Differential Tracking based on Spatial-Appearance Model (SAM)

Ting Yu and Ying Wu
Department of Electrical Engineering and Computer Science
Northwestern University
2145 Sheridan Road, Evanston, IL, 60208
tingyu,yingwu@ece.northwestern.edu

Abstract

A fundamental issue in differential motion analysis is the compromise between the flexibility of the matching criterion for image regions and the ability of recovering the motion. Localized matching criteria, e.g., pixel-based SSD, may enable the recovery of all motion parameters, but it does not tolerate much appearance changes. On the other hand, global criteria, e.g., matching histograms, can accommodate dramatic appearance changes, but may be blind to some motion parameters, e.g., scaling and rotation. This paper presents a novel differential approach that integrates the advantages of both in a principled way based on a spatial-appearance model (SAM) that combines local appearances variations and global spatial structures. This model can capture a large variety of appearance variations that are attributed to the local non-rigidity. At the same time, this model enables efficient recovery of all motion parameters. A maximum likelihood matching criterion is defined and rigorous analytical results are obtained that lead to a closed form solution to motion tracking. Very encouraging results demonstrate the effectiveness and efficiency of the proposed method for tracking non-rigid objects that exhibit dramatic appearance deformations, large object scale changes and partial occlusions.

1. Introduction

One of the major challenges of appearance-based tracking lies in the large variations of the visual appearances, which may be caused by many reasons, such as non-rigid deformations, and partial occlusions, etc. Such large uncertainties in the visual appearance significantly complicate the matching of the visual appearances. Inappropriate matching results in the inability of motion recovery and tracking failure.

Existing solutions to appearance-based tracking have different treatment and exploitation of the spatial structure of the appearance. Two opposite extremes are template matching that requires a fine localized match [9, 12, 2, 11], and histogram matching that completely discards the spatial structure [5, 4, 14, 10].

Template tracking with SSD measure requires strict pixel-wise alignments between the object template and the candidate object region [9]. This is fine to handle rigid objects, while having a very limited power to handle non-rigid objects. To allow more appearance variations, improvements have been made by generalizing the template to be a template manifold, which can be linearly expanded by a set of eigenvectors [3], or support vectors [2]. Such a template manifold has to be learned off-line.

Histogram, on the other hand, completely discards the spatial information, thus allows dramatic appearance changes. Histogram-based tracking methods have demonstrated their superb performance in handling the non-rigid deformation, pose change and partial occlusions [5, 14]. However, the ignorance of the spatial layout also brings difficulties, e.g., less discriminative to appearance changes and thus less sensitive to certain motions. For example, the mean-shift tracker is awkward to handle scaling and rotation. Improvements have been made by using multiple kernels [10, 8].

This paper presents a novel differential approach based on a spatial-appearance model (SAM) that combines local appearances variations and global spatial structures, thus integrating the advantages of both. SAM is in the form of a Gaussian mixture model. This model can capture a large variety of appearance variations that are attributed to the local non-rigidity. At the same time, this model enables efficient recovery of all motion parameters. A maximum-likelihood estimation is defined for tracking, and is solved by a proposed variant of Expectation-Maximization (EM) algorithm. The analytical derivations lead to a closed-form solution for motion estimation. The proposed EM iterations guarantee the continuous increase of the likelihood, and result in a differential approach to motion recovery. The physical meaning of our solution indicates that the exact pixel-wise alignment is relaxed and the pixels in the candidate object region are weighted by their nearby spatial-appearance

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

Gaussian components in motion estimation.

Besides the ability of handling the appearance variations of non-rigid objects, another advantage of the proposed method is its ability of estimating various motions (e.g., translation, rotation, scaling, and affine) in a unified and principled manner, rather than having different mechanisms to handle them individually. It is actually a very appealing property comparing with mean-shift that only copes with translation in a principal manner. The new method proves very powerful to handle non-rigid objects.

The proposed method is different from some recent approaches that also make use of spatial and appearance models. For example, a model based on the pixel spatial-color features is proposed and is constructed by kernel density estimation [7]. An entropy-based similarity measure between two kernel densities is used for matching. A recent study [17] showed that this approach might not be suitable for the handling of complex motions and the entropy-based similarity measure is difficult to compute. Our approach differs greatly in the matching criteria, the analysis and thus the solutions.

2. Spatial-Appearance Model (SAM)

Recall the two extremes of appearance modelling vary from the approaches that strictly obey the spatial layout of object appearance (rigid template representation) to the ones where spatial locations of appearance features are completely discarded (histogram-based representation). Both of the modelling approaches have their merits and limitations. We choose to seek a tradeoff between the two approaches, and arrive at an intermediate level appearance modelling, which not only maintains a rough global spatial structure of object appearance as in template representation, but also preserves the simplicity of the histogram-based representation by only keeping some dominant feature values in the object region.

Given an initial object region $R_0 = \{x_i, i = 1, \ldots, N\}$, selected manually or automatically, a d dimensional spatial-appearance feature vector is extracted from each pixel and denoted by x_i. N is the total number of pixels within the initial object region. A K-component Gaussian mixture model (GMM) is adopted to fit to the collected data points, leading to a spatial-appearance model characterized by GMM with parameters $\theta = (p_k, \mu_k, \Sigma_k), k = 1, \ldots, K$. p_k, μ_k, Σ_k represent the prior probability, mean and variance of Gaussian component k in the mixture model. Each Gaussian component is denoted by $g(x; \mu_k, \Sigma_k)$. The likelihood of a pixel x within a candidate object region is simply the mixture probability as:

$$p(x|\theta) = \sum_{k=1}^{K} p_k g(x; \mu_k, \Sigma_k) \quad (1)$$

Depending on different features, the model dimension d could take different values with the first two dimensions occupied by the pixel spatial coordinate features (u, v). For example, we may take $d = 3$ by augmenting the spatial features with the intensity feature, or when color features are preferred, we may add dimensions with pixel feature values from (r, g, b) color channels.

Similar to the de-correlation strategy of spatial-appearance features as in [7, 15], we assume the spatial and appearance dimensions of the GMM model are decoupled, i.e., the covariance matrix of the Gaussian component takes the block diagonal form, $\Sigma_k = \begin{pmatrix} \Sigma_{k,s} & 0 \\ 0 & \Sigma_{k,c} \end{pmatrix}$, where s and c stand for spatial and appearance features respectively. Thus the joint feature x of each pixel can be written as $x = (x_s, c(x_s))$, with the spatial x_s and the appearance $c(x_s)$ features of a pixel at the location x_s. Each GMM Gaussian component then has the following factorized form:

$$g(x; \mu_k, \Sigma_k) = g(x_s; \mu_{k,s}, \Sigma_{k,s}) g(c(x_s); \mu_{k,c}, \Sigma_{k,c}) \quad (2)$$

The appearance feature $c(x_s)$ is actually the function of pixel location x_s, implying the intrinsic correlations between the spatial and appearance features although the decoupled Gaussian distribution.

Figure 1. The fitted spatial-appearance Gaussian mixture model to the object region.

An illustrative example of fitting the spatial-appearance model to the object region (a kid face) with 40-component mixture model is shown in Figure 1, where the left image is the original video frame, and in the right image each red ellipse represents a spatial Gaussian component fitted using Expectation-Maximization (EM) [6].

3. Expectation-Maximization (EM) Tracking

3.1. Maximum-Likelihood Formulation

Assume the object undergoes a motion transform characterized by a general motion model $T(x_s; a_t)$. a_t is the transform parameter at time t that warps a pixel at location x_s from reference frame to the location $T(x_s; a_t)$ in the current frame. Without losing the generality, we can assume the considered motion model having a general linear form

as follows:

$$T(x_s; a_t) = \begin{pmatrix} a_{1,t} & a_{2,t} \\ a_{3,t} & a_{4,t} \end{pmatrix} x_s + \begin{pmatrix} a_{5,t} \\ a_{6,t} \end{pmatrix} = A_t x_s + B_t \quad (3)$$

which can actually cover a broad spectrum of object motions, such as translation, scaling, rotation, and affine motion, etc.

With the SAM object model initialized in the reference frame, the likelihood of an object pixel x_i, warped from the reference frame to the current frame by motion transform $T(x_s; a_t)$, is evaluated as:

$$\begin{aligned} & p(T(x_i; a_t)|\theta) \\ & = p(T(x_{i,s}; a_t), c(T(x_{i,s}; a_t))|\theta) \\ & = \sum_{k=1}^{K} p_k g(T(x_{i,s}; a_t); T(\mu_{k,s}, \Sigma_{k,s}; a_t)) \\ & \qquad \times g(c(T(x_{i,s}; a_t)); \mu_{k,c}, \Sigma_{k,c}) \\ & = \sum_{k=1}^{K} p_k g(x_{i,s}; \mu_{k,s}, \Sigma_{k,s}) g(c(T(x_{i,s}; a_t)); \mu_{k,c}, \Sigma_{k,c}) \end{aligned} \quad (4)$$

Note from Eq. 4 that not only the pixel spatial coordinate $x_{i,s}$ is transformed to $T(x_{i,s}; a_t)$, but also the Gaussian parameter values on the spatial dimension are changed from $(\mu_{k,s}, \Sigma_{k,s})$ to $T(\mu_{k,s}, \Sigma_{k,s}; a_t)$. With the general linear motion model defined in Eq. 3, such that $T(x_{i,s}; a_t) = A_t x_{i,s} + B_t$, and $T(\mu_{k,s}, \Sigma_{k,s}; a_t) = (A_t \mu_{k,s} + B_t, A_t \Sigma_{k,s} A_t^T)$, this generally leads to a cancelled out effect on the Gaussian function evaluations on the spatial GMM components. However, since the appearance features of each pixel are coupled with the transformed position of the pixel in the current frame, i.e., $c(T(x_{i,s}; a_t))$, it essentially correlates the pixel likelihood evaluation with the unknown object motion estimation a_t.

To ease the derivations, define the data component probability $q(k, x_i; a_t)$ as

$$q(k, x_i; a_t) = p_k g(x_{i,s}; \mu_{k,s}, \Sigma_{k,s}) g(c(T(x_{i,s}; a_t)); \mu_{k,c}, \Sigma_{k,c}) \quad (5)$$

We propose a matching criterion to recover the object motion a_t based on the integration of the pixel data logarithm likelihood over the object region.

$$\begin{aligned} E(a_t; \theta) &= \sum_{x_i \in R_0} \log p(T(x_i; a_t)|\theta) \\ &= \sum_{x_i \in R_0} \log\{\sum_{k=1}^{K} q(k, x_i; a_t)\} \end{aligned} \quad (6)$$

This joint data likelihood term measures the data fitness of a candidate object region R_t at current time t, warped from the reference object region R_0, to the object SAM model characterized by model parameter θ. Thus the problem of object tracking becomes an essential optimization problem, where the objective is to look for an optimal value a_t^* that maximizes the joint likelihood energy function $E(a_t; \theta)$, i.e.,

$$a_t^* = \max_{a_t} E(a_t; \theta) \quad (7)$$

3.2. Closed-Form Tracking with EM

Treating the motion transform parameter a_t as the only unknown value in the above maximum likelihood estimation of Eq. 6, the Expectation-Maximization (EM) algorithm is well suitable to be adopted here to recover the unknown value of a_t for the current frame, with simultaneous achievement of energy function maximization.

Unlike the general EM algorithm for the parameter fitting of GMM model, where the objective is to find an optimal model parameter set θ^* that best explains the training data set. Here we assume that the GMM model parameter θ remains unchanged during this optimization process, while only deriving a solution to incrementally update the motion parameter a_t embedded into the EM iterations. We should clarify that our assumption that the GMM model parameter stays constant during this one frame EM iteration is a quite valid assumption. It actually has been intrinsically utilized by most existing tracking approaches, where the object model, once firstly initialized, will generally remain fixed during the whole tracking sequence, unless some online updating mechanism is adopted in order to handle the non-stationary visual process [12, 11, 16].

In fact, it is interesting to point out that our mixture framework does allow a straight-forward incorporation of an online updating process to handle the problem of tracking non-stationary object appearance. Although the current version of the algorithm does not take such a further step, we leave this issue for the future improvements. All the experiments reported in this paper do not take an online updating step, while still achieving very encouraging tracking results.

Similar to the general EM algorithm, an initial value for the unknown parameter must be specified in order to start the EM iterations. In our case we simply take the recovered motion estimation a_{t-1}^* from previous frame as the initialization of a_t, i.e., $a_t^{(0)} = a_{t-1}^*$. The superscript indexes the EM algorithm iteration. Assume that we have already obtained an estimation of a_t during the j_{th} EM iteration, i.e., $a_t^{(j)}$, the E-step involves the computation of pixel assignment probability to each Gaussian component as

$$p^{(j)}(k|x_i; a_t^{(j)}) = \frac{q(k, x_i; a_t^{(j)})}{\sum_{m=1}^{K} q(m, x_i; a_t^{(j)})} \quad (8)$$

with the data component probability $q(k, x_i; a_t^{(j)})$ defined in Eq. 5.

From the Jensen's inequality, we have the following lower bound to the original energy function $E(a_t; \theta)$:

$$\begin{aligned}
&E(a_t;\theta) \\
&= \sum_{x_i \in R_0} \log\{\sum_{k=1}^{K} q(k, x_i; a_t)\} \\
&= \sum_{x_i \in R_0} \log\{\sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \frac{q(k, x_i; a_t)}{p^{(j)}(k|x_i; a_t^{(j)})}\} \\
&\geq \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \log \frac{q(k, x_i; a_t)}{p^{(j)}(k|x_i; a_t^{(j)})} \qquad (9) \\
&= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \log q(k, x_i; a_t) - \\
&\sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \log p^{(j)}(k|x_i; a_t^{(j)}) \\
&= E^{(j)}(a_t; \theta)
\end{aligned}$$

Maximizing $E(a_t; \theta)$ can be achieved by maximizing the lower bound function $E^{(j)}(a_t; \theta)$, and subsequently maximizing the first term of $E^{(j)}(a_t; \theta)$ in Eq. 9, since the old pixel assignment probabilities $p^{(j)}(k|x_i; a_t^{(j)})$ are known provided the value $a_t^{(j)}$ at j_{th} EM iteration, thus the second term is unrelated to the objective function maximization over a_t.

we define the first term of lower bound function by $\widetilde{E}^{(j)}(a_t; \theta)$,

$$\widetilde{E}^{(j)}(a_t; \theta) = \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \log q(k, x_i; a_t) \qquad (10)$$

Iteratively maximizing $\widetilde{E}^{(j)}(a_t; \theta)$ by finding an updated estimation $a_t^{(j+1)}$ has the same effect on the incremental maximization of the original objective function $E(a_t; \theta)$. Rather than the logarithm of a sum as in $E(a_t; \theta)$, the derived $\widetilde{E}^{(j)}(a_t; \theta)$ only contains a linear combination of K logarithms, which breaks the coupling of the equations when setting the derivatives of $\widetilde{E}^{(j)}(a_t; \theta)$ over the parameter a_t to zero.

We take an incremental updating form by assuming that $a_t^{(j+1)} = a_t^{(j)} + \Delta a_t$, then the above maximization can be written as

$$\begin{aligned}
&\max_{\Delta a_t} \widetilde{E}^{(j)}(\Delta a_t; \theta) \\
&= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; a_t^{(j)}) \log q(k, x_i; a_t^{(j)} + \Delta a_t)
\end{aligned} \qquad (11)$$

Taking the partial derivative of $\widetilde{E}^{(j)}(\Delta a_t; \theta)$ over Δa_t and setting it to zero, we can obtain a series of linear updating equations to incrementally maximize the objective function depending on what motion model is used.

To ease the exposition, we firstly show the updating equations for the simple case, where a translational motion model is adopted, and object appearance feature is simply the pixel intensity. Then we generalize the discussions to handle more complex motion model, including scaling, rotation, or affine transform, and multi-dimension appearance features such as pixel values in (r, g, b) color channel are also considered there.

Recall that the motion parameter $a_t = \{A_t, B_t\}$ as defined in Eq. 3, when translational motion model is taken, A_t becomes an Identity matrix, we only need to consider the second term, i.e., $a_t = B_t$. By taking the incremental updating form, we have $\Delta a_t = \Delta B_t$.

$$\begin{aligned}
&\widetilde{E}^{(j)}(\Delta B_t; \theta) \\
&= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) \log q(k, x_i; B_t^{(j)} + \Delta B_t)
\end{aligned} \qquad (12)$$

Taking the partial derivative over ΔB_t

$$\begin{aligned}
&\frac{\partial \widetilde{E}^{(j)}(\Delta B_t; \theta)}{\partial \Delta B_t} \\
&= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) \frac{\partial \log q(k, x_i; B_t^{(j)} + \Delta B_t)}{\partial \Delta B_t} \\
&= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) \\
&\qquad \times \frac{\partial \log g(c(x_{i,s} + B_t^{(j)} + \Delta B_t); \mu_{k,c}, \Sigma_{k,c})}{\partial \Delta B_t}
\end{aligned} \qquad (13)$$

where the spatial component probability $g(x_{i,s}; \mu_{k,s}, \Sigma_{k,s})$ and component priori p_k in $q(k, x_i; B_t^{(j)} + \Delta B_t)$ disappear due to their uncorrelation with motion update ΔB_t. However, please note that their effects on the motion estimation do reflect on the computation of pixel assignment probability $p^{(j)}(k|x_i; B_t^{(j)})$.

Following the small motion assumption, $c(x_{i,s} + B_t^{(j)} + \Delta B_t)$ can be linearized by taking the first order Taylor expansion as

$$c(x_{i,s} + B_t^{(j)} + \Delta B_t) = c(x_{i,s} + B_t^{(j)}) + {H_{i,t}^{(j)}}^{\tau} \Delta B_t \qquad (14)$$

where $H_{i,t}^{(j)}$ is the Jacobian matrix of the appearance feature over motion estimation evaluated at its current value $B_t^{(j)}$.

When the appearance feature is simply the pixel intensity, $H_{i,t}^{(j)}$ takes the form as

$$H_{i,t}^{(j)} = \begin{pmatrix} c_u(x_{i,s} + B_t^{(j)}) \\ c_v(x_{i,s} + B_t^{(j)}) \end{pmatrix} \quad (15)$$

where $(c_u(x_{i,s} + B_t^{(j)}), c_v(x_{i,s} + B_t^{(j)}))$ are the horizontal and vertical intensity gradients at location $x_{i,s} + B_t^{(j)}$ of the current frame.

Recall that the probability distribution in appearance dimension $g(c(x_{i,s} + B_t^{(j)} + \Delta B_t); \mu_{k,c}, \Sigma_{k,c})$ also takes the Gaussian form, in combination with the linearized form of $c(x_{i,s} + B_t^{(j)} + \Delta B_t)$ in Eq. 14, the partial derivative of the objective function $\widetilde{E}^{(j)}(\Delta B_t; \theta)$ over ΔB_t can be eventually reached as

$$\begin{aligned} &\frac{\partial \widetilde{E}^{(j)}(\Delta B_t; \theta)}{\partial \Delta B_t} \\ &= \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) \\ &\qquad \times H_{i,t}^{(j)} \Sigma_{k,c}^{-1} [(c(x_{i,s} + B_t^{(j)}) - \mu_{k,c}) + H_{i,t}^{(j)^\tau} \Delta B_t] \\ &= 0 \end{aligned} \quad (16)$$

i.e., the following linear system equation can be derived to solve ΔB_t

$$\begin{aligned} U \Delta B_t &= V \\ \Delta B_t &= U^{-1} V \end{aligned} \quad (17)$$

where matrix U and V are defined as follows

$$U = \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) H_{i,t}^{(j)} \Sigma_{k,c}^{-1} H_{i,t}^{(j)^\tau} \quad (18)$$

$$V = - \sum_{x_i \in R_0} \sum_{k=1}^{K} p^{(j)}(k|x_i; B_t^{(j)}) H_{i,t}^{(j)} \Sigma_{k,c}^{-1} (c(x_{i,s} + B_t^{(j)}) - \mu_{k,c}) \quad (19)$$

The form of linear system equation implies that the contribution of each pixel to motion estimation is weighted by its nearby spatial-appearance Gaussian components, through assignment probability $p^{(j)}(k|x_i; B_t^{(j)})$, and appearance mean $\mu_{k,c}$ and variance $\Sigma_{k,c}$. Thus exact pixel-wise alignment between initial object region R_0 and warped candidate R_t is relaxed, leading to a more flexible framework of tolerating large appearance deformation during tracking. The extent of deformation tolerance is governed by the variance coverage of each mixture component. The contributions of all pixels to motion estimation are combined and voted for the optimal solution of motion update.

With the estimated motion update ΔB_t solved from Eq. 17, a new circle of EM iteration starts with the updated estimation of the motion parameter $B_t^{(j+1)}$ as

$$B_t^{(j+1)} = B_t^{(j)} + \Delta B_t \quad (20)$$

In summary, the proposed EM tracking approach takes the following two-step iterative procedure.

E-Step: compute the mixture component assignment probability for each pixel x_i by Eq. 8.

M-Step: obtain a motion update estimation Δa_t by solving the linear system equation as in Eq. 17.

The above EM iterations are iteratively computed to increase the joint data likelihood until convergence.

3.3. Tracking under General Motion Transform

The proposed EM tracking procedure could be easily generalized to handle more complex motion model, and incorporate more informative appearance features, while the same M-Step updating equation as in Eq. 17 could still be derived. The only difference between these variations lies on the computations of Jacobian matrix of the appearance feature over motion estimation, i.e., $H_{i,t}^{(j)}$. For example, for similarity motion model, handling translation, scaling, and rotation, the 4-dimensional motion vector $a_t = (a_{1,t}, a_{2,t}, a_{3,t}, a_{4,t})^\tau$ has the following form:

$$A_t = \begin{pmatrix} a_{1,t} & -a_{2,t} \\ a_{2,t} & a_{1,t} \end{pmatrix}, B_t = \begin{pmatrix} a_{3,t} \\ a_{4,t} \end{pmatrix} \quad (21)$$

the corresponding Jacobian matrix $H_{i,t}^{(j)}$ with intensity feature is defined as

$$H_{i,t}^{(j)} = \begin{pmatrix} c_u(A_t^{(j)} x_{i,s} + B_t^{(j)}) u_{i,s} + c_v(A_t^{(j)} x_{i,s} + B_t^{(j)}) v_{i,s} \\ -c_u(A_t^{(j)} x_{i,s} + B_t^{(j)}) v_{i,s} + c_v(A_t^{(j)} x_{i,s} + B_t^{(j)}) u_{i,s} \\ c_u(A_t^{(j)} x_{i,s} + B_t^{(j)}) \\ c_v(A_t^{(j)} x_{i,s} + B_t^{(j)}) \end{pmatrix} \quad (22)$$

When color appearance features are used, the Jacobian matrix $H_{i,t}^{(j)}$ becomes multi-columns with each column having the same form as in Eq. 22 but in a different color channel. More complex motion model, such as affine transform, can be derived in a similar manner, therefore we omit the discussion here.

It is clear that our framework allows tracking object under any general linear motion transforms that are solved in a unified way. The more complex motion recovery puts no more computation overhead than the simple ones. It is actually a very appealing property in comparison with the kernel-based tracking approaches using mean-shift, where only object translation is principally handled.

As guaranteed by the Jensen's inequality in Eq. 9, the lower bound optimization in the proposed EM iterations will subsequently lead to a continuous maximization of the original objective function, i.e., the data likelihood, thus driving the motion estimation towards the optimal candidate object region. Compared with the Kernel-based tracking [5], where a line search procedure is usually required for the optimal step length decision of mean shift iteration, our closed-form linear solution derived in Eq. 17 enjoys a Newton-style iteration as in template matching [9] and Kernel-based tracking with SSD [10]. It reaches a local optimum in an one-step jump, thus avoiding the tedious process of line search.

Figure 2 shows an illustrative example of one-frame EM iterations. The left figure represents the iterative motion estimations, illustrated by a series of colorized quadrangles overlapping on the original frame, with pure red to pure yellow depicting this sequential iteration procedure. The right figure clearly demonstrates the continuous increase of the data logarithm likelihood, as provably guaranteed in Eq. 9. In this example, 10 EM iterations are performed to reach the algorithm convergence, where we declare a convergence when there is no significant change between the motion estimations in two consecutive iterations.

Figure 2. Logarithm likelihood evaluation of the candidate object region during one frame iterations.

4. Experiments

In this section, we present extensive experiments tested under challenging real-world sequences. A differential tracker based on the proposed approach is implemented, capable of handling object translation, scaling, and rotation. Comparisons are made with simple template tracker and Kernel-based tracker, demonstrating the very encouraging performance of our unified approach for tracking non-rigid objects under dramatic appearance deformations, large object scale changes and partial occlusions.[1]

Depending on the availability of color channels from the input video, the appearance features in the SAM model vary from intensity feature to color features in the RGB color space. The number of mixture components used to model objects may take different values depending on the relative size of objects. It is actually a trade-off factor to govern the model flexibility to appearance deformation, where more components imply more localized component coverage, thus more strict observance of rigid structure assumption, while less components allow more relaxed alignment between the candidate region and object model. Our experience shows that 20-40 components usually work well for a broad spectrum of non-rigid objects we are testing on. We leave the investigation on optimal number of components selection for future study. Some related work along this direction includes [1, 11]. To speed up the model initialization, we take the tracked object regions in the first 50 frames of each sequence to update the mixture model, with one frame one EM iteration to obtain the model. After that, the model is fixed without further updating, and used for tracking the rest of video frames. The current unoptimized C++ implementation of the algorithm runs comfortably around 5-10 fps on average on Pentium 3G.

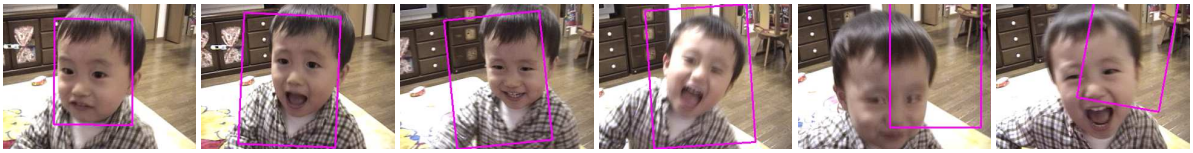
4.1. Large Appearance Deformation

Figure 3 shows the tracking results over a home video sequence, where a kid presents significant expression changes, thus dramatic appearance deformations. Considering the relative large size of object, a 40-component mixture model is adopted here to initialize the differential tracker with similarity transform motion model. The first row gives the result from a template matching tracker. It loses a tight tracking of the kid face at the early stage of the sequence, when the kid starts to behave his exaggerating expression and simultaneously shows the significant head movement. The second row of the Figure 3 shows the iterative motion estimations in each frame via the proposed differential tracker, with colorized quadrangles from pure red to pure yellow depicting the series of updating as before. The thickened boundaries due to multiple iterations clearly reflect the large motion effects, which are not only from translation, but also through rotation and scaling. Albeit the difficulties, the proposed differential tracker successfully keeps localizing the non-rigid face with correct motion estimations until the kid completely turns his head to the right side, thanks to the intrinsic deformation tolerance of the proposed approach.

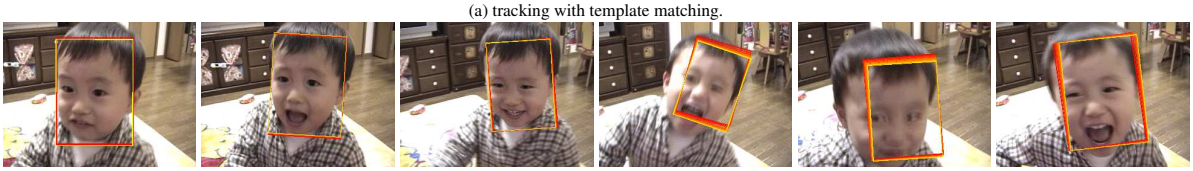
Figure 4 demonstrates our tracking results on the famous but challenging *Dudek* sequence[2], which has been tested over several approaches addressing online tracking adaptation [12, 13]. The person in this sequence presents not only large appearance variations by changing pose during movement, but also several short periods of severe occlusions. Without counting on the online adaptation, which is acknowledged hard to find the balance between the model adaptability and resistance to noise [11], our approach still achieves very encouraging results, that the improved robust-

[1] Please see the supplemental video for the detailed tracking results.

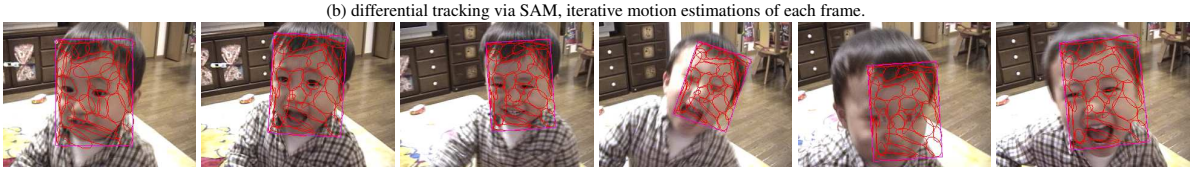
[2] We acknowledge Dr. El-Maraghi [12] for allowing us to download this sequence from his website for testing.



(a) tracking with template matching.



(b) differential tracking via SAM, iterative motion estimations of each frame.



(c) differential tracking via SAM, final tracking result of each frame overlapped by spatial mixture components.

Figure 3. Tracking a kid face under large appearance deformation. (560 Frames)

(a) differential tracking via SAM, iterative motion estimations of each frame.

(b) differential tracking via SAM, final tracking result of each frame overlapped by spatial mixture components.

Figure 4. Tracking a human face under large scale change and severe occlusions (1145 Frames).

ness to partial occlusions could be attributed to the some extent model tolerance on spatial-appearance misalignments in the SAM model.

4.2. Large Scale Change

Figure 5 shows a real-world surveillance video to demonstrate our tracker capacity of handling large object scale change[3]. A person enters the scene distantly with a quite small scale. Our tracker is initialized on this small object region, and robustly tracks the person for the remaining 1000 frames. Note the accurate scale estimations of the person during most of the tracking period. Two severe occlusions happen when the person is coming across with other pedestrians, which shortly affects the tracker's scale estimations during occlusion. After the person re-appearance from occlusion, the tracker recovers itself and starts to report the accurate motion estimations again.

Figure 5. Tracking a pedestrian under large scale change and partial occlusions with the proposed differential tracker via SAM. Results overlapped by spatial mixture components. (1230 Frames)

[3]We acknowledge the source data is provided from the EC Funded CAVIAR project/IST 2001 37540, found at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

The last example in Figure 6 also shows a home video filming the same kid as in Figure 3. Now the kid demonstrates a dramatic scale change, and also brings the trouble to the tracker by intentionally presenting serious occlusion. Our tracker again robustly tracks the kid face with the correct scale estimations for the whole sequence as shown in Figure 6 (b). In comparison, the results obtained from a color-based mean-shift tracker in Figure 6 (a) reports an incorrect scale estimation, and consequently loses tracking the object.

(a) mean shift tracking.

(b) differential tracking via SAM, iterative motion estimations of each frame.

Figure 6. Tracking a kid face under large scale change. (690 Frames)

5. Conclusion

In summary, this paper presents a novel differential approach for non-rigid object tracking under the general motion transform. A spatial-appearance model (SAM) is introduced to model both the object appearance variations and its global spatial structures. A maximum likelihood matching criterion is defined and rigorous analytical results are obtained through Expectation-Maximization (EM) algorithm, leading to a closed form solution to motion tracking. The derived linear system equation also suggests us to take a new view to look at the connections between the two standard tracking paradigms, template tracking and Kernel-based tracking. Our ongoing research will mainly focus on a deeper investigation on the intrinsic relations of the proposed approach with them, and their more recent advances, such as [10, 8].

6. Acknowledgements

This work was supported in part by National Science Foundation Grants IIS-0347877, IIS-0308222 and Northwestern faculty startup funds.

References

[1] O. Arandjelovic and R. Cipolla. Incremental learning of temporally-coherent gaussian mixture models. In *In Proc. British Machine Vision Conference (BMVC)*, 2005.

[2] S. Avidan. Support vector tracking. *In IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, pages 1064–1072, 2004.

[3] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *In Int'l Journal of Computer Vision (IJCV)*, 26(1):63–84, 1998.

[4] R. T. Collins. Mean-shift blob tracking through scale space. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Madison, Wisconsin, June 2003.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *In IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2003.

[6] A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *In Journal of the Royal Statistical Society, Series B (Methodological)*, 1(39):1–38, 1977.

[7] A. Elgammal, R. Duraiswami, and L. S. Davis. Probabilistic tracking in joint feature-spatial spaces. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Madison, Wisconsin, June 2003.

[8] Z. M. Fan, Y. Wu, and M. Yang. Multiple collaborative kernel tracking. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 502–509, San Diego, CA, 2005.

[9] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *In IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 20(10):1025–1039, 1998.

[10] G. D. Hager, M. Dewan, and C. V. Stewart. Multiple kernel tracking with ssd. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, D. C., June 2004.

[11] B. Han and L. Davis. On-line density-based appearance modeling for object tracking. In *In Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, Beijing, China, Oct. 2005.

[12] A. D. Jepson, D. J. Fleet, and T. El-Maraghi. Robust on-line appearance models for visual tracking. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume I, pages 415–422, Kauai, 2001.

[13] J. W. Lim, D. Ross, R. S. Lin, and M. H. Yang. Incremental learning for visual tracking. In *In Proc. Neural Information Processing Systems 17 (NIPS)*, 2005.

[14] Y. Wu and T. S. Huang. A co-inference approach to robust visual tracking. In *In Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, pages 26–33, 2001.

[15] C. J. Yang, R. Duraiswami, and L. Davis. Efficient mean-shift tracking via a new similarity measurement. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005.

[16] M. Yang and Y. Wu. Tracking non-stationary appearances and dynamic feature selection. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1059–1066, San Diego, CA, 2005.

[17] H. H. Zhang, W. M. Huang, Z. Y. Huang, and L. Y. Li. Affine object tracking with kernel-based spatial-color representation. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005.