International Journal of Computer Vision, Special Issue on Object Tracking and Classification Beyond the Visible Spectrum

-

Multi-sensory and Multi-modal Fusion for Sentient Computing

Christopher Town
University of Cambridge Computer Laboratory
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
cpt23@cam.ac.uk

Abstract

This paper presents an approach to multi-sensory and multi-modal fusion in which computer vision information obtained from calibrated cameras is integrated with a large-scale sentient computing system known as "SPIRIT". The SPIRIT system employs an ultrasonic location infrastructure to track people and devices in an office building and model their state. Vision techniques include background and object appearance modelling, face detection, segmentation, and tracking modules. Integration is achieved at the system level through the metaphor of shared perceptions, in the sense that the different modalities are guided by and provide updates to a shared world model. This model incorporates aspects of both the static (e.g. positions of office walls and doors) and the dynamic (e.g. location and appearance of devices and people) environment.

Fusion and inference are performed by Bayesian networks that model the probabilistic dependencies and reliabilities of different sources of information over time. It is shown that the fusion process significantly enhances the capabilities and robustness of both sensory modalities, thus enabling the system to maintain a richer and more accurate world model.

Keywords Multi-sensory Fusion, Multi-modal Fusion, Sentient Computing, Object Tracking, Bayesian Networks

1 Introduction

1.1 Problem Definition and Context

Efforts in ubiquitous computing are increasingly focused on providing a model of computing in which the proliferation of relatively cheap communications, sensors, and processing devices is leveraged in such a way as to make the resulting systems aware of aspects of their environment and the interactions which take place within it. The goal of what is termed *sentient*[1] [20] or *context-aware* computing is to enable systems to perceive the world and relate to it in much the same way as people do, thereby creating the illusion of a shared perception that carries with it an implicit understanding of *context*. Indeed it can be argued that deriving an accurate representation of context

[1] From *sentient:* having the ability to perceive via the senses.

is a “holy grail” of human computer interaction [9], as it would allow people to interact much more naturally with computer systems in a way which is pervasive and largely transparent to the user.

Sentient computing thus aims to model aspects of the context within which human-computer interactions take place in order to better infer and anticipate user intentions and requirements. This is achieved by integrating information from a range of networked sensors and processors distributed throughout a (typically indoor) space in order to maintain an internal representation, or *world model*, of that environment. Applications utilise the world model in order to obtain implicit knowledge of user context. To realise the goal of shared perception, the robustness and accuracy of sensory data and its interpretation must approximate that of human beings in the chosen domain [10], and the world model must maintain an accurate up-to-date representation of context.

The sentient computing system considered in this paper (see section 1.3) uses ultrasound to track tagged devices such as computers and phones within an office. It is currently the most accurate large-scale wireless tracking system of its kind. Nevertheless, systems of this kind have a number of limitations arising from the fact that they are largely restricted to tracking the 3D location of ultrasonic tags which must be attached to objects of interest. On the other hand, visual information holds the promise of delivering richer representations of the world without an inherent need to tag salient entities. Computer vision offers a range of capabilities such as detection, classification, and tracking, which are important prerequisites of a context-aware computing system [6]. However, apart from the need to deploy sufficient numbers of cameras to ensure adequate coverage, machine vision remains hampered by problems of *generality* and *robustness*, which reduce its suitability as a primary (or sole) sensory modality. There is clearly much scope for work that integrates information from these disparate sources.

1.2 Related Work

Efforts in ubiquitous and context-aware computing have employed a range of sensor systems such as accelerometers, touch sensitive surfaces, and more commonly sound (microphones, ultrasound) or light (infrared, radio) at different frequencies in order to detect or track people and devices. Most indoor location systems such as the infrared-based Active Badge system [18] only allow positioning at room accuracy. The Cricket system [27] uses a series of ultrasonic beacons placed throughout a building at known locations, which periodically emit both ultrasonic pulses and a radio signal. Devices can use the time difference of the two pulses to determine the closest beacon and estimate their distance from it to within 5cm to 25cm. [14] describes a method for obtaining positioning accuracies of up to 50cm using link quality measurements associated with Bluetooth connections.

The role of computer vision in practical office-scale sentient systems has largely been restricted to the detection of tagged objects [21], although vision-based systems for gesture recognition and motion analysis have been under development for some time [13, 37, 4]. Approaches relying on probabilistic integration of different sources of visual evidence such as face detectors and models for motion, shape, and colour have shown particular promise (e.g. [28, 31]). However, the difficulties of deploying perceptual user interface mechanisms on the basis of vision alone are exacerbated by problems such as brittleness and lack of real-time performance and generality. Many vision-based systems have consequently been developed for fairly circumscribed control tasks where a limited vocabulary of pointing and selection gestures is appropriate. A truly pervasive system however requires a richer basis for interactions and a notion of context strong enough to recognise when no intended interaction with the system occurs.

These problems have led researchers to investigate fusion of vision with other sensory modalities [37]. Most such systems rely on integration with sound in the audible range via microphone assemblies, which has proven particularly popular for videoconferencing applications [32, 6]. There are also systems that have integrated vision as a secondary modality to an existing system, for example [3], which deploys visual gait recognition as an additional identifying cue to a system based on a pressure sensitive "Active Floor". Multi-modal user localisation is also an important topic in robotics research [11], where information from stereo or omni-directional cameras mounted on a moving robot is often combined with sensors such as laser range finders.

Bayesian approaches to multi-modal fusion have been gaining prominence in the vision and other research communities. The approach presented in this paper is related to work by Sherrah and Gong [28], which shows how multiple sources of evidence (split into necessary and contingent modalities) for object position and identity can be fused in a continuous Bayesian framework using indicator variables to model observation reliabilities. Tracking algorithms which perform concurrent probabilistic integration of multiple complementary and redundant cues have been shown to be much more robust than those that utilise only a single cue [25, 26]. Tracking on the basis of multiple sources of information is also demonstrated by Choudhury et al. [5], who present a system that fuses auditory and visual cues for speaker detection, while Torralba et al. [33] describes work that highlights the importance of context for object and place recognition in situated vision systems.

1.3 The SPIRIT System

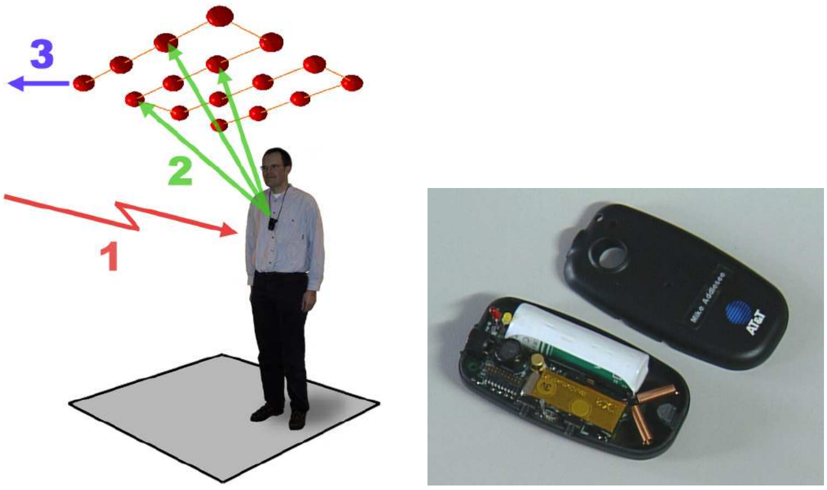

Figure 1: *Left:* **Function of the SPIRIT location system. A Bat sensor is triggered by radio (1), emits an ultrasonic pulse (2), and time-of-flight of the pulse is measured by receivers in the ceiling (3) to compute 3D position.** *Right:* **One of the wireless tracking tags known as a "Bat". The device is about the size of a matchbox and features two buttons for input and two LEDs and a sound chip for output. It contains a radio receiver, ultrasonic transmitter, microchip with a unique 48bit ID, and a AA lithium battery supplying enough power for up to 12 months of operation under normal conditions.**

The SPIRIT[2] [1, 17] system was originally developed at AT&T Laboratories Cambridge, where it was in continuous operation by 50 staff members. The system is currently deployed throughout the Laboratory of Communication Engineering at Cambridge University (www-lce.eng.cam.ac.uk). As shown in figure 1, the system uses mobile ultrasonic sensor devices known as "Bats" and a receiver infrastructure to gather high-resolution location information for tagged objects such as people and machines. Such information is used to maintain a sophisticated world model

[2] Originally an acronym for "SPatially Indexed Resource Identification and Tracking".

of the office environment where it has been deployed. Applications can register with the system to receive notifications of relevant events to provide them with an awareness of the spatial context of user interactions. The achieved spatial granularity is better than 3cm for over 95% of Bat observations (assuming only small motion), and Bats may be polled using radio base stations with a variable quality of service to give update frequencies of up to 25Hz (shared among all Bats assigned to a given radio base station) while remaining scalable to hundreds of tagged people and devices in a large office. The Bats are equipped with two buttons, two LEDs and a sound chip to allow them to be used as portable input-output devices.

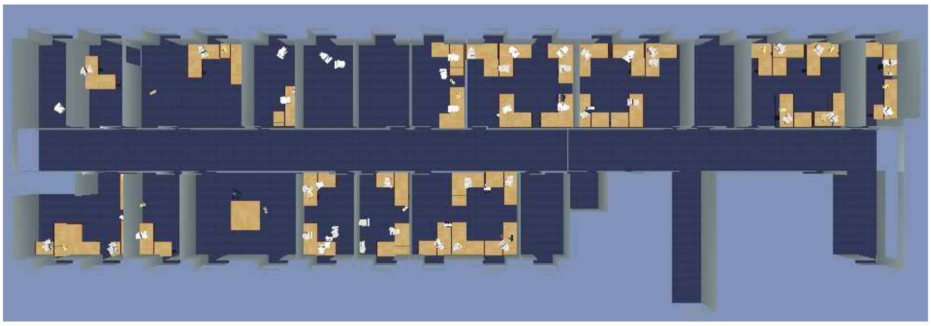

Figure 2: **The "LabView" application displays a 3D real-time map representing the state of the world model. The bird's eye view shown provides an overview of the sentient office (at the Laboratory of Communications Engineering) and objects such as furniture, devices and people.**

The driving paradigm is that of "computing with space" [20, 19], i.e. physical location and spatial context (typically expressed in terms of containment and proximity) together with the attributes and capabilities of entities and devices present at a given time drive the behaviour of applications built upon the system. Some applications such as "LabView" shown in figure 2 allow users to navigate and browse the world model itself, while others respond to particular configurations of interest. Co-location and spatial composition can be used to infer aspects of context (e.g. "user A has entered office O","user B is using his Bat as a 3D mouse to control the scanner in corridor C", "user B has picked up PDA P") which can influence or trigger application behaviour, hence space itself becomes part of the user interface. Current SPIRIT applications include "follow me" event notification, personnel and resource localisation, office visualisation, user authentication, desktop teleporting, virtual 3D interfaces, and location support for augmented reality.

2 Integration of Visual Information

2.1 Motivation

Although the SPIRIT system has proven effective in providing fairly fine-grained spatial context upon which sentient computing applications can be built, difficulties remain [23]. Bat system spatial observations are limited to the location of the Bat sensor, which is polled sporadically by a central base station. Each Bat has an associated identity (e.g. "Digital camera 1", "User J.Smith"), which may carry associated semantics (e.g. digital cameras must be operated by a person, people can exhibit certain patterns of movement). However, only objects tagged with Bats can be tracked, and the model of the environment is static unless other sensors (e.g. light switches and temperature dials) provide information on it.

Computer vision methods can provide multi-modal human-computer interfaces with transparent detection, recognition, and tracking capabilities, but on their own suffer from a lack of robustness and autonomy in real world interaction scenarios. The integration of distinct sources of information about the world in light of application specific constraints holds great promise for building systems that can optimally leverage different sensory capabilities and failure characteristics. Vision offers the possibility of acquiring much richer representations of entities in terms of their orientation, posture, and movements. It can also detect and to some extent classify and track additional features of the static (e.g. furniture) and dynamic (e.g. people and portable devices not equipped with Bats) environment. It may also be used to smooth over some of the difficulties inherent in an ultrasonic location infrastructure, thereby making it more robust. Information from the SPIRIT world model can in turn be used to provide constraints to the fusion process, to (re)initialise computer vision modules, and to act as a focus of attention mechanism.

2.2 Spatial and Temporal Correspondence

In order to fuse data from the visual and SPIRIT modalities, one must translate between their underlying representations. This requires translation between the 3D SPIRIT and 2D image reference frames and synchronisation of SPIRIT events with corresponding video frames acquired by a particular camera. Further details are available in [34].

2.2.1 Frame of reference

Visual input is acquired from cameras placed at known locations within the SPIRIT world frame. Both the positions of the cameras and their intrinsic and extrinsic parameters were calibrated carefully. Intrinsic parameters were estimated using a chessboard calibration pattern and the Matlab toolbox developed by Jean-Yves Bouguet [2]. Several images (typically 20-30) of the chessboard were analysed, and camera parameters were estimated through correspondence analysis of positions of the corner points to establish planar homographies.

Camera position, view area, and extrinsic parameters were determined by means of a surveying application running on top of the SPIRIT system. This allowed feature points such as the position of the camera, the corners of its field of view, and calibration points visible by it to be localised very accurately in the 3D SPIRIT coordinate system. These steps make it possible to determine which objects should be visible (in the absence of occlusions) from a given camera, and to calculate the projection of 3D Bat system coordinates onto the image plane of that camera with a mean error of a few pixels.

2.2.2 Synchronisation

SPIRIT location events need to be precisely synchronised with associated video frames. The synchronisation can be initialised manually using the buttons on the Bat device. Arrival events for people entering the view area of a camera can be used to perform automatic re-synchronisation by using the visual tracking method and a motion history window to interpolate locations and correlate these to Bat sensor sightings. Together with the camera calibration process described above, this enables data in both the spatial and temporal domain to be translated between the SPIRIT system and the visual information captured by the cameras.

2.2.3 Quality of service

Location information captured by a sensor infrastructure is a limited resource, and variable rates of quality of service are therefore imposed by the SPIRIT scheduler to determine the rate at which location events are generated for a given Bat. The frequency at which Bats are polled is reduced when the device is stationary. An internal accelerometer allows sampling rates to be increased when the Bat is in motion, or if it is to be used as a "3D mouse" to drive particular applications. However, there is some latency before an increased polling rate comes into effect.

2.2.4 Accuracy

The accuracy of the SPIRIT location events is primarily affected by the properties of the sensor technology and the quality of the camera calibration and frame synchronisation. Ultrasound imposes intrinsic limits on update frequency and resolution due to its propagation characteristics and the filtering that is applied to dampen echoes and remove spurious (e.g. multi-path) observations.

2.2.5 Visibility

The SPIRIT world model contains the locations of walls, doors, and windows, thereby making it possible to determine constraints on the environment viewed by each camera to deduce which objects and entities known to the system are likely to be visible by a given camera. A certain amount of occlusion reasoning may also be performed on this basis by computing a spatial ordering and predicting likely regions of overlap between tagged objects. However, there are many aspects of the world model such as furniture and the state of doors (open vs. closed) that are not fully modelled and must therefore be inferred by other means.

2.3 Vision Techniques for Detection, Identification, and Tracking

This section provides an overview of the vision techniques which have been implemented to provide additional information on the sentient computing environment and objects within it.

2.3.1 Skin and face detection

Human skin colour is modelled as a region in HSV space [12]. Histogram equalisation is applied to the entire image (or a target region predicted by other means) and candidate pixels that lie within the HSV subspace are clustered into regions using morphological operators to remove noise. Then face detection methods are applied to candidate regions identified by means of skin colour classification across the whole image or selectively to regions likely to contain faces as determined by the other modalities (i.e. head regions predicted by SPIRIT person observations or blob tracker based appearance models). In the former case, ellipse fitting is applied to the skin clusters and clusters may be split based on how elongated they are. Face detection is applied to PCA-transformed sub-windows of the candidate region at multiple scales.

Two face detection methods were trained: the first uses a generative mixture of Gaussians model trained using Expectation Maximisation and the second consists of polynomial kernel SVM classifiers. In both cases, the classifiers are arranged in a two-level cascade with the first classifier acting as a fast rejection filter for the second classifier, which was trained by incorporating test set mis-classifications into the training set for the second stage. The two classification schemes are combined using simple disjunction of their binary classification decisions. This may increase false positive rates but ensures that fewer faces are missed.

2.3.2 Background modelling and foreground detection

As described [35], the system maintains a background model and foreground motion history which are adapted over time. A motion history matrix M_t is used to identify a background image bim_t of pixels undergoing sufficiently slow change which can then be used to reliably update the background model B_t and estimate its variance. Pixels are deemed to be part of the dynamic foreground if they exceed a difference threshold that is a multiple of the background variance σ_t^B, and if they are not deemed to be part of a shadow as determined by the DNM1 algorithm (see [35, 34]). Following these steps, candidate foreground pixels are subjected to morphological operations (dilation and erosion) to reduce noise in the final estimate.

2.3.3 Blob analysis and tracking

Foreground pixels are clustered using connected component analysis to identify moving regions ("blobs"). These are then parameterised using shape (bounding box, centre of gravity, major axis orientation) and appearance measures as described in section 3. Blobs are tracked using a Kalman filter or Condensation tracker with a second order motion model. Tracked objects are matched to detected blobs using a weighted dissimilarity metric which takes into account differences in predicted object location vs. blob location and changes in shape and appearance.

2.3.4 Occlusion reasoning

To make tracking more robust, the object to blob assignment stage features a Bayesian network for reasoning about occlusions and object interactions based on observed or predicted overlap of object bounding boxes and failures of object assignment. Dynamic occlusions can also be disambiguated by using 3D SPIRIT data to predict spatial ordering of tracked objects, while static occlusions, object arrivals, and object departures can often be resolved with reference to the world model.

2.3.5 Static scene segmentation and region classification

The region segmentation facilitates correspondence analysis between the world model and the scene viewed by a camera for environmental analysis and constrained tracking. The segmentation method due to [29] is applied to video frames at infrequent intervals. This method segments images into non-overlapping regions by computing a Canny-style colour edge detector and generating Voronoi seed points from the peaks in the distance transform of the edge image. Regions are grown agglomeratively from seed points with gates on colour difference with respect to the boundary colour and mean colour across the region. A texture model based on discrete ridge features is also used to describe regions in terms of texture feature orientation and density. Sets of properties for size, colour, shape, and texture are computed for each region. These properties are fed into artificial neural network classifiers which have been trained to classify regions into "wood", "cloth", "carpet", and "internal walls". The classifiers were trained on images taken in the SPIRIT office and were found to perform well in identifying areas which would otherwise have been mislabelled (e.g. skin rather than wood) and in identifying furniture and wooden doors.

3 Vision-based Adaptive Appearance Modelling

This section describes how fusion of three different appearance models enables robust tracking of multiple objects on the basis of colour information and by using the visual tracking framework described in section 2.3. In this paper, short-term variation in object colour is modelled non-parametrically using adaptive binning histograms.

Appearance changes at intermediate time scales are represented by semi-parametric (Gaussian mixture) models, while a parametric subspace method (Robust Principal Component Analysis, RPCA [7]) is employed to model long term stable appearance. Fusion of the three models is achieved through particle filtering and the Democratic integration method. It is shown how robust estimation and adaptation of the models both individually and in combination results in improved visual tracking accuracy.

3.1 Adaptive Binning Colour Histogram

Non-parametric density estimation techniques such as histograms assume no functional form for the underlying distribution and are robust to changes in orientation, relative position and occlusion of objects. Their simplicity and versatility make them suitable for modelling appearance over short time scales and during the initialisation phases of GMM and subspace estimation.

The optimal number and width of histogram bins is determined by means of k-means clustering with colour differences computed in the CIELAB space using the CIE94 distance d_{kp}. The clustering is repeated n times or until no pixels are left unclustered. Matching of tracked objects with candidate blobs is performed using weighted correlation. The similarity between two histogram bins is calculated by using a weighted product of the bin counts $H[i]$ and $H[j]$, where the weight w_{ij} is determined from the volume of intersection V_s between the two bins. Since the CIELAB space is perceptually uniform, histogram bins are spherical with radius r and so $V_s = V - \pi r^2 d + \frac{\pi}{12} d^3$ and $V = \frac{4}{3}\pi r^3$ where d is the distance between the bin centroids, therefore

$$w_{ij} = \frac{V_s}{V} = \begin{cases} 1 - \frac{3}{4}\frac{d}{r} + \frac{1}{16}(\frac{d}{r})^3 & \text{if } 0 \leq \frac{d}{r} \leq 2 \\ 0 & \text{otherwise} \end{cases} \quad ; \quad w_{ij} \in [0,1] \tag{1}$$

Histogram dissimilarity is then given by $D_{pq} = 1 - \sum_{i=1}^{n} \sum_{j=1}^{n'} w_{ij} H_p[i] H_q[j]$ and normalised such that $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} H_p[i] H_p[j] = \sum_{i=1}^{n'} \sum_{j=1}^{n'} w_{ij} H_q[i] H_q[j] = 1$ where n and n' are the number of bins. In order to incorporate some longer term appearance variation and smooth over fluctuations, the histograms are adapted using exponential averaging. Given a colour histogram H_t calculated for a blob at frame t and a smoothed object colour histogram S_{t-1} from frame $t-1$, the new smoothed object colour histogram S_t for frame t is given by $S_t = \alpha H_t + (1-\alpha) S_{t-1}$ where $\alpha = 1 - e^{-\frac{1}{\lambda}}$ determines the rate of adaptation. This is set to increase with increasing object speed in order to keep track of rapidly moving objects.

3.2 Gaussian Mixture Model

The conditional density for a pixel $\boldsymbol{\psi}$ belonging to an object O can be represented by a mixture of M Gaussians: $P(\boldsymbol{\psi}|O) = \sum_{j=1}^{M} P(\boldsymbol{\psi}|j)\pi(j)$; $\sum_{j=1}^{M} \pi(j) = 1$; $0 \leq \pi(j) \leq 1$ where the mixture parameters $\pi(j)$ give the prior probability that O was generated by the jth component and each mixture component is a Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$.

In this paper, mixture modelling is performed in Hue-Saturation (HS) space to gain a degree of illumination invariance. Model estimation is performed on blob pixels using subsampling for efficiency and discarding samples whose intensity value is very low or close to saturation. Components are estimated using k-means with priors computed from the proportion of samples in each cluster. The parameters of the Gaussians (mean and covariance) are calculated from the clusters. Model order selection is performed using cross validation on a training and validation set randomly selected from the pixel samples. The training set is used to train a number of models of different order M by means of the Expectation Maximisation algorithm. The process terminates once a maximum in the likelihood function is found or the maximum number of iterations has been exceeded.

Adaptation of the GMM over time is performed using the approach suggested in [24]. Matching of blobs to objects is performed by calculating the blob's normalised data log-likelihood $\mathcal{L} = \frac{1}{N^{(t)}} \sum_{\xi \in X^{(t)}} \log P(\xi|O)$ with respect to the object's GMM from a sample of blob pixels $X^{(t)}$ in the current frame. The log-likelihood threshold g for accepting a match is adapted over time to take into account current and previous log-likelihoods. Given an array of n most recent data log-likelihoods calculated for the previous n frames, the threshold is set to $g = \upsilon - k\sigma$, where υ is the median and σ is the standard deviation of the previous n data log-likelihood values.

3.3 Robust Principal Component Analysis

In order to acquire a stable model of object appearance over longer timescales, an extension of the Robust Principal Component Analysis (RPCA) method proposed by De la Torre and Black [8] is applied. Given a matrix $\mathbf{D} = [\mathbf{d}_1\mathbf{d}_2 \dots \mathbf{d}_n]$ whose column vectors $\mathbf{d}_i$ represent images each containing d pixels, the purpose of PCA is to find a lower dimensional subspace of k principal components $\mathbf{B} = [\mathbf{b}_1\mathbf{b}_2 \dots \mathbf{b}_k]$ $(k \leq n)$ such that each image $\mathbf{d}_i$ can be approximated by $\mathbf{d}_i^{\sim} = \mathbf{B}\mathbf{B}^T\mathbf{d}_i = \mathbf{B}\mathbf{c}_i$, where $\mathbf{c}_i$ are linear coefficients obtained by projecting the data onto the subspace, i.e. $\mathbf{C} = [\mathbf{c}_1\mathbf{c}_2 \dots \mathbf{c}_n] = \mathbf{B}^T\mathbf{D}$. PCA can be formulated as least-squares estimation of the basis images $\mathbf{B}$ by minimising

$$E_{pca}(\mathbf{B}) = \sum_{i=1}^{n} e_{pca}(\mathbf{e}_i) = \sum_{i=1}^{n} \left\| \mathbf{d}_i - \mathbf{B}\mathbf{B}^T\mathbf{d}_i \right\|_2 = \sum_{i=1}^{n}\sum_{p=1}^{d} \left(d_{pi} - \sum_{j=1}^{k} b_{pj}c_{ji} \right)^2 \tag{2}$$

where $c_{ji} = \sum_{t=1}^{d} b_{tj}d_{ti}$, $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, and $e_{pca}(\mathbf{e}_i) = \mathbf{e}_i^T\mathbf{e}_i$ is the reconstruction error of $\mathbf{d}_i$.

RPCA enhances standard PCA by means of a pixel outlier process using M-estimators. the error equation above is reformulated to obtain RPCA robust mean $\boldsymbol{\mu}$, bases $\mathbf{B}$, and coefficients $\mathbf{C}$:

$$E_{rpca}(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} e_{rpca}\left(\mathbf{d}_i - \boldsymbol{\mu} - \mathbf{B}\mathbf{c}_i\ ,\ \ \boldsymbol{\sigma}\right) = \sum_{i=1}^{n}\sum_{p=1}^{d} \rho\left(d_{pi} - \boldsymbol{\mu}_p - \sum_{j=1}^{k} b_{pj}c_{ji}\ ,\ \ \sigma_p \right) \tag{3}$$

where ρ is the Geman-McClure error function $\rho(x,\ \sigma_p) = \frac{x^2}{x^2+\sigma_p^2}$, and σ_p is a scale parameter that controls convexity and hence determines which residual errors are treated as outliers. To compute the mean and the subspace spanned by the first k principal components robustly, equation 3 is minimised using gradient descent with a local quadratic approximation.

To ensure adequate performance for tracking, RPCA has been extended in this work using a robust incremental subspace learning technique to efficiently re-compute the Eigenspace (see below). In addition, rather than computing RPCA over image intensity alone, RPCA was applied to one-dimensional colour statistics histograms derived from the colour distribution of each object in HSV space. Following Hanbury [15], the saturation-weighted hue mean histogram $H_{S\ell}$ (where hue H is measured as an angle in the range $\{0°, 1°, \dots, 360°\}$) is calculated at each sample luminance level $\ell \in \{0, 1, 2, \dots, N\}$:

$$H_{S\ell} = \arctan\left(\frac{\sum_x S_x\ \sin\ H_x \delta_{L_x \ell}}{\sum_x S_x\ \cos\ H_x \delta_{L_x \ell}} \right) \tag{4}$$

where H_x, S_x, and L_x are the hue, saturation, and luminance values at pixel location x.

The number of pixels in the sample sets for Eigenspace computation was normalised by sub-sampling (and if necessary re-sampling) object pixels or through normalisation of the colour statistics histograms. Re-estimation of the RPCA coefficients is performed incrementally by adapting the method proposed in [30]. Given the current RPCA robust mean $\boldsymbol{\mu}^{(t)}$, bases $\mathbf{B}^{(t)}$, coefficients $\mathbf{C}^{(t)}$, and data sample $\mathbf{x}$, then at each frame t the algorithm proceeds as follows:

1. Project the data sample $\mathbf{x}$ into the current Eigenspace $\mathbf{B}^{(t)}$ and form the reconstruction $\mathbf{y}$ of the data: $\mathbf{c} = \mathbf{B}^{(t)\,T}(\mathbf{x} - \boldsymbol{\mu}^{(t)})$; $\quad \mathbf{y} = \mathbf{B}^{(t)}\mathbf{c} + \boldsymbol{\mu}^{(t)}$

2. Compute the residual vector $\mathbf{r} = \mathbf{x} - \mathbf{y}$, which is orthogonal to $\mathbf{B}^{(t)}$, and form matrices $\mathbf{B}_e$ and $\mathbf{C}_e$: $\mathbf{B}_e = \left[\mathbf{B}^{(t)}\ \frac{\mathbf{r}}{||\mathbf{r}||}\right]$; $\quad \mathbf{C}_e = \begin{bmatrix} \mathbf{C}^{(t)} & \mathbf{c} \\ \mathbf{0} & ||\mathbf{r}|| \end{bmatrix}$

3. Compute Robust PCA on $\mathbf{C}_e$, and obtain the updated robust mean $\boldsymbol{\mu}_s$ and robust bases $\mathbf{B}_s$. Discard the least significant Eigenvector of the new basis $\mathbf{B}_s$ and obtain the coefficient matrix for frame $t+1$: $\mathbf{C}^{(t+1)} = \mathbf{B}_s^T(\mathbf{C}_e - \boldsymbol{\mu}_s \mathbf{1}_{1\times(t+1)})$ where $\mathbf{1}_{m\times n}$ denotes a matrix whose elements have the value 1.

4. Calculate the new basis matrix $\mathbf{B}^{(t+1)} = \mathbf{B}_e\mathbf{B}_s$ and new mean $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \mathbf{B}_e\boldsymbol{\mu}_s$ for frame $t+1$.

In order to compute the match distance between a candidate blob represented by a column vector $\mathbf{e}$ of pixel samples and an object represented by RPCA parameters ($\mathbf{B}$, $\mathbf{C}$, $\boldsymbol{\mu}$), $\mathbf{e}$ is projected into the RPCA subspace. This is achieved by computing projection coefficients $\tilde{\mathbf{c}}$ which minimise $E(\tilde{\mathbf{c}}) = \sum_{j=1}^{d} \rho\left(e_j - \mu_j - \sum_{i=1}^{k} b_{ji}\tilde{c}_i\ ,\ \ \sigma_j\right)$, where ρ is the Geman-McClure error function. The object-blob distance is then defined as the minimum of the Euclidean distances between the blob coefficients $\tilde{\mathbf{c}}$ and each column of the object RPCA coefficient matrix $\mathbf{C}$.

3.4 Integration through Condensation

Particle filtering algorithms such as Condensation (conditional density propagation, [22]) pose the problem of tracking as estimation of states $\mathbf{X}$ from observations $\mathbf{Z}$ using the recursion

$$P(\mathbf{X}_t|\mathbf{Z}_t) \propto \mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \int P(\mathbf{X}_t|\mathbf{X}_{t-1})P(\mathbf{X}_{t-1}|\mathbf{Z}_{t-1})d\mathbf{X}_{t-1}$$

where the dynamical model $P(\mathbf{X}_t|\mathbf{X}_{t-1})$ describes state evolution and the observation likelihood model $\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t)$ gives the likelihood of any state in light of current observations.

The posterior probability distribution is then represented by a weighted set of "particles" $P(\mathbf{X}_t|\mathbf{Z}_t) = \{s_t^{(n)}, \pi_t^{(n)}\}$, where $s_t^{(n)}$ is the nth sample and $\pi_t^{(n)}$ is the corresponding weight such that $\sum_n \pi^{(n)} = 1$. At each step of the Condensation algorithm the evolution of the weighted sample set is calculated by applying the dynamical model to the set. The observation likelihood function is then used to correct the prediction by calculating the weight π_t of each element in the set, i.e. $\pi_t \propto \mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$. N samples are then drawn with replacement by choosing a particular sample with probability $\pi^{(n)} = P(Z_t|X_t = s_t^{(n)})$. The mean state vector of an object in frame t is then modelled as the expectation $E[S] = \sum_{n=1}^{N} \pi^{(n)} s^{(n)}$.

Here, the observation density is modelled by a function that contains Gaussian peaks where the observation density is assumed to be high, that is, where an object could have generated a set of blobs with high probability. Each Gaussian peak corresponds to the position of a blob, and the peak is scaled by the object-blob distance. The likelihood $\mathcal{L}$ for a particle is computed as $\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \propto e^{-k\times\mathrm{dist}^2}$, where dist is a distance under one of the appearance models of the local image patch at a given particle and the object under consideration, and k is a constant. Likelihoods are calculated for each particle for each of the three appearance modelling schemes above and combined using:

$$\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \propto [\mathcal{L}_{rpca}(\mathbf{Z}_t|\mathbf{X_t})]^{\alpha_1} [\mathcal{L}_{chist}(\mathbf{Z}_t|\mathbf{X}_t)]^{\alpha_2} [\mathcal{L}_{gmm}(\mathbf{Z}_t|\mathbf{X}_t)]^{\alpha_3} \tag{5}$$

where $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ are the reliability weights for each appearance model, initialised to $\frac{1}{3}$.

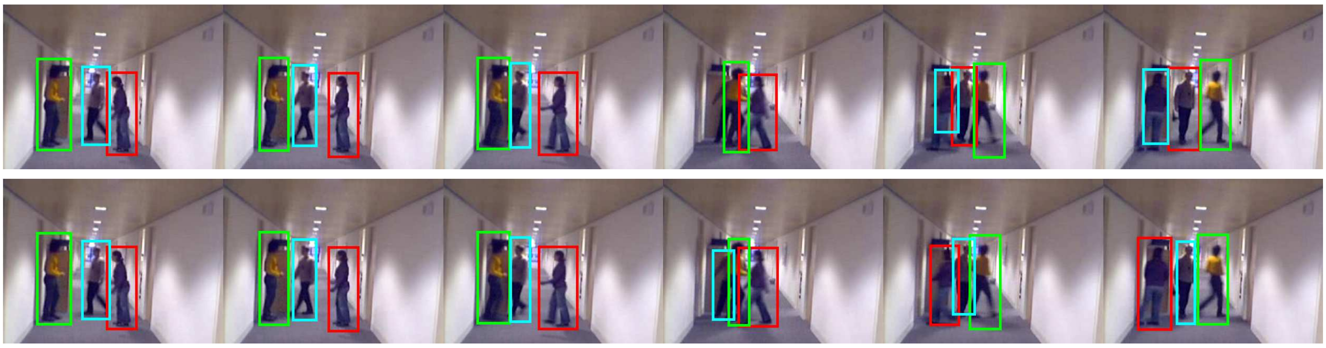

Figure 3: **Indoor tracking results using vision cues only. The rectangles indicate bounding boxes of tracked objects and their colour indicates object identity assignment throughout the sequence.** ***Top*****: tracking using only blob features and distances.** ***Bottom*****: tracking using the robust fusion of adaptive appearance models as described in section 3. Note how this allows identity of tracked entities (indicated by bounding box colour) to be maintained during and across occlusions.**

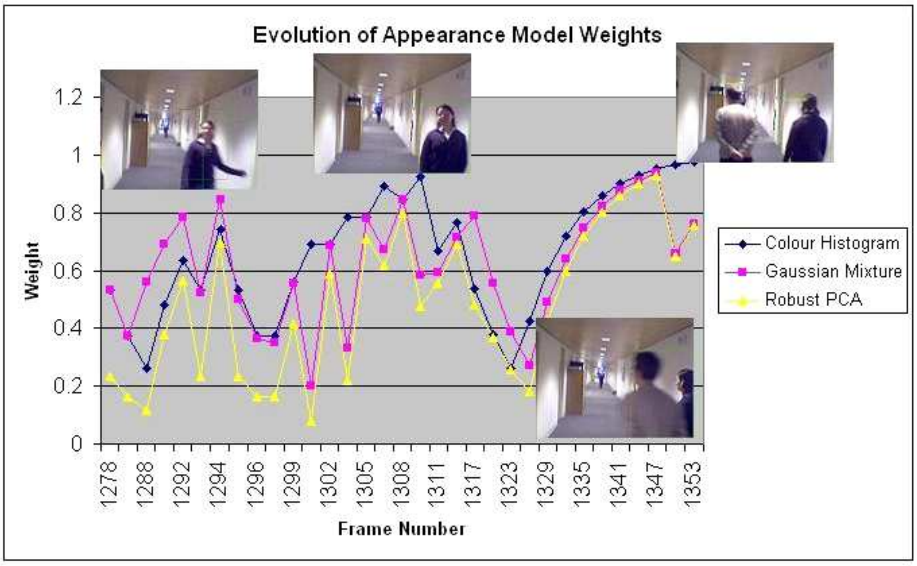

Figure 4: **Evolution of reliability weights and object colour appearance models.** ***Left*****: graph plotting the reliabilities of the appearance model cues for the woman shown in the test sequence. There is an initial rise in the reliability of all models due to the clear visibility of the woman. The large fall in reliability at frame 1320 onwards is due to occlusion by the man entering the scene. After the occlusion the appearance models successfully recover and their reliability increases very rapidly. Note the lag of the RPCA (and in some cases the Gaussian mixture) model behind the colour histogram model due to their slower adaptation.**

3.5 Adaptation of Cue Weights by Democratic Integration

Adaptation of the weights in equation 5 is performed dynamically during tracking by extending the idea of Democratic integration [31] to the Condensation framework. Four separate observation likelihoods are computed: one for the joint appearance model, and three for each of the RPCA, adaptive histogram and GMM appearance cues. Condensation is performed separately for each of the observation functions, resulting in four hypotheses, R_{fused}, R_{rpca}, R_{chist}, and R_{gmm}, which are regions where the object is thought to be in the current frame. Each region centroid is obtained by computing the expectation of the respective particle sets for each cue.

The Euclidean distances $E_{k,t}$ between the centroid of R_{fused} and the centroids of R_{rpca}, R_{chist}, R_{gmm} at time t are then calculated. Since the joint observation function is assumed to exhibit the best performance, appearance cues that result in relatively large values of $E_{k,t}$ are considered less reliable in the current frame, and their reliability weight is lowered accordingly. A score $\gamma_{k,t}$ is computed for each cue k as follows:

$$\gamma_{k,t} = \frac{\tanh(-aE_{k,t} + b) + 1}{2} \tag{6}$$

where a, b are constants (set to 2 and 5 respectively) and tanh is the hyperbolic tangent function. Given $\gamma_{k,t}$, the weights $\alpha_{k,t}$ for each cue k are then adapted using first order exponential averaging, i.e. $\alpha_{k,t+1} = \beta\gamma_{k,t} + (1-\beta)\alpha_{k,t}$, where β controls the rate of adaptation (setting $\beta = 0.75$ was found to give good results in most sequences). Performing the Condensation algorithm four times during each frame was found not to be a bottleneck since most of the computation time is required for the particle distances (which need only be computed once per frame).

3.6 Results

To evaluate the adaptive appearance models and the fusion mechanism discussed above, testing was carried out on a number of indoor surveillance sequences which were acquired in the same office in which the SPIRIT system has been deployed (see section 4.2). The tracking conditions are especially demanding due to the presence of intermittent bright lighting, skin coloured walls, motion blur and occlusions as the people interact. Figure 3 shows how the fusion framework makes tracking robust with respect to occlusions and movements of people (the results shown are for sequence *S7* discussed in section 4.2). In figure 4 it is shown how the appearance modelling improves accuracy in light of erroneous blob hypotheses generated by the background differencing and blob detection framework. Quantitative results for multi-hypothesis tracking using visual cues alone are presented in section 5 and contrasted with those achievable through integration of SPIRIT information as described in section 4.

4 Multi-modal and Multi-sensory Fusion

4.1 The World Model

The SPIRIT system maintains an internal dynamic representation of the office environment including objects and events within it. This world model [16] comprises a static part consisting of those aspects of the environment which are not monitored using the ultrasonic location system, and a dynamic part consisting of those objects and devices which are. The former includes the location and spatial extent of rooms, walls, windows, doors, and items of furniture such as desks and shelves that were manually added to the model. The latter tracks personnel, visitors, portable devices, and other assets that have been tagged with one or more of the Bats.

The role of the sensor systems is to keep the model consistent and accurate. As shown in figure 5, applications

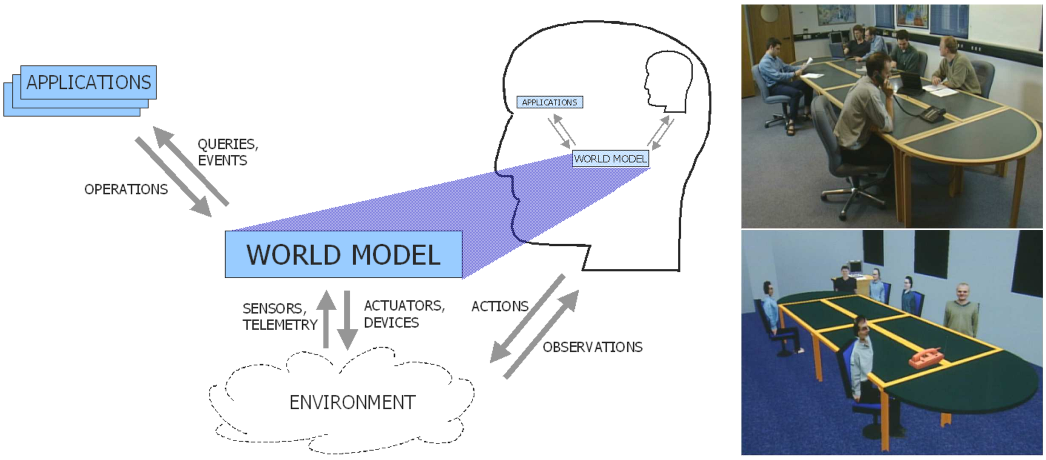

Figure 5: *Left:* **Diagrammatic overview of the world model maintained by the sentient computing system.** *Right:* **The world perceived by (top) users and (bottom) the sentient computing system (LabView visualisation).**

see a description of the environment that is abstracted away from the sensor level. The interpretation of such information is application dependent, for example routing a phone call may require finding the phone that is closest to a given person (provided they are in the same room), whereas a "follow-me" desktop session would need to know something about the user's orientation relative to available screens in order to select the best one for display.

4.2 Experimental Setup

To facilitate the integration of visual information into the world model, cameras were deployed in various parts of the sentient office (see map in figure 2), namely one at the western entrance, one facing east along the corridor, and two in the meeting room (3rd room from the left at the bottom part of the map). Additional experimental sequences were also taken in some of the other rooms of the office. The cameras used were two standard Philips webcams yielding a picture resolution and frame rate of (320x240 pixels, 12fps) and (640x480 pixels, 15fps) respectively.

A number of sequences (S1,..., S10) featuring a total of 7 individuals were acquired to experiment with different scenarios and activities, such as one or more people entering or leaving a room, holding a meeting, drawing on the whiteboard, walking in front of each other, etc.. Several thousand frames from 10 sequences were manually annotated by marking the bounding boxes of peoples' bodies and heads visible within each frame. For reasons of time, not every sequence was exhaustively annotated. In most cases the annotations are limited to a subset of the actual footage and/or only label every 5th frame, which is the also the rate at which the visual analysis is usually performed.

4.3 Multi-hypothesis Bayesian Modality Fusion

A viable multi-modal fusion method must generate reliable results that improve upon the individual modalities, while maintaining their fidelity and uncertainty information for higher-level processing and adaptation of the fusion strategy. The approach taken here is essentially a development of Bayesian Modality Fusion [36, 28] for multi-object tracking. It uses a Bayesian graphical network (shown in figure 6) to integrate information from the different sources. Discrete reliability indicator variables (R_S, R_F, R_D, R_C, and R_B) are used to model how reliable each modality is at the current time. At present each variable may take on one of the values "low", "normal", and "high". The network serves as a shared template from which individual tracking hypotheses are derived. Hypotheses are instantiated by SPIRIT observations or the blob tracking framework, thus allowing tracking of people who are not tagged with a functioning Bat device or who are not currently visible by a given camera. Other visual cues such as skin colour and face detection serve as supporting modalities. Spatial and object-specific ontologies from the world model or the region segmentation and classification methods provide contextual constraints and guide the generation of hypotheses.

Reliabilities are adapted on the basis of manually specified rules over reliability indicators, such as motion and appearance variation, and performance feedback measures, such as consistency and log-likelihood of the observations under each modality. Posterior probabilities for each hypothesis can then be computed by integrating all available information using the fusion network. The position and spatial extent of tracked people is computed by reliability-weighted interpolation of the object bounding box deduced from the SPIRIT observation (if available for the current observation) and blob tracker appearance model.

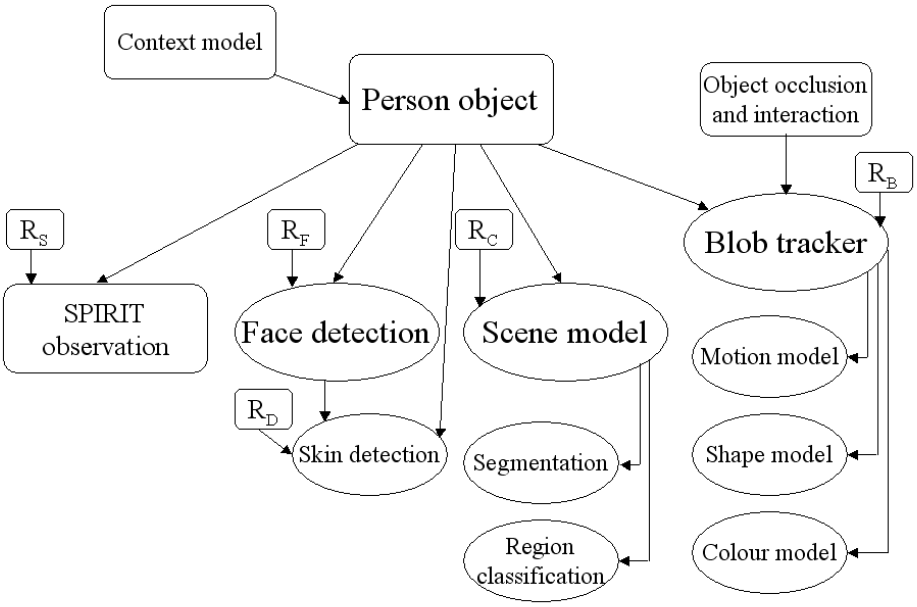

Figure 6: **Bayesian graphical model for multi-modal fusion. Reliability variables allow adaptive integration of different sources of information.**

Each hypothesis maintains its own set of reliability variables and competes for observations with other hypotheses. The conditional probabilities (including the dependency on reliability variables) of the underlying network structure were initially set by hand, but have since been re-estimated by means of the EM algorithm on statistics gathered from manually labelled training sequences consisting of over 3000 frames. Temporal evolution of the model occurs via a Kalman or particle filter applied to the colour blob tracker and through the modification of reliability variables in light of current observations. This update stage introduces a coupling of the observation models for the individual modalities. Some results from a meeting room sequence are shown in figure 8.

SPIRIT observations consist of projected 3D positions of Bat sensor locations together with information on the type and identity of the observed object as available from the SPIRIT world model. The world model contains

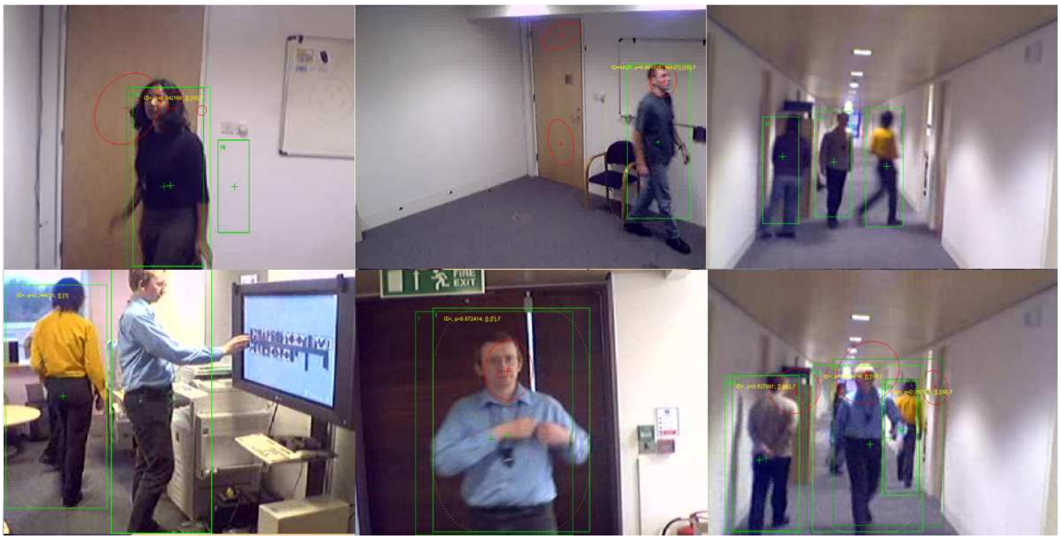

Figure 7: **Example images from the test sequences. Green rectangles indicate bounding boxes of blob tracker objects, green polygons are the convex hulls of corresponding blob regions. Red ellipses are skin clusters, those with a red cross were deemed to contain a face. The cyan and magenta coloured rectangles are predicted object locations based on SPIRIT observations (magenta indicates an older observation). Yellow dotted ellipses and numbers indicate hypotheses resulting from the fusion process described below.**

information about people's height and the probable position of their Bat on their body, and hence the projective mapping of the person's likely silhouette onto the camera's image plane can be calculated. Location events are generally quite accurate but are assigned a reduced reliability if they are not well synchronised with the current frame or if the Bat has been undergoing rapid motion.

The reliability of tracked blobs depends on the correspondence between predicted and actual position and appearance dissimilarity. Face detection can be a powerful cue for head position but becomes unreliable when there is too much variation in appearance due to movement, occlusions, or changes in posture. Evidence for false positives consists of detections in regions of high motion energy or areas where there is no expectation of faces being observed, i.e. where the other modalities fail to hypothesise the appearance of a person. This is particularly the case for areas of skin colour (such as a wooden door or table) where one or more faces have been detected, but which are unlikely to coincide with the appearance of a human head due to their position or shape. Conversely, face detections in regions where head presence is predicted by other modalities lead to an increase in reliability of the face cue for the given hypothesis. Skin detections may be used to model other body parts such as hands and legs. The scene model serves to disambiguate skin detections by dictating low reliability in regions that are likely to lead to false detection, e.g. wood. The scene model consists of the static region segmentation of each frame and the neural network classifications of each region. Areas of high motion energy lead to blur which degrades the reliability of the segmentation and classification. Segmentation is also unreliable when similarly coloured objects overlap.

The integration process computes a probability for each tracked object given the current observations and reliabilities. Person position is computed by weighted interpolation of the object bounding box deduced from the SPIRIT observation and blob tracker object currently associated with a given hypothesis, taking into account their respective reliabilities. Skin colour, face detections, and the scene model serve as supporting modalities, whereas the output of the SPIRIT and blob tracker maintain object identity and can serve as instantiating modalities, i.e. a new object hypothesis must be supported by either a tracked colour blob or a SPIRIT observation (see figure 7). In the latter case both location events for people and devices assigned to a particular person can be used. People generally wear their Bat sensor at a particular calibrated position on their body, and together with the known distance and projective properties of the camera this can be used to instantiate an expected 2D occupancy region for the person, even if no useful blob tracker can be assigned to the hypothesis at present. Face detections contribute to the combined probability if they occur within the upper third of the bounding box, and skin colour contributes if it is found anywhere within this region. Objects that are tracked only on the basis of SPIRIT information but don't appear to be visible in the current frame continue (for a while) to be represented by an hypothesis whose likelihood is adjusted according to the occlusion reasoning described above.

Hypotheses can be viewed as competing for observations, since each visual cue and SPIRIT location event may only be used to support one tracked object in a given frame. Hypotheses are removed after their probability has dropped below a threshold for a certain number of frames, but may be revived if a new SPIRIT location event tagged with the same object ID occurs. New hypotheses are instantiated in response to blob or Bat observations. Due to the relative brittleness of the visual cues alone, new hypotheses are given a low initial probability until they have been "confirmed" by a SPIRIT observation or have proven stable over several frames. This allows people who are not wearing a functioning Bat device to be tracked. SPIRIT data is also particularly valuable in maintaining object identity across occlusions (although in some cases this is also possible on the basis of blob appearance and predicted motion), and to generate expectations (expressed as hypotheses) for people who are about to enter the visual field of a camera. Hence, the Bat system and the federation of visual cues may each serve to guide the combined system's focus of attention by instantiating hypotheses and generating expectations.

5 Enhanced Tracking

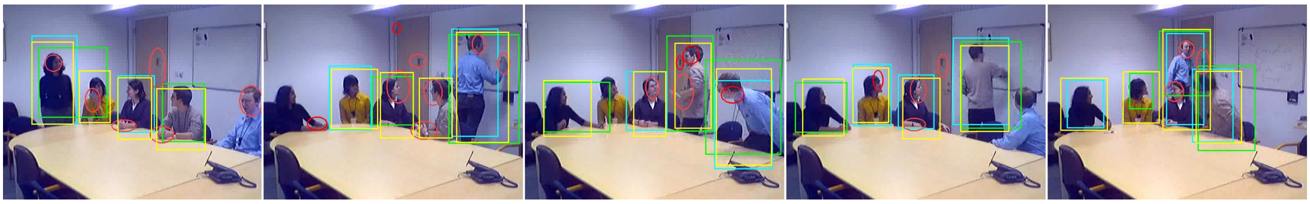

Figure 8: **Examples of tracking results obtained for sequence S5: Rectangles denote object hypotheses derived from the two modalities (green: visual tracking, blue: SPIRIT observations) and the resulting fused hypothesis (yellow). Red ellipses indicate face detections.**

Using the combined tracking framework, position estimates can be made more robust and accurate. As described in section 4.3, this can be achieved through Bayesian multi-modal fusion. Figure 8 shows sample results for a meeting scenario with multiple participants. As indicated below, additional information apart from a person's location can be inferred through the joint integration of the various perceptual cues.

As described in section 4.2, several sequences were manually annotated with ground truth information in order to analyse tracking performance. Figures 9, 10, 11 and 12 show performance data for sequences S2, S3, S5, and S6 respectively. For each sequence, results are shown which compare performance when tracking is performed using the two modalities on their own (i.e. only vision or only SPIRIT information) and for the fusion method described

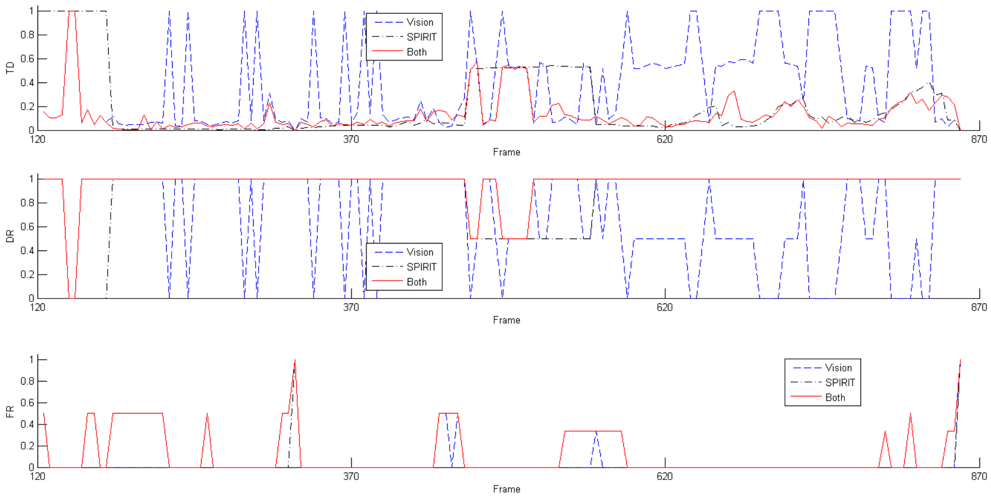

Figure 9: **Comparative tracking results for test sequence S2 when using the two modalities, both in isolation and combined by means of the fusion process.** *Top:* **mean distance-from-track** TD; *Middle:* **detection rate** DR; *Bottom:* **false positive rate** FR. **In each case, the solid red line shows the value of the given performance measure for the outcome of the fusion method while the blue dashed and black dotted lines indicate results when using the vision and SPIRIT modalities in isolation respectively.**

above. The performance measures plotted are the mean distance-from-track TD, the detection rate DR, and the false positive rate FR, computed for each frame in the sequence. Consequently, a value of DR close to 1 indicates that all objects are being tracked in a given frame while FR close to 0 means that there are few false positives (spurious instances of objects which do not correspond to any objects marked in the ground truth for that frame). The measure TD characterises the mean accuracy of object tracks in terms of the distance between the centres of gravity of tracked objects and ground truth objects (which takes a value of 1 if no such correspondence can be established).

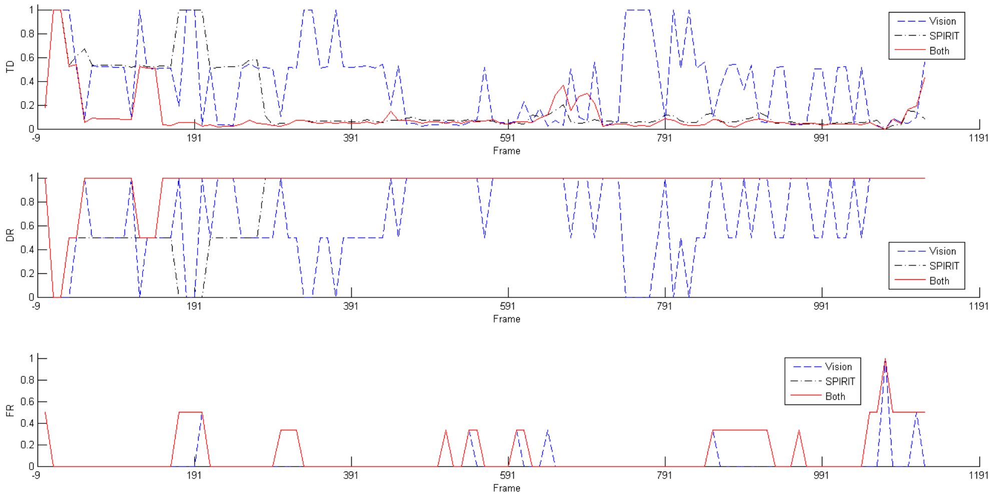

Figure 10: **Comparative tracking results for test sequence S3.**

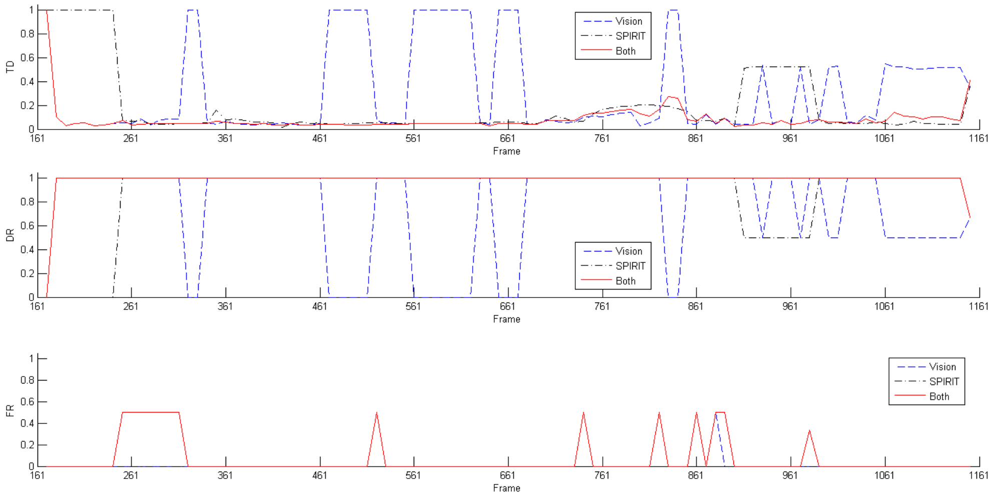

Figure 11: **Comparative tracking results for test sequence S5.**

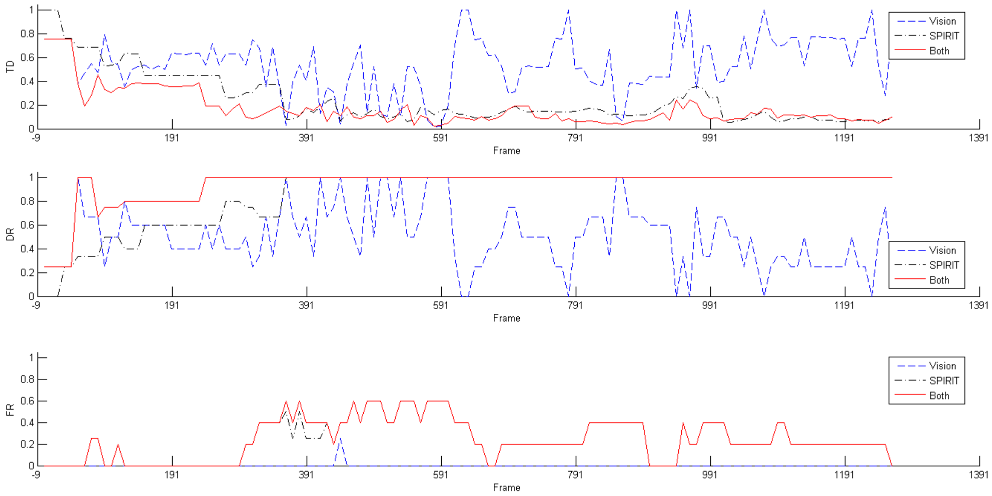

Figure 12: **Comparative tracking results for test sequence S6.**

In order to summarise these results, figures for overall recall and precision are shown in table 1, where $\mathit{Recall} = \text{mean}(\text{DR})$ and $\mathit{Precision} = \text{mean}(N_{\text{tp}}/(N_{\text{tp}} + N_{\text{fp}}))$ (N_{tp} is the number of true positives and N_{fp} the number of false positives for each processed frame in the sequence). As can be seen from the results, tracker accuracy and performance are generally enhanced by the combined fusion process compared to the outcome using vision modules or only SPIRIT observations, especially in difficult situations such as object-object occlusions. The system can exploit the multi-modal redundancies to successfully track objects that are only detected by one of the tracking components. Reliability indicators allow the system to discount modalities that fail in particular circumstances and rely on those which are likely to give accurate results, thus ensuring that the fusion process delivers results that are as good and sometimes better than those of the modality which performs best at a given time.

However, these results show that the fusion method sometimes incurs in a slightly increased false positive rate compared to either of the component modalities, which may lead to a reduction in precision alongside significant

Modality	*Vision*		*SPIRIT*		*Fusion*	
Sequence	Recall	Precision	Recall	Precision	Recall	Precision
S2	0.674	0.907	0.857	0.976	0.963	0.852
S3	0.673	0.931	0.845	0.933	0.960	0.868
S5	0.729	0.948	0.875	1.000	0.987	0.906
S6	0.501	0.996	0.860	0.731	0.943	0.747

Table 1: **Overall mean recall and precision for test sequences S2, S3, S5, and S6. For each sequence, recall and precision were computed by comparing tracking results obtained using vision and SPIRIT information (both in isolation and combined through Bayesian fusion) with manually labelled ground truth annotations.**

Sequence	Recall	Precision	p_thresh	p_retention	fobj_retention
S2	0.979	0.974	0.75	0.20	10
S3	0.956	0.940	0.70	0.20	5
S5	0.882	0.992	0.75	0.20	10
S6	0.952	0.730	0.40	0.30	5

Table 2: **Recall achieved by the fusion method at a comparable level of precision as that yielded by the SPIRIT modality on its own as shown in table 1. The values of the three variables affecting the fusion process that were optimised to achieve these results are also shown.**

increases in recall. It is possible to optimise the precision-recall trade-off to best suit the requirements of particular applications of the sentient computing environment. For example, security applications are likely to require high recall whereas tracking for human computer interaction may need high precision.

To analyse how this can be done, some of the parameters affecting the fusion process were optimised with the goal of assessing the level of recall achievable by the fusion method at the same level of precision as that of the SPIRIT system as shown in table 1. No modifications were made to the Bayesian fusion network or any other aspects of the system, but 20 different combinations of values were evaluated for the three internal software variables "p_thresh", "p_retention", and "fobj_retention". In the case of results shown in table 1, these variables had been set to 0.40, 0.30, and 10 respectively. The parameter "p_thresh" specifies the minimum probability (as calculated by the Bayesian network in figure 6) that a hypothesis must satisfy in order to be regarded as a tracked object. In order to enable tracking of objects whose hypotheses have temporarily dropped below this threshold, "p_retention" specifies the minimum probability which must be associated with a tracked object hypothesis in order for it to continue to be tracked. "fobj_retention" specifies the maximum number of frames during which a hypothesis may be retained in this way before it is discarded, unless its probability once again rises above "p_thresh".

Table 2 shows resulting recall and precision values together with the values of the three aforementioned variables that were selected to bring the fusion system's precision as close as possible to that of the SPIRIT modality for each of the four test sequences. It can be seen that the fusion method exhibits greater accuracy (as measured by recall) than the Bat system for comparable levels of precision.

6 Summary

As computer vision continues to mature, it is likely to play an increasingly important role in the rapidly growing field of ubiquitous computing. This paper presents a novel approach to harnessing the benefits of computer vision within the context of a sentient computing system deployed throughout an office space. It is shown how different computer vision methods such as tracking algorithms and appearance models can be fused with information from an ultrasonic tracking system to significantly augment the capabilities and robustness of the system's world model.

The sentient computing system provides a variable granularity spatial model of the environment and a reliable device tracking facility that can be used to automatically (re)initialise and re-focus vision modules whenever an event or scene context of interest is observed by a camera. Changing reliabilities of different sources of information are handled robustly in a unified inference framework based on Bayesian networks. A hierarchy of object and environmental properties is used to integrate different hypotheses about the perceived context. The resulting world model serves as a shared representation of context that is made available to users and applications.

Unlike related approaches to Bayesian Modality Fusion [36, 28] for multi-object tracking, this approach is capable of incorporating both 2D and 3D visual and non-visual data and does not rely on an expensive space partitioning. The number of tracked objects and the size of the modelled environment are not fixed, object arrivals and departures are handled robustly, and the approach scales smoothly as additional sensor resources (cameras or Bat tags) are added or become unavailable. Unlike fusion methods such as Kalman filtering or Condensation, high-dimensional inputs can be modelled and both symbolic structural information and continuous data can be integrated into the Bayesian network.

Nevertheless, as explained in section 3, the Condensation algorithm proved effective for fusion of the visual appearance models and to integrate such information with target positions estimated by the vision-based tracker. Democratic Integration is an efficient yet simple method for adapting the weights used in uni-modal fusion of the visual cues. However, these methods are only used to perform vision-based fused to maintain hypotheses associated with particular objects as required by particular applications such as a security camera near the entrance of the office. Unlike the Bayesian framework described in section 4, they cannot easily handle differences reliabilities and accuracies associated with different sensor modalities and are too slow to be suitable for constrained real-time tracking in a large-scale distributed multi-sensory environment.

A number of applications of the sentient computing technology can in turn benefit from the video interpretation framework through the fusion of the ultrasonic and visual modalities. As further described in [34], such applications include pose and orientation estimation, biometric security, recognition of static elements of the environment such as whiteboards and furniture, and enhanced visualisation for applications such as teleconferencing and shared augmented reality environments. Having visual information as an additional sensory modality is useful when the SPIRIT system has trouble detecting a person (e.g. they are not wearing a Bat or it is temporarily concealed), or when an application requires additional information about a person's posture, direction of gaze, gestures, interactions with devices and other people, or facial expression.

To ensure sufficient performance to enable real-time processing, the fusion of individual perceptual modalities can be set up as a hierarchy where inexpensive detectors (e.g. finding the rough outline of a person) narrow down the search space to which more specific modules (e.g. a face spotter or gesture recogniser) are applied. The sensor fusion and information integration adds value to both the visual and ultrasonic modality by complementing their capabilities and adapting to error characteristics exhibited by the different sources of information at different times.

Acknowledgements

The author gratefully acknowledges financial support from the Royal Commission for the Exhibition of 1851.

References

[1] M. Addlesee, R. Curwen, S. Hodges, J. Newman, P. Steggles, A. Ward, and A. Hopper. Implementing a sentient computing system. *IEEE Computer*, 34(8):50–56, 2001.

[2] J.-Y. Bouguet. Matlab calibration toolbox. `http://www.vision.caltech.edu/bouguetj`.

[3] P. Cattin, D. Zlatnik, and R. Borer. *Biometric System using Human Gait.* Mechatronics and Machine Vision in Practice (M2VIP), 2001.

[4] M. Cerney and J. Vance. Gesture recognition in virtual environments: A review and framework for future development. Technical report, Human Computer Interaction Center, Iowa State University, 2005.

[5] T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection. In *Proc. Int. Conference on Pattern Recognition*, 2002.

[6] J. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual components for context aware computing. In *Proc. Ubicomp 2002*, 2002.

[7] F. De la Torre and M. Black. Robust principal component analysis for computer vision. In *Proc. International Conference on Computer Vision*, 2001.

[8] F. De la Torre and M. Black. Robust parameterized component analysis: Theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 2003.

[9] A. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001.

[10] T. Erickson. Some problems with the notion of context-aware computing. *Communications of the ACM*, 45(2):102–104, 2002.

[11] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plotz, G. Fink, and G. Sagerer. Multi-modal anchoring for human-robot interaction. *Robotics and Autonomous Systems*, 43(2), 2003.

[12] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.

[13] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[14] Alessandro Genco. Three Step Bluetooth Positioning. In *LNCS 3479: Location- and Context-Awareness*, 2005.

[15] A. Hanbury. Circular statistics applied to colour images. *8th Computer Vision Winter Workshop*, 2003.

[16] R. Harle. *Maintaining World Models in Context-Aware Environments.* PhD thesis, University of Cambridge Engineering Department, 2004.

[17] R. Harle and A. Hopper. Deploying and evaluating a location-aware system. In *Proc. MobiSys2005*, 2005.

[18] A. Harter and A. Hopper. A distributed location system for the active office. *IEEE Network*, 8(1), 1994.

[19] M. Hazas, J. Scott, and J. Krumm. Location-aware computing comes of age. *IEEE Computer*, pages 95–97, February 2004.

[20] A. Hopper. Sentient Computing - The Royal Society Clifford Paterson Lecture. *Philosophical Transactions of the Royal Society of London*, 358(1773):2349–2358, 2000.

[21] D. Ipina and A. Hopper. TRIP: a low-cost vision-based location system for ubiquitous computing. *Personal and Ubiquitous Computing*, 6(3):206–219, 2002.

[22] M. Isard and A Blake. Condensation – conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[23] K. Mansley, A. Beresford, and D. Scott. The Carrot Approach: Encouraging use of location systems. In *Proceedings of UbiComp*. Springer, 2004.

[24] S. McKenna, Y. Raja, and S. Gong. Object tracking using adaptive color mixture models. In *Proc. Asian Conference on Computer Vision*, pages 615–622, 1998.

[25] K. Nummiaro, E. Koller-Meier, and L.V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21:99–110, 2003.

[26] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

[27] N. Priyantha, K. Allen, H. Balakrishnan, and S. J. Teller. The cricket compass for context-aware mobile applications. In *Mobile Computing and Networking*, pages 1–14, 2001.

[28] J. Sherrah and S. Gong. Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects. In *Proc. International Conference on Computer Vision*, 2001.

[29] D. Sinclair. Smooth region structure: folds, domes, bowls, ridges, valleys and slopes. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 389–394, 2000.

[30] D. Skocaj and A. Leonardis. Robust continuous subspace learning and recognition. In *Proc. Int. Electrotechnical and Computer Science Conference*, 2002.

[31] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Lecture Notes in Computer Science*, 2095:93–106, 2001.

[32] S. Stillman and I. Essa. Towards reliable multimodal sensing in aware environments. In *Proc. Perceptual User Interfaces Workshop, ACM UIST 2001*, 2001.

[33] A. Torralba, K. Murphy, W. Freeman, and A. Mark. Context-based vision system for place and object recognition. In *Proc. International Conference on Computer Vision*, 2003.

[34] C.P. Town. *Ontology based Visual Information Processing.* PhD thesis, University of Cambridge, 2004.

[35] C.P. Town. Ontology-driven Bayesian networks for dynamic scene understanding. In *Proc. Int. Workshop on Detection and Recognition of Events in Video (at CVPR04)*, 2004.

[36] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proc. Asian Conference on Computer Vision*, 2000.

[37] M. Turk. Computer vision in the interface. *Communications of the ACM*, 47(1), 2004.