

# A Novel Brain-Based Approach for Multi-Modal Multi-Target Tracking in a Mixed Reality Space

Zenon Mathews<sup>1</sup>, Sergi Bermúdez i Badia<sup>1</sup>, Paul F.M.J. Verschure<sup>1,2,3</sup>

(1) Institut Universitari de l'Audiovisual (IUA), Universitat Pompeu Fabra, Barcelona, Spain

(2) ICREA, Barcelona, Spain

(3) Foundation Barcelona Media, Spain

E-mail: zenon.mathews@upf.edu

## Abstract

*We propose an architecture for multi-modal multi-target tracking, for integration of multi-sensory input and its top-down modulation through active deployment of sensors and effectors based on Bayesian inference. The implementation of our architecture, which is inspired by the superior colliculus (SC), engages a method for dynamical allocation of sensors and effectors to enhance tracking and to resolve conflicting multi-modal data. We suggest to use the joint probabilistic data association method as the basis for facilitating the top-down modulation of bottom-up sensory data. High level knowledge of the current state and interaction scenarios of a mixed reality space allows for modulation of bottom-up sensory data. We test our implementation for multi-sensory data fusion to tackle the real-time multi-person tracking problem in a human accessible mixed reality environment called XIM (eXperience Induction Machine).*

**Keywords:** Multimodality, Tracking, Mixed Reality, Presence, Attention

## 1 Introduction

The integration of inputs from multiple sensory systems is mastered by the brain at an unparalleled level of perfection. Association of different sensory cues with external objects or events, their registration, processing and the subsequent generation of motor commands are indeed critical for survival of animals. Neurophysiological research suggests that the superior colliculus (SC) is one of the primary areas for sensory data association and appropriate motor action generation for orienting response toward the source of stimulation (Stein and Meredith 1993). It has been shown that the SC possesses sensory maps for individual sensors, from which motor maps for motor action generation are formed (Stein and Meredith 1993; Anastasio et al 2000; MA et al 2006). Also Bayes' rule has been successfully used to model multi-sensory fusion as exhibited by the SC (Stein and Meredith 1993). High-level tuning/enhancement of sensory information could possibly be a key aspect in sensor data processing using limited resources. Nevertheless, the mechanisms for top-down modulation of bottom-up sensory information, i.e. using already available knowledge to prune or modulate sensory input, are relatively unknown and a matter of intense research (Navalpakkam and Itti 2007). We address the multi-modal multi-target tracking problem from this perspective and propose a complete model called A-BUTDT (Active Bottom-Up Top-Down Tracking), for integration of multi-sensory input and its top-down modulation through active deployment of sensors and effectors based on Bayesian inference. We implement a generic multi-modal tracking framework that engages an SC based method for dynamical recruitment of sensors and effectors to enhance tracking and to resolve conflicting multi-modal data. By sensor recruitment we understand

---

active deployment of reliable sensors and effectors to enrich tracking by collecting and comparing object attributes. We implement and test our SC model for multi-sensory data fusion to tackle the multi-person tracking problem in a human accessible mixed reality environment called XIM (eXperience Induction Machine).

## 2 XIM: A Mixed Reality Space

XIM is the physical space, which is part of the PVC (Persistent Virtual Community) where groups of real, remote and synthetic characters interact with each other, making accurate tracking of real objects in the mixed reality environment a requirement for meaningful interaction scenarios. XIM comprises a pressure sensitive floor, overhead, infrared and movable cameras, moving lights (light fingers), triples of microphones for sound recognition and localization, projection screens, and also ambient and spatialized sonification. On the three projection screens the virtual world of persistent virtual community (PVC) is made visible to the real visitors of the XIM. The tracking system A-BUTDT gets input from the XIM sensors. But A-BUTDT can also actively deploy some of these effectors and sensors for its attentional mechanisms (see figure 1).

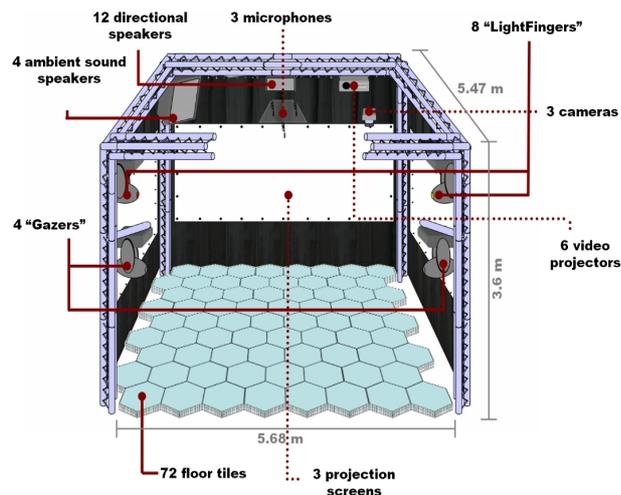


Figure 1. The Mixed Reality Space eXperience Induction Machine (XIM)

The different sensors are used for multi-modal sensory input to the tracking system. XIM is about 25 square meters and allows several humans to be active in it simultaneously. This often causes clutter in sensory data, which is challenging for the data association mechanism used for tracking. Multi-target tracking in XIM is a challenging task as it is a very dynamic environment, making single-modal tracking infeasible.

### 2.1 Multimodal Sensors

Single sensor modalities are prone to various errors and outfalls. For a tracking system to work reasonably it should therefore not be dependent on a single modality. To ensure precise tracking despite the limitations of single sensors we need to employ several sensor modalities simultaneously. To facilitate this, A-BUTDT allows for easy and dynamic acquisition of new sensors. In the following subsections the two sensor modalities we use at the moment are introduced in detail.

### ***Pressure Sensitive Floor***

We use a pressure sensitive tactile floor, which delivers loads of persons (Delbrück et al, 2007). The floor sends the detected loads (as two dimensional Gaussians) to A-BUTDT using a UDP socket. A graphical user interface is available which displays the current state of the floor tiles (figure 1, left). The colour of each floor tile can be set to any RGB value. The GUI displays the current state of the floor in real-time and also the loaded tiles (dots on tiles in figure 1, left).

### ***Overhead Tracking: Using the Efference Copy***

Overhead visual tracking is very difficult as the illumination conditions changes dynamically since the XIM contains the floor tiles and other changing sources of light like moving light fingers and projection screens (see figure 1, right). Object detection in dynamically changing lighting conditions is a huge field of research for itself in computer vision and image processing.

We employ the idea of subtracting the efference copy to be able to detect people in the input from the overhead cameras. In biological systems, efference refers to the totality of motor impulses necessary for a movement. And efference copy is an *image/copy* of the efference residing in the central nervous system. This lets the animal for example ignore the changes in the visual stimuli caused by its own movements.

To be able to subtract the “efference copy”, we chose to employ two overhead cameras at the same time, instead of just one. The two overhead cameras have the same view field and are mounted on a rig in the XIM. One of them has an infra-red filter and the other one delivers a greyscale image (figure 2 ,right). The image from the infra-red camera contains the only lighted tiles of the floor, as this emits most heat. The image from the greyscale camera contains everything in its view field (figure 2, right). After some image processing using edge extraction, subtraction from a predefined edge mask, and combining both the images from the two different cameras, people are located (figure 2, right). Two dimensional Gaussian distributions are calculated using the number of pixels detected as part of a person in the image. We use Intel's open source computer vision library (Open CV) for all the image processing. The overhead tracking also sends the computed Gaussians for detected people to A-BUTDT via an UDP socket.

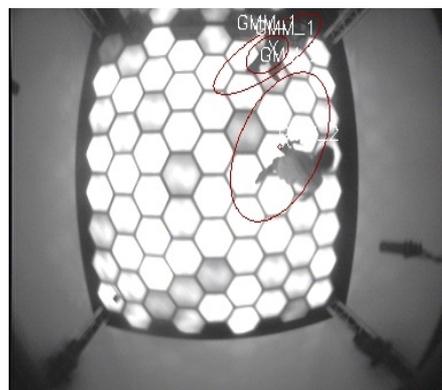


Figure 2. Left: View of Tactile Floor GUI. Right: Person Detection Using Overhead Tracking

### 3 Active Bottom-Up Top-Down Tracking

Using several sensor modalities at the same time inherently leads to the problem of data association and data fusion. For solving the data association problem the brain employs top-down modulation mechanisms. We are interested in employing a sound mathematical framework to facilitate top-down modulation of bottom-up sensory data in the case of XIM.

Recent psychophysical research suggests that humans perform near-optimal Bayesian inference in solving different tasks such as multi-sensory integration, decision-making and motor control (Massaro 1987; Stein and Meredith 1993; Ma et al 2006). Further, Bayes' rule has been proposed to model multi-sensory enhancement in the SC (Anastasio et al 2000). In our work we suggest the use of joint probabilistic data association (JPDA) for multi-sensory data association, since the JPDA filter is a suboptimal single-scan approximation to the optimal Bayesian filter (Bar-Shalom 1987), and more importantly, it provides a suitable framework for the top-down modulation of the acquired sensory data. This is facilitated by the JPDA as it operates with association events between data and targets, the probability of which can be computed separately. Moreover, multi-sensory enhancement can be achieved in the framework of JPDA as more accurate sensor data (i.e. with less measurement error) is probabilistically preferred. Owing to the association probabilities in the JPDA formulation, high-level information can be used to tune sensory input. This can be done by biasing the computation of the association probabilities between targets and events. All the same, the number of association events in JPDA grows exponentially with the number of targets and exact computation of the association probabilities is NP-hard. For this reason we use a Markov chain Monte Carlo data association (MCMCDA) algorithm, that finds an approximate solution to JPDA in polynomial time, as suggested in (Sastry 2005).

We propose the A-BUTDT framework which employs the above idea of top-down modulation of bottom-up sensory data. The world-model which contains many information such as target positions, target features, coherence map of sensors and interaction scenario of the mixed reality space, delivers the high-level information to the data fuser. The data fuser also receives bottom-up sensory data. Data Fuser uses JPDA for data association and fusion. The fused data is then used to update the world model but also to generate motor actions in order to recruit sensors/effectors to verify the world model (figure 3).

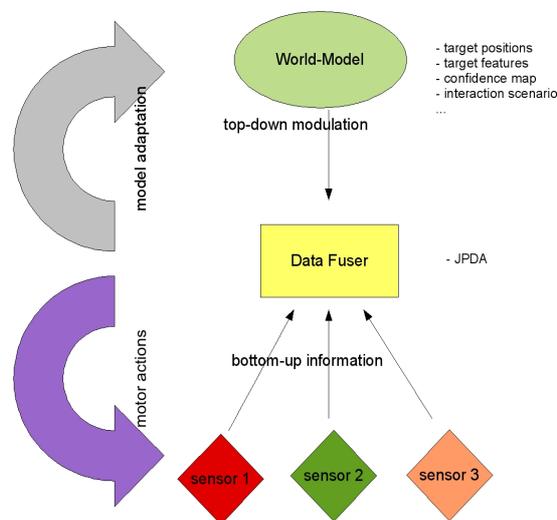


Figure 3. A-BUTDT Architecture

Assuming that in a particular interaction scenario in the XIM, for example if it can be deduced from the given circumstances that a target (real person 2 in figure 4) would not go to the right, then the event association probabilities can be modulated, pushing the data association to prefer the sensor data to the target's *left*, if any. The knowledge that “the target will not go to the right” is an input from the world-model which is continuously informed about the interaction scenario. In figure 4, the data (represented by the black dot, which falls inside the validation gates of both persons 1 and 2) will then be associated only to real person 1.

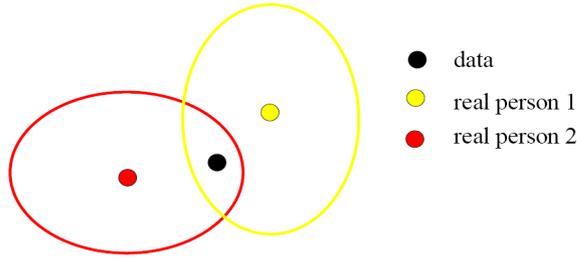


Figure 4. Validation Gates for Data Association

### 3.1 Joint Probabilistic Data Association (JPDA)

Joint probabilistic data association (JPDA) is a powerful tool for solving data association problems, which arises in many applications such as computer vision, surveillance, mobile robots etc. JPDA is a single-scan approximation to the optimal Bayesian filter, which associates latest observations/data to known targets sequentially. JPDA enumerates all possible associations between data and target at each time step and computes the association probabilities  $\beta_{jk}$ , which is the probability that the  $j$ -th observation/data is from the  $k$ -th target. Given such an association, the target state is estimated by Kalman filtering and this conditional expectation of the state is weighed by the association probability. Let  $x_t^k$  indicate the state of target  $k$  at time step  $t$ ,  $\omega_{jk}$  the association event where the observation  $j$  is associated to target  $k$  and  $Y_{1:t}$  stays for all the observations/data from time step 1 to time step  $t$ . The state of the target can be estimated as

$$E(x_t^k | Y_{1:t}) = \sum_{\omega} E(x_t^k | \omega, Y_{1:t}) P(\omega | Y_{1:t}) \quad (1)$$

$$\hat{=} \sum_j E(x_t^k | \omega_{jk}, Y_{1:t}) P(\omega_{jk} | Y_{1:t}) \quad (2)$$

where  $\omega_{jk}$  denotes the association event where observation  $j$  is associated to target  $k$  and  $\omega_{0k}$  denotes the event that no observation is associated to target  $k$ . Therefore the event association probability is

$$\beta_{jk} = P(\omega_{jk} | Y_{1:t}) \quad (3)$$

JPDA computes a validation gate for each target (ellipses around the real persons in figure 4) using the innovation of new observations. It only considers observations inside the validation gate for each target. For further mathematical details of JPDA see (Bar-Shalom 1987).

The computation of  $\beta_{jk}$  requires a summation over the posteriors and its exact calculation is NP-hard and is a major drawback of JPDA (Collins and Uhlmann 1992). The number of

association events rise exponentially in relation to the number of observations. We therefore implemented a Markov Chain Monte Carlo method to compute  $\beta_{jk}$  in polynomial time as suggested in (Sastry 2005).

### 3.2 Markov Chain Monte Carlo (MCMC) Approximation of $\beta_{jk}$

MCMC is used to estimate the association event probabilities  $\beta_{jk}$  in polynomial time and with good fidelity as shown in (Sastry 2005). For this the problem is reformulated as a bipartite graph. Consider the bipartite graph  $G = (U, V, E)$ , where  $U$  is a vertex set of predicted observations,  $V$  is vertex set of observations and  $E$  is mapping predicted observations (or in other words a target) to an observation. We thereby only consider feasible mappings, i.e. the ones that respect the validation gate criteria for the JPDA. The algorithm starts with one such feasible mapping and a Markov chain is generated as described in (Sastry 2005). For details of the MCMC approximation of  $\beta_{jk}$ , its convergence and stability see (Sasty 2005). The polynomial time complexity with respect to the number of targets allows MCMC to be used for computing  $\beta_{jk}$  in real-time.

## 4 Active Sensor/Effector Allocation Mechanism

The multi-sensor integration described above delivers input to the *world-model* maintained by the system, which includes, among others, exact positions of real visitors of XIM. Given this information, and further high level information about the visitors (e.g. the dress colour) or their behaviour, the sensory input can be tuned to enhance tracking performance. This *just-in-need* recruitment, as also adopted by the SC through its motor-maps, complies totally with the limited resources constraint, which is a reasonable choice when we have large number of real visitors in mixed reality spaces like the XIM.

Such sensor recruitment helps in verifying the world-model and correcting it when needed. With the moving cameras, which are recruited using the data fusion result to check the world-model for its knowledge of position of real persons in XIM, the Ids of persons can be restored if they get swapped. The ID swapping problem occurs when we have high clutter, meaning people are very close to each other. We use the Haar classifier from (Open CV) for human body detection (figure 3). The colour histograms of individual persons are collected and compared to ensure the integrity of the world-model.



Figure 3. Moving Cameras for Human Torso Detection

## 5 Control Experiments

Our implementation of A-BUTDT using JPDA and MCMC was done in a C++ Linux environment. The unoptimized version of the implementation of A-BUTDT runs at around 10Hz on an AMD Athlon 1.2 GHz machine.

We tested the need for multimodality by comparing the tracks of persons created by using single sensor modalities individually and by combining them as described above using JPDA. Figure 4 shows the reconstruction of two tracks when we used two sensor modalities (floor and overhead tracking). The vertical axis denotes time in seconds and the horizontal ones the space. the tracked persons were performing a zick-zack trajectory with two crossings. We see that the tracks are still reconstructed very well (figure 4).

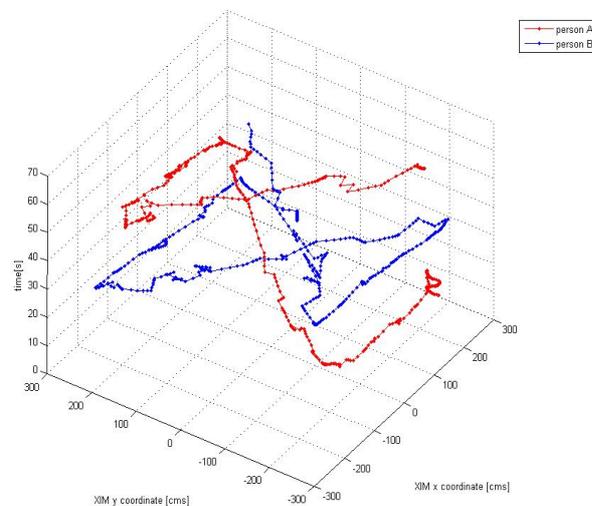


Figure 4. Visualization of Fused Tracks for Persons in XIM Using Multimodal Data

Using single modalities, A-BUTDT is not able to track the persons as the single modalities may not deliver data at every time step and also has false alarms. The zick-zack trajectory cannot be followed and A-BUTDT loses track after some time. Figure 5 depicts the case when using just the floor modality and figure 6 the overhead camera tracking.

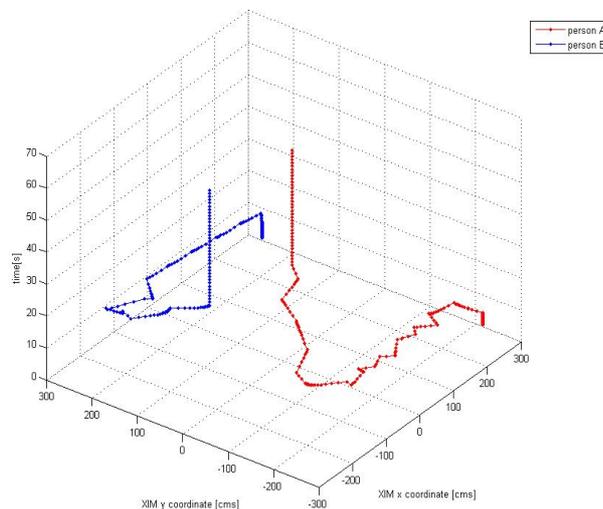


Figure 5. Visualization of Fused Tracks for Persons in XIM Using Single Modality: Floor

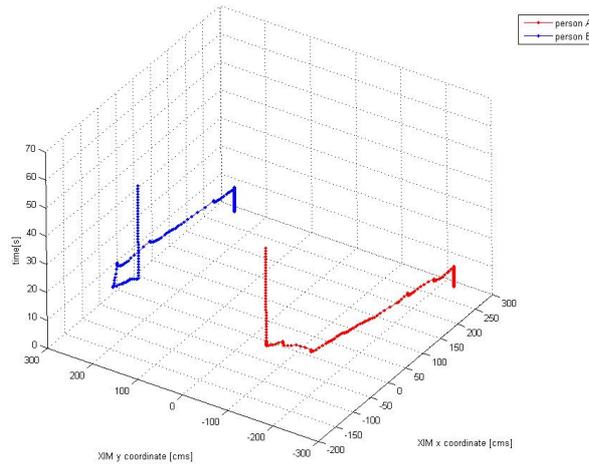


Figure 5. Visualization of Fused Tracks for Persons in XIM Using Single Modality: Camera

We use the information from the data fuser to recruit movable pan-tilt cameras and light fingers to direct themselves to positions where there are persons according to the world-model. The pan-tilts then extract colour histograms of the persons to verify the world-model. We see an example of such colour histogram extraction in figure 6. This information is used for example to restore the Ids of persons, in case they got swapped during crossings.

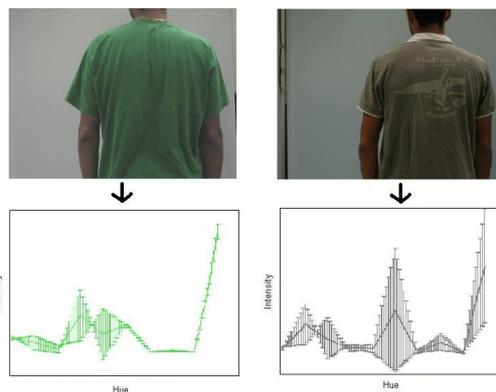


Figure 6. Color Histogram Extraction Using Movable Pan-Tilt Cameras

## 6 Conclusion and Outlook

We introduced the first version of A-BUTDT and conducted preliminary experiments to validate our architecture. Our architecture is capable of modulating bottom-up sensory data

---

using top-down information. Further the experiments show how multimodal data improves tracking considerably. Some issues like precise colour histogram extraction using the moving pan-tilt cameras have to be resolved more efficiently. In future work we plan to develop algorithms to for incorporating all the knowledge contained in the world-model to modulate the JPDA event probabilities. Further we envisage also applying learning models to enable A-BUTDT to learn its recruitment actions by itself using learning models as the one suggested in (Verschure and Althaus 2003).

## 7 References

- Anastasio, T.J., Patton, P. E., Belkacem-Boussaid, K. Using Bayes' Rules to Model Multisensory Enhancement in the Superior Colliculus. *Neural Computation* 12, 1165-1187, © 2000 MIT
- Bar-Shalom, Y. Tracking and Data Association. *Academic Press Professional Inc.* San Diego CA USA 1987
- Collins, J., Uhlmann, J., Efficient gating in data association with multivariate distributed states. *IEEE Trans. Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 909-916, 1992
- Delbrück, T. and Whatley, A. M. and Douglas, R. and Eng, K. and Hepp, K. and Verschure, P. F. M. J. A tactile luminous floor for an interactive autonomous space , *Robotics and Autonomous Systems*, 55:(6) 433--443, Jun, 2007
- Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A. Bayesian Inference with Probabilistic Population Codes. *Nature Neuroscience*, Vol. 9, Nr. 11, November 2006
- Massaro, D. W., Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry, *Hillsdale, N.J.: Lawrence Erlbaum Associates* 1987
- Navalpakkam, N., Itti, L., Search Goal Tunes Visual Features Optimally, *Neuron* 53, 605-617, © 2007 Elsevier Inc.
- Oh, S., Sastry, S., A Polynomial-Time Approximation Algorithm for Joint Probabilistic Data Association. in *Proc. of the American Control Conference (ACC)*, Portland, OR, June 2005
- Open CV, Intel Open Source Computer Vision Library, [http:// www.intel.com/ technology / computing/ opencv /](http://www.intel.com/technology/computing/opencv/)
- Stein, B.E., Meredith, M.A. The Merging of the Senses. *MIT Press Cambridge MA* 1993
- Verschure, P.F.M.J., Althaus, P., A real-world rational agent: unifying old and new AI. *Cognitive Science* 27 2003, 561..590
-