Continuous Global Evidence Based Bayesian Modality Fusion for Simultaneous Tracking of Multiple Objects

Sherrah, Jamie; Gong, Shaogang

QUEEN MARY
AND WESTFIELD COLLEGE
UNIVERSITY OF LONDON

Department of Computer Science

Research Report No. RR-00-06 ISSN 1470-5559 December 2000

Continuous Global Evidence Based Bayesian Modality Fusion for Simultaneous Tracking of Multiple Objects

Jamie Sherrah & Shaogang Gong

Continuous Global Evidence-Based Bayesian Modality Fusion for Simultaneous Tracking of Multiple Objects

Abstract

Robust, real-time tracking of objects from visual data requires probabilistic fusion of multiple visual cues. Prevtous approaches have either been ad hoc or relied on a Bayesian network with discrete spatial variables which suffers from discretisation and computational complexity problems. We present a new Bayesian modality fusion network that uses continuous domain variables. The network architecture distinguishes between cues that are necessary or unnecessary for the object's presence. Computationally expensive and inexpensive modalities are also handled differently to minimise cost. The method provides a formal, tractable, exact and robust probabilistic method for simultaneously tracking multiple objects.

1. Introduction

Robust tracking of a single object under occlusion from visual data is difficult due to ambiguity and noise in the sensors, uncertainty in the trajectory of the object, and variations in the appearance of the object over time. The problem of noisy sensors, or sensors that generate ambiguous output from distracting objects, can be addressed through a process of *Bayesian modality fusion* [9, 10]. Bayesian modality fusion (BMF) uses a Bayesian network to probabilistically combine the outputs of several complementary modalities. A reliability indicator for each modality is incorporated in the network. The use of complementary modalities overcomes the problem of ambiguity if a distracting object in one modality is not present in another. The problems of noisy or failing sensors are addressed through the use of probabilities and reliabilities. Uncertainty in object trajectory is generally an unsolvable problem since often we cannot know the intentions of the object (if it is a person, for example). All that can be done is to impose a general temporal model, and use global searching of the spatial domain for focusing cues rather than local searching. Unfortunately local searching is often used because global searching is necessarily computationally prohibitive. Finally the varying appearance of the object must be approached by making the chosen modalities invariant to these appearances. For example, motion and colour are generally consistent over varying appearance.

One difficulty with modality fusion is that the existing implementation [9] uses discrete variables to model the spatial domain. Each spatial variable X can take the values $1, \ldots, N$, where N is the number of pixels in the image. Hence marginalisation over conditional probability distributions involving spatial variables has an undesirable $\mathcal{O}(N^2)$ complexity. The problem can be managed to some extent by excessively sub-sampling the image domain. The consequences are not as bad as one might first think since the probability values in the sub-sampled domain are continuous. However, Toyama and Horvitz [9] do not address the problem of choosing a sufficiently accurate level of subsampling. Furthermore, the discretisation of the spatial domain prohibits extensibility to larger spatial domains and tracking of multiple objects.

Another limitation of the BMF approach in [9] is that observations of object positions are entered as specific localised evidence from an isolated tracker, *ie:* the observations are uni-modal. However, this is undesirable given that the combination of uncertain object trajectory and ambiguity in the modalities can result in multiple feasible observations. Much information is discarded at an early stage that could have been valuable later. Contemporary tracking approaches such as CONDENSATION [3] suffer from the same problem in a different way: only a sub-set of the current observations are used for tracking, that sub-set being determined by a temporal model. This approach contravenes the recent wisdom that successful vision requires both data-driven and model-driven processing simultaneously. To that end, the full set of observations needs to be considered simultaneously, combined with prior information, and the most likely joint hypothesis inferred, provided it is computationally tractable.

To track multiple objects, an exclusion principle must be applied on the observations so that multiple object trackers do not continually claim responsibility for the same observation [5]. There is generally a combinatorial explosion in the number of matching possibilities over time. Previous approaches at explicitly tracking multiple objects [6, 2, 8, 4] have generally used heuristic approaches to deal with this complexity.

We propose a new Bayesian modality fusion, *Continuous Global Evidence-Based Bayesian Modality Fusion* (CBMF), that makes four novel contributions but is also computationally tractable: 1) Continuous sampling: the formerly suggested discrete domain spatial variables [9] are turned into continuous variables to assuage the complexity of inference. 2) Global evidence: all observations from a single modality are considered during inference rather than a single position decided upon at an early and premature stage. 3) Distinct Modality Types: a distinction is made between modalities that are necessary for the presence of an object and those that only hint at its presence. 4) Selective Computation: computationally expensive modalities are treated differently from inexpensive modalities to improve performance. The network architecture is modularly expanded to simultaneously track multiple objects and impose an exclusion principle in a theoretically principled manner that exploits Bayesian "explaining away" [7].

2. CBMF for a Single Object

Rather than entering observations of a 2-dimensional spatial variable $Z = [z_1, z_2]$ as a specific value, evidence is entered as a likelihood over variable values. The likelihood of the observational evidence e_Z on the variable Z is modelled as a mixture of K Gaussians: [1]

$$p(e_Z|z) = \sum_{k=1}^{K} \alpha_k G(z;\mu_k,\sigma_k); \quad \sum_{k=1}^{K} \alpha_k = 1 \qquad (1)$$

Clearly evidence must be discarded as irrelevant at some stage in the process to avoid high computational cost. However we allow this discarding process to be driven by the observations rather than some prior and possibly misconceived hypothesis. Gaussian mixtures are only defined for spatial regions in which the modality yields a non-zero response. Therefore in general, K will vary from observation to observation.

Conditional probabilities between continuous variables X and Y are modelled using a continuous 2-dimensional Gaussian distribution:

$$p(y|x) = G(y;x,\sigma) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(y_1-x_1)^2}{2\sigma_1^2} - \frac{(y_2-x_2)^2}{2\sigma_2^2}} \qquad (2)$$

Therefore marginalisation over large tables of discretised Gaussian distributions is avoided through the simplified analytical form of Gaussian convolution.

The general architecture of the CBMF network is shown in Figure 1. The chief inferred node X represents the position of the object. It is a continuous variable whose distribution is generally a mixture of Gaussians. X nodes are conditionally dependent over time to allow for a temporal dynamic model. Modalities are divided into two classes: *necessary* and *contingent*. Necessary modalities, Y_k, must be present when an object is present, and these nodes form the set of child nodes of X. Contingent modalities are represented by U_k and may or may not be present when X is present. They form the set of parents of X. Each modality has a continuous spatial variable which is a mixture of Gaussians. Each modality also has an associated reliability node and sub-network that measures the reliability of the modality. Each reliability variable has a set of child indicator variables which serve as external information alluding to the current reliability of the modality. For example, for a motion-based modality, a suitable indicator may be the instantaneous motion energy in the image. If the energy were to drop to zero, this would indicate that the motion modality is unreliable. The reliability R_k and associated indicators $I_{k,1}, \dots, I_{k,n_k}$ are all discrete variables. Virtual evidence is entered into each modality node using a dummy child node e_k. We exploit the global independence of variables by applying local propagation rules to determine the belief distribution for X given the observations, $P(X|e)$.

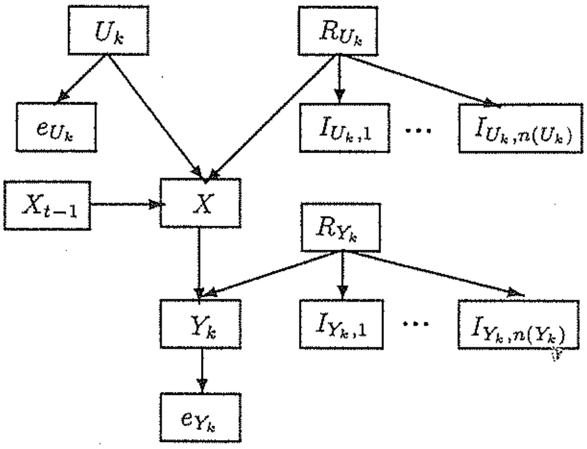

Figure 1: Continuous Global Evidence-based Modality fusion network. Only one necessary modality Y_k and one contingent modality U_k are shown for simplicity.

In the remainder of this section, we derive a tractable solution to the proposed CBMF network for tracking. Given M necessary modalities $Y_1, \dots, Y_M$ and N contingent modalities, $U_1, \dots, U_N$ in a network [2], evidence $e_Y = \{e_{Y_1}, e_{I_{1,1}}, \dots, e_{I_{1,n(1)}}, \dots, e_{Y_K}, e_{I_{K,1}}, \dots, e_{I_{K,n(K)}}\}$ (similarly for e_U) is entered via specific values $e_{I_{k,j}}$ for the indicators **I**, and likelihoods e_{Y_k} for the Y_k (see Eqn.(1)). The posterior distribution for X is:

$$\begin{aligned} P(X|e) &= P(e|X)P(X)/P(e) \\ &= \beta P(e_X^-|X)P(X|e_X^+) \end{aligned} \qquad (3)$$

[1] In this paper, all 2D Gaussians have diagonal covariance, and the functional parameters z, μ and σ are 2-vectors.

[2] Note that only one necessary modality Y_k and one contingent modality U_k are shown in Figure 1 for simplicity.

where e_X^+ is all evidence contained in the parent sub-trees of X, e_X^- is evidence contained in child sub-trees of X, and $\beta = P(e_X^+)/P(e)$, an irrelevant constant. We have exploited the independence of the parent and child evidences. Now we can determine the two contributions separately:

$$P(e_X^-|x) = \prod_{i=1}^{M} P(e_{XY_i}^-|x) \tag{4}$$

where $e_{XY_i}^-$ is the evidence down the ith sub-tree only. Considering Y_i and dropping the i subscript for simplicity:

$$P(e_{XY}^-|x) \quad = \quad \int_{r,y} p(e_{RY}^+|r)\, p(e_Y^-|y)\, p(y|r,x)p(r)\, dr\, dy$$

where we have exploited the conditional independence of e_{RY}^+ from X and Y given R and of e_Y^- from X and R given Y, and the marginal independence of R and X.

$$P(e_{XY}^-|x) \quad = \quad \gamma \int_r p(r|e_{RY}^+) \int_y p(e_y^-|y)\, p(y|r,x)\, dr\, dy$$

where $\gamma = p(e_{RY}^+)$ and $p(e_Y^-|y)$ is our entered evidence at Y. e_{RY}^+ is all evidence in R and its child sub-trees excluding Y. By further exploiting the independence of the reliability indicator evidences $e_{RL_j}^-$:

$$p(r|e_{RY}^+) \quad = \quad \zeta\, p(r) \prod_{j=1}^{n} p(e_{RI_j}^-|r); \quad \zeta = 1/p(e_{RY}^+)$$

Note that ζ can be calculated by normalising $p(r|e_{RY}^+)$ over r to sum to one. Each indicator contributes:

$$p(e_{RI_j}^-|r) \quad = \quad \int_{i_j} p(e_{i_j}|i_j)\, p(i_j|r)\, di_j$$

where $p(e_{i_j}|i_j)$ is our entered evidence. If discrete indicator evidence e_{i_j} is entered, and defining a reliability weighting function $F(\cdot)$ for convenience yields:

$$F(r) = p(r|e_{RY}^+) = \zeta\, p(r) \prod_{j=1}^{n} p(i_j = e_{i_j}|r) \tag{5}$$

Applying the observational mixture of Gaussians Eqn. (1) and the continuous conditional probability Eqn. (2) gives:

$$P(e_{XY}^-|x) = \gamma \sum_r F(r) \int_y \sum_{k=1}^{K} \alpha_k G(y; \mu_k, \sigma_k) G(y; x, \sigma_r) dy$$

where σ_r is the standard deviation corresponding to the given level of reliability, *ie:* low reliability needs high standard deviation in position for this modality. Given the 2-dimensional Gaussian functions on the domain $[x_1, x_2]$ with mean vector $[\mu_{i,1}, \mu_{i,2}]$ and diagonal covariance $diag[\sigma_{i,1}^2, \sigma_{i,2}^2]$, we apply the identity: $\int_x G(x;\mu_1,\sigma_1)\, G(x;\mu_2,\sigma_2)\, dx = G\left(\mu_2 - \mu_1; 0, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$which yields a mixture of Gaussians:

$$P(e_{XY}^-|x) = \gamma \sum_r F(r) \sum_{k=1}^{K} \alpha_k G(x; \mu_k, \sqrt{{\sigma_k}^2 + {\sigma_r}^2}) \tag{6}$$

At the X node, these necessary nodes result in a product of mixtures of Gaussians. Now we address incorporation of the contingent modalities. Let us define $\mathbf{U} = \{U_1, \ldots, U_N\}$ and $\mathbf{R} = \{R_1, \ldots, R_N\}$. Then:

$$\begin{aligned} P(x|e_X^+) \quad = \quad & \int_{\mathbf{U},\mathbf{R},x_{t-1}} p(x|\mathbf{u},\mathbf{r},x_{t-1})p(x_{t-1}|e_{x_{t-1}}^+) \\ & \prod_{i=1}^{N} p(u_i|e_{XU_i}^+)p(r_i|e_{XR_i}^+)\, d\mathbf{u}\, d\mathbf{r}\, dx_{t-1} \end{aligned} \tag{7}$$

Since the modalities are not necessarily present at the object location, we need a noisy-OR type rule. Given the modality locations, an object could really be anywhere, but is more likely to be found where the modalities occur. As modalities are superimposed, the likelihood of finding the object at a given location should increase since we have more evidence to that effect. These considerations can be represented as:

$$p(x|\mathbf{u},\mathbf{r},x_{t-1}) = \delta + w\, p(x|x_{t-1}) + w \sum_{i=1}^{N} p(x|u_i, r_i) \tag{8}$$

where δ is a constant signifying that the object could be anywhere, $w = (1-\delta)/(N+1)$ is a weighting giving equal favour to all modalities, and $p(x|x_{t-1})$ represents the temporal model for evolution of object position. Note that the distribution's expectation over X must equal one, hence the weightings w. δ represents the extent to which no contingent modality can give indication as to the object's whereabouts. For example, for a motion-based modality, δ would represent the proportion of image frames in which no motion occurs. For a face detection modality, δ would represent the on-line failure rate of the face detector in a typical sequence. Such a parameter can be estimated off-line from data. Substituting Eqn. (8) into Eqn. (7) gives:

$$\begin{aligned} & P(x|e_X^+) = \\ & \int_{\mathbf{U},\mathbf{R},x_{t-1}} \left(\delta + w\, p(x|x_{t-1}) + w \sum_{i=1}^{N} p(x|u_i, r_i) \right) \\ & p(x_{t-1}|e_{x_{t-1}}^+) \prod_{i=1}^{N} p(u_i|e_{XU_i}^+)p(r_i|e_{XR_i}^+)\, d\mathbf{u}\, d\mathbf{r}\, dx_{t-1} \end{aligned}$$

Now through nested integration, many of these terms integrate to unity. Assuming uniform priors on U_i, defining

Δx_{t-1} and σ_{t-1} to be the position offset and dispersion specified by the temporal model, and considering the prior distribution of X as a mixture of Gaussians:

$$p(x_{t-1}|e^{+}_{x_{t-1}}) = \sum_{k=1}^{T} \alpha_{t-1,k} G(x_{t-1}; \mu_{t-1,k}, \sigma_{t-1,k}),$$

we obtain:

$$\begin{aligned} &P(x|e^{+}_{X}) = \delta + \\ &w \sum_{k=1}^{T} \alpha_{t-1,k} G\left(x; \mu_{t-1,k} + \Delta x_{t-1}, \sqrt{\sigma^2_{t-1} + \sigma^2_{t-1,k}}\right) \\ &+ w \sum_{r_i} F(r_i) \sum_{i=1}^{N} \sum_{k=1}^{K_i} \alpha_{i,k} G\left(x; \mu_{i,k}, \sqrt{\sigma^2_{r_i} + \sigma^2_{i,k}}\right) \quad (9) \end{aligned}$$

The final result is obtained by substituting Eqns. (4) and (9) into (3), and is a product of mixtures of Gaussians. The consequent exponential growth in the number of Gaussian terms is characteristic of a method that evaluates multiple joint hypotheses.

3. Querying Expensive Modalities

Modalities such as frame differencing and skin colour classification are inexpensive to compute and can be acquired for each pixel in the image. Other modalities, such as face detection and ellipse fitting, are not only expensive to compute, but rely on a size parameter that adds a search to the computation. It would be computationally infeasible to compute these expensive cues at each pixel for real-time applications. The modality fusion approach here can be used to selectively calculate the more expensive modalities. It is a property of Bayesian networks that evidence need only be entered in a sub-set of variable nodes at any given time. We can begin by entering evidence e_C for the inexpensive modalities, resulting in a posterior $P(x|e_C)$. Now if there is a clear maximum in the distribution, no further computation is required. However if there is some ambiguity as to the maximum of $P(x|e_C)$, evidence e_F can be gathered from the expensive modalities at the candidate locations. Propagation of this evidence should disambiguate the result. The criterion used in this work to query expensive modalities is to calculate at the set of local optima on $P(x|e_C)$. The new expensive evidence e_F is then propagated to yield $P(x|e_T) = P(x|e_C, e_F)$, where e_T is the total evidence.

4. CBMF for Multiple Objects

Now consider the case in which there are L objects to be tracked in the scene. Here we present an Extended CBMF network to track multiple objects simultaneously. When tracking the objects from visual stimulus there will generally be two types of modalities: those that indicate the presence of all objects (*eg:* motion), and those that identify a single object (*eg:* appearance). The architecture is shown in Figure 2. The variable X is taken from the single-object network discussed previously, and represents modalities that are common to all objects. The posterior distribution obtained from the CBMF, $P(X|e_T)$, is treated as the observation for this network. It is a simplifying assumption to treat X as an isolated variable in this case. X has a set of parents $\mathbf{A} = \{A_1, \ldots, A_L\}$ which are continuous variables each representing the position of an object. The figure shows the relevant variables for A_i, the position of the ith object. In similar fashion to the previous CBMF network, each A_i has a set $\mathbf{Y}_{A_i} = \{Y_{i,1}, \ldots, Y_{i,n(A_i)}\}$ of object-specific modalities and associated reliabilities that are instantiated with mixtures of Gaussians and reliability indicator observations. The conditional probabilities associated with these object-specific modalities are the same as in the previous network. However, to facilitate explaining away by the object variables, the conditional probability table for X is different, using a Noisy-OR rule [7]:

$$p(x|\mathbf{A}) = \eta(\mathbf{A}) \left(1 - \prod_{i=1}^{L} (1 - \delta(x; a_i))\right) \qquad (10)$$

where $\eta(\mathbf{A})$ is a normalising constant ensuring that the distribution integrates to unity for a specific configuration of $\mathbf{A}$, and $\delta(x; y)$ is the unit delta function:

$$\delta(x; y) = \begin{cases} 1 & \text{if } x = y; \\ 0 & \text{otherwise} \end{cases}$$

For now, we ignore the normalising constant and let $\eta(\mathbf{A}) = 1$. The constant will only be different from $1/L$ for configurations of $\mathbf{A}$ in which $a_i = a_j$ for some i and j in $[1, L]$. Ignoring these cases means that hypotheses regarding occluding objects are incorrectly weighted. However, ignoring η greatly simplifies the analysis.

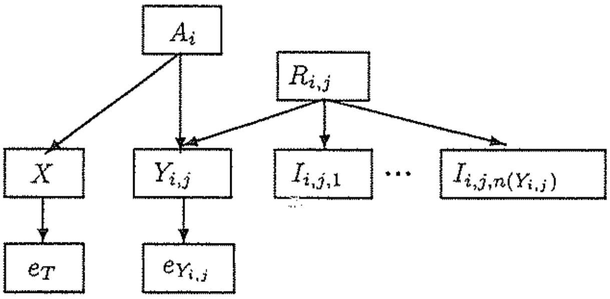

Figure 2: Unit of multi-object tracking BBN. In general there are L objects, $A_1, \ldots, A_L$, and each object i has $n(A_i)$ associated modality sub-networks. For simplicity only one object and one associated modality are shown here.

Inference using Eqn. (10) is now somewhat more complex because to infer the distribution of object A_i, information must be gathered from all $A_{j \neq i}$ through X. In this way objects are able to claim evidence at X through the Noisy-OR rule. Let the evidence entered into the multi-object network be e_S. We can determine the posterior $P(a_i|e_T, e_S)$ for each A_i as:

$$P(a_i|e_T, e_S) = \alpha\, p(a_i)\, p(e^-_{A_i X}|a_i) \prod_{j=1}^{n(A_i)} p(e^-_{Y_{i,j}}|a_i)$$

where $\alpha = 1/p(e_T, e_S)$. The evidences $p(e^-_{Y_{i,j}}|a_i)$ from the object-specific modalities are computed in analogous manner to Eqn. (6). The prior $p(a_i)$ is important and can be taken as the posterior position distribution from the previous time frame, which may be a single Gaussian for example. The difficult term is $p(e^-_{A_i X}|a_i)$ which relies on the other objects being tracked:

$$p(e^-_{A_i X}|a_i) = \beta \int_x \int_{A_j : j \neq i} p(e^-_X|x)\, p(x|\mathbf{a}) \prod_{j \neq i} p(a_j|e^+_{A_j X})\, dA_j\, dx$$

where $\beta = p(e^+_{A_j : j \neq i})$ is a normalising constant, $p(e^-_X|x) = p(x|e_T)$ is the posterior from the CBMF network, treated here as evidence and generally a mixture of Gaussians, and $p(a_j|e^+_{A_j X})$ is the evidential support provided by object j:

$$p(a_j|e^+_{A_j X}) = \frac{p(a_j) \prod_{k=1}^{n(A_j)} p(e^-_{A_j Y_k}|a_j)}{p(e^+_{A_j X})} \qquad (11)$$

Substituting Eqn. (10) into the above expression and simplifying yields:

$$p(e^-_{A_i X}|a_i) = \beta \left[1 + p(e^-_X|x = a_i) \frac{\prod_{j=1}^{L} m_j(a_i)}{m_i(a_i)} - \int_x p(e^-_x|x) \frac{\prod_{j=1}^{L} m_j(x)}{m_i(x)}\, dx \right] \qquad (12)$$

where $m_j(x) = 1 - p(a_j = x|e^+_{A_j X})$ is the *object map* for object j.

5. The Issue of Tractability

Although the general theory for fusing quantities probabilistically has been presented, there are several issues that must be considered for computational tractability. The observations must be presented as a mixture of Gaussians. In the simplest case, the observation could be a single Gaussian, as was the case in [9]. A more general method could use individual modality trackers to obtain an economical mixture of Gaussians using traditional techniques such as K-means clustering or the EM algorithm [1]. In the examples presented here, observations are represented as a dense mixture of Gaussians to overcome the problems of fitting mixtures of Gaussians to data. The given modality is thresholded to remove pixels with low probability. The remaining pixels are each instantiated as the mean of a single Gaussian, with variance arbitrarily set to 1 in the x- and y-directions, and a weighting coefficient in proportion to the modality strength. The weightings are normalised to sum to 1 so that the observation likelihood is a true distribution.

Possibly the most significant issue is interpretation and manipulation of the posterior distribution of X. The final distribution on X is the product of mixtures of Gaussians, resulting in a combinatorial explosion in the total number of Gaussians. For example, the contingent modalities result in $R.N.\bar{K}$ Gaussians, where $\bar{K}$ is the average number of observation Gaussians per modality, and R is the number of discrete reliability values. The necessary modalities produce the product of M mixtures each containing $R\bar{K}$ Gaussians. Therefore the overall complexity is $\mathcal{O}(N.(R.\bar{K})^{M+1})$. For our chosen observations, $\bar{K}$ may be on the order of 1000. In implementation, we circumvent this problem by discretising the belief distribution of X and accumulating the products over the Gaussian mixtures. Hence the complexity is reduced to $\mathcal{O}((N+M).R.K)$. The price paid is that the analytical Gaussian mixture representation is lost. The most plausible use of $P(X|e)$ is to find the value of X that maximises the distribution: $x^* = \underset{x}{\text{argmax}}\ P(x|e)$. However, there is no straight-forward way to maximise a superposition of Gaussians.[3] Given the discretised function, however, x^* can be easily determined.

The propagation of evidence over time through the temporal connection between X nodes would result in the endless proliferation of Gaussians. To assuage this problem, each optimum in the posterior distribution of X at time t is used as a centre in a Gaussian mixture to represent $p(x_{t-1}|e^+_{x_{t-1}})$ at the next time instant.

It is worth noting the comparison in computational expense between our approach and other approaches. Let D be a measure of the extent of the spatial domain being modelled. For instance, $D = w \times h$ would be the number of pixels in an image. In the case of tracking a single object, the complexity is $\mathcal{O}(M\bar{K}DR)$, where M is the total number of modalities. Therefore computation is linear in the number of modalities, observation units and domain size. Compare this with the original BMF framework which used discrete spatial variables. In this case, $\bar{K} = 1$ because only one observation hypothesis was used per modality. The com-

[3] In general the local maxima of the superposition will occur when the sums of derivatives of terms equals zero:

$$\frac{dF(x)}{dx} = 0 = \sum_i \alpha_i \frac{dG(x; \mu_i, \sigma_i)}{dx}$$

plexity of inference is $O(M\bar{K}D^2R)$, which is quadratic in domain size! The saving has come about through exploitation of the simple analytic form of the convolution of two Gaussians. The other improvement is that our method allows $\bar{K} > 1$ so that valuable low-level information is not discarded during high-level inference. Here we have assumed that the implementation uses discretised forms of the mixtures of Gaussians. The computational complexity may be reduced further in special cases. In particular, if the observations consist of only one or a few Gaussians, then the analytical form can be tractably used throughout inference.

For the case of tracking L objects, the complexity is $O(L\bar{M}\bar{K}DR)$ where $\bar{M}$ is the average number of object-specific modalities. Therefore the complexity is the same as for our single-object case, but scales linearly with L. This is a profoundly important property for simultaneous tracking of multiple objects: the usual combinatorial explosion in joint object location hypotheses is avoided by communication through the X node. By comparison, other approaches such as [4] retain the L^2 complexity and assume tractability due to a small number of objects. Our approach can be compared with the partitioned sampling method of [5], in which a hierarchical model of object independence is exploited to avoid L^2 complexity. However, our approach is deterministic, does not suffer from sparse sampling problems and has fixed computational complexity.

For multi-object tracking, note that the last term in Eqn.(12) is a constant over a_i. Therefore a discretised object map can be calculated for each object. A combined map $\prod_{j=1}^{L} m_j(x)$ for X can then be computed. For a specific object, the maps are combined with the observation at X to determine $p(e^-_{A_i X}|a_i)$.

6. Experimental Results

We tested the CBMF approach on the problem of tracking an individual's head in a video sequence. Three modalities were used: skin colour (necessary), frame differencing (contingent), and ellipse fitting (necessary). While skin and motion are cheap to compute, the fitting of an ellipse to an edge image is expensive since the head position and size must be first hypothesised. Therefore the network used had $\mathbf{Y} = \{Y_1, Y_2\}$ where Y_1 is skin colour and Y_2 is the ellipse fit, and $\mathbf{U} = \{U_1\}$ where U_1 is the motion estimate. The ellipse fitting modality was queried as an expensive modality. We used a broad Gaussian distribution for the temporal model to specify the object's expected position at the next time step, with $\Delta x_{t-1} = 0$. Three discrete reliability values, *low*, *medium* and *high*, are used in the network. Similarly all reliability indicators are discretised to one of the three values low, medium and high.

The cues were calculated as follows. The frame difference is the absolute difference between consecutive greyscale images. The skin image was computed using a single multi-dimensional Gaussian for classification in normalised RG-colourspace, where $R = r/N$, $G = g/N$, and $N = r + g + b$. The Gaussian parameters were estimated off-line using user-selected image regions. In the cases of motion and skin colour, these real-valued images were then thresholded to obtain a binary classification. At the queried image locations, the ellipse fit was obtained on a blurred edge image at multiple sizes on the range of 20 to 60 pixels in width. An ellipse aspect ratio of $x : y = 1 : 1.2$ was assumed. The criterion used is $f = s/n$, where s is the number of non-zero edge pixels under the ellipse perimeter, and n is the total number of pixels along the perimeter. The ellipse size with the highest criterion value at that position was used.

The reliability indicator for the motion cue was the number of moving pixels in the image, the rationale being that when there is either virtually zero or a great deal of motion present, that cue is unreliable for identifying the head. For skin, two indicators were used. The number of skin pixels was used in similar manner to the number of motion pixels. The second indicator is here termed *pearling*, or patchiness of the skin image. It is computed as the average variance of the binary skin image in 3×3 tiles. The more patchy the skin image is, the less reliable this modality. No reliability indicator was used for the ellipse fitting modality. δ in Eqn. (8) was arbitrarily set to 0.1.

A sample frame from results on a test sequence is shown in Figure 3. The figure shows (from left to right, top to bottom) the original image, the motion image, the skin image, the motion modality observation as a mixture of Gaussians, the skin modality observation, the prior distribution of X, the intermediate posterior $P(X|e_C)$, the expensive ellipse fitting modality observation instantiated at the appropriate locations, and the final fused distribution $P(X|e)$. It can be seen that many hypotheses for the position of the head are considered by the tracker. In the initial fusion result there are two competing peaks, one corresponding to the hand and the other to the head. The expensive ellipse fit modality is queried at the local maxima of the fused distribution and propagated to yield the final fused posterior. The expensive modality has successfully disambiguated the head with a clear peak at the proper location.

In the second example, CBMF is used to simultaneously track the heads of three people in the scene under occlusion. The same experimental configuration as the first example was used to obtain a distribution on X, based on skin colour classification, frame differencing and ellipse fitting. The skin colour model used was based on training pixels from all three individuals. The single object-specific modality for each object was a skin classification based on a person-specific colour model. Reliabilities were used for the object-specific colour models as in the previous exper-

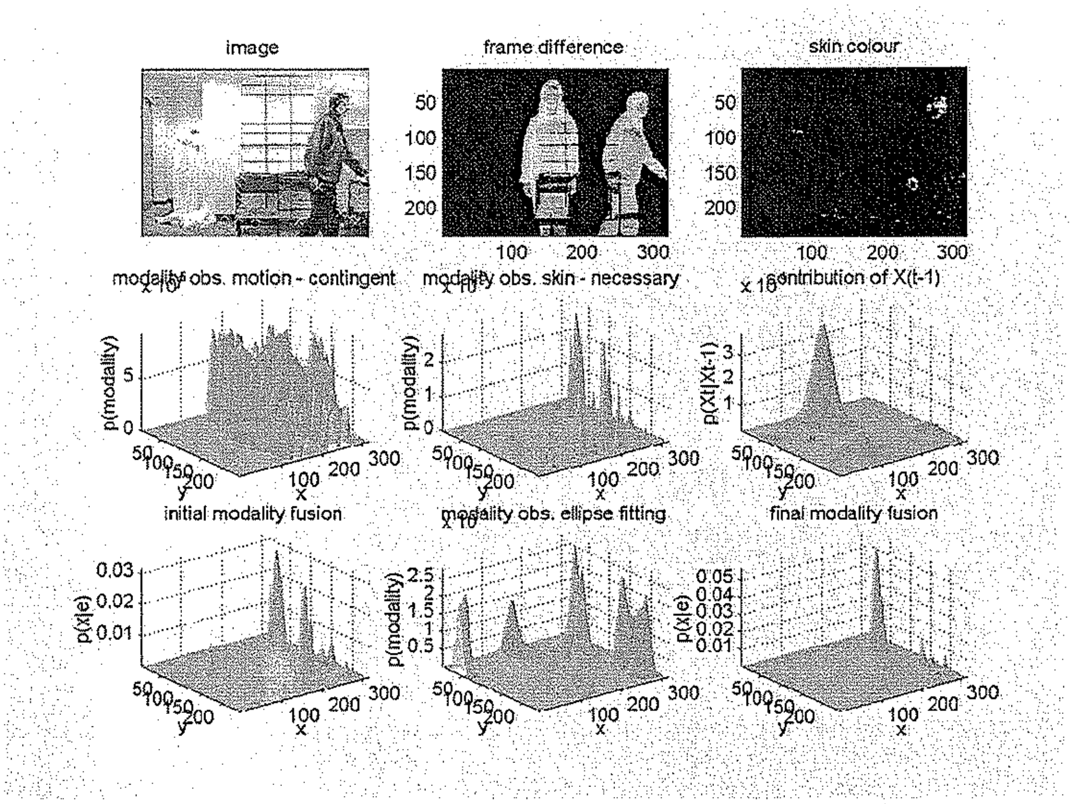

Figure 3: Continuous Global Evidence-Based Bayesian Modality Fusion results from a sample frame in a sequence. The figure shows from left to right, top to bottom: the original image, the motion image, the skin image, the motion modality observation as a mixture of Gaussians, the skin modality observation, the prior distribution of X, the intermediate posterior $P(X|e_C)$, the expensive ellipse fitting modality observation,and the final fused distribution $P(X|e_T)$.

iment. The sample frame is shown in Figure 4, and the CBMF results are shown in Figure 5. The figure shows the object-specific skin colour modalities as mixtures of Gaussians on the first row, and the final posterior distribution for each object position on the second row. The results are quite startling. The modes in the posterior distributions match the positions of the correct faces for the respective skin colour model. However examining the figure it can be seen that only the modality distribution for the first object is very distinctive for that object. Nevertheless, the mechanism of Bayesian "explaining away" has ensured that the second and third objects cannot be found at the distinctive position of the first object.

7. Conclusion

We have presented a theoretically sound, computationally tractable, comprehensive probabilistic framework for continuous-valued, global evidence-dependent Bayesian modality fusion to track multiple objects in space. For tracking multiple objects simultaneously, the model complexity grows linearly with the number of objects rather than quadratically as for some existing techniques. The method uses exact inference, is deterministic, and combines information globally from all observations with prior infor-

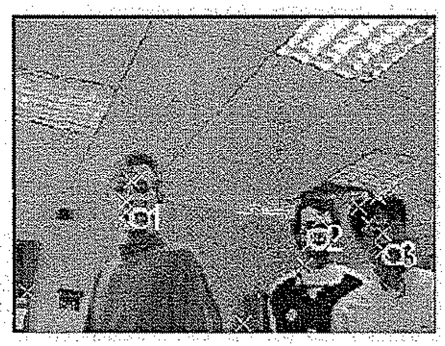

Figure 4: Sample frame from multi-object tracking example. Crosses show positions of local optima in $P(X|e_T)$, and labelled circles show estimated object positions.

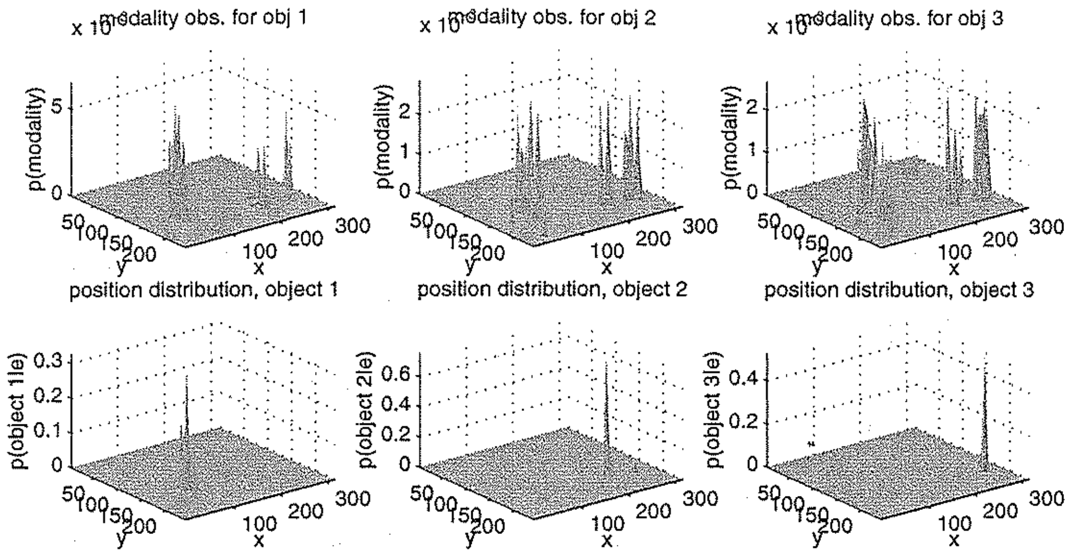

Figure 5: CBMF tracking three heads simultaneously; results from a sample frame out of a sequence. The top row shows these observations as mixtures of Gaussians. The bottom row shows the posterior distribution of position for each object after inference. The corresponding image frame is shown in Figure 4.

mation propagated over time. The method can distinguish between necessary and contingent modalities, and between computationally expensive and cheap visual cues.

This method is a recent development and can be improved in a number of ways. A more efficient implementation may be developed that is able to quickly find the local maxima of a mixture of Gaussians. This would remove the reliance of the computational complexity on the spatial domain size. Certain parameters in experiments have been selected in an ad hoc manner, but could be estimated from data. Currently the CBMF network does not explicitly handle the case that the object leaves the field of view. The architecture will need to be modified to handle this case.

The temporal dependence between object location distributions in the multi-object tracker has been simplified by copying the posterior at time t to be the prior at time $t+1$ because inclusion of these connections would overcomplicate inference. It remains to be seen whether proper inclusion of these dependencies can improve tracking. Another issue with the multi-object tracker is that the number of objects being tracked is currently fixed over time. We are experimenting with the following solution to this problem. Tracking begins with a single object. Using Bayesian model selection techniques, the addition or removal of an object can be hypothesised periodically. Three networks are periodically tested: one with L objects, one with $L-1$ and one with $L+1$. A difficult issue occurs with removal of an object, since knowing which object to remove may require consideration of L new networks.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition.* Cambridge University Press, 1995.

[2] I. Haritaoglu, D. Harwood, and L. Davis. W^4: Real-time surveillance of people and their activities. *IEEE PAMI*, 22(8):809–830, August 2000.

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

[4] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE PAMI*, 22(8):758–767, August 2000.

[5] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *ICCV*, volume 1, pages 572–578, Corfu, Greece, September 1999.

[6] S. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In *IEEE FG*, pages 348–353, Grenoble, France, 2000. IEEE Computer Society.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[8] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE PAMI*, 22(8):747–758, August 2000.

[9] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proceedings of the Fourth Asian Conference on Computer Vision*, Tapei, Taiwan, January 2000.

[10] J. Triesch and C. von der Malsburg. Self-organized integration of adaptive visual cues for face tracking. In *IEEE FG*, pages 102–107, Grenoble, France, March 2000. IEEE Press.