MOTEXATION: Multi-Object Tracking with the Expectation-Maximization Algorithm

Yonggang Jin and Farzin Mokhtarian
Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
{y.jin, f.mokhtarian}@surrey.ac.uk

Abstract

The paper proposes a new edge-based multi-object tracking framework, MOTEXATION, which deals with tracking multiple objects with occlusions using the Expectation-Maximization (EM) algorithm and a novel edge-based appearance model. In the edge-based appearance model, an object is modelled by a mixture of a non-parametric contour model and a non-parametric edge model using kernel density estimation. Visual tracking is formulated as a Bayesian incomplete data problem, where measurements in an image are associated with a generative model which is a mixture of mixture models including object models and a clutter model and unobservable associations of measurements to densities in the generative model are regarded as missing data. A likelihood for tracking multiple objects jointly with an exclusion principle is presented, in which it is assumed that one measurement can only be generated from one density and one density can generate multiple measurements. Based on the formulation, a new probabilistic framework of multi-object tracking with the EM algorithm (MOTEXATION) is presented. Experimental results in challenging sequences demonstrate the robust performance of the proposed method.

1 Introduction

Visual tracking is an important research area of computer vision. Previous work on edge-based contour tracking includes contour tracking with Kalman filtering [3] or particle filtering [9], contour tracking with the EM algorithm [14], which are all for single object tracking. Some similar previous work on joint tracking of multiple objects was presented in [12, 10, 17]. In [10, 17] multi-object tracking with particle filtering was proposed. However the number of samples will grow exponentially with the number of objects, and usually the depth order of multiple objects is needed or needs to be jointly estimated. In [12] Joint Probabilistic Data Association (JPDA) with the exclusion principle is applied to multiple contour tracking in comparison with Probabilistic Data Association (PDA) for single contour tracking in CONDENSATION [9]. Due to the complexity of enumerating all feasible events, the extension to track more than two objects is computationally expensive and also the depth order needs to be estimated and used in the likelihood. On the other hand, many iterative algorithms were proposed for color-based tracking(though only for single object tracking), including mean-shift algorithm with color histogram [6],

kernel-based tracking with spatial-color non-parametric model [8], EM-like tracking with spatial-color Gaussian mixture model [18].

This paper proposes a new edge-based multi-object tracking framework, MOTEXATION, which deals with tracking multiple objects with occlusions using the EM algorithm and a novel edge-based appearance model. The proposed approach differs from previous similar work on contour tracking [3, 9, 12] mainly in three aspects: object model, likelihood and inference used. In the edge-based appearance model, an object is modelled by a mixture of a non-parametric contour model and a non-parametric edge model using kernel density estimation similar to that for color-based non-parametric model [8]. Visual tracking is formulated as a Bayesian incomplete data problem where measurements in an image are associated with a generative model which is a mixture of mixture models including object models and a clutter model and unobservable associations of measurements to densities in the generative model are regarded as missing data. A likelihood for tracking multiple objects jointly with an exclusion principle is presented where it is assumed that 1. one measurement can only be generated from one density 2. one density can generate multiple measurements. The first assumption incorporates the same exclusion principle essential to track objects during occlusion as that of [12], based on JPDA, whereas the second assumption is relaxed like that of Probabilistic Multi-Hypothesis Tracker (PMHT) [15] to allow one density to generate multiple measurements rather than one measurement only. This significantly reduced the complexity of enumerating all feasible events in comparison with JPDA. Tracking multiple objects jointly will increase the dimensionality of state space and often the likelihood will become sharply peaked [16], which makes tracking with particle filtering difficult. The iterative EM algorithm is employed for multi-object tracking due to its monotonicity property which can seek the mode of the likelihood or the posterior despite high dimensional state space and sharply peaked likelihood. In addition it is also possible to combine edge features with color features using the iterative algorithm, for more robust tracking.

The organization of the paper is as follows. Tracking is formulated in Sec. 2; Multi-object tracking with the EM algorithm is presented in Sec. 3; Results are given in Sec. 4 and the paper is concluded in Sec. 5.

2 Tracking formulation

State vector is denoted as $\mathbf{x}(t) = [x(t)\ y(t)\ a(t)\ b(t)]^T$ where $[x(t)\ y(t)]^T$ is the spatial position of the object centre, $a(t)$ and $b(t)$ are the width and height of the object respectively. A second order auto-regressive model is employed as the dynamical model, $\mathbf{x}(t) = \mathbf{A}_1\mathbf{x}(t-1) + \mathbf{A}_2\mathbf{x}(t-2) + \mathbf{B_0}\mathbf{w}(t)$ where $\mathbf{w}(t)$ is Gaussian noise $\mathcal{N}(\mathbf{w}(t); \mathbf{0}, \mathbf{I})$.

2.1 Gating and clustering

Edge measurements are first detected by Canny edge detector [5]. The gating procedure of PDA is then applied. A validation region is computed based on the predicted state vector using dynamical model for each object so only measurements from within the validation region of the predicted state vector are used [1].

The clustering procedure from JPDA is also employed [1] for multi-object tracking. Multiple objects are first grouped into clusters and then are tracked jointly in each cluster. It often occurs that more than one object are grouped into the same cluster if there are occlusions between objects. After clustering, measurements in validation regions of all

objects in a cluster are used for jointly tracking multiple objects in that cluster. Measurements in a cluster are denoted as $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$, where N is the number of measurements in a cluster, $\mathbf{z}_i = \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}$, $\mathbf{u}_i = [x_i, y_i]^T$ and $\mathbf{v}_i = \theta_i \in [0, 2\pi)$ are the spatial position and orientation of ith edge measurement respectively.

2.2 Object model

The edge-based object appearance model $p_l(\mathbf{z})$ is a mixture of a non-parametric contour model $p_{con}(\mathbf{z})$, which consists of contour sample points, and a non-parametric edge model $p_{edge}(\mathbf{z})$, which consists of edge pixels inside the object contour, $p_l(\mathbf{z}) = \pi_{con} p_{con}(\mathbf{z}) + \pi_{edge} p_{edge}(\mathbf{z})$ where π_{con} and π_{edge} is the mixture weight of contour model and edge model respectively, $\pi_{con} + \pi_{edge} = 1$.

For the non-parametric contour model,

$$p_{con}(\mathbf{z}) = \frac{1}{M_{con}} \sum_{j=1}^{M_{con}} \mathcal{K}_{con}(\mathbf{z}; \mathbf{m}_{con,j}, \Sigma) = \frac{1}{M_{con}} \sum_{j=1}^{M_{con}} \mathcal{N}(\mathbf{u}; \mathbf{u}_{con,j}, \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},con}(\mathbf{v}; \mathbf{v}_{con,j}, \Sigma_{\mathbf{v}})$$

where $\mathbf{m}_{con,j} = \begin{bmatrix} \mathbf{u}_{con,j} \\ \mathbf{v}_{con,j} \end{bmatrix}$, $\mathbf{u}_{con,j}$ and $\mathbf{v}_{con,j} = \theta_{con,j} \in [0, \pi)$ are the spatial position and orientation of the normal of jth contour sample respectively, $\Sigma = \begin{bmatrix} \Sigma_{\mathbf{u}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathbf{v}} \end{bmatrix}$, $\Sigma_{\mathbf{u}}$ and $\Sigma_{\mathbf{v}} = \sigma_\theta^2$ are the fixed covariance of spatial position and orientation respectively, $\mathcal{K}_{\mathbf{v},con}(\mathbf{v}; \mathbf{v}_{con,j}, \Sigma_{\mathbf{v}}) \propto e^{-\frac{d_{con}^2(\theta, \theta_{con,j})}{2\sigma_\theta^2}}$ and $d_{con}(\theta, \theta_{con,j}) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Object contour is expressed parametrically by $\mathbf{m}_{con} = f(s, \mathbf{x})$ where s is the contour parameter. An ellipse can be used for head tracking and more complex contours can be represented by B-spline [4].

For the non-parametric edge model,

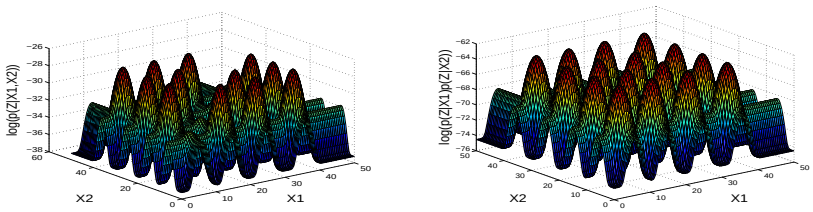
$$p_{edge}(\mathbf{z}) = \frac{1}{M_{edge}} \sum_{j=1}^{M_{edge}} \mathcal{K}_{edge}(\mathbf{z}; \mathbf{m}_{edge,j}, \Sigma) = \frac{1}{M_{edge}} \sum_{j=1}^{M_{edge}} \mathcal{N}(\mathbf{u}; \mathbf{u}_{edge,j}, \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},edge}(\mathbf{v}; \mathbf{v}_{edge,j}, \Sigma_{\mathbf{v}})$$

where $\mathbf{m}_{edge,j} = \begin{bmatrix} \mathbf{u}_{edge,j} \\ \mathbf{v}_{edge,j} \end{bmatrix}$, $\mathbf{u}_{edge,j}$ and $\mathbf{v}_{edge,j} = \theta_{edge,j} \in [0, 2\pi)$ are the spatial position and orientation of jth edge pixel inside the object contour respectively, $\mathcal{K}_{\mathbf{v},edge}(\mathbf{v}; \mathbf{v}_{edge,j}, \Sigma_{\mathbf{v}})$ $\propto e^{-\frac{d_{edge}^2(\theta, \theta_{edge,j})}{2\sigma_\theta^2}}$ and $d_{edge}(\theta, \theta_{edge,j}) \in [-\pi, \pi]$.

Note that contour model $p_{con}(\mathbf{z})$ can be regarded as a "stable" component and edge model $p_{edge}(\mathbf{z})$ as a "wandering" component in the object model [11]. Rewrite $p_l(\mathbf{z})$ as $p_l(\mathbf{z}) = \sum_{j=1}^{M} \omega_j \mathcal{N}(\mathbf{u}; \mathbf{u}_j, \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},j}(\mathbf{v}; \mathbf{v}_j, \Sigma_{\mathbf{v}})$ where $\{\omega_j\}_{j=1}^{M} = \left\{ \left\{ \frac{\pi_{con}}{M_{con}} \right\}_{j=1}^{M_{con}}, \left\{ \frac{\pi_{edge}}{M_{edge}} \right\}_{j=1}^{M_{edge}} \right\}$, $\{\mathbf{m}_j\}_{j=1}^{M} = \left\{ \{\mathbf{m}_{con,j}\}_{j=1}^{M_{con}}, \{\mathbf{m}_{edge,j}\}_{j=1}^{M_{edge}} \right\}$, $M = M_{con} + M_{edge}$ and later on for brevity, it will not be specified whether a density is from contour model or edge model.

2.3 Clutter model

A clutter model $p_c(\mathbf{z})$ is used to assimilate the measurements not from objects. It also corresponds to a "lost" component [11]. Uniform density is used so $p_c(\mathbf{z}) = p_c = \frac{1}{V_{\mathbf{u}} \times V_{\mathbf{v}}}$



(a) (b)

Figure 1: Comparison of (a) joint tracking likelihood $p(\mathbf{Z}|\mathbf{x}_1,\mathbf{x}_2)$ with exclusion principle and (b) separate tracking likelihood $p(\mathbf{Z}|\mathbf{x}_1)p(\mathbf{Z}|\mathbf{x}_2)$.

where $V_{\mathbf{u}}$ and $V_{\mathbf{v}}$ are the volume of validation region and of feature space without validation respectively [1].

2.4 Likelihoods

To explain measurements of a cluster with more than one object, the generative model is a mixture of mixture models including transformed mixture models of all objects in that cluster and a clutter model. The generative model can be written as $p(\mathbf{z}|\mathbf{x}) = \pi_c p_c(\mathbf{z}) + \sum_{l=1}^{L} \pi_l p_l(\mathbf{z}|\mathbf{x}_l)$, where $\mathbf{x} = \{\mathbf{x}_l\}_{l=1}^{L}$ includes state vectors of L objects in a cluster, π_l and π_c are the mixture weight of the lth object model and clutter model respectively and $\pi_c + \sum_{l=1}^{L} \pi_l = 1$, $p_l(\mathbf{z}|\mathbf{x}_l) = \sum_{j=1}^{M_l} \omega_{l,j} \mathcal{N}(\mathbf{u}; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l), \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},l,j}(\mathbf{v}; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}})$ is the transformed lth object model assuming unchanged orientation feature vector, M_l and $\omega_{l,j}$ are the number of densities and jth mixture weight in the lth object model respectively.

Assuming measurements $\mathbf{Z}$ are drawn independently from the generative model $p(\mathbf{z}|\mathbf{x})$, the likelihood given the incomplete data $\mathbf{Z}$ is

$$p(\mathbf{Z}|\mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{z}_i|\mathbf{x}) = \prod_{i=1}^{N} \left[\pi_c p_c + \sum_{l=1}^{L} \pi_l p_l(\mathbf{z}_i|\mathbf{x}_l) \right] \tag{1}$$

Despite its simplicity, the same exclusion principle as that in [12] is included in the likelihood 1 in comparison with likelihood of tracking multiple objects separately

$$\mathscr{L}(\mathbf{x}) = \prod_{l=1}^{L} p(\mathbf{Z}|\mathbf{x}_l) = \prod_{l=1}^{L} \prod_{i=1}^{N} [\pi_c p_c + (1.0 - \pi_c) p(\mathbf{z}_i|\mathbf{x}_l)] \tag{2}$$

Fig. 1 illustrates a 1D example with 4 measurements and 2 objects with 1 density each as that in [12].

In practice the assumption of independent measurements is not valid if measurements are close to each other as there are strong correlations between measurements [16]. A more practical likelihood is to incorporate measurement weights described in section 2.5,

$$p(\mathbf{Z}|\mathbf{x}) = \prod_{i=1}^{N} \left[\pi_c p_c + \sum_{l=1}^{L} \pi_l p_l(\mathbf{z}_i|\mathbf{x}_l) \right]^{\alpha_i} \tag{3}$$

where α_i is weight for ith measurement.

From the viewpoint of the Bayesian incomplete data problem, the missing data of association of measurements with densities are introduced and denoted as $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^{N}$ and $\mathbf{k}_i = \{k_i^1, k_i^2\}$ where $k_i^1 \in \{1, \cdots, L, c\}$, $k_i^1 = c$ indicates the association with clutter, and $k_i^1 = l, l \in \{1, \cdots, L\}$ association with object l; $k_i^2 \in \{1, \cdots M_{k_i^1}\}$ gives the association with one of the mixture densities in k_i^1th model. Assuming that 1. a measurement can have only one source 2. more than one measurement can originate from a density, where the first assumption is the same as that of JPDA known as exclusion principle in [12] and the second assumption is relaxed similar to that of PMHT, there are $N_e = (\sum_{l=1}^{L} M_l + 1)^N$ feasible events $\{\chi_n\}_{n=1}^{N_e}$. The likelihood given the complete data is

$$p(\mathbf{Z}, \mathbf{K} = \mathbf{K}(\chi_n)|\mathbf{x}) \propto \prod_{i:k_i^1(\chi_n)=c} \pi_c p_c \prod_{i,l,j:\substack{k_i^1(\chi_n)=l \\ k_i^2(\chi_n)=j}} \pi_l \omega_{l,j} \mathcal{N}(\mathbf{u}_i; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l), \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},l,j}(\mathbf{v}_i; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}}) \tag{4}$$

For comparison, JPDA can also be viewed in light of Bayesian incomplete data problem with a slightly different assumption that 1. a measurement can have only one source 2. no more than one measurement can originate from a density, so there are $\sum_{n=0}^{\min(N,M)} \frac{M!N!}{(M-n)!(N-n)!n!}$ feasible events. Denote $N_0(\chi_n)$ as number of densities which have no allocated measurements and $N_1(\chi_n)$ as number of densities which have only one allocated measurement in a feasible event χ_n, the likelihood given complete data in JPDA is

$$p(\mathbf{Z}, \mathbf{K} = \mathbf{K}(\chi_n)|\mathbf{x}) \propto {p_c}^{N-N_1(\chi_n)} \mu_F(N - N_1(\chi_n))(1 - P_D P_G)^{N_0(\chi_n)} (P_D)^{N_1(\chi_n)} \frac{(N-N_1(\chi_n))!}{N!}$$
$$\times \prod_{i,l,j:\substack{k_i^1(\chi_n)=l \\ k_i^2(\chi_n)=j}} \mathcal{N}(\mathbf{u}_i; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l), \Sigma_{\mathbf{u}}) \mathcal{K}_{\mathbf{v},l,j}(\mathbf{v}_i; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}})$$

where P_D is the detection probability, P_G is the probability that the true measurement will fall in the validation region, $\mu_F(n)$ is the probability mass function of the number of false measurements [1].

After marginalization of equation (4), $p(\mathbf{Z}|\mathbf{x})$ is factorized to N terms in equation (1) in comparison with $\sum_{n=0}^{\min(N,M)} \frac{M!N!}{(M-n)!(N-n)!n!} \gg N$ terms in marginalized likelihood of JPDA.

2.5 Measurement weighting

Histogram back-projection is used to incorporate background information. A background edge orientation histogram $\{h_i\}_{i=1}^{N_B}$ with N_B bins of orientation is built by using the edge pixels in a rectangular window surrounding each object. The background histogram is adapted online by weighted sum of previous background histogram and background histogram built given current object state estimation.

A ratio histogram $\{r_i\}_{i=1}^{N_B}$ is computed by $r_i = \min\left(\frac{\hat{h}}{h_i}, 1\right)$ where $\hat{h} = \min_{i:h_i>0}(h_i)$. Measurement weight α_i is computed from the ratio histogram as $\alpha_i = \frac{r_{b(\mathbf{z}_i)}}{\sum_{i=1}^{N} r_{b(\mathbf{z}_i)}} \times \frac{1}{2\sigma^2}$ where $b(\mathbf{z}_i)$ denotes the bin to which $\mathbf{z}_i$ belongs and σ is a constant.

(a) Initial estimation (b) Final estimation (c) Lower bound

Figure 2: Iterative update of EM algorithm where edge measurements are marked in yellow: (a) initial estimation, (b) final estimation, (c) lower bound increasing monotonically.

Measurements with orientations occurring most commonly in the background will have the lowest weight and measurements with orientations which are not in the background will have the highest weight. If the ratio histogram is uniform, it degenerates to the case that each measurement has the same weight $\alpha_i = \frac{1}{N} \times \frac{1}{2\sigma^2}$.

3 Multi-object tracking with the EM algorithm

State vector $\mathbf{x}(t)$ is estimated by either Maximum Likelihood (ML) estimation $\hat{\mathbf{x}}(t) = \arg\max_{\mathbf{x}(t)} p(\mathbf{Z}(t)|\mathbf{x}(t))$ or Maximum a Posteriori (MAP) estimation $\hat{\mathbf{x}}(t) = \arg\max_{\mathbf{x}(t)} p(\mathbf{x}(t)|\mathscr{Z}(t))$, where $\mathscr{Z}(t) = \{\mathbf{Z}(j)\}_{j=0}^{t}$, using the EM algorithm [7] and its generalization [13].

From Jensen's inequality it can be shown that

$$\log p(\mathbf{Z}|\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \log \left[\frac{\pi_c p_c(\mathbf{z}_i)}{q_{i,c}} q_{i,c} + \sum_{l=1}^{L} \sum_{j=1}^{M_l} q_{i,l,j} \frac{\pi_l \omega_{l,j} \mathscr{N}(\mathbf{u}_i; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l), \Sigma_{\mathbf{u}}) \mathscr{K}_{\mathbf{v},l,j}(\mathbf{v}_i; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}})}{q_{i,l,j}} \right]$$
$$\geq \sum_{i=1}^{N} \alpha_i \left[q_{i,c} \log \frac{\pi_c p_c(\mathbf{z}_i)}{q_{i,c}} + \sum_{l=1}^{L} \sum_{j=1}^{M_l} q_{i,l,j} \log \frac{\pi_l \omega_{l,j} \mathscr{N}(\mathbf{u}_i; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l), \Sigma_{\mathbf{u}}) \mathscr{K}_{\mathbf{v},l,j}(\mathbf{v}_i; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}})}{q_{i,l,j}} \right]$$

where $q_{i,c} = p(k_i^1 = c)$, $q_{i,l,j} = p(k_i^1 = l, k_i^2 = j)$ are the probabilities of missing data $\mathbf{K}$ and $q_{i,c} + \sum_{l=1}^{L} \sum_{j=1}^{M_l} q_{i,l,j} = 1$. So the lower bound of likelihood $J_{ML}(\mathbf{Q}, \mathbf{x}(t))$ for ML estimation and lower bound of posterior $J_{MAP}(\mathbf{Q}, \mathbf{x}(t))$ for MAP estimation are

$$J_{ML}(\mathbf{Q}, \mathbf{x}(t)) = \sum_{i=1}^{N} \alpha_i \left[q_{i,c} \log \frac{\pi_c p_c(\mathbf{z}_i)}{q_{i,c}} + \sum_{l=1}^{L} \sum_{j=1}^{M_l} q_{i,l,j} \log \frac{\pi_l \omega_{l,j} \mathscr{N}(\mathbf{u}_i; T_{\mathbf{u}}(\mathbf{u}_{l,j}, \mathbf{x}_l(t)), \Sigma_{\mathbf{u}}) \mathscr{K}_{\mathbf{v},l,j}(\mathbf{v}_i; \mathbf{v}_{l,j}, \Sigma_{\mathbf{v}})}{q_{i,l,j}} \right] \tag{5}$$

$$J_{MAP}(\mathbf{Q}, \mathbf{x}(t)) = J_{ML}(\mathbf{Q}, \mathbf{x}(t)) + \log p(\mathbf{x}(t)|\mathscr{Z}(t-1)) \tag{6}$$

where $\mathbf{Q} = \left\{ q_{i,c}, \left\{ \left\{ q_{i,l,j} \right\}_{j=1}^{M_l} \right\}_{l=1}^{L} \right\}_{i=1}^{N}$.

The prior is given by

$$p(\mathbf{x}(t)|\mathscr{Z}(t-1)) = \prod_{l=1}^{L} p(\mathbf{x}_l(t)|\mathscr{Z}(t-1)) = \prod_{l=1}^{L} \mathscr{N}(\mathbf{x}_l(t); \tilde{\mathbf{x}}_l(t), \tilde{\mathbf{P}}_l(t)) \tag{7}$$

Algorithm 1 Multi-Object Tracking with the EM Algorithm (MOTEXATION)

1. Predict by equation (7)
2. EM algorithm
   $k=1$, $\mathbf{x}^{(0)}(t)=\tilde{\mathbf{x}}(t)$
   (i) E-step by equation (8)
   (ii) M-step by equation (9) or equation (10)
   if $\left\|\mathbf{x}_l^{(k)}(t)-\mathbf{x}_l^{(k-1)}(t)\right\|<\varepsilon, l=1,\cdots,L$ **then**
   $\hat{\mathbf{x}}(t)=\mathbf{x}^{(k)}(t)$ and stop
   else
   $k=k+1$ go to (i)
   end if

where $\tilde{\mathbf{x}}_l(t)=\mathbf{A}_1\hat{\mathbf{x}}_l(t-1)+\mathbf{A}_2\hat{\mathbf{x}}_l(t-2)$ and $\tilde{\mathbf{P}}_l(t)\approx\mathbf{B}_0\mathbf{B}_0^T$ are the predicted state vector and covariance of lth object respectively,

In E-step, given fixed $\mathbf{x}^{(k-1)}(t)$, maximize $J_{ML}(\mathbf{Q},\mathbf{x}(t))$ or $J_{MAP}(\mathbf{Q},\mathbf{x}(t))$. Let $T_{\mathbf{u}}(\mathbf{u}_{l,j},\mathbf{x}_l(t))$ $=\mathbf{W}_{l,j}\mathbf{x}_l(t)$ where $\mathbf{W}_{l,j}$ is Jacobian of the transformation. At iteration k, $\mathbf{Q}^{(k)}$ is

$$\begin{aligned}
&q_{i,c}^{(k)}\propto\pi_c p_c(\mathbf{z}_i)\\
&q_{i,l,j}^{(k)}\propto\pi_l\omega_{l,j}\mathcal{N}(\mathbf{u}_i;\mathbf{W}_{l,j}\mathbf{x}_l^{(k-1)}(t),\Sigma_{\mathbf{u}})\mathcal{K}_{\mathbf{v},l,j}(\mathbf{v}_i;\mathbf{v}_{l,j},\Sigma_{\mathbf{v}})\\
&q_{i,c}^{(k)}+\sum_{l=1}^{L}\sum_{j=1}^{M_l}q_{i,l,j}^{(k)}=1, i=1\cdots N
\end{aligned}\tag{8}$$

In M-step, given $\mathbf{Q}^{(k)}$, maximize $J_{ML}(\mathbf{Q},\mathbf{x}(t))$ or $J_{MAP}(\mathbf{Q},\mathbf{x}(t))$. At iteration k, $\mathbf{x}(t)$ is given by

$$\mathbf{x}_{l,ML}^{(k)}(t)=\left[\sum_{j=1}^{M_l}\mathbf{W}_{l,j}^T\tilde{\Sigma}_{l,j}^{(k)^{-1}}\mathbf{W}_{l,j}\right]^{-1}\left[\sum_{j=1}^{M_l}\mathbf{W}_{l,j}^T\tilde{\Sigma}_{l,j}^{(k)^{-1}}\tilde{\mathbf{u}}_{l,j}^{(k)}\right]\tag{9}$$

or

$$\mathbf{x}_{l,MAP}^{(k)}(t)=\left[\sum_{j=1}^{M_l}\mathbf{W}_{l,j}^T\tilde{\Sigma}_{l,j}^{(k)^{-1}}\mathbf{W}_{l,j}+\tilde{\mathbf{P}}_l^{-1}(t)\right]^{-1}\left[\sum_{j=1}^{M_l}\mathbf{W}_{l,j}^T\tilde{\Sigma}_{l,j}^{(k)^{-1}}\tilde{\mathbf{u}}_{l,j}^{(k)}+\tilde{\mathbf{P}}_l^{-1}(t)\tilde{\mathbf{x}}_l(t)\right]\tag{10}$$

where $\tilde{\mathbf{u}}_{l,j}^{(k)}=\frac{\sum_{i=1}^{N}\alpha_i q_{i,l,j}^{(k)}\mathbf{u}_i}{\sum_{i=1}^{N}\alpha_i q_{i,l,j}^{(k)}}$ is the synthetic measurement and $\tilde{\Sigma}_{l,j}^{(k)}=\frac{\Sigma_{\mathbf{u}}}{\sum_{i=1}^{N}\alpha_i q_{i,l,j}^{(k)}}$ is the synthetic covariance.

The main stages of multi-object tracking with the EM algorithm are given in algorithm (1) and the iterative update of MAP estimation is shown in Fig. 2 where the lower bound of posterior is also verified to be increased monotonically.

4 Results

The experiments are carried out in challenging test sequences with heavy occlusions. With unfully optimized C++ code, it runs comfortably at average $0.071s$ per object per frame

t=5210 t=5286 t=5380 t=5420 t=5430

Figure 3: Tracking results of *"office"* sequence. ©Mitsubishi Electric ITE 2005.

t=425 t=433 t=450 t=460 t=475

Figure 4: Tracking results of *"head"* sequence.

t=1060 t=1180 t=1280 t=1296 t=1380

Figure 5: Tracking results of *"Caviar OneShopOneWait2cor"* sequence.

on 3GHz Pentium IV. Note that to illustrate joint tracking of multiple objects in a cluster, white lines show the links between objects which are tracked jointly in the same cluster.

Three results of multiple head tracking are shown and the size of head also varies from small ones to large ones. Fig 3 shows multi-object tracking results on the *"office"* sequence, in which there are dramatic appearance changes, scale changes and four heavy occlusions. The light green ellipse occluded dark green ellipse from frame 5280 to 5320, from frame 5340 to 5370 and from frame 5380 to 5410. The red ellipse occluded both light green and dark green ellipses from frame 5410 to 5424.

The results of *"head"*[1] are then given in Fig. 4 where there are two heavy occlusions from frame 420 to 442 and from frame 452 to 468.

Fig. 5 shows the results of *"Caviar* [2] *OneShopOneWait2cor"* sequence where the size of target heads are quite small and there are two heavy occlusions from frame 1166 to 1176 and from frame 1276 to 1292.

To track more complex contours, a B-spline contour model is learned as that of [2, 4]. Results of *"Caviar EnterExitCrossingPaths1cor2"* sequence are given in Fig 6 where there are large appearance changes, scale changes and one heavy occlusion from frame 86 to 100.

Fig. 7 presents the results of *"Caviar OneStopMoveEnter1cor2"* sequence, a very crowded and cluttered scene involving large appearance changes, scale changes and also one heavy occlusion from frame 256 to 272.

[1] The sequence is from http://vision.stanford.edu/ birch/headtracker/.

[2] The EC Funded CAVIAR project/IST 2001 37540, see http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

t=0 t=80 t=90 t=150 t=350

Figure 6: Tracking results of *"Caviar EnterExitCrossingPaths1cor2"* sequence.

t=0 t=262 t=276 t=300 t=480

Figure 7: Tracking results of *"Caviar OneStopMoveEnter1cor2"* sequence.

(a) (b) (c) (d)

Figure 8: Examples of tracking failure. (a)(b) tracking multiple objects separately using the EM algorithm, (c) contour tracking with CONDENSATION, (d) mean-shift tracking with color histogram.

It should be noted that if multiple objects are tracked separately using the EM algorithm with likelihood 2, which does not have exclusion principle, objects may be lost during occlusion as shown in Fig. 8(a)(b). The proposed method has also been compared with contour tracking using CONDENSATION [9], color tracking using mean-shift [6] and both failed when there are heavy occlusions. Examples of tracking failure are shown in Fig. 8(c)(d).

5 Conclusions

The paper proposes a new edge-based multi-object tracking framework, MOTEXATION, which deals with tracking multiple objects with occlusions using the EM algorithm and a novel edge-based appearance model. In the edge-based appearance model, an object is modelled by a mixture of a non-parametric contour model and a non-parametric edge model using kernel density estimation. Visual tracking is formulated as a Bayesian incomplete data problem where measurements in an image are associated with a generative model which is a mixture of mixture models including object models and a clutter model and unobservable associations of measurements to densities in the generative model are regarded as missing data. A likelihood for tracking multiple objects jointly with an exclusion principle is presented. Based on the formulation, a new probabilistic framework

of multi-object tracking with the EM algorithm (MOTEXATION) is presented. Results in challenging sequences demonstrate the robust performance of the proposed method.

Acknowledgements

The support of the Visual Information Laboratory (VIL), Mitsubishi Electric Information Technology Centre Europe (ITE) and Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey is gratefully acknowledged. We would also like to thank Dr O'Callaghan, Dr Bober and Dr Ratliff of VIL for helpful discussions and for providing the *"office"* sequence.

References

[1] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[2] A.M. Baumberg and D.C. Hogg. Learning flexible models from image sequences. In *Proc. ECCV*, pages 299–308, 1994.

[3] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *IJCV*, 11(2):127–145, 1993.

[4] A. Blake and M. Isard. *Active Contours*. Springer, 1998.

[5] J. Canny. A computational approach to edge detection. *IEEE PAMI*, 8(6):679–698, Nov. 1986.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. CVPR*, pages 142–149, 2000.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximal likelihood from incomplete data via the EM algorithm. *RoyalStat*, B 39:1–38, 1977.

[8] A. Elgammal, R. Duraiswami, and L. Davis. Probabilistic tracking in joint feature-spatial spaces. In *Proc. IEEE CVPR*, pages 781–788, 2003.

[9] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. ECCV*, pages 343–356, 1996.

[10] M. Isard and J. MacCormick. BraMBLe: a Bayesian multiple-blob tracker. In *Proc. ICCV*, pages 34–41, 2001.

[11] A. D. Jepson, D. J. Fleet, and T. F. Ei-Maraghi. Robust online appearance models for visual tracking. *IEEE PAMI*, 25(10):1296–1311, October 2003.

[12] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. ICCV*, pages 572–578, 1999.

[13] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental, sparse and other variants. *Learning in Graphical Models*, pages 355–368, 1998.

[14] A.E.C. Pece and A.D. Worrall. Tracking with the EM contour algorithm. In *Proc. ECCV*, pages 3–17, 2002.

[15] R. Streit and T. Luginbuhl. Maximum likelihood method for probabilistic multi-hypothesis tracking. *Proc. SPIE*, 2235:394–405, 1994.

[16] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by Bayesian correlation. In *Proc. ICCV*, pages 1068–1075, 1999.

[17] Y. Wu, T. Yu, and G. Hua. Tracking appearances with occlusions. In *Proc. CVPR*, pages 789–795, 2003.

[18] T. Yu and Y. Wu. Differential tracking based on spatial-appearance model (SAM). In *Proc. CVPR*, pages 720–727, 2006.