A
Dissertation
on

Multi Modal Tracking of Multiple Objects Based on Speech

Submitted in Partial fulfillment of the requirement
For the award of the degree of

Master of Technology
in
Signal Processing & Digital Design

By:

Kanika Jain
2K13/SPD/07

Under The Guidance Of

DR. RAJIV KAPOOR
Professor, ECE
Delhi Technological University

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

2013-2015

CERTIFICATE

DEPARTMENT OF ELECTRONICS & COMMUNICATION
DELHI TECHNOLIGICAL UNIVERSITY
DELHI - 110042

This is to undertake that the work presented in this M.Tech. dissertation titled **“Multi Modal Tracking of Multiple Objects Based on Speech ”** is an authentic record of my own work carried out under the supervision of **Dr. Rajiv Kapoor, Professor,** Department of Electronics & Communication Engineering. It is submitted in partial fulfillment of the requirements for the award of the **Master of Technology in Signal Processing & Digital Design** at Department of Electronics & Communication Engineering, **Delhi Technological University**. The matter presented in this dissertation has not been submitted by me for the award of any other degree elsewhere.

KANIKA JAIN
2K13/SPD/07

This is to certify that the above statement made by the candidate is correct to the best of my knowledge and dissertation be forwarded for external evaluation.

Date:

Dr. Rajiv Kapoor
Professor, ECE

ACKNOWLEDGEMENT

I would like to acknowledge all those people who gave me an opportunity and their valuable guidance to work on this project. My sincere regards and gratitude to my project supervisor **Dr. Rajiv Kapoor**, Professor, Department of Electronics & Communication Engineering, Delhi Technological University, for providing me this stupendous opportunity of working in this area Signal Processing.

Again, I would like to thank **my project guide Dr. Rajiv Kapoor** for his efforts and ideas that helped me incalculably during the course of my work. I also sincerely thank for his timely help and guidance throughout the project. The project was a great challenge for me and without the help and support of mentor, it would not have been possible to go about this task.

Finally, I again wish to place on my records my heartfelt thanks and gratitude to all the senior students and my classmates at Delhi Technological University without whose sincere help, guidance, encouragement and cooperation this project could not have been successfully completed.

Kanika Jain
2K13/SPD/07
M.Tech (SPDD)
(2013-2015)

ABSTRACT

Object tracking has always been a very interesting and important research area in the field of Computer Vision and Artificial Intelligence. Tracking an object of interest is an application that can benefit from multiple sensing modalities. If the object of interest emits sound then information from both audio and video sensors can be fused together to remove effects of clutter and background noise. Therefore, the use of same visual and audio interface modalities that humans take for granted can make indoor spaces more intelligent. Audio and visual modalities complement each other when background noise impairs a single modality.

This work presents a new approach for modeling and processing data from audio and visual sensors for tracking multiple objects simultaneously. This approach is based on graphical model for visual data and Time Delay of Arrival (TDOA) analysis for sound cue. Then both the cues are modeled by a data likelihood function. Finally, Particle Filtering for multiple target tracking and Dezert-Smarandache theory (DSmT) for fusing the information provided by audio-visual cues are combined.

For modeling the visual cue, initially dominant motion is detected from the video frames. Then some dominant motion points are selected for depicting movements of target object. When some occlusion occurs, these motion points estimate the object position from a graphical model.

As for modeling the sound cue, the Time Delay of Arrival (TDOA) which occurs between the two audio signals received by the two different microphones kept a fixed distance apart from each other provides an indication of the position of the sound source(s) relative to the microphone pair. This provides an estimate of the horizontal position of object in the image.

KEYWORDS: Tracking multiple objects, multiple sensing modalities, Particle filtering, Dezert-Smarandache theory (DSmT).

CONTENTS

TITLE *PAGE NO.*

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

TDOA	Time Delay of Arrival
DSmT	Dezert-Smarandache theory
SMCM	Sequential Monte Carlo Methods
DST	Dempster-Shafer
DSmC	Classical DSm
Gbbm	Generalized Basic Belief Mass
KLT	Kanade–Lucas–Tomasi
GCCF	Generalized Cross-Correlation Function
IFFT	Inverse Fast Fourier Transform
FFT	Fast Fourier Transform

1. INTRODUCTION

1.1 Motivation

Object tracking has always been a very interesting and important research area in the field of Computer Vision and Artificial Intelligence. Tracking an object of interest is an application that can benefit from multiple sensing modalities. If the object of interest emits sound then information from both audio and video sensors can be fused together to remove effects of clutter and background noise. Therefore, the use of same visual and audio interface modalities that humans take for granted can make indoor spaces more intelligent.

Audio and visual modalities complement each other when background noise impairs a single modality. Tracking an object of interest is an application that can benefit from multiple sensing modalities. If the object of interest emits sound then information from both audio and video sensors can be fused together to remove effects of clutter and background noise. Therefore, the use of same visual and audio interface modalities that humans take for granted can make indoor spaces more intelligent. Audio and visual modalities complement each other when background noise impairs a single modality.

This work presents a new approach for modeling and processing data from audio and visual sensors for tracking multiple objects simultaneously. This approach is based on graphical model for visual data and Time Delay of Arrival (TDOA) analysis for sound cue. Then both the cues are modeled by a data likelihood function. Finally, Particle Filtering for multiple target tracking and Dezert-Smarandache theory (DSmT) for fusing the information provided by audio-visual cues are combined. For modeling the visual cue, initially dominant motion is detected from the video frames. Then some dominant motion points are selected for depicting movements of target object. When some occlusion occurs, these motion points estimate the object position from a graphical model.

As for modeling the sound cue, the Time Delay of Arrival (TDOA) which occurs between the two audio signals received by the two different microphones kept a fixed distance apart from each other provides an indication of the position of the sound source(s) relative to the microphone pair. This provides an estimate of the horizontal position of object in the image.

Overview of the system can be seen in fig. 1.

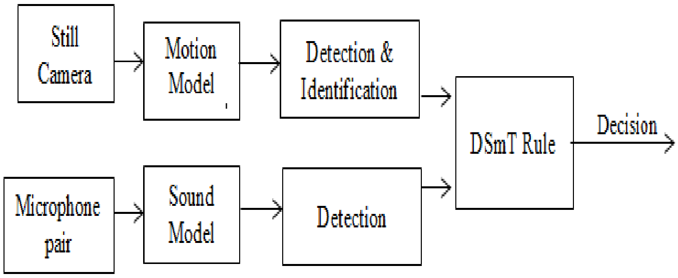


Figure 1. Audio-Visual fusion using DSmT in Particle filtering process

1.2 Literature Review

Recently Particle Filters, also known as Sequential Monte Carlo Methods (SMCM) [1], [2], [3] and [4] are popularly used as a tool to solve the tracking problem. Their popularity comes from their properties like : simplicity of the framework, flexibility and adaptability, their ease of implementation etc. Isard and Blake [3] proved the usefulness of particle filters for visual tracking and then coined the term CONDENSATION. This led to a vast body of literature in [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] and [15] based on SMCM tracking which is not attempted to be reviewed here.

But one advantage of the SMCM based framework that makes it useful in multimodal tracking is that it provides provision for the fusion of information from different measurement in its framework in a principled manner that is, multiple sensing modalities can be exploited for tracking purpose using particle filters. Although this property of SMCM has also been exploited by various trackers before, but it has not been fully explored for multi target tracking within a visual tracking framework.

Also tracking based on data fusion with SMCM based trackers has been mostly limited to applications for face and hand tracking [16], [17], [18], [19].

But in this work a visual tracker that fuses data from audio and visual sensors in a novel way (Fig. 2.2) for tracking multiple objects has been proposed. The generic objective is such approach is to successfully track multiple objects or regions of interest in a given sequence of images (video) and sound samples captured by a stationary camera and microphones.

Numerous methods [20, 21] have proved that fusion of information from multiple cues do increase the visual robustness of tracking algorithm especially in case of complex scenes resulting in occlusions. Also numerous probabilistic [20–23] and non-probabilistic methods [24–27] have been implemented for integration/fusion of the data from multiple cues for object representation. If we have the prior distributions along with their conditional probabilities, some probabilistic methods such as the Bayesian inference can provide a most simple, complete, scalable as well as theoretically justifiable modeling approach for data fusion. But, due to various problems such as occlusions, background noise, camera calibration and illumination problems in real time scenarios with multiple targets, that it becomes difficult to obtain such complete knowledge.

A probabilistic approach based solely on Bayesian inference has been proposed in [28] for data fusion, that is the Dempster-Shafer theory (DST). In DST, the uncertainty and imprecision of a knowledge source is represented in terms of confidence values that are committed to a single or a union of hypotheses. But certain inherent limitations of DST such as not being able to handle conflicting information sources pose a problem for its use as a fusion technique in real-world tracking scenarios. To overcome this limitation, recently the Dezert-Smarandache Theory (DSmT) has been proposed [29]. DSmT is considered as a generalized version of the DST.

In this work, a novel method for multiple targets using audio and visual cues is proposed. The audio-visual cues are combined using DSmT. When the target is partially or fully occluded, the conflict that arises between the cues is analyzed using DSmT and then exploited in the tracking process. This proposed scheme is simple, novel and provides accurate tracking results even in cluttered scenes.

1.3 Objective and Scope of the Project

The objective of this project is to track multiple objects/targets in a given video sequence by utilizing both audio and visual information collected from a still camera and a microphone pair. This approach is based on using a graphical model for visual data for detecting some dominant motion points and TDOA analysis for sound cue. Then both the cues are modeled by a data likelihood function. Finally, Particle Filtering for multiple target tracking and Dezert-Smarandache theory (DSmT) for fusing the information provided by audio-visual cues are combined.

Using this approach, this work aims at successfully tracking all the objects of interests simultaneously in a given video sequence even in presence of background noise and partial/full occlusion. For modeling the visual cue, initially dominant motion is detected from the video frames. Then some dominant motion points are selected for depicting movements of target object. When some occlusion occurs, these motion points estimate the object position from a graphical model.

As for sound model, Time Delay of Arrival (TDOA) between the audio signals arriving at the two different microphones provides indication of the position of the sound source relative to the microphone pair. This provides an estimate of the horizontal position of object in the image.

Finally information from both sound and motion model is fused together using DSmT data fusion technique and combined with Particle filtering.

1.4 Organization of Thesis

This dissertation has been divided into five chapters. A brief overview of each chapter is presented below.

Chapter 2: Object Tracking Overview

It explains the basics of object tracking. It first presents the details about Particle filtering for tracking objects. Then it explains the meaning of multimodal object tracking using multiple cues. Then it provides the method for tracking more than one object of interest that is, multi target tracking. Finally it explains the DSmT theory used in this dissertation for data fusion.

Chapter 3: System Overview

This chapter provides the complete details about the proposed system for multi modal multi target tracking. It explains the motion model adopted for processing of data from visual cue along with sound model used for audio cue. It also describes the final tracking including data fusion with Particle filter.

Chapter 4: Results

In this section the performance of our tracking algorithm will be demonstrated. It shows the results that are obtained by implementing the proposed approach to track multiple (two in this experiment) targets walking in a video sequence captured by a single stationary camera and a pair of stereo microphones kept apart at a fix distant. This chapter then justifies the proposed approach in contrast to the already existing algorithms for tracking.

Chapter 5: Conclusion and Future work

This chapter concludes this thesis by discussing the overall contribution of this dissertation in the research field of object tracking. It also enlists the limitations of the approach and points to future research directions.

2. OBJECT TRACKING OVERVIEW

Object tracking has always been a very interesting and important research area in the field of Computer Vision and Artificial Intelligence. In general sense object tracking basically refers to first detecting moving object of interest in a sequence of images and then tracking that detected object from frame to frame in that sequence over a period of time. Object tracking can then also be used for analyzing the tracks of the objects in order to estimate and predicting their behavior in cluttered sections of the video sequence. Therefore tracking of objects can be applied in numerous online and offline image and video processing based systems, such as motion-based recognition (a system where automatic human identification is done based on automatic object detection), traffic monitoring (a real-time (online) application for analysis of traffic data from live traffic video feed to direct traffic flow), automated surveillance and anomaly detections (monitoring video sequence for suspicious activities and anomalous behaviors), video indexing (basically meaning automatic searching and retrieving of the videos from online multimedia databases, gesture recognition based human-computer interaction, vehicle navigation systems.

2.1 Particle Filtering

SMCM based particle filters have been widely used for visual tracking [1], [2], [3], [4]. Here we are discussing a brief summary of its framework, and some details about the architectural variations that are require due to the presence of multiple measurement sources.

If x_n and y_n are the hidden state of the object of interest and the measurements at discrete time step n, respectively, then $p(x_n/y_{1:n})$ gives the tracking posterior distribution of object of interest, called the filtering distribution, where $y_{1:n} = y_1, y_{2,} \dots, y_n$ gives all the measurement observations up to the current time step n. Using Bayesian Sequential estimation, the filtering distribution is computed according to the two step recursion:

Prediction:

$$p(x_n/y_{1:n}) = \int p(x_n/x_{n-1})\, p(x_{n-1}/y_{1:n-1})\, dx_{n-1} \qquad (1)$$

Filtering:

$$p(x_n/y_{1:n}) \propto p(y_n/x_n)p(x_n/y_{1:n}) \quad (2)$$

The above recursion requires a dynamic model describing the state evolution, $\mathrm{p}(\mathrm{x_n}/\mathrm{x_{n-1}})$ and a likelihood model providing the likelihood of any state in the light of the current observation, $p(y_n/x_n)$. After initializing the sequence of filtering distributions, the point estimates of the state x_n is obtained using any appropriate loss function such as the Maximum a Posteriori (MAP) estimate (arg max max_{x_n} $p(x_n/y_{1:n})$ or the Minimum Mean Square Error (MMSE) estimate.

The basic idea behind particle filters is very simple. As the name suggests, starting with a weighted set of particles (samples) $S_{n-1}^{(i)} = \{x_{n-1}^{(i)}, w_{n-1}^{(i)}\}, i = 1,2,..,N_P$ which are normalised and are distributed according to $p(x_{n-1}/y_{1:n-1})$ approximately. At each time step new samples are generated according to a distribution depending on the old state and the new measurements i.e., $S_n^{(i)} \sim q(x_n/x_{n-1}^{(i)}, y_n),\ i = 1,2,..,N_P$. To maintain a reliable set of samples at the end of each recursion step, the new importance weights are set calculated and the target's estimate as calculated as:

$$w_n^{(i)} \propto w_{n-1}^{(i)} \frac{p(y_n/x_n^{(i)})\, p(x_n^{(i)}/x_{n-1}^{(i)})}{q(x_n/x_{n-1}^{(i)}, y_n)},$$

and

$$E[S_n] = \sum_{i=1}^{N_P} x_n^{(i)} \, . \, w_n^{(i)} \quad (3)$$

Now, the new particle set $S_n^{(i)} = \{x_n^{(i)}, w_n^{(i)}\}$ is distributed according to $p(x_n/y_{1:n})$. Also, to avoid the problem of degeneracy of the weights of the particles, it is necessary to resample [1] the particles, that is the concentration of most of the weight on a single particle from time to time. The resampling procedure basically involves multiplication of particles with high importance weights, meanwhile discarding those with low weights. This procedure can be optionally applied after each time step, or when a measure of the "quality" of the importance weights falls below a certain threshold value. A detailed discussion about degeneracy and resampling can be studied in [1].

Now the main challenge that affects the performance of particle filtering for tracking application is to find an efficient proposal distribution. Special consideration has been given to this issue in the design of our tracker in Section II(C) where DSmT, a data fusion technique has been exploited for designing the proposal distribution function.

For multiple sensing sources this general particle filtering framework can be used, by exploiting the relation between the structure of the model and the information from the measurement modalities. If we have M information sources, then the instantaneous measurement vector can be given by $y = (y^1, \dots, y^M)$. Also as measurements are assumed to be conditionally independent, so the likelihood can be factorized as

$$p(y/x) \ = \prod_{m=1}^{M} p(y^m/x) \tag{4}$$

Also, the weight update step involving *M* likelihood evaluations can be according to (4).

2.2 The Dezert-Smarandache Theory

DSmT proposed in [29] for the sole purpose of information fusion is a generalized version of the classical data fusion theory that is, Dempster-Shafer theory (DST). DST initially became popular for providing a formal framework for combining information arising from multiple independent but potentially highly conflicting, paradoxical, uncertain and imprecise sources. But as the complexities grew in terms of the conflicting sources, some inherent limitations of DST became evident. Therefore DSmT was proposed in [29]. DSmT is able to successfully solve complex data fusion problems where DST normally fails, especially in the case where conflicts between sources become high. Data fusion techniques based on this recent DSmT has also been exploited for fusing data in multiple target tracking systems. In this section, firstly a quick review of DST is presented before we go into the details of the DSmT.

2.2.1 Dempster-Shafer Theory

Dempster's rule of data fusion (DST) is used for making inferences from conflicting, uncertain, imprecise and incomplete knowledge by combining several sources of confidence, for example in the case of partially conflicting and contradictory sensors. It represents the uncertainty and imprecision in the measurements from multiple sources in terms of confidence values. One of the advantages that led to its popularity is that it does not require any knowledge of the prior probabilities. Also Bayesian theory of partial belief as included in DST a special case.

Dempster's rule of data fusion basically combines the given imprecise measurements from multiple sources and from them provides a reliable assessment of the uncertainty. But it is successful only in case of low conflicts between the sensors/sources. As when the conflict between the sources grows to a high level, Dempster's rule of combination results in false conclusions and therefore cannot be trusted for providing a reliable fusion result at all.

While DST strictly considers the participated sources/sensors as a set of mutually exclusive elements (that is all the information sources should be completely mutually exclusive), DSmT relaxes this exclusion condition and allows for intersecting, interdependent and overlapping hypotheses (sources need not be fully exclusive). This also requires DSmT to quantify the conflicts that might result between the overlapping sources/sensors due to noise and occlusion.

2.2.2 DSmT

Let us consider a set of some n potentially overlapping elements represented by $\Theta = \{ \theta_1, \theta_2, \dots., \theta_n \}$. Then, by using the operators intersection ($\bigcap$) and union ($\bigcup$) on the set Θ, the hyper-power set D^{Θ} is defined as a set of the entire composite hypothesis such that:

i) $\phi, \theta_1, \theta_2, \dots., \theta_n \epsilon D^{\Theta}$

ii) If $A, B \epsilon D^{\Theta}$ then $(A \bigcup B) \epsilon D^{\Theta}$ and $(\mathrm{A} \bigcap B) \epsilon D^{\Theta}$.

iii) Except the elements defined in i) and ii), D^{Θ} contains no other element.

As given in [29] by F. Smarandache and Dezert, if the cardinality of Θ equals n, the cardinality of D^{Θ} in terms of n is given by $2^{2^{\mathrm{n}}}$. Actually, the Dedekind's problem for enumeration of Boolean functions defines the generation of D^{Θ} in [39]. Also for any finite set Θ if $|D^{\Theta}| \geq |2^{\Theta}|$, D^{Θ} then is the hyper-power set of the set Θ.

DSmT assigns a generalized basic belief mass (gbbm) function m(A) for each hypothesis A in D^{Θ} (similar to DST). This gbbm function m(A) is defines a map m (.) : D^{Θ} → *[0, 1],* satisfying the conditions expressed in i) , ii) and iii)**.**

Bel (A) and Pls (A *)* representing the belief function and the plausibility function respectively are expressed in DSmT in the same way as for the DST. They are given as:

$$Bel\,(A) = \sum_{\substack{A_i \in D^{\Theta} \\ A_i \subseteq A}} m(A_i)\ , \tag{5}$$

$$Pls\,(\,A\,) = \sum_{\substack{A_i \in D^{\Theta} \\ A_i \cap A \neq \phi}} m\,(A_i) \tag{6}$$

Generally the Classical DSm (DSmC) rule for source combination is used by DSmT for combining the gbbms assigned to the information sources, while assuming a free DSm model. So, the DSmT rule for combining overlapping (conflicting), uncertain and imprecise sources of information is given by:

$$m\,(A\,) = \sum_{\substack{A_1, A_2, \ldots. A_d \in D^{\Theta} \\ A_1 \cap A_2 \cap \ldots \cap A_d \neq A}} \prod_{i=1}^{d} m_i\,(A_i) \tag{7}$$

Now, for combining d number of independent sources, the hybrid DSm (DSmH) rule of combination is applied. DsmH is again a generalized version of DSmC and is given by:

$$m(A) = \phi(A).\,[S_1(A) + S_2(A) + S_3(A)]\ , \tag{8}$$

where

$$S_1(A) = \sum_{\substack{A_1, A_2, \ldots. A_d \in D^{\Theta} \\ A_1 \cap A_2 \cap \ldots \cap A_d \neq A}} \prod_{i=1}^{d} m_i\,(A_i)\ ,$$

$$S_2(A) = \sum_{\substack{A_1, A_2, \ldots. A_d \in \phi \\ A_1 \cap A_2 \cap \ldots \cap A_d \neq A}} \prod_{i=1}^{d} m_i\,(A_i)$$

$$S_3(A) = \sum_{\substack{A_1, A_2, \ldots. A_d \in D^{\Theta} \\ A_1 \cup A_2 \cup \ldots \cup A_d = A \\ A_1 \cap A_2 \cap \ldots \cap A_d \in \phi}} \prod_{i=1}^{d} m_i\,(A_i)$$

2.3 DSmT Based Tracking

As mentioned in section 2.1, main challenge that affects the performance of particle filtering for tracking application is to find an efficient proposal distribution. So, now we will discuss the fusion of audio and visual cue using DSmT for designing the proposal distribution function for calculating the new importance weights of the particle set ($S_n^{(i)}$).

This work essentially aims for multimodal multi-target tracking. So, let the number of targets and the number of cues be τ and c. $\{\theta_j\}, j = 1, .., j$ be the tracks associated with each target up to time step n–1. At the next time step n, as an image and sound sample frame is obtained from the video and audio sequence, a number of measurements are obtained from all the c cues for each target candidate. Thus, we need to combine these measurements for estimating the best track for each target candidate. It is important to note that a target candidate here refers to a particle sample. As discussed in the previous section, the hyper-power set D^{Θ} represents the set of the hypotheses for which the different cues have provided some confidence levels. The possible hypotheses can correspond to: (1) individual tracks θ_j , (2) union of tracks $\theta_P \cup \theta_S$ (representing ignorance), (3) intersection of tracks $\theta_P \cap \theta_S$ (representing conflict) or (4) any tracks combination obtain by $\cup$ and $\cap$ operators. Now, if $m_{n,l}^{i}(A)$ gives the confidence level with which cue l associates particle i to hypothesis A at time n , then a single map function $m_n^i(.)$ providing confidence for particle i at time n can be expressed using the DSmT combinational rule (eq. 7) as follows:

$$m_n^i(A) = m_{n,1}^i(A) \oplus m_{n,2}^i(A) \oplus \ldots . \oplus m_{n,l}^i(A) \qquad (9)$$

As all the target are associated with individual tracks, only single hypotheses (tracks) are considered for decision making using the notions of the belief or plausibility functions :

$$Bel_n^i\,(\theta_j) = \sum_{\substack{A_i \in D^{\Theta} \\ A_i \subseteq A}} m_n^i(A)\ ,$$

$$Pls_n^i\,(\theta_j) = \sum_{\substack{A_i \in D^{\Theta} \\ A_i \cap A \neq \phi}} m_n^i(A) \qquad (10)$$

The confidence levels in (10) quantify the weight of the candidate (or particle) as a sample of the state posterior distribution to $p(x_n/y_{1:n})$ that is,

$$w_{n,j}^{(i)} = Bel_n^i\,(\theta_j)\ or\ w_{n,j}^{(i)} = Pls_n^i\,(\theta_j) \qquad (11)$$

The DSmT based particle filtering algorithm employed in this paper for multimodal multi-target tracking can be summarized as below:

Step 1: Initialize

– Set n = 1, N=100, j=1

– Generate N samples $S_{n,j} = \{ x_{n,j}^{(i)},\ w_{n,j}^{(i)} \}$, i=1,…, N_P for each target j independently, with $w_{n,j}^{(i)} = 1/N_P$.

Step 2: Propagate

– $x_{n,j}^{(i)} = H.x_{n-1,j}^{(i)} + W_{n-1,j}^{(i)}$

where H is a square matrix indicating the deterministic component of the target's motion model and W_n is a random component of the target's motion model.

Step 3: Observe (for each particle i)

– Compute $m_{n,l}^{(i)}(A)$ for A ∈ D^{Θ} from sound model. $(l = 1)$.

If a new pitch frequency found, the set $\mathrm{j} = \mathrm{j} + 1$ & initialize a new track for it.

– Compute $m_{n,l}^{(i)}(A)$ for A ∈ D^{Θ} from motion model $(l = 2)$

(initialize the KLT tracker for new target using its estimated position from sound model)

– Compute $m_n^{(i)}(A)$ according to (7)

– Calculate the particle weight as in (11) :

$$w_{n,j}^{(i)} = Bel_n^i\,(\theta_j)\ or\ w_{n,j}^{(i)} = Pls_n^i\,(\theta_j)*$$

– Normalize the weights.

Step 4: Estimate

–Each Target j is given by $E[S_{n,j}] = \sum_{i=1}^{N_P} x_{n,j}^{(i)} .\ w_{n,j}^{(i)}$.

Step 5: Resample (for each target)

– Generate $S_{n,j} = \left\{ x_{n,j}^{(i)} .\ w_{n,j}^{(i)} \right\}$,, n $= 1, .., N_P$ by resampling N_P times from $S_{n^*,j}$

where $p(x_{n,j}^{(i)} = x_{n^*,j}^{(k)}) = w_{n^*,j}^{(k)}$.

Step 6: Increment

– set $n = n + 1$, then go to step 2.

3. SYSTEM OVERVIEW

Now we will provide the details of all the components of our tracking algorithm based on dominant motion and sound. First, the system setup is presented along with the object model, and then we proceed to discuss the motion & sound cues and their impact on the tracking algorithm in more detail thereby summarizing the proposed tracking algorithm.

3.1 Audio-Visual System Setup

The setup consists only of a single stationary camera and pair of spatially separated microphones (kept at a fixed distance part from each other). The placement of camera and microphone pair is such that the line that connects the microphone pair is orthogonal to the optical axis of camera and also goes through the optical center of the stationary camera. The simplicity of the setup can be seen from the fact that performance of this system depends only a small number of calibration parameters, namely: the spatial separation 'd' between the microphone pair, the optical focal length 'f' of the camera used, the actual width of the image plane captured by the camera, denoted by '$\widehat{W}$' (expressed in meters), and the width of the digital image captured by camera in in pixels denoted by 'W'. All these parameters can either by measured manually (d & $\widehat{W}$) while capturing the dataset or can be easily obtained (f & W) .

As the inaccuracies in setup and calibration parameters are probabilistic in nature, our tracking algorithm is robust to these errors by accommodating them in the likelihood models for motion and sound by explicitly modeling the measurement uncertainty.

The objective of this work is to successfully and accurately track multiple objects or regions of interest in the video sequence captured by the still camera. The data available to us here is the raw measurements of camera & microphone in terms of the image frames and audio samples.

3.2 Multi-target Multi-modal Tracking

As described previously, the DSmT based particle filter is now applied for multi-target tracking using dominant motion and sound. Here the complete tracking algorithm is concluded by summarizing the framework of our multimodal multi-target tracking algorithm. Although the proposed framework is capable of tracking any number of targets, for simplicity, we are considering the case for two individual objects (A and B).Now our multi modal fusion problem can be characterized by the frame of discernment:

$$\Theta = \{\theta_1, \theta_2, \overline{\theta_1 \cup \theta_2}\}, \tag{12}$$

where the hypothesis θ_1 indicates target 1, $\theta_2$denotes target 2 whereas $\overline{\theta_1 \cup \theta_2}$ is the false alarm hypothesis, originating from the background clutter (as it refers to the rest of scene, which tends to change during). Also, in this model $\theta_1 \cap \theta_2 \neq \emptyset$ as it refers to the possible occlusion and $\theta_j \cap \overline{\theta_1 \cup \theta_2} \neq \emptyset$ for j=1, 2.

3.3 Sound Model

Since the actual tracking is performed in the video sequence, the discrete time index *n* corresponds to the video frame number captured by the still camera. As opposed to the video sequence which is naturally discretized, the audio samples arrive in a continuous manner, and there is no concept of natural audio frames. But for the purposes of the tracking algorithm, however, the n^{th} audio frame is defined as a window of N_s audio samples centred around the sample corresponding to the n^{th} video frame. If T_{video} and T_{audio} denote the sampling period for video frames and audio samples, respectively, then the centre of the audio frame corresponding to the n^{th} video frame can be computed as:

$$n_s = [(n-1)\,T_{video}\ /T_{audio} + 1] \tag{13}$$

where [.] denotes the rounding operation. In the audio frame N_s the number of samples taken is normally such that the duration of the audio frame is roughly 50ms. Now the next step in developing sound model is to extract the Time Delay of Arrival (TDOA) features from the audio data. Then a likelihood model is derived for the TDOA measurements. After that, finally an efficient TDOA based proposal is developed for the Kalman filter, based on an inversion of the likelihood model. This proposed method is especially useful for initialization and recovering of tracks when they are lost during brief periods of partial or full occlusion.

3.3.1 Frequency Analysis

The main requirement for performing frequency analysis is that this approach is based on multiple target tracking. Therefore in order to identify more than one object and relate it to a particular track, frequency analysis is done.

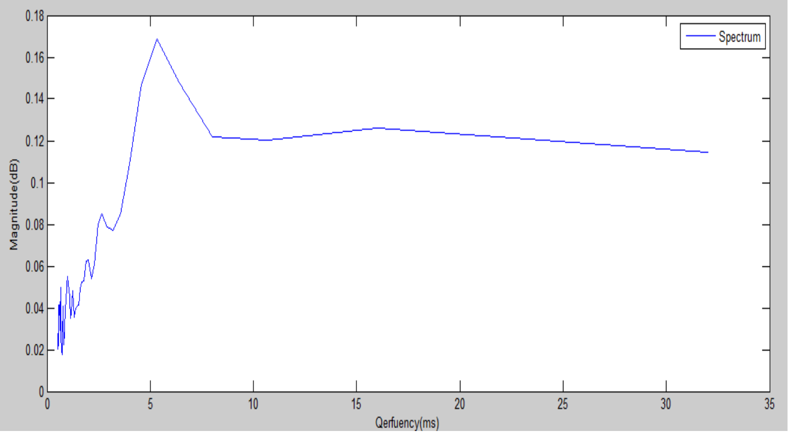


Figure 2 Cepstrum analysis of a window

This part first starts with cepstral analysis of speech frames obtained in the above section. The basic aim is to identify some sound parameters through which individual objects can be distinguished. Here we are using pitch as one such parameter.

Pitch (frequency f_0) can be defined in simple terms as the base of the speech signal. It is also the frequency of oscillation of vocal cords present in our vocal tract. This is a characteristic of the vocal tract and has different values for different individuals in different conditions. The pitch however is found to vary within a small range for the same person while producing different sounds. Children have the highest pitch frequencies in the range of 300 Hz followed by adult females who have frequencies around 225 Hz which are still more than adult males lying in the range of 120 Hz. This pitch however is variable within individuals as well and varies with different sounds. The pitch of a speech window maybe determined by finding the location of the first autocorrelation peak of the speech signal, alternately this may be found out by the use of cepstral analysis of speech.

For finding the pitch value, cepstral analysis is done on the voiced part of the signal (as unvoiced part is only a type of noise signal). So, if a signal frame window is found to be voiced then its cepstrum is formed. From this cepstrum, pitch period corresponding to the cepstral peak in the signal, post 4 milliseconds, is calculated. The cepstrum is calculated for each frame. Mathematically, cepstrum is given as:

$$c(t) = \mathrm{ifft}(\log(\mathrm{abs}(\mathrm{fft}(x_s)))) \qquad (14)$$

where x_s is the sound frame window (obtained after windowing). Magnitude of c(t) is plotted with qerfuency (inverse of frequency). Cepstrum before 4ms is neglected (this part is used for calculation of formant frequencies). Post 4ms part is scanned and peak value is selected. The index corresponding to this peak gives the pitch period. This is demonstrated in fig. 2 (Peak gives the pitch period).

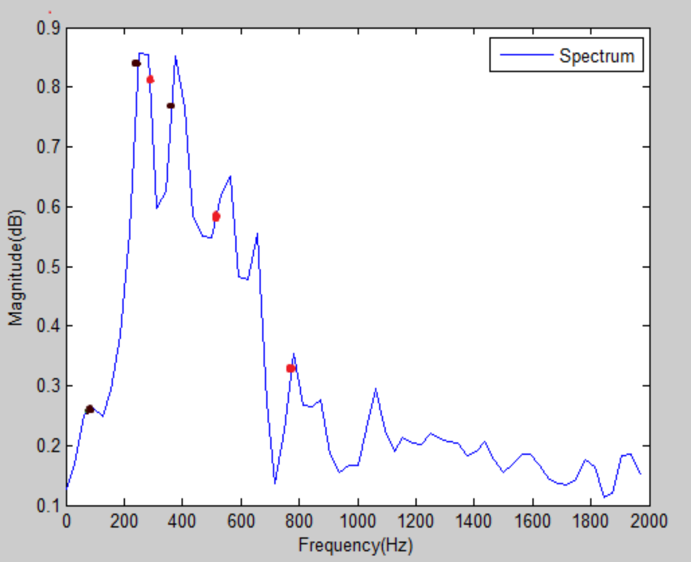


Figure 3 Analysis of an overlapping sample, containing two peaks.

Therefore if a speech signal corresponds to a single source, its pitch contour will have a single visible peak (as shown in above figure). But if it contains voices from two or more sources; it will contain multiple peaks equal to the number of sources present. In order to separate samples from those two or more, narrow band filtering is done in the cepstral domain. For each peak, in the cepstrum, an active narrow band filter (in time domain) is used with central frequency f_c equal to that peak frequency with cut-off equal to $f_c/2$. Then the segmented samples are processed separately using TDOA.

But as the main of frequency estimation was identification of the sound sources, therefore, as each audio frame is processed, first the pitch present in that frame is calculated (before performing TDOA). If a single pitch peak is found, it is identified as an individual object and a track θ_j is associated with. Then TDOA is performed for tracking that object with track θ_j. If two or more peaks are present (fig. 3), then segmentation of overlapping sound sample is done to separate the samples corresponding to different pitch values. Although fig.3 shows the cepstrum plot but it has been plot against frequency (inverse of qerfuency) to illustrate the presence to two peak frequencies.

Then those pitch values are checked. If the values are already identified in any previous frame, then their tracks from TDOA are related to the previous tracks (θ_j)

corresponding to those pitch values (with some allowance for variation), else new tracks (corresponding to new objects in frame) are initialized.

Therefore using pitch analysis, individual objects can be distinguished and tracked accurately.

3.3.2 TDOA Measurement

For object tracking application using sound samples, one of the commonly used approaches is calculate the TDOA which occurs between the two audio signals received by the two different microphones kept a fixed distance apart from each other. This provides an indication of the location of object in the image frame.

The strategy that we have used in this research work for measuring TDOA [30], [31] is the maximization of the GCCF (Generalized Cross-Correlation Function). [30], [31] provides the complete details of GCCF.

As it suggests, GCCF is a correlation function calculated between the audio samples received from each other by the two microphones (kept spatially apart). Then the peaks of this GCCF signal are found, as positions of these peaks indicate an estimate of the TDOAs of the objects (sound sources). Background noise can also infect the sound samples, thereby leading to some false peaks in the GCCF. Therefore, these false audio peaks are filtered out by, eliminating all the peaks below a predefined threshold as candidates for the true TDOA values. Therefore, sound vector for any audio sample is given by Y_s = (D_1, $D_2, \ldots., D_{N_s}$), where, $D_i \in [-D_{max}, D_{max}]$, ∀ i = 1, 2,…, N_s .

The threshold TDOA value for noise filtering that is D_{max} is calculated from the setup parameters: microphone separation d and speed of sound c (normally 342m/s). Thus the maximum TDOA that should be measured is defined as: $D_{max} = d/c$.

For handling the uncertainty due to the presence of multiple active sources of sounds (multiple-targets) for the true TDOA value, we have used a multi-hypothesis likelihood model, which is described in the next section.

3.3.3 Likelihood Model

To obtain the likelihood of presence of an active sound source(s) at a position x from the true TDOA values for sound model, a likelihood model is generated. The modelling starts with processing the TDOA measurements (obtained from filtered peaks of GCCF depending only on the (horizontal) position of the source in the image (video frame) that is the x position of source. So, the object position x in the image frame, can be estimated from a deterministic hypothesis from TDOA, which is computed as follows:

$$D_x = f(x) = f_3 .\, f_2 .\, f_1 (x); \qquad (15)$$

with $\hat{x} = f_1 (x) = \widehat{W} (x/W - 0.5)$

$\theta = f_2 (\hat{x}) = arctan (f/\hat{x})$

$D_x = f_3 (\theta) = D_{max}\ Cos\, \theta$ (16)

These functions are described now. Function f_1 is a linear mapping relating the x (horizontal position) of the source in the image frame to the corresponding $\hat{x}$ position (output of f_1) in the actual image plane captured by the system's still camera. As mentioned earlier, the actual width of the image plane '$\widehat{W}$' is expressed in meters, and the width of the digital image W captured by camera in in pixels. Function f_2 is a mapping relating the $\hat{x}$ position (output of f_1) to the sound source (it terms of angular position) using camera focal length f. Finally, the function f_3 relates the sound source location to the hypothesized TDOA using the Fraunhoffer approximation. Now the likelihood of presence of a source at x can written as $p(y^s/x) = p(y^s/D_x)$.

Of the entire TDOA measurements at most only one peak is related to the true sound sources, whereas all the remaining measurements are due to noise. For separating these two cases, a classification label l_i is used such that $\mathrm{l_i} = \mathrm{T}$ if D_i is comes from the true audio source and $l_i = C$ if D_i is related with clutter. The conditional likelihood that is, for a TDOA measurement associated with a true source is given by:

$$p(D_i /D_x , l_i = T) = \mathcal{N}(D_i /D_x, \sigma_D)\, \mathbb{I}_D(D_i) \qquad (17)$$

where $\mathbb{I}_D$ (.) representing the indicator function on the set D, assuming that an additive Gaussian noise is corrupting the true TDOA with noise of deviation σ_D . For measurements associated with clutter, the likelihood is calculated as

$$p(D_i /c_i = C = \mathcal{U}_D(D_i) \qquad (18)$$

Thus, for N_S TDOA measurements, (N_S + 1) hypotheses are possible. In other words these hypotheses can be written as,

$$H_0 = \{l_i = C : i = 1,2, \dots N_S \}$$

$$H_i \;\; = \{l_i = T, l_j = C \;\; : j = 1,2, \dots N_S \, , j \neq i\} \qquad (19)$$

with $i = 1,2, \dots N_S$. Also, note the for each sound sample, we already have a track $\theta_j$associated with it. So, now the likelihoods can be expressed as:

$$p(y_s/H_0) = \mathcal{U}_{D\_N_S}(y_s).p(y_s/D_x, H_i \, , \theta_j \,)$$

$$= l_x \, \mathcal{N}(D_i \, /D_x, \sigma_D) \; \mathbb{I}_D(D_i \,) \; \mathcal{U}_{D_{N_S}-1}(y_s - i) \qquad (20)$$

where $\mathcal{U}$ is indicating a uniform distribution within the allowed interval and independent of the TDOA measurement from the true sound source. However, the correct hypothesis and therefore the final likelihood is expressed as the summation of all the $(N_S + 1)$ possible hypotheses i.e.

$$p(y_s/D_x, \theta_j \,) = \sum_{i=0}^{N_S} p(H_i \, /D_x) p(y_s/D_x, H_i \, , \theta_j \,) \qquad (21)$$

where $p(H_i \, /D_x)$ gives the i^{th} hypothesis's prior probability. If a situation occurs where no source is present (i.e. no TDOA measurement) the likelihood is then set to $\mathrm{p(y_S/D_x)} \propto 1$.

3.3.4 TDOA based Proposal Distribution

Now, it is possible to develop an efficient proposal distribution function for the particle filter using the sound localization cues. This can be done by designing a proposal distribution for the object position x \ incorporating the TDOA measurements while they are available. The inverse of the mapping in (15) can be obtained easily. Then passing the TDOA measurements through the resulting inverse mapping f^{-1} yields a plausible x position for the object. Therefore object position x can be proposed using the information in the TDOA measurements *as:*

$$\tilde{x} = q^s \left(\frac{x_n}{x_{n-1}}, {y^s}_n, \theta_j \right) = \beta_{RW} \mathcal{N} \left(\frac{x_n}{x_{n-1}}, \sigma_x \right) + \frac{1 - \beta_{RW}}{N_S} \left| \frac{d_f(x_n)}{d(x_n)} \right| \sum_{i=0}^{N_S} \mathcal{N}(f(x_n) | \, D_{i,n,} \, \sigma_D) \qquad (22)$$

The first part with β_{RW} in the above equation represents the Gaussian component used in the state evolution model for the *x* component. In the second component of the equation a mixture is represented including the TDOA measurement values, obtained from inversion of the non-clutter part of the likelihood model. The derivative of the mapping f is obtained by the chain rule. But it should be noted that the TDOA process is only bale to estimate the x positions of the sources only.

So, at each time step *n*, for each particle *i*, located at some ($x_n^{(i)}$, $y_n^{(i)}$), the x position ($\tilde{x}$) returned by TDOA along with the associated track θ_j is used to calculate the likelihood of that the particle belongs to track target *j=1,2.* as a Gaussian pdf as:

$$p_{t,j}^{(i)} = p\left(y_s/D_x, \theta_j\right) . \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(x_n^{(i)} - \tilde{x})^2}{2\sigma_s{}^2}} , \mathrm{j} = 1{,}2 \tag{23}$$

In the case where only one sound source is active, we get only likelihood $p\left(y_s/D_x, \theta_j\right)$ for only j=1 or 2 from sound model. In that case, the likelihood for the missing track is set to a very small value (~0) to indicate that sound model has not actually detected that target. And likelihood that the particle belongs to false alarm ($\overline{\theta_1 \cup \theta_2}$) directly becomes:

$$p_{t,FA}^{(i)} = p(y_s/H_0) \tag{24}$$

Now, the mass functions of particle *i* depending on its x location are defined as:

$$m_{n,1}^{(i)}(\theta_j) = \frac{p_{t,j}^{(i)}}{p_{t,1}^{(i)} + p_{t,2}^{(i)} + p_{t,FA}^{(i)}} , \quad \mathrm{j} = 1{,}2 \tag{25}$$

$$m_{n,1}^{(i)}(\overline{\theta_1 \cup \theta_2}) = \frac{p_{t,FA}^{(i)}}{p_{t,1}^{(i)} + p_{t,2}^{(i)} + p_{t,FA}^{(i)}} \tag{26}$$

Also, note that

$$m_{n,1}^{(i)}(\emptyset) = 0$$

Here subscript 1 represents the cue 1 which is the sound cue.

3.4 Motion Model

The motion model used in this work starts with detecting some dominant [32] motion points in the video frames, which is a very challenging and important task. A system is developed for automatically identifying the dominant motion in terms of some points in a image frame/scene. Then the statistics of these points are used for accurately tracking individual targets in scenes when it is difficult to track them otherwise due to full or partial occlusions and clutter.

Our approach begins by initially tracking some low-level features using the optical flow methodology. But most of the point feature tracks are unreliable. So this problem can be eliminated by clustering them into dominant motion points using some distance measure.

Then using graph theory, a method of association is adopted, where some statistical measures of these motion points are calculated and are feed into a Bayesian network. The purpose of using Bayesian network is to predict these motion points even when they are occluded due to object overlapping. Then these predict point locations help in tracking the object accurately.

3.4.1 Dominant Motion

As proposed in [32] for detecting dominant motion points, first task is to identify some low-level point features by manually selecting the region of interests in the initial frame using the famous Shi–Tomasi–Kanade detector [32] , a fast algorithm for finding corner points. Then the selected dominant motion points in the form of low-level features are tracked in each frame using optical flow by the Kanade–Lucas–Tomasi algorithm (KLT) [33] and are associated with a particular object in the frame. To decrease computational time and load, new features points are detected for each object only in every fifth frame. Then these feature points are clustered into dominant motion points for each object present in the frame using eign distance measure. That is, if new point is spatially too close to an already existing point, then it is discarded, else it is retained. Then all the feature points are again tracked over time. Fig. 4 shows some feature points (representing dominant motion) identified in an example frame.

3.4.2 Method of Association

These detected motion points are associated to each other using some statistical measure like Mahalanobis distance. For each frame, the feature points detected for each object identified in that frame are averaged together to find an approximate centroid location of that object. The variance and standard deviation between the points are calculated. Next, at each time step , these measures are calculated and used in Bayesian inference and updating of the target’s location. Fig. 4 shows the detected motion points connected together along with a centroid point. If no occlusion occurs in the next time step, the procedure we just explained is followed. But when occlusion occurs, most of the dominant get hidden. But, if any one of the dominant point is detected, then Bayesian prediction is employed, which estimates the posterior probability of target’s location using the prior probability and a "likelihood function" derived from a statistical model for the observed statistical data. Bayesian inference computes the posterior probability according to Bayes' theorem.

3.4.3 Likelihood Model

Now we propose to embed the information obtained from the statistical measures of dominant points (after applying method of association) in a probability likelihood model in a manner similar to the followed in sound model for the sound measurements.

Let at time *n-1*, $X_{n-1,i} = (x_{n-1,i}\,, y_{n-1,i})$ denote the position of the centroid providing the estimate of location of track associated with candidate j (track θ_j) and $d_{n,j}$ be the observed data vector containing statistical measures indicating the distribution of dominant motion points associate with track θ_j at next time step n. That means $X{\sim}p(X/d)$. If α, is the hyper parameter for the parameter d, then the posterior predictive distribution likelihood of the of a new data point j, marginalized over the posterior is given by:

$$p(X_{n,j}/\alpha) = \int p(X_{n-1,j}/d_{n-1,j})\, p(d_{n-1,j}/\alpha)\, d(d_{n-1,j}) \quad (27)$$

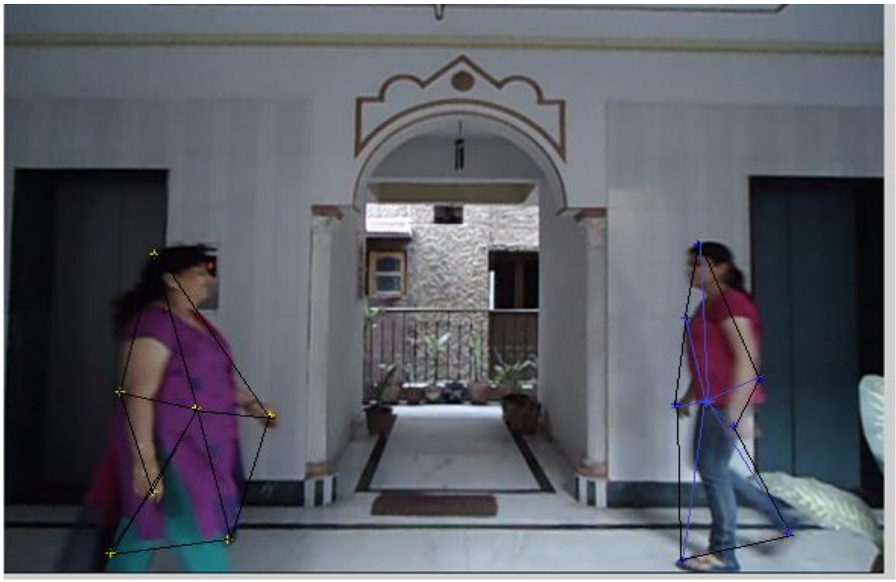


Figure 4 An example frame showing the dominant motion points detected in that frame.

This new data point $X_{n,j} = (x_{n,j}\,, y_{n-1,j})$ gives the estimate of target’s location at time n associated with track θ_j,. Now the likelihood that a particle i, located at some ($x_n^{(i)}$, $y_n^{(i)}$), belongs to track target j=1,2 is defined as a Gaussian pdf as:

$$p_{t,j}^{(i)} = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x_n^{(i)} - x_{n,j})^2 + (y_n^{(i)} - y_{n,j})^2}{2\sigma_m{}^2}} \tag{28}$$

Here σ_m is a bandwith parameter. Siilarly, the likelihood that the particle does not belongs to target 1 & 2 or θ_1 and θ_2, implies that it belongs to false alarm $\overline{(\theta_1 \cup \theta_2})$ hypothesis. The measure of likelihood that the particle belongs to false alarm hypothesis is given by:

$$p_{t,FA}^{(i)} = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x_n^{(i)})^2 + (y_n^{(i)} - y_{n,j})^2}{2\sigma_m{}^2}} \tag{29}$$

where d_m is the radius of the circles centered on the midpoint of line joining the centroid of target 1 & 2 and contains all the particles used in tracking at previous time step $n-1$. And $d_{n,j}$ is the separation between the particle and the midpoint of the estimated positions of target 1& 2 :

$$d_{n,j}{}^2 = (x_n^{(i)} - (x_{n,1} + x_{n,2})/2)^2 + (y_n^{(i)} - (y_{n,1} + y_{n,2})/2)^2 \tag{30}$$

Now, again the mass functions of particle i depending on its location are defined as:

$$m_{n,2}^{(i)}(\theta_j) = \frac{p_{t,j}^{(i)}}{p_{t,1}^{(i)} + p_{t,2}^{(i)} + p_{t,FA}^{(i)}} \text{ , j=1,} \tag{31}$$

and

$$m_{n,2}^{(i)}(\overline{\theta_1 \cup \theta_2}) = \frac{p_{t,FA}^{(i)}}{p_{t,1}^{(i)} + p_{t,2}^{(i)} + p_{t,FA}^{(i)}} \tag{32}$$

Also

$$m_{n,2}^{(i)}(\emptyset) = 0 \tag{33}$$

Here subscript 2 represents the second motion cue.

3.5 Tracking Architecture

Here have already presented the complete tracking algorithm by summarizing the framework of our multimodal multi-target tracking algorithm in this section. What all is left is the final cue combination for defining the mass functions required for allocating importance to the particles.

Now at time step n, the final mass function of particle i providing the confidence level that it belongs to track $\theta_j$using DSmT is given by:

$$m_n^{(i)}(\theta_1\,) = \; m_{n,1}^{(i)}(\;\theta_1)\;.\;m_{n,2}^{(i)}(\theta_1\,) \qquad (34)$$

$$m_n^{(i)}(\theta_2\,) = \; m_{n,1}^{(i)}(\theta_2)\;.\;m_{n,2}^{(i)}(\theta_2\,) \qquad (35)$$

$$m_n^{(i)}(\theta_1 \cap \theta_2) = \; m_{n,1}^{(i)}(\theta_1\,).m_{n,2}^{(i)}(\theta_2\,) + m_{n,1}^{(i)}(\theta_2)\;.m_{n,2}^{(i)}(\theta_1) \qquad (36)$$

$$m_n^{(i)}(\overline{\;\theta_1 \cup \theta_2}) = \; m_{n,1}^{(i)}(\overline{\;\theta_1 \cup \theta_2})\;.m_{n,2}^{(i)}(\overline{\;\theta_1 \cup \theta_2}) \qquad (37)$$

Equation (34) & (35) expresses the confidence level values with which both the cues associate particle i at time step n to target 1 and 2 respectively. Equation (36) represents the conflict between the cues for association of the particle to target 1 or 2. Equation (37) gives the confidence level values with which both the cues agree that the particle belong to false alarm.

For the frame of discernment (Θ = {θ_1, θ_2, $\overline{\theta_1 \cup \theta_2}$}) defined in this case for tracking two targets , the plausibility and belief functions (expressed by equation 10) results in identical values. They are now used to calculate the weight of particles as:

$$w_{n,j}^{(i)} = \; Bel_n^i\;(\theta_j) = \; Pls_n^i\;(\theta_j) = m_{n,}^{(i)}\big(\theta_j\,\big) + m_n^{(i)}(\theta_1 \cap \theta_2) \quad ,j\,=\,1{,}2 \qquad (38)$$

The estimate of target $j\,=\,1{,}2$ is calculated as defined in (3):

$$E\big[S_{n,j}\big] = \; \textstyle\sum_{i=1}^{N_P}\,w_{n,j}^{(i)}\;.\,x_{n,j}^{(i)} \qquad (39)$$

4. RESULTS

In this chapter, the actual tracking results of our approach developed in this dissertation to track multiple (two in this experiment) targets (persons) are the presented. The video sequence used for the experiment is captured by a single camera (still) and a stereo microphone pair (kept at a fixed distance part from each other). So, In this section the performance of our tracking algorithm will be demonstrated. Also our results will be compared with a general Kalman Filter tracking method where no motion and sound cues were used. The aim here is to show the uniqueness and accuracy of our approach.

The values of the fixed parameters used in the likelihood models are presented in Table I for the motion and sound cues.

TABLE I

Model Parameters for The Tracking Experiments.

Parameters	**Symbol used**	**Value**
Speed of sound	c	342 m/s
Microphone separation	d	1.5 m
Optical Focal length of camera	f	0.5 m
Width of the actual image plane	$\widehat{W}$	3 m
Width of the digital image	W	640 pixels
Sampling rate for video frames	T_{video}	1/30 SEC
Sampling rate for audio sample	T_{audio}	16 KHz
Additive Gaussian noise deviation	σ_D	0.0001

Before tracking, both motion and sound frames were preprocessed for synchronization and cleaned to avoid misalignment problems. Also in motion model, as the purpose of this work is tracking and simply not just detection, all the regions of interest were selected manually for the first frame in dominant motion based cue. For the sake of simplicity and clarity, this experimental video contains only two moving targets.. Although our method is able track any number of targets successfully. A set of N_P = 30 particles are randomly generated around each target. To prove the efficiency of our multimodal tracker, we will first present the behavior of the tracker when using each of the cues in isolation, and after that, we will show how the shortcomings of such single modality trackers is eliminated by fusion of information from multiple cues.

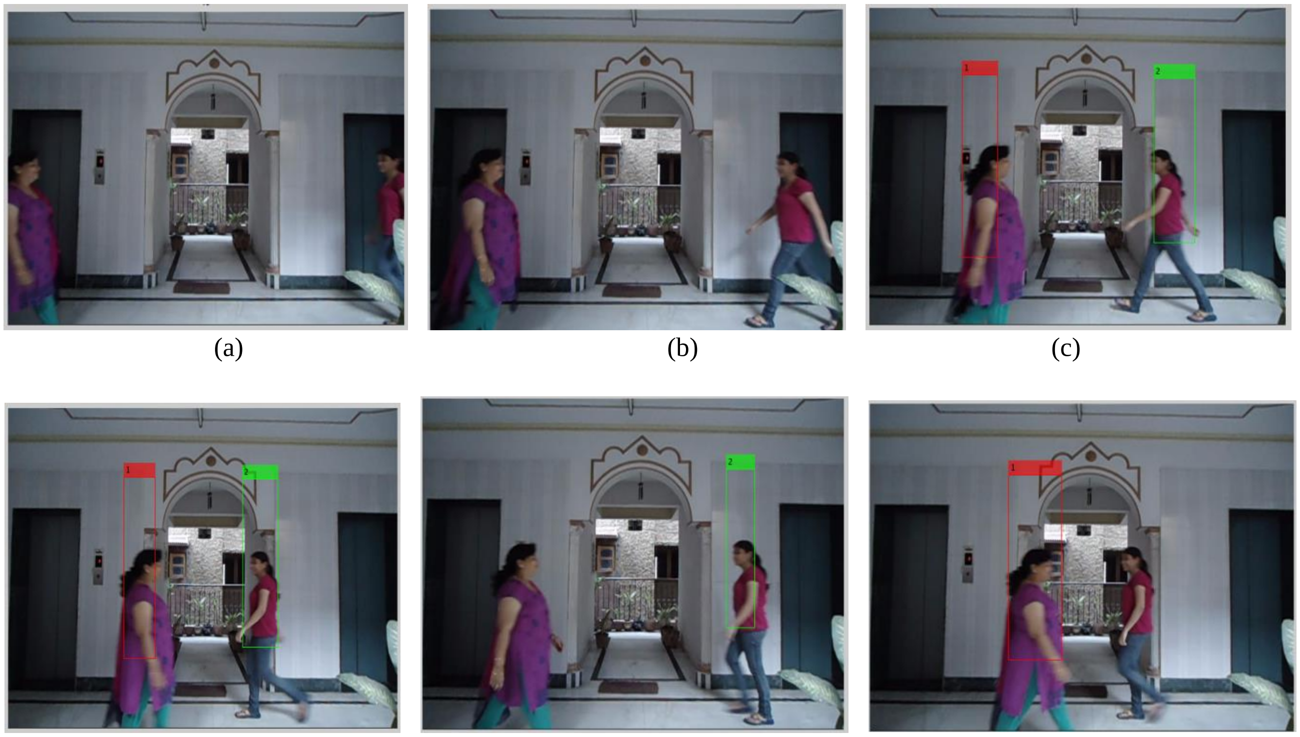
4.1 Single Modality Tracker

4.1.1 Sound Tracker

In this section the behavior of the sound only tracker is presented. But as mentioned earlier, that sound cue estimates the horizontal position only, so here only the horizontal position of a active sound source(s) in the image is tracked, based on the TDOA measurements obtained from the stereo microphone pair. Also, it will be evident that this cue is able to track (focus on) only the active sound sources present in the frame even if one or more silent targets are present.

We have presented three sequences, each featuring two targets in the video frames. In the first sequence the both the targets are silent. Therefore no tracking is performed as there is no active sound source. Fig.5 (a), (b) shows snapshots of the tracking result for this sequence. In the second sequence, both the targets generate sound samples simultaneously (although for a short duration only). Here the sound cue successfully tracks both the targets. Fig.5 (c), (d) shows snapshots of the tracking result for this sequence. In the third sequence, both the target are present, but only one is an active source of sound. So, here the target which is active, is tracked. Fig.5 (e), (f) shows snapshots of the tracking result for this sequence.

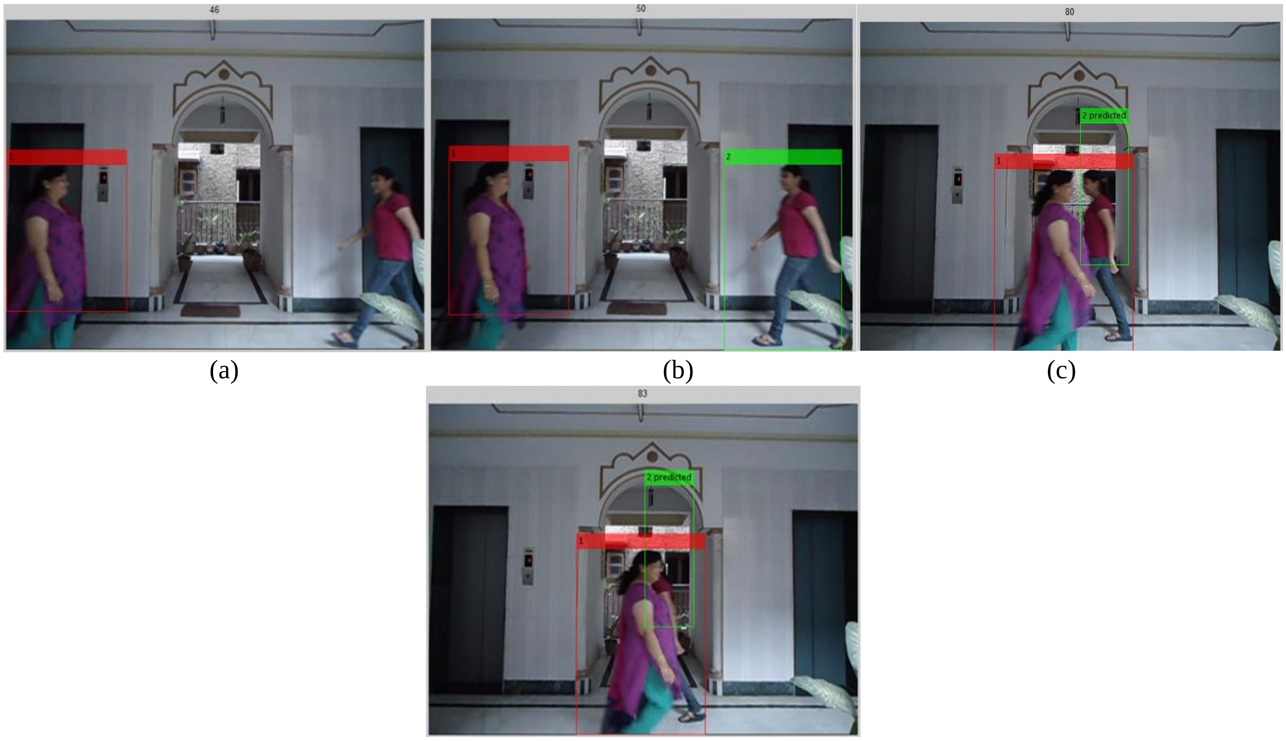
Since the sound cue lacks accuracy and persistence, either due to only horizontal detection or due to the absence of speech, the sound based tracker is unable to provide consistent tracking over the complete period of time. But we will see in section V-B how the fusion of sound with motion will solve all these problems**.**



(a) (b) (c)

(d) (e) (f)

Figure 5 Sound only tracker results. (a), (b) shows no tracking as both targets are silent. (c), (d) shows tracking of both the targets successfully as both targets generate sound samples simultaneously. (e), (f) both the target are present, but the only target which is an active source of sound is tracked.



(a) (b) (c)

(d)

Figure 6 Motion only tracker results. (a), (b) the second target is not tracked initially (in frame 46), but only after 4 frames (in frame 50). (c), (d) shows the results of tracking during occlusion.

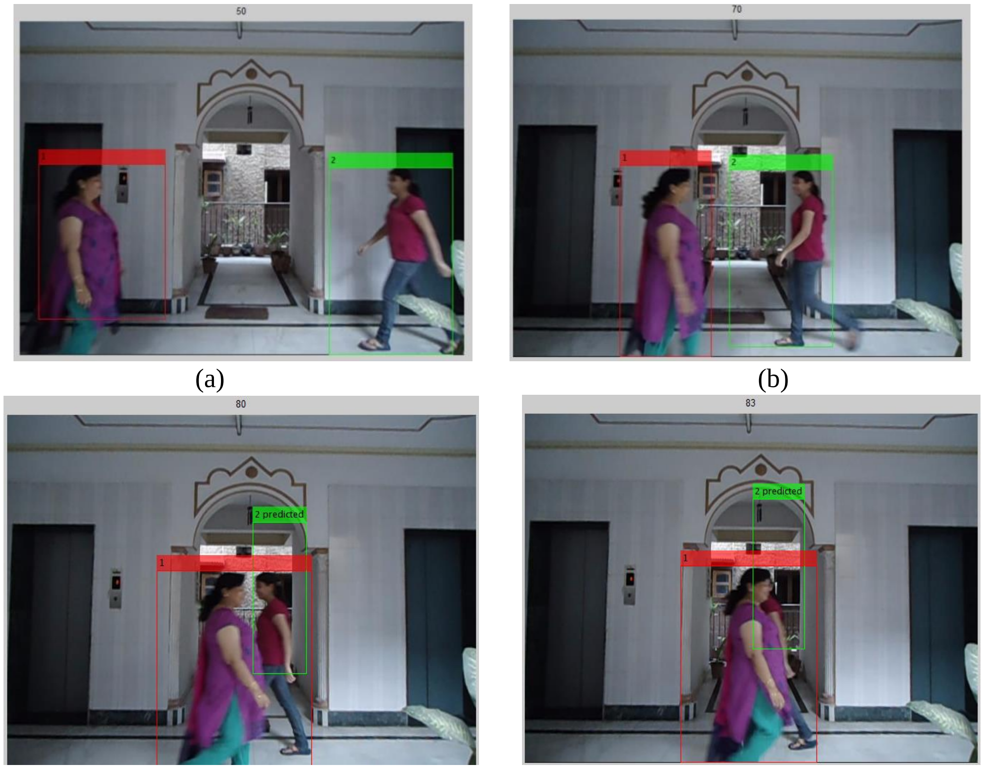
4.1.2 Motion Tracker

Motion of an object is most persistent feature as compared to speech. Only motion based tracker, (using dominant motion) is able to robustly track objects even during occlusion. Fig.6 shows snapshots of the tracking result for this motion tracker. But the KLT feature detector used in motion model requires, either manual selection of region of interests corresponding to a target in the first frame (done using the estimate of the target from sound model, if available) or it uses some initial frames for automatic detection of dominant motion points corresponding to a target. Therefore as the new target enters the scene, it is tracked, only some frames. This is exemplified by the sequence of tracking snapshots presented in Fig. 6 (a), (b) where the second target is not tracked initially, but only after 3-4 frames. Fig. 6 (c), (d) shows the results of tracking during occlusion.

4.2 Multiple Modality Tracking

Here by fusing the sound and motion cues the accuracy is greatly increased. As the main aim of this approach is to accurately track multiple targets, even in the presence of partial and full occlusion, therefore the tracked video results are shown in 4 parts.
Section/Phase 1 contains the pre-occlusion part of the video, phase 2 corresponds to the occluded part of the video, and phase 3 is the post-occlusion sequence. In phase1, the tracking is done on the initial section of the video where no occlusion was present is shown. That is, there was no overlapping between multiple regions of interest. The tracking results from this section of video are shown in figure 7(a), (b). Here both the targets are spatially separated and are successfully tracked. For visual clarity, the target on the LHS, denoted as target 1 is tracked with a red rectangular bounding box and the target on RHS, denoted as target 2 is tracked with the green rectangle.

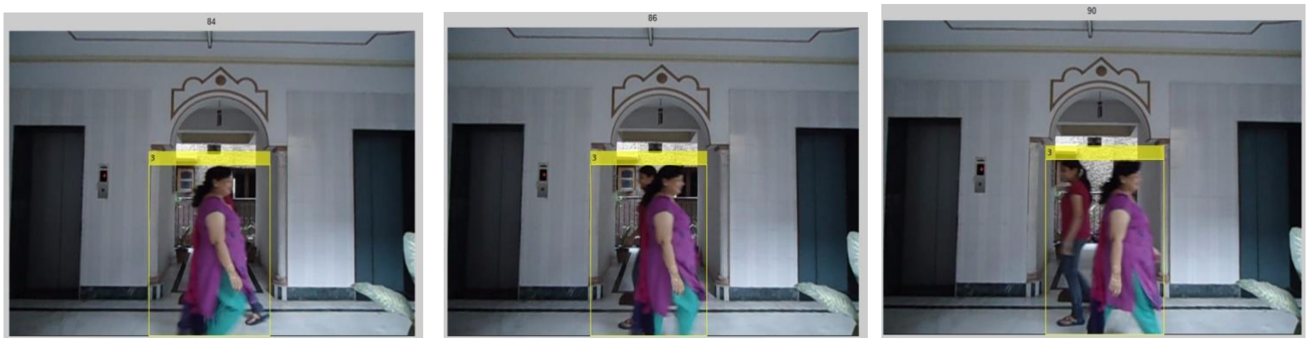
In the phase 2, the targets cross each other, occluding each other as they walk. Initially this sequence starts with partial occlusion, but as the video proceeds, full occlusion occurs and the again the amount of occlusion decreases. The tracking results from this phase are shown in figure 7(c), (d). They demonstrate that the approach presented in this work is successfully able to deal with the cluttered scenes, involving partial or total occlusion.



(a) (b)

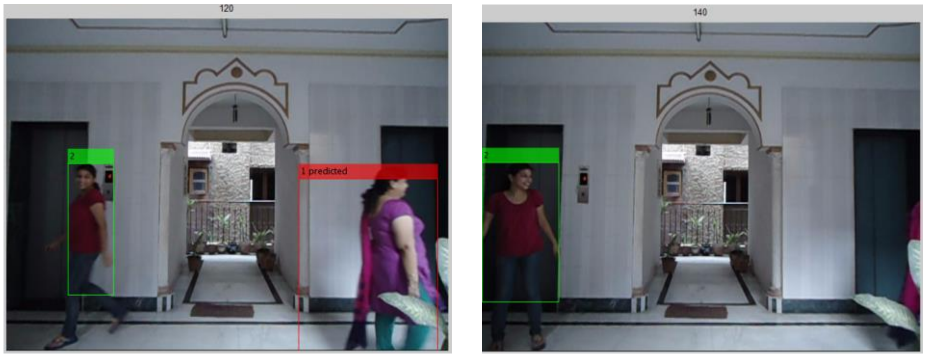
(c) (d)

Figure 7. Audio-Visual tracking results. (a), (b) Result from phase 1(pre-occlusion) where both the targets are spatially separated and are successfully tracked. (c), (d) Result from phase 2 (occlusion) where both the targets are spatially occluded.



(a) (b) (c)

Figure 8. Tracking results of a general Particle Filter. (a), (b) shows tracking during full occlusion between target 1& 2 only one combined target is detected, (c) shows even during partial occlusion only one target is detected instead of two.



(a) (b)

Figure 9. Tracking results from phase 3 (post occlusion). (a), (b) tracking where again target 1& 2 are spatially separate.

Although tracking in this occlusion phase spatial overlapping of the targets make tracking particularly challenging by infecting the cue measurements and leading to false identifications otherwise, but due to our dominant motion and sound model, our algorithm is able to overcome these problems while tracking.

In order to test the performance of our algorithm during this phase of occlusion, we have also presented the results (in Figure 7 of a simple particle filtering approach for multi target tracking without our motion and sound cues. As mentioned above that the spatial closeness and overlapping of the targets make tracking in this occlusion phase is challenging, corrupting the cues. The single location cue used in Particle filter gradually fails to separate target 1 and 2 and detects them as a single moving entity. From figure 8 (a), (b), it can be seen that during occlusion, this tracking algorithm fails to distinguish between multiple targets and only a single combined target is detected.

In phase 3, the tracking is done on the post occlusion section of the video where no occlusion was present. That is, there was no overlapping between multiple regions of interest. Fig. 9 shows the results of this sequence.

It is evident from the tracking results shown in this chapter that our proposed method is able to track multiple targets accurately in all the scenarios, especially in partial and fully occluded scenes.

5. CONCLUSION & FUTUREWORK

This chapter concludes the work presented in this thesis and also the future aspects of the proposed multimodal multi target tracking algorithm. Other researchers can extend the work upon this research work and particularly on this proposed tracking methodology. These all aspects are discussed here.

5.1 CONCLUSION

In this work we introduced a novel technique for multi target tracking using for data fusion within particle filtering. This particle filter based tracker combines sound and motion cues in a novel way using DSmT. A set of particles is used to track each target using both the cues. Sound and motion model evaluates a likelihood for every particle. Then the DSmT model assigns confidence level values for the membership of each particle, while considering the conflict between the cues, thus resulting in a better tracking performance.

The experimental results presented in section IV illustrate the accuracy and effectiveness of our algorithm in case of multiple targets.

These cues also allow detection and initialization of multiple tracks for the particle filter and aid the recovery of lock following periods of partial or complete occlusion. For multi-object tracking such event based proposals are essential for the detection of new objects when they appear in the scene.

5.2 FUTURE WORK

In future this work can be extended by performing a number of modifications to have better results. Let us discuss some points which can be implemented in future:

- More information can be embedded in our proposed algorithm by integrating other type of cues. For example, shape cues can be included for tracking a predefined class of object.

  Also this work can be deployed in some real-time platforms to handle the real time issues like traffic monitoring, anomaly detection and surveillance systems.

REFERENCE

[1] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing, 10(3):197–208, 2000.

[2] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non- Gaussian Bayesian state estimation. IEE Proceedings-F, 140(2):107–113, 1993.

[3] M. Isard and A. Blake. CONDENSATION–conditional density propagation for visual tracking. Int. J. Computer Vision, 29(1):5–28, 1998.

[4] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association, 93:1032–1044, 1998.

[5] F. Dellaert, W. Burgard, D. Fox, and S. Thrun. Using the Condensation algorithm for robust, vision-based mobile robot localization. In Proc. Conf. Comp. Vision Pattern Rec., pages II: 588–594, Fort Collins, CO, June 1999.

[6] K. Choo and D.J. Fleet. People tracking using hybrid Monte Carlo filtering. In Proc. Int. Conf. Computer Vision, pages II: 321–328, Vancouver, Canada, July 2001.

[7] A. Doucet, N. de Freitas, and N. Gordon, editors. Sequential Monte Carlo Methods in Practice. Springer-Verlag, New-York, 2001.

[8] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J. Computational and Graphical Stat., 5(1):1–25, 1996.

[9] E. Koller-Meier and F. Ade. Tracking multiple objects using the Condensation algorithm. Journal of Robotics and Autonomous Systems, 34(2-3):93–105, 2001.

[10] J. Konrad. Motion detection and estimation. In A. Bovik, editor, Handbook of Image and Video Processing, pages 207–225. Academic Press, 2000.

[11] V. Philomin, R. Duraiswami, and L.S. Davis. Quasi-random sampling for Condensation. In Proc. Europe Conf. Computer Vision, pages 134–149, Dublin, Ireland, June 2000.

[12] Y. Rui and Y. Chen. Better proposal distributions: Object tracking using unscented particle filter. In Proc. Conf. Comp. Vision Pattern Rec., pages II:786–794, Kauai, Hawaii, December 2001.

[13] H. Sidenbladh, M.J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In Proc. Europ. Conf. Computer Vision, pages I: 784–800, Copenhagen, Denmark, May 2002.

[14] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In Proc. Conf. Comp. Vision Pattern Rec., pages I:447–454, Kauai, Hawaii, December 2001.

[15] C. Sminchisescu and B. Triggs. Hyperdynamics importance sampling. In Proc. Europ. Conf. Computer Vision, pages I: 769–783, Copenhagen, Denmark, May 2002.

[16] M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In Proc. Europe Conf. Computer Vision, pages 893–908, 1998.

[17] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. In Int. Workshop on Computer Vision Systems, Vancouver, Canada, July 2001.

[18] Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Realtime tracking of the human body. IEEE Trans. Pattern Anal. Machine Intell., 19(7):780–785, 1997.

[19] Y. Wu and T.S. Huang. A co-inference approach to robust visual tracking. In Proc. Int. Conf. Computer Vision, pages II: 26–33, Vancouver, Canada, July 2001.

[20] Frank, O., Nieto, J., Guivant, J., Scheding, S.: Multiple target tracking using sequential Monte Carlo methods and statistical data association. In: Proc. of the IEEE Int. C. on Intell. Rob. and Syst., vol. 3, pp. 2718–2723 (2003)

[21] Schultz, D., Burgard, W., Fox, D., Cremers, A.B.: Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In: Proc. of IEEE Int. C. on Rob. And Auto., vol. 1, pp. 1665–1670 (2001)

[22] Chen, C., Lin, X., Shi, Y.: Moving object tracking under varying illumination conditions. Patern Recogn. Lett. 27(14), 1632–1643 (2006)

[23] Ozyilidiz, E., Krahnstover, N., Sharma, R.: Adaptive texture and color segmentation for tracking moving objects. Patern Recogn. 35(10), 2013–2029 (2002).

[24] McCane, B., Galvin, B., Novins, K.: Algorithmic fusion for more robust feature tracking. Int. J. Comput. Vis. 49(1), 79–89 (2002)

[25] Dewasurendra, D.A., Bauer, P.H., Premaratne, K.: Evidence filtering. IEEE Trans. Signal Process. 55(12), 5796–5805 (2007)

[26] Ramasso, E., Panagiotakis, C., Pellerin, D., Rombaut, M.: Human action recognition in videos based on the transferable belief model: application to athletics jumps. Pattern Anal. Appl. 11(1), 1–19 (2008)

[27] Ramasso, E., Rombaut, M., Pellerin, D.: Forward-Backword Viterbi procedures in the transferable belief model for state sequence analysis using belief functions. Lect. Notes Artif. Intell. 4724, 405–417 (2007)

[28] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

[29] Smarandache, F., Dezert, J.: Applications and Advances of DSmT for Information Fusion. Am. Res. Press, Rehoboth (2004)

[30] G. Carter. Coherence and time delay estimation. Proceedings of the IEEE, 75(2):236–255, 1987.

[31] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24(4):320–327, 1976.

[32] J. Shi and C. Tomasi, “Good features to track,” in Proc. IEEE Conf. Comput. Vision Pattern Recognition, 1994, pp. 593–600.

[33] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proc. 7th Int. Joint Conf. Artificial Intell., 1981, pp. 674–679.