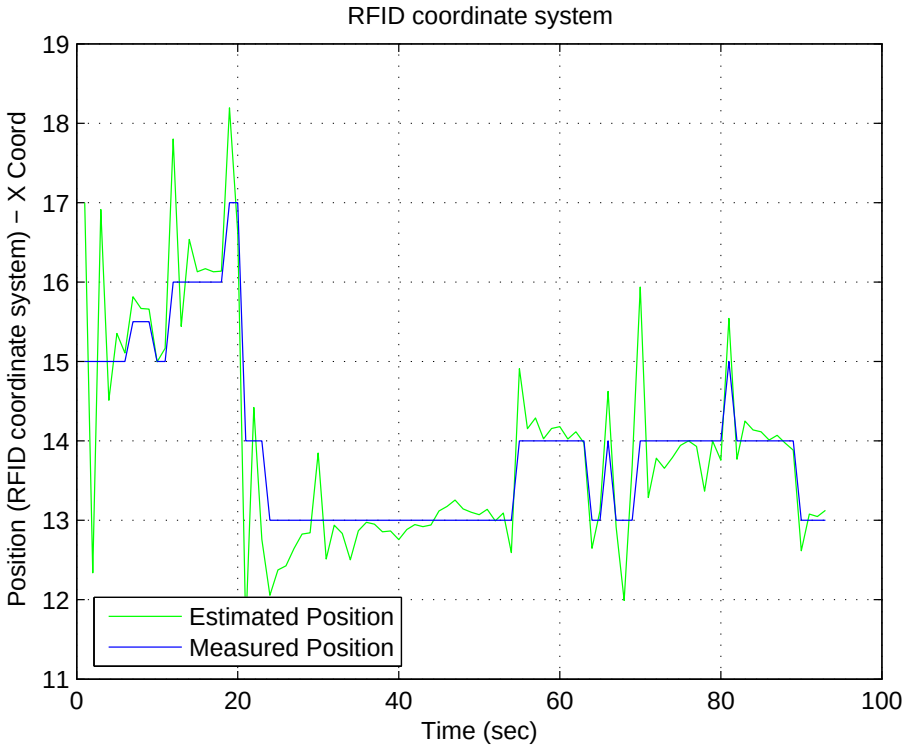
Figure 5.6 shows the kalman filter prediction for the two dimensional coordinates obtained from the wireless tracker. Figure 5.7 shows the results of using an extended kalman filter over 25 frames.

In Figure 5.8 we show the effect of occlusion when occlusion lasts for a longer duration like 30 frames in a video. This might be the case when the subject being tracked is walking behind a wall. In a real video we simulate the occlusion between frames 22 and 53 by not using any new measurements for the visual tracker. The $x$ axis of the plots show the time instants and the $y$ axis show the position of the subject being tracked in the horizontal direction and the vertical direction based on the coordinate system of the floor plan. The joint tracker scheme provides better prediction during occlusion than the measured wireless tracker position and close the the actual position measured using the background subtraction on each frame.

In Figure 5.9 we show the effect of occlusion when occlusion lasts for a shorter duration like 10 frames in a video. This might be the case when the subject is occluded by another subject. The $x$ axis of the plots show the time instants and the $y$ axis show the position of the subject being tracked in the horizontal direction and the vertical direction in pixels on the image plane. For a shorter duration of occlusion the prediction from the joint wireless and visual tracker is much more closer to the actual position.




(a) X coordinate vs Time



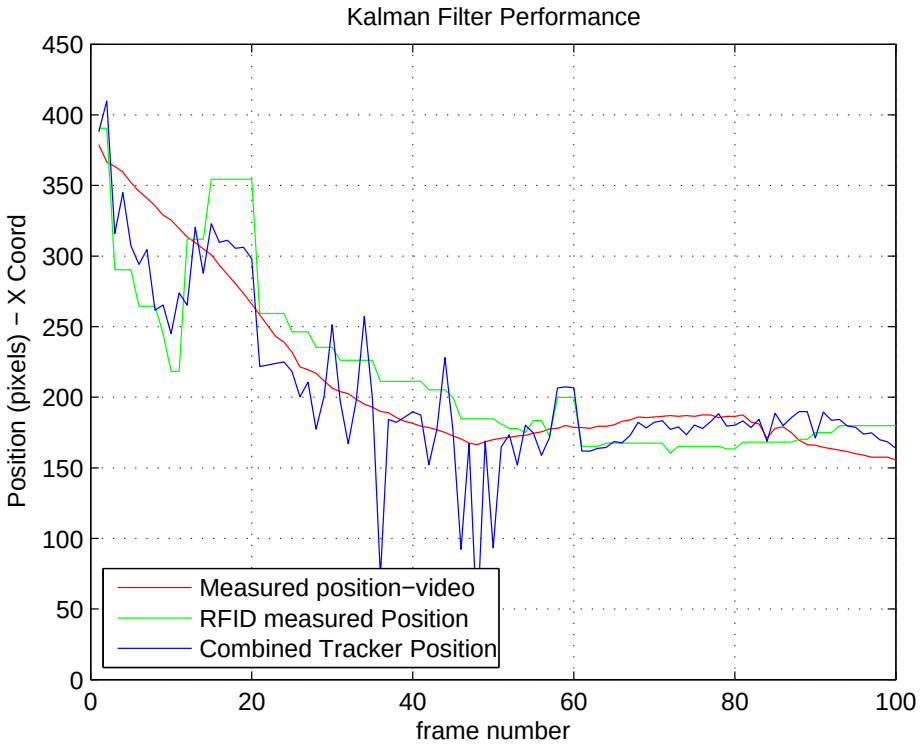

(b) Y coordinate vs Time

Figure 5.6: Prediction for real data from wireless tracker

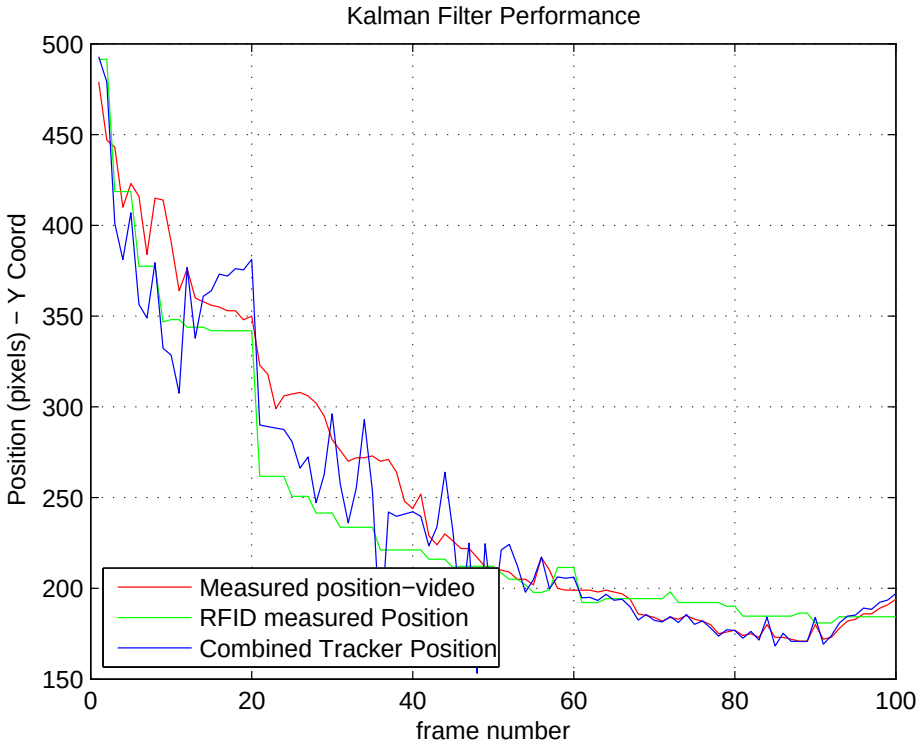

(a) X coordinate vs Time


(b) Y coordinate vs Time

Figure 5.7: Visual Tracking using Real Video

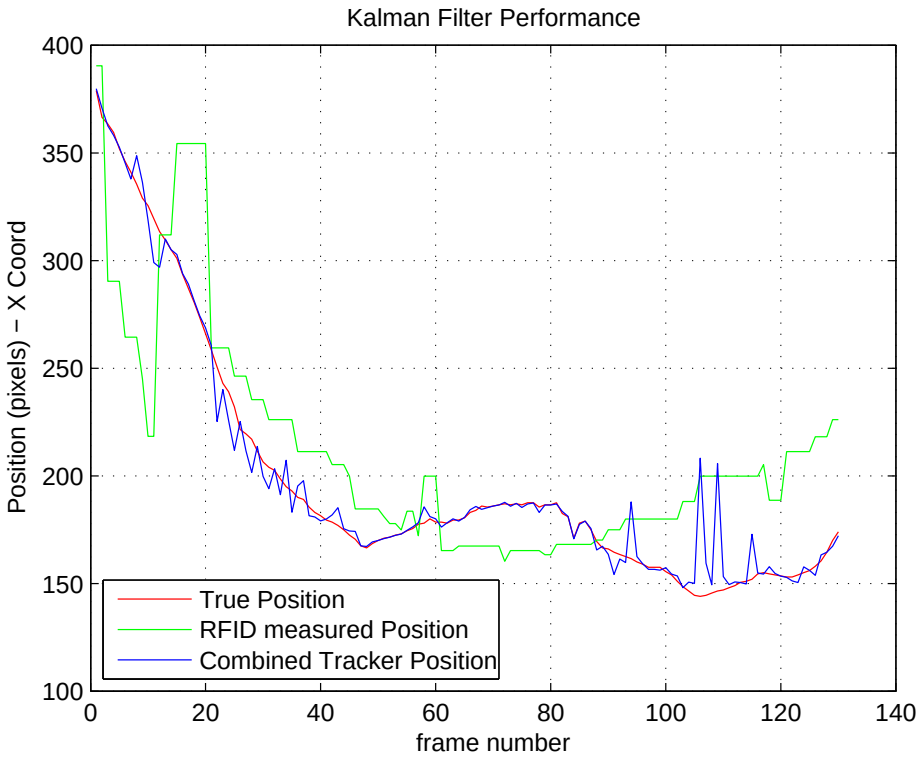



(a) X coordinate vs Time

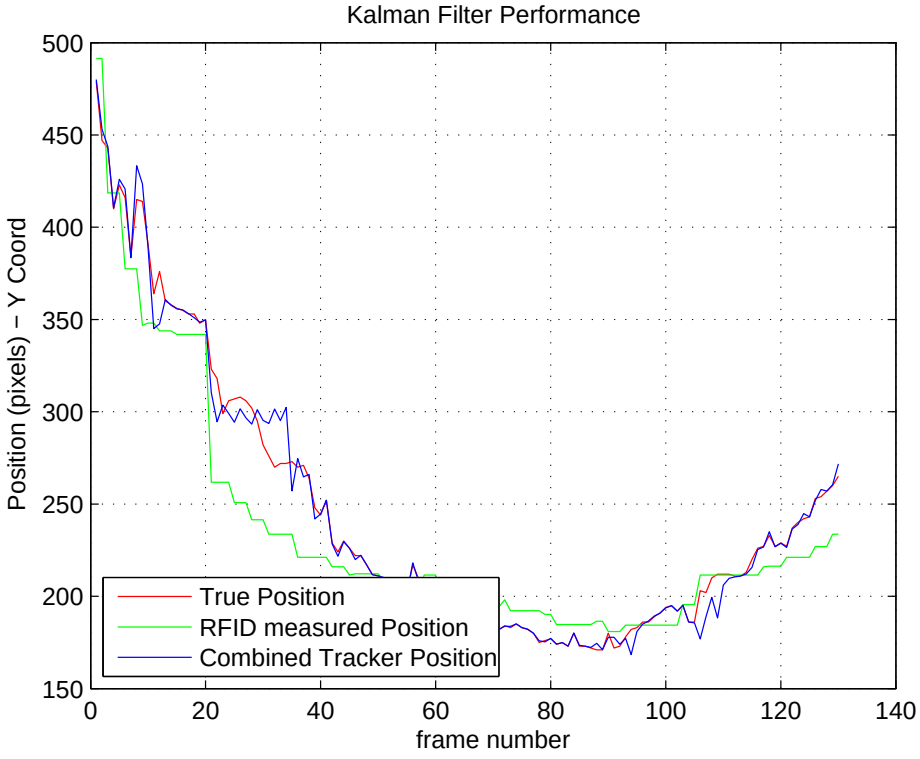



(b) Y coordinate vs Time

Figure 5.8: Occlusion for longer duration-30 Frames




(a) X coordinate vs Time




(b) Y coordinate vs Time

Figure 5.9: Occlusion for short duration-10 Frames

### 5.2.1 Background Subtraction Results

In this section we discuss the results of the background subtraction algorithm used to obtain the measurements of the foot and the head coordinates for the extended kalman filter.

(a) Single Person Sequence

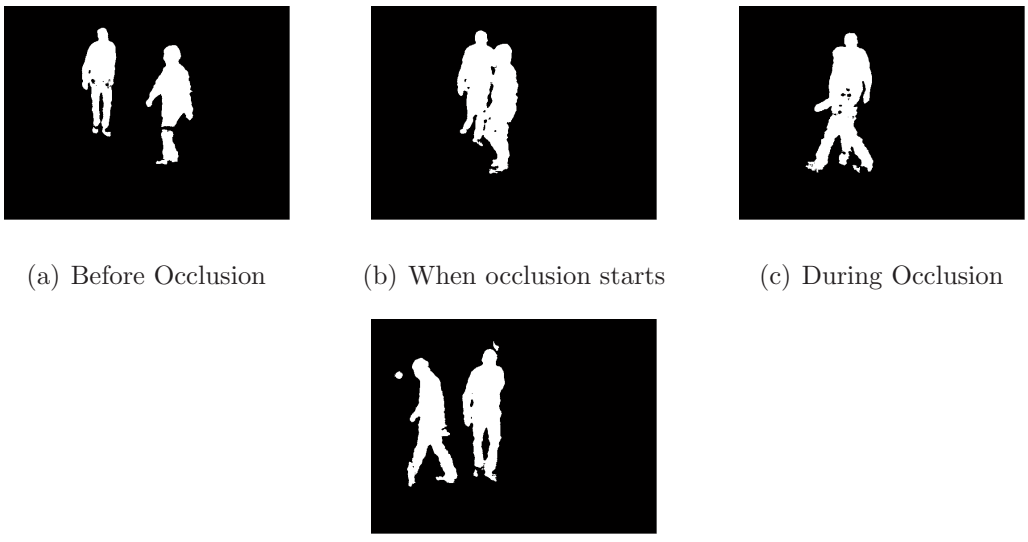
(b) Two Person Sequence

Figure 5.10: Background Subtraction

In Figure 5.10 there are two images. One image shows the background subtraction results for a single person where a single person is walking in the field of view of the camera and is the subject being tracked. In this image, the smaller blob comprises from the right hand's wrist to the fingers while the larger blob comprises the rest

of the subject being tracked. Since we are considering the blob with the largest area to obtain the image coordinates of the foot of the subject, the measured image coordinate of the foot of the subject is reasonably accurate. On the other hand the second image shows the background subtraction results for a two person sequence where two persons are walking in the field of view of the camera and one subject is being tracked.

In this section the results of occlusion detection discussed in the previous chapter are discussed. In Figure 5.11 there are four images, showing blobs before occlusion, when the occlusion starts, during the occlusion and after occlusion. In an ideal scenario there would be just two blobs in a two person sequence and when the two blobs combine to form one single blob i.e during occlusion it can be used as occlusion detector. As in Figure 5.11 the blobs when the occlusion starts and during the occlusion combine together to form a single blob.



(a) Before Occlusion (b) When occlusion starts (c) During Occlusion

(d) After Occlusion

Figure 5.11: Occlusion Detection

### 5.2.2 Tracking by the joint visual and wireless tracker

In this section the results of the joint visual and wireless tracker to track the foot coordinates of the subject are discussed. In Figure 5.12 we show two images of the subject being tracked with a small $x$ marking the estimated foot coordinate of the subject at two different time instants.

The head plane homography obtained using the scheme proposed in chapter four was erroneous. Using the head plane homography when an RFID-tracker coordinate was mapped back on to the image plane the error between the actual image coordinate and the re-mapped image coordinate was found to be more than 20 pixels in some cases. The reason for this might have been that during the training sequence the subject being tracked was walking in a single straight line and not covering the entire field of view of the camera. On the other hand in a wider environment with the subject being tracked traversing throughout the environment during the training sequence would result in a homography much more accurate.

Figure 5.12: Tracking the foot coordinate-No Occlusion

# Chapter 6

## Conclusions

In this research work we propose a joint RFID and Visual tracking system. The proposed work is a multi sensor data fusion scheme targeted towards applications where there are lots of occlusions. The joint wireless and visual tracking scheme is developed using a probabilistic framework based on graphical models. The joint RFID and visual tracking system has been evaluated in a small room where the occlusion behind a wall has been simulated, and also occlusion by multiple persons have been simulated. Further work in improving the proposed scheme would be to develop a real time implementation of the joint RFID and visual tracking system, obtain a more accurate homography for the head plane and obtain the bounding box to track the subject. At the same time the wireless tracking system could be made more robust with an improved calibration. One of the disadvantages of the proposed scheme is the wireless tracking system works robustly over a relatively small area. A key challenge would be to extend the scheme and test it over a larger area.

## Bibliography


[1] Rama Chellappa Ted.J.Broida. Estimation of object motion parameters for noisy images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, PAMI-8, NO.1:90–99, 1986.

[2] Romer Rosales and Stan Sclaroff. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *IEEE conference on Computer Vision and Pattern Recognition*, 1999.

[3] Chris J Needham and Rojer D Boyle. Tracking multiple sports players through occlusion, congestion and scale. In *British Machine Vision Conference*, 2001.

[4] Aggarwal Chang. 3d structure reconstruction from an ego motion sequence using statistical estimation and detection theory. In *Workship in Visual Motion*, 1991.

[5] M.J.Black H.Sidenbladh and D.J.Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision*, 2000.

[6] Dimitrios Tzovaras Konstantinos Moustakas and Michael G Strintzis. A non causal bayesian framework for object tracking and occlusion handling for the synthesis of stereoscopic video. In *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, 2004.

[7] N. Otsuka, K. Mukawa. Multiview occlusion analysis for tracking densely populated objects based on 2-d visual angles. In *Computer Vision and Pattern*

*Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

[8] Noh.J Huttenlocher.D and Rucklidge.W. Tracking non rigid objects in complex scenes. In *International Conference in Computer Vision*, 1993.

[9] Peter Meer Dorin Comaniciu, Visvanathan Ramesh. Kernel based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:564–577, 2003.

[10] Mubarak Shah Khurram Shafique. A non iterative greedy algorithm for multiple frame point correspondence. *IEEE Transactions on Pattern Analysis and Motion Intelligence*, 27:51–65, Jan 2005.

[11] Khurram Shafique Mubarak Shah Omar Javed, Zeeshan Rasheed. Tracking across multiple cameras with disjoint views. In *International Conference on Computer Vision*, 2003.

[12] A. MacCormick, J. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999.

[13] W. Freeman E. Sudderth, M. Mandel and A. Willsky. Distributed occlusion reasoning for tracking with non parametric belief propagation. In *Neural Information Processing Systems Conference*, December 2004.

[14] Ying-Ti Lian Lisa Brown Sharath Pankanti Andrew Senior, Arun Hamapur and Ruud Bolle. Appearance models for occlusion handling. In *Image and Vision Computing*, Volume 24, Issue 11, November 2006.

[15] H.T. Nguyen and A.W.M. Smeulders. Fast occluded object tracking by a robust appearance filter. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26:1099–1104, 2004.

[16] Justus H Piater Pierre F. Gabriel, Jacques G. Verly and Andre Genon. The state of the art in multiple object tracking under occlusion in video sequences. In *Advanced Concepts for Intelligent Vision systems*, 2003.

[17] Z.Duric S.Mckenna, S.Jabri and H. Wechsler. Tracking groups of people. In *Computer Vision and Image Understanding*, 2000.

[18] F. Bremond and M.Thonnat. Tracking multiple non rigid objects in video sequences. *IEEE Transactions on Circuits and Systems*, 8:585–591, 1998.

[19] Edward Dickerson Dickey Arndt and Jianjun Ni. Ultra wideband two cluster angle of arrival tracking system design for space exploration. In *Third IEEE Workshop on Local Area Networks*, 2001.

[20] Maxim Foursa. Real time infrared tracking for virtual environments. In *Virtual Reality Continuum And Its Applications Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*, 2004.

[21] H. Schantz and R. Depierre. System and method for near-field electromagnetic ranging. Technical report, U.S. Patent, 2005.

[22] *The Q-Track Corporation*. http://www.q-track.com.

[23] Stephen J.Maybank Nils T.Siebel. Fusion of multiple tracking algorithms for robust people tracking. In *Proceedings of the 7th European Conference on Computer Vision Part IV*, 2002.

[24] Martin Spengler and Bernt Schiele. Towards robust multi cue integration for visual tracking. In *Proceedings of the Second International Workshop on Computer Vision Systems*, 2001.

[25] Christoph Von Der Malsburg Jochen Triesch. Self-organized integration of adaptive visual cues for face tracking. In *Fourth International Conference on Automatic face and gesture recognition*, 2000.

[26] J. BLAKE A. PEREZ, P. VERMAAK. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92:495–513, March 2004.

[27] Jianru Xue Xiaopin Zhong and Nanning Zheng. Graphical model based cue integration strategy for head tracking. In *British Machine Vision Conference*, 2006.

[28] M. Rivlin E. Leichter, I. Lindenbaum. A probabilistic framework for combining tracking algorithms. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

[29] Ying Wu and Thomas S Huang. Robust visual tracking by integrating multiple cues based on co inference learning. *International Journal of Computer Vision*, 58:55–71, 2004.

[30] Oruc I.1; Maloney L.T.; Landy M.S. Weighted linear cue combination with possibly correlated error. *Vision Research*, 43:2451–2468, October 2003.

[31] S. Dept. of Comput. Sci. & Eng. UC San Diego La Jolla CA USA Branson, K. Belongie. Tracking multiple mouse contours(without using too many samples). In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005.

[32] Michael Jordan. *An Introduction to Probabilistic Graphical Models*. 2002.

[33] Larry S Davis Thanarat Horprasert, David Harwood. A statistical approach for real time robust background subtraction and shadow detection. In *IEEE Framrate Applications Workshop*, 1999.

[34] Jian Zhao M.Vijay Venkatesh, Sen-Ching Samson Cheung. Efficient object-based video inpainting. *Pattern Recognition Letters*, 30:168–179, 2009.

# VITA

Name: Viswajith Karapoondi Nott

Bachelor's in Electrical and Electronics Engineering

Anna University, Chennai, India

Date of birth: May 22, 1984

Place of birth: Chennai, India

Positions held:

1. Assistant Systems Engineer trainee, TCS, India

2. Research Assistant, University of Kentucky

Publications: Ultra Folded high speed architectures for Reed Solomon Decoders at the 19th International Conference for VLSI and Embedded Systems, January 2006, Hyderabad, India.