Are You There?

Active Attention for Person Tracking in Mixed Reality Environments

Zenon Mathews[1], Sergi Bermúdez i Badia[1], Paul F.M.J. Verschure[1 2 3]

(1) Institut Universitari de l'Audiovisual(IUA), Universitat Pompeu Fabra, Barcelona, Spain
(2) ICREA, Barcelona, Spain
(3) Foundation Barcelona Media, Spain

{zenon.mathews@upf.edu, sergi.bermudez@upf.edu, paul.verschure@upf.edu}

Abstract

How to convince human participants in a mixed reality environment that they themselves, the virtual world around them and synthetic characters in the virtual world exist as separate entities? The human accessible mixed reality environment called XIM (eXperience Induction Machine) is primarily being developed to investigate such questions about the notion of presence in mixed reality environments. More realistic interaction scenarios contribute to enhanced presence and for this real-time knowledge of position of individual human participants inside the mixed reality space becomes indispensable. We propose a multi-modal tracking system which fuses data from different sensors in XIM and employs top-down modulation of bottom-up sensory data. It actively reallocates XIM's attention to verify its world-model. Our approach employs selective attentional mechanisms, bestowing the mixed reality space the appearance of a single autonomous entity.

Keywords--- **multimodal tracking, mixed-reality, presence, human computer interaction, data fusion**

1. Introduction

Mixed reality (MR) environments enable interactions between real, synthetic and virtual characters. These interactions can be of diverse nature and quality depending on the technologies available. Many interaction scenarios, e.g. interactive games, virtual societies etc. require accurate tracking of the physical entities in the MR space. From the perspective of a real person in an MR space, a sense of presence inside the MR world is generated from the feeling that he/she exists within the space but as a separate entity. The experience of self can be enhanced if other beings that exist in the virtual world appear to recognize that one exists [9, 10]. Subjective personal presence is a measurement of the extent to which and the reasons why a person feels like he/she is in a virtual world [9]. Social presence refers to the extent to which other beings (real or synthetic) also exist in the world and appear to react to a person in the MR space [9]. Social presence derives from conversing with other humans or from interacting with synthetic entities in the MR space. If there is *someone* or *something* that recognizes that a person is there, it is easier for himself/herself to believe that he/she is there [10]. The same argument goes for environmental presence, which refers to the extent to which the environment itself appears to know that the human entity is there and to react to him/her [11].

All the three notions of presence described above are feasible only if information about the position of individual human visitors of the MR space is available at any given time, making real-time person tracking a necessity. Furthermore the reactions of the space itself to the actions of individual human entities in the space are of immense importance to the derivation of environmental presence. The mixed reality space XIM (eXperience Induction Machine) includes controllable pan-tilt cameras and moving lights, which can be oriented towards a particular position. XIM turns its attention to a particular human participant for example by directing the moving lights towards her/him. We exploit such *selective attentional* mechanisms to enhance the tracking performance, as discussed in the following sections. But also this active deployment of sensors and effectors makes the human participant feel that the space *knows* where he/she is and that it would like to *know* him/her better. This is very similar to real world interaction scenarios: imagine a group of several people

talking to each other. Each one of them selectively pools his attention to one (or may be two) of the participants at a time when he/she wants to listen to what that particular person is saying. Such display of *attention allocation* considerably contributes to enhancement of social presence, as this clearly displays that the opposite entity (synthetic or real) is looking at or listening to the human participant.

Often, as is in the case of XIM, single modal person tracking in mixed reality spaces is very difficult as each of the different sensors have high errors [2]. Multi-modal tracking is therefore necessary to reliably track human participants in dynamic mixed reality spaces. We propose a tracking mechanism which employs the above mentioned sensor/effector allocation mechanisms to track individual real entities in a MR space. Our approach is based on brain mechanisms for solving data association and fusion tasks.

The integration of inputs from multiple sensory systems is mastered by the brain at an unparalleled level of perfection. The association of different sensory cues with external objects or events, their registration, processing and the subsequent generation of motor commands are 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 [7]. It has been shown that the SC possesses sensory maps for individual sensors, from which motor maps for motor action generation are formed [7]. Also Bayes' rule has been successfully used to model multi-sensory fusion as exhibited by the SC [1, 3, 7]. High-level modulation of sensory information could possibly be a key aspect in sensor data processing using limited resources [7]. All the same, the brain 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 [5]. We tackle the multi-modal multi-target tracking problem in XIM from this perspective and introduce 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 have already an completed the first version of 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 data, as shown in [12]. We use the term *sensor recruitment* to refer to active deployment of reliable sensors and effectors to enrich tracking by collecting and comparing object attributes. This involves issuing motor commands after top-down modulated data fusion (see figure 2). 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).

In this work, the key focus will be on achieving the goal of inducing the feel of a separate autonomous entity in the mixed reality space. Tracking enables social presence in the PVC (Persistent Virtual Community), which surrounds the XIM, as the virtual characters need to know where and when the human participants are inside the XIM.

2. XIM: A Mixed Reality Space as an Attentive Single Entity

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 (see figure 1). On the three projection screens the virtual world of persistent virtual community (PVC) is made visible to the real visitors of the XIM (see figure 1).

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 unfeasible.

Mixed Reality environments like the XIM facilitate testing our model of selective attention for multi-modal tracking, as it contains several sensors, some of which are actively deployable. Active deployment of sensors and effectors can be used to collect attributes of individual entities in the space. Further, the interaction scenarios in a mixed reality space highly facilitate the use of high-level knowledge to enhance the tracking performance.

Mixed reality spaces are not the only environments which provide this rich diversity of data, but they surely make a good test bed for the selective attentional model for tracking. As mentioned, XIM deploys its effectors and also sensors when it needs to retrieve more data about particular individuals in order to enhance the tracking performance. Moreover this gives the human visitor the feeling of being observed, which again boosts his/her environmental presence feeling [9].

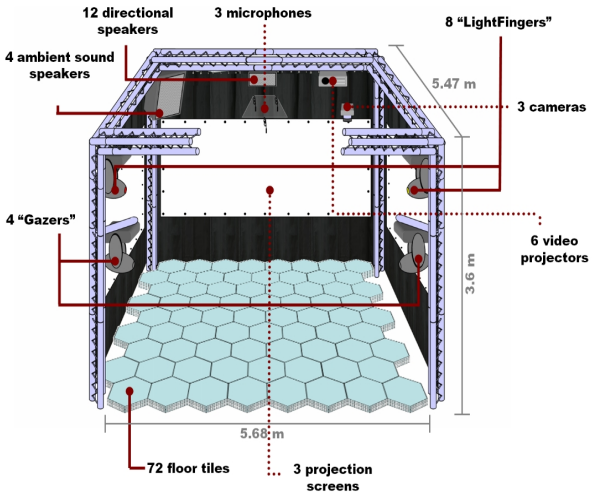

Figure 1. The mixed reality space called eXperience Induction Machine (XIM) consists of a pressure sensitive floor, controllable pan-tilt and overhead cameras, triples of microphones, moving lights (called light fingers), ambient and directional speakers. The tracking system A-BUTDT gets input from these sensors. But A-BUTDT can also actively deploy some of these effectors and sensors for its attentional mechanisms. XIM further has three projection screens, where the virtual world of the persistent virtual community (PVC) is made visible to the real visitors of the XIM.

3. World-Model as Top-Down Modulator

The superior colliculus (SC) is considered as the primary domicile of sensory data association and appropriate motor action generation in animals. It has been shown that the SC contains a "sensory map" for each sensor, on which the whole sensory space is represented [7]. The presence of a motor map, which is in alignment with the sensory-evoked orienting movements, has also been confirmed in physiological experiments. Moreover the generation and maintenance of these maps and the modulation of the animal's responses by using high-level information from other parts of the brain is still subject of ongoing research. Additionally 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 [3, 4, 5, 7]. Further, Bayes' rule has also been proposed to model multi-sensory enhancement in the SC [1].

Our work is based on an SC-motivated tracking mechanism which is deployed in XIM for tracking its physical visitors as proposed in [12]. Multi-modal multi-target tracking is a tested for any implementation, which aims at tackling resource allocation and sensor data processing inspired by the SC. The architecture we propose to facilitate high-level modulation of bottom-up sensory information uses the knowledge from a dynamic world-model, which is updated using the result of the data fusion and contains for example the position information of the real entities in the XIM. This world-model is used to tune the sensory information received from the different sensors. The world-model also contains information about the sensors and effectors (figure 2). The generated motor actions attests to XIM the mannerisms of an autonomous attentive entity.

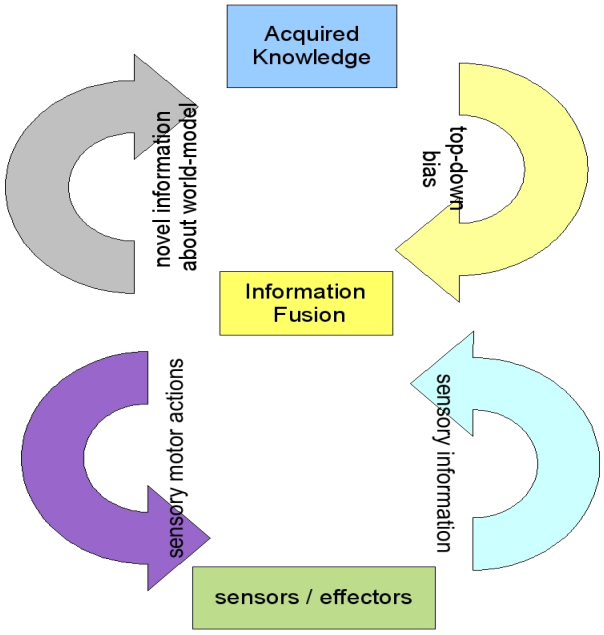

Figure 2. Top-down modulation of bottom-up sensory information in A-BUTDT: the individual sensors deliver data to the data fuser. The data fusion process is modulated by the input from the world-model. The result of the data fusion is then used to update the world-model and also to actively deploy sensors and effectors. Such deployment of motor actions lends XIM the characteristics of an attentive entity, fostering the feeling of environmental presence for the XIM visitors.

4. Event Based Multi-Sensory Integration

Data association problem arises in many applications such as surveillance, air-traffic, computer vision and mobile robots. In target tracking data association problem means the problem of determining which observation is generated by which target.

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 [2], and more importantly, it provides a suitable framework for the top-down modulation of the acquired sensory data. JPDA was developed to solve multi-target tracking especially in clutter and it is an approximation to the optimal Bayesian filter, in which the associations between the "known" targets and the latest observations are

made sequentially. JPDA enumerates all possible associations between targets and observations inside a validation gate (see figure 3) and computes the probabilities for each such enumeration (also called *events*). This is facilitated by 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. The detailed formulation details of JPDA can be found in the original JPDA work [2] and in [12].

The above JPDA formulation sets up the basis of our tracking framework, which tries to prune the JPDA events using high-level knowledge. For example if it can be deduced from the given circumstances that a real entity is involved in a spatially and temporally constrained interaction (possibly with a virtual entity), then this information from the world-model can be used to modulate the data associations and the generation of JPDA *events* (see figure 3). Consider the three example events: *a)* observation was caused by person 1. *b)* observation was caused by person 2 *c)* none of the persons caused that observation (false alarm). Now the world-model can be used to prune the generated events. E.g. if the world-model says that person 1 is involved in a stationary interaction with a virtual entity, then event *a* can be pruned.

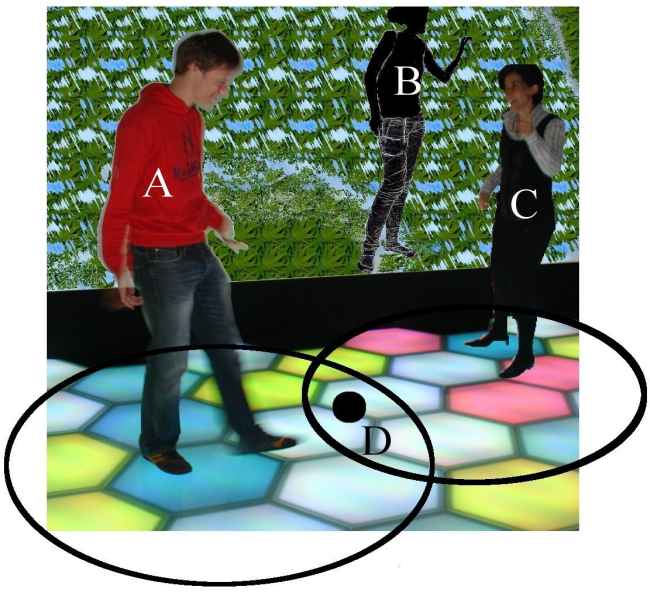

Figure 3. Pruning JPDA events using information from the world-model: A and C are two real people interacting in the XIM. B represents a virtual character. Characters B and C are engaged in a spatially static interaction. The ellipses indicate the validation gates for characters A and C. D and the corresponding black dot on the floor indicates a new sensory data point. An observation-to-person enumeration or event is only generated if the observation falls inside the validation gate of that person. In the figure, the observation falls into the validation gate of both persons and therefore different JPDA events are generated for each possible event. Nevertheless the information that character C is involved in a spatially static interaction can help to prune the JPDA events. In this case the data point D will be associated exclusively for target A.

5. XIM as an Attentive Entity for People Tracking

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 position information, and further high level information about the visitors (e.g. the dress color) or their behavior, the sensory input can be tuned to enhance tracking performance. This sensor recruitment, which is employed just when needed, is comparable to the strategy adopted by the SC through its motor-maps. This 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.

A-BUTDT uses controllable pan-tilt cameras and the moving lights to actively collect attributes of humans in the XIM. The world-model contains a hypothesis about where individual people are inside the XIM. A-BUTDT tests this hypothesis by actively collecting attributes: e.g. the moving lights are directed onto the real entity and the movable color camera extracts the hue of the person's clothes. Thus by collecting and comparing hues A-BUTDT tests the hypothesis provided by the world-model and if necessary corrects it (see figure 4).

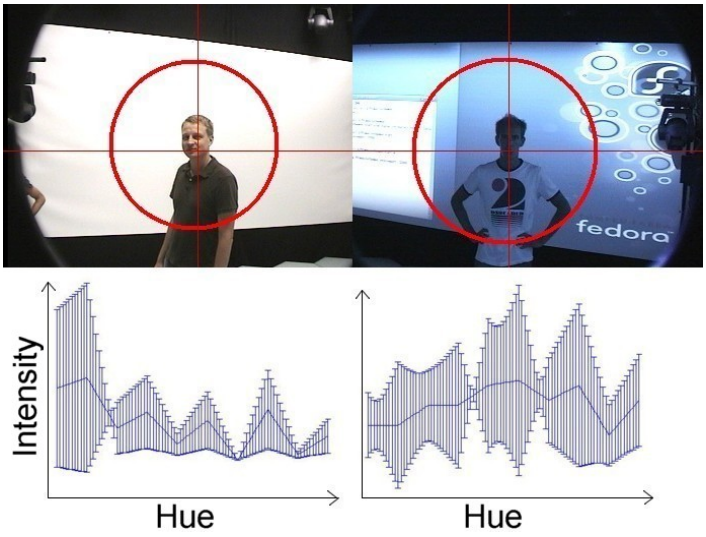

Figure 4. World-model adaptation through color histogram analysis: images of tracked objects of various colors (top) have different color histograms (bottom). The bottom panel shows the color distribution in the hue space and their standard deviations. These color histograms are collected from the torsos of the detected persons. Using a Kolmogorov-Smirnoff test it can be shown that the two distributions are significantly different and this information can be useful in identifying persons.

One use of such active attentional mechanism is ID management. This means that by collecting and comparing hue attributes of human participants, the identity information contained in the world-model can be verified for consistency with data fusion result. In other words A-BUTDT compares color histograms of humans, collected using active deployment of movable pan-tilts, with the color histogram information of the concerned person stored in the world-model. After this it can either correct the world-model or modulate the data fusion result (figure 2).

The pan-tilt cameras and the moving lights are just an example of sensor/effector deployment that XIM can undertake. For example the floor lighting could be changed to appropriate conditions to light up an area in a specific color if required. Our implementation is generic enough to facilitate easy addition of new sensors and effectors.

6. Control Experiments

A-BUTDT was implemented using the C++ programming language under the Linux environment. For communication between different applications we use UDP sockets. Visualizations of the tracked area, data and detected targets are available. The tracking data is then sent to a Torque game engine, which visualizes the virtual world surrounding the XIM. The physical visitors of XIM can therefore see their virtual representations on the projections. Further they can also experience synthetic characters that interact with them.

In our control experiments we implement and test A-BUTDT for real-time tracking of real visitors in the XIM and demonstrate how e.g. the movable color cameras can be recruited to actively collect target attributes such as hue or height and how this is used to enhance the tracking performance.

For our experiments we use multi-modal data, the single modes being visual tracking (using two overhead cameras) and the pressure sensitive floor. The visual tracking combines the information from an infrared and a grayscale camera mounted on the ceiling to deliver information about presence of humans in the XIM. The pressure sensitive floor similarly delivers loads of people in the space. Note that both these modalities suffer from characteristic errors of their own. The visual tracking has above all distortion errors along the periphery. The floor has delays as its major drawback. Both of them might deliver false alarms and also missed hits.

We track multiple humans in the XIM and evaluate the single modal and the fused data. One of the visitors follows a zick-zack trajectory which is easy to recognize in the data plots. We show that multi-modal information improves tracking considerably and that the trajectory can be reliably reconstructed (figures 5). By using just either the visual camera tracking or the floor data, reliable tracking is not possible (figure 5). Further the ID management is improved using the active deployment of sensors and effectors described above. The hue extraction mechanism should nevertheless be considerably improved to efficiently adapt to vastly changing lighting conditions.

Further we show how high-level knowledge about the sensors can automatically be generated from the data fusion result (figure 6). This is done by computing the error between the sensor data and the fused data and interpolating it to fit the tracked area. In the case of the visual tracking, we can observe that the data is more erroneous along the periphery, mainly due to the lens distortion the camera data suffers from (figure 6). This being just one example of high-level data contained in the world-model can be very useful to weigh the sensor data.

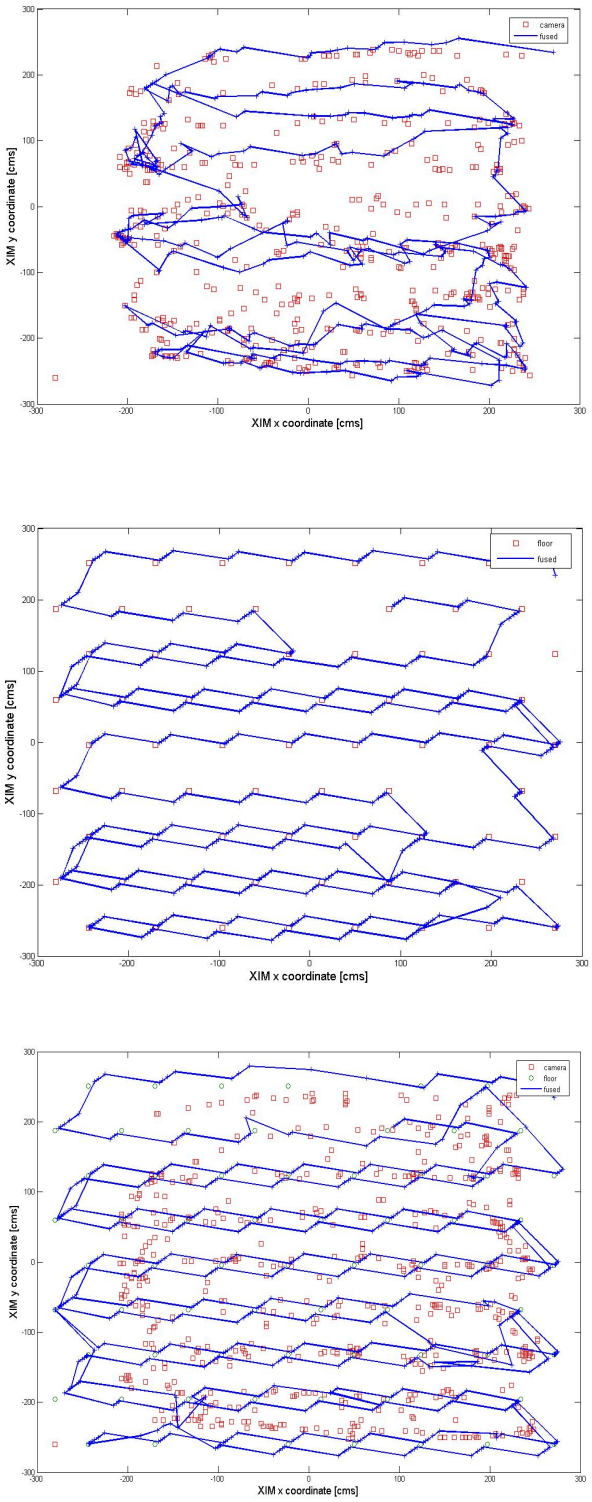

Figure 5. Top: The plot depicts single modal data from the visual tracking system (which consists of two overhead cameras, one with an infrared filter and the other one not) and the trajectory which the tracking system could reconstruct from this data using the described JPDA approach. Note that the trajectory is not lost throughout the session as the visual tracking delivers higher resolution than the floor data. Nevertheless the visual tracking has distortion errors especially along the periphery of the tracked area. Compare with figure 6. Middle: The plot shows single modal floor data and the trajectory the tracking system could infer from it. The trajectory is lost at some points because the floor data delivery suffers from delays and other errors, especially if the person moves fast and if there are multiple people at the same time in the XIM. Bottom: The plot shows multi-modal data, from the visual tracking and the pressure sensitive floor. The JPDA data association mechanism employed by A-BUTDT is used to construct the trajectory for this person and the performed zick-zack trajectory is correctly reconstructed. Note that by using multi-modal input we do get rid of the errors of the single modalities.

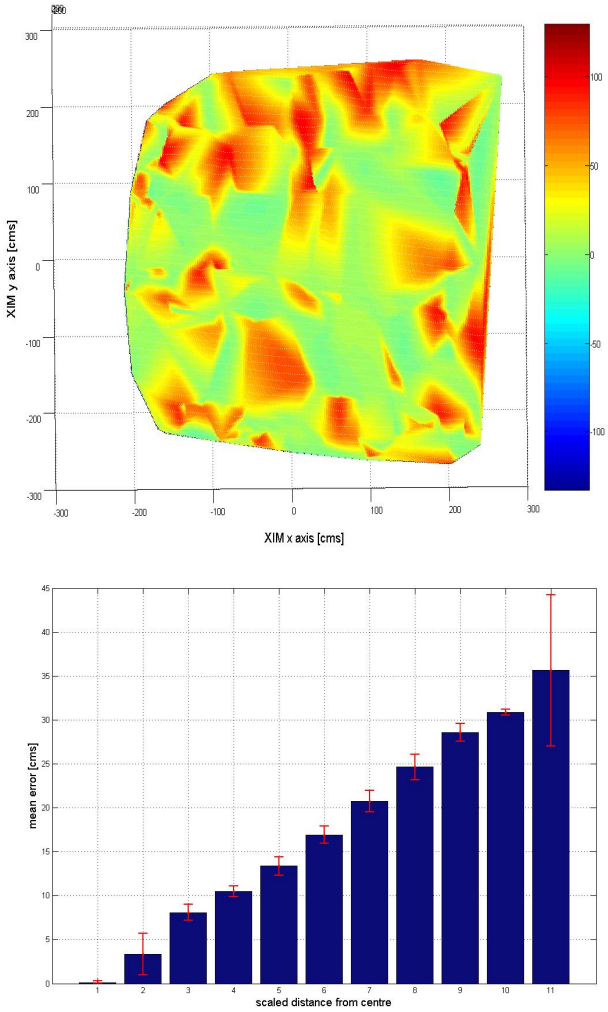

Figure 6. Top Panel: Confidence map generated automatically for the visual tracking system. The color map shows the absolute error in the sensor data. The visual tracking has more error along the periphery due to lens distortions. Bottom Panel: Distance from the midpoint of the floor and the error in the visual tracking data, when compared to the fused data. This error rises with the distance from the center. The above confidence map is an example of high-level information contained in the world-model, which is then used to weigh the data from this sensor. Such confidence maps can automatically be generated for each sensor modality and used to modulate bottom-up sensory data.

7. Conclusions

We showed how a brain-based framework for multi-modal data fusion in a mixed reality environment can be used to enhance the feeling of presence for the real visitors of the MR space. Our framework facilitates top-down modulation of bottom-up sensory data using the world-model which is automatically built up during the tracking process. By actively deploying sensors and effectors when needed the system is able to acquire attributes of physical visitors of a mixed reality space. Such active attention mechanisms support the enhancement of feeling of presence in the mixed reality space by giving the human participant the feeling that the space allocates its attention to him or her. Thus, XIM is perceived by the human participants in the mixed reality space as a single entity, which allocates its attention to him/her whenever needed, dependent on the context in which it operates.

8. Future Work

In future work, we intend to apply further selective attentional mechanisms, effective data processing (e.g. in the human visual cortex) and working memory models proposed from recent functional brain imaging studies. Also, we plan to enable A-BUTDT to learn how its own actions can influence the movements of XIM visitors, using adaptive memory-based learning models such as suggested in [8]. Beginning with random sensor/effector allocations A-BUTDT should be then able to learn which sensor allocations lead to better coherence of multi-modal data. This would enable it to adapt automatically to dynamic sensors and environments.

Acknowledgments

This project is supported by the European PRESENCCIA (IST-2006-27731) project.

References

[1] 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

[2] Bar-Shalom, Y. *Tracking and Data Association.* Academic Press Professional Inc. San Diego CA USA 1987

[3] 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

[4] Massaro, D. W., *Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry,* Hillsdale, N.J.: Lawrence Erlbaum Associates 1987

[5] Navalpakkam, N., Itti, L., *Search Goal Tunes Visual Features Optimally*, Neuron 53, 605-617, © 2007 Elsevier Inc.

[6] 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

[7] Stein, B.E., Meredith, M.A. *The Merging of the Senses.* MIT Press Cambridge MA 1993

[8] Verschure, P.F.M.J., Althaus, P., *A real-world rational agent: unifying old and new AI.* Cognitive Science 27 2003, 561..590

[9] Heeter, C. *Being There: The subjective experience of presence.* Presence: Teleoperators and Virtual Environments, MIT Press, fall, 1992.

[10] Bricken, M. (1991). *Virtual Worlds: No Interface to Design.* Tech. Rep., Seattle: University of Washington, Human Interface Technology Laboratory.

[11] Held, R. and Durlach, N. (1992). *Telepresence*. Presence: Teleoperators and Virtual Environments, 1:1, 109-112.

[12] Mathews, Z., Bermúdez i Badia, S., Verschure, P.F.M.J. *A Novel Brain-Based Approach for Multi-Modal Multi-Target Tracking in a Mixed Reality Space*, 4th INTUITION International Conference and Workshop, 4-5 October Athens, Greece, (in Press)