The Dense Estimation of Motion and Appearance in Layers

Hulya Yalcin
Division of Engineering
Brown University
Providence, RI 02912
hy@lems.brown.edu

Michael J. Black
Computer Science
Brown University
Providence, RI 02912
black@cs.brown.edu

Ronan Fablet
LASAA
IFREMER
Plouzane, France 29280
rfablet@ifremer.fr

Abstract

Segmenting image sequences into meaningful layers is fundamental to many applications such as surveillance, tracking, and video summarization. Background subtraction techniques are popular for their simplicity and, while they provide a dense (pixelwise) estimate of foreground/background, they typically ignore image motion which can provide a rich source of information about scene structure. Conversely, layered motion estimation techniques typically ignore the temporal persistence of image appearance and provide parametric (rather than dense) estimates of optical flow. Recent work adaptively combines motion and appearance estimation in a mixture model framework to achieve robust tracking. Here we extend mixture model approaches to cope with dense motion and appearance estimation. We develop a unified Bayesian framework to simultaneously estimate the appearance of multiple image layers and their corresponding dense flow fields from image sequences. Both the motion and appearance models adapt over time and the probabilistic formulation can be used to provide a segmentation of the scene into foreground/background regions. This extension of mixture models includes prior probability models for the spatial and temporal coherence of motion and appearance. Experimental results show that the simultaneous estimation of appearance models and flow fields in multiple layers improves the estimation of optical flow at motion boundaries.

1. Introduction

The recovery of 2D image motion, or optical flow, has a long history and current methods have proven useful in fields as diverse as graphics and biology. In particular, the accuracy of dense optical flow techniques [9] has improved to the point where there are now more pressing issues to address before optical flow methods are more widely adopted. Consider, for example, domains such as surveillance where dense (pixelwise) motion estimation may be very useful. In such domains, optical flow algorithms must run continuously and automatically adapt to changes in lighting and motion over both short and long time frames. In such an application one might be willing to trade absolute accuracy for stability, dependability, and full automation. Traditional optical flow methods that rely on a simple assumption of brightness constancy are at a disadvantage in such applications as they have no “memory” about the motion in the scene over time and no “model” of the objects in the scene and their appearance. In contrast, model-based tracking methods achieve high accuracy and reliability by exploiting a rich appearance representation; dense optical flow techniques have no such model. Summarizing: optical flow methods lack an appearance model of what is being “tracked” and they lack any sort of explicit model of scene structure or segmentation. We argue that both of these are necessary for stable optical flow estimation and that the description of appearance and scene structure must adapt over time.

We propose a Bayesian framework for estimating dense optical flow over time that explicitly estimates and exploits a persistent model of image appearance. The approach assumes that the scene can be described by a number of layers but that the motion of each layer is highly flexible. The approach also exploits prior models that express how motion and appearance may change over time. To achieve this, we extend mixture model methods to the case of dense (rather than parametric) flow estimation and derive a mixture model formulation that includes explicit spatial and temporal priors.

The key contributions of this method are: 1) it is a straightforward extension of standard robust optical flow methods; 2) it estimates dense, subpixel-accurate, flow fields; 3) it produces an estimate of foreground and

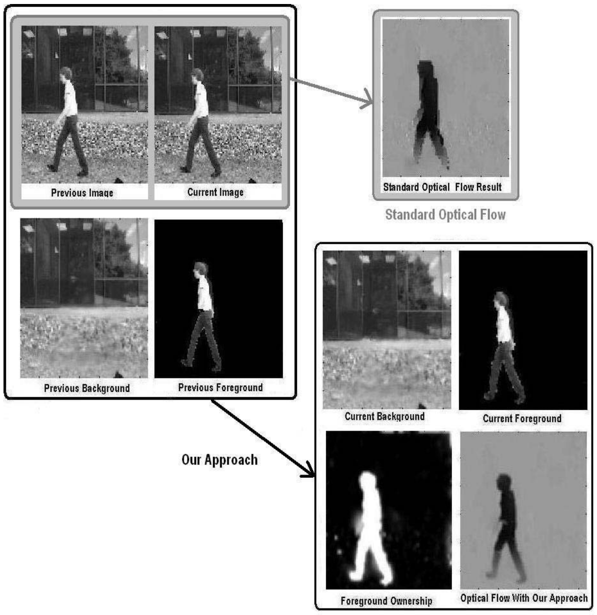


Figure 1. Estimating dense motion in layers for an outdoor sequence. Comparison between the horizontal flow obtained with a standard optical flow technique (sharp motion boundaries but poor detail) and the horizontal component of flow field estimated with our approach (notice the fine details associated with the head and foot). Right bottom box shows the results of our approach: A layer mask indicating the foreground ownership and appearance models for two layers.

background appearance; 4) it can be used to segment the scene into layers; 5) the layers help localize motion boundaries and reduce over-smoothing.

In an Expectation-Maximization framework, we alternate between solving for the layer "ownership" weights and the estimates of motion and appearance. These appearance and motion models are simultaneously estimated in a maximum a posteriori (MAP) framework. First appearance models are estimated, holding motion flow fields fixed and then appearance models are fixed and flow fields are refined. Figure 1 illustrates the method.

In the current model, there is no complex reasoning about depth ordering and consequently the method may not produce assignments of motion and appearance to layers that are intuitive. Regardless, the resulting motion estimates, however, benefit from the stabilizing effects of the appearance model and the motion boundaries more accurately correspond to the object boundaries in the scene.

2. Related Work

While the obvious goal of motion estimation is to compute how things move in images, the goal is also to discern something about the structure of the scene (i.e. "what goes with what"). It has been suggested that knowing the image motion would facilitate segmentation of the scene into physically meaningful regions. Conversely, knowing the segmentation can facilitate the accurate estimation of optical flow [4]. We address the problem of coupling these processes in a unified framework.

Layered models of optical flow have been one of the key paradigms for simultaneously segmenting the scene and estimating its motion [1, 10, 20]. In particular, mixture model frameworks make a soft assignment of pixels to layers. Unfortunately, this segmentation does not typically enforce spatial coherence between neighboring pixels and may, hence, be quite sparse. Additionally, these methods are typically limited to parametric motion models or highly constrained motions [20]. These methods also typically lack any explicit models of spatial coherence between pixels. Here we extend the mixture model framework to estimate dense optical flow in image layers. Each pixel can belong to one of a number of layers and consequently may have multiple interpretations. Traditional optical flow priors are extended to this framework and result in spatially coherent segmentations.

One of the perennial problems of optical flow estimation is the precise localization of motion boundaries. We argue that this problem is due, in part, to the lack of an appearance-based segmentation. Image segmentation itself is a hard problem however and mistakes in segmentation may affect the optical flow estimates. Consequently, we seek a coupled solution.

There have been attempts to couple the motion and appearance segmentation problems in various ways; for example, by exploiting explicit boundary contour processes and the statistics of image regions [17]. We take a different approach that draws from the tracking literature. Optical flow is typically viewed as a 2-frame (or n-frame) problem where the assumption of brightness constancy between the frames is exploited to compute motion; when the next frames come along the previous ones are forgotten. This is in contrast to tracking approaches that model the appearance of an object and can, hence, track its motion over many frames. We argue that optical flow estimation suffers from not having any persistent appearance model characterizing what is being "tracked".

Recent tracking work has exploited mixture models to track regions over many frames while incrementally estimating a model of the region's appearance [11]. We extend these ideas to the problem of dense flow estimation and adaptively estimate a pixelwise appearance model in multiple layers. Having such an appearance model serves a number of purposes. First, (if it is correct) it provides additional constraints on the optical flow that help reduce the effects of noise. Second, the temporal persistence of appearance aids in segmenting the scene into coherent regions. This latter point aids in the precise localization of motion boundaries.

Previous approaches have had somewhat similar goals. Tao *et al.* [18] estimate layered parametric motion with foreground regions being modeled by Gaussian blobs. They learned a model for the appearance of the layers and estimated this over time in a Bayesian framework. This has recently been extended to explicitly model the ordering of the layers and their occlusion [7]. While our approach does not attempt to reason explicitly about depth ordering, we go beyond previous work to model general, dense, optical flow and to formulate the problem of layered appearance and motion estimation in a unified mixture model framework.

Similar goals have been pursued by [6, 12, 13]. In contrast to their work, we formulate the appearance estimation and layer recovery problem in a gradient-based optical flow framework. This allows us to exploit traditional techniques such as coarse-to-fine estimation, robust regularization, and continuous optimization and thus to compute dense estimates of optical flow and appearance in a relatively straightforward fashion.

Finally, we should note that our approach exploits a temporal coherence prior for both motion and appearance. Previous approaches have exploited temporal coherence of flow [2, 3, 16, 19] but they did not also model appearance. We also formulate the Bayesian problem in a robust way and the optimization then extends previous approaches to robust optical flow estimation [8, 10, 15, 21].

3. Problem Formulation

In this section, we first model dense motion estimation in a Bayesian framework and then extend it to appearance estimation and finally develop a unified Bayesian framework to simultaneously estimate the appearance of multiple image layers and their corresponding dense flow fields from image sequences. Both the motion and appearance models adapt over time and the probabilistic formulation can be used to provide a segmentation of the scene into foreground/background regions. This extension of mixture models includes priors for the spatial and temporal coherence of motion and appearance.

3.1. Standard Bayesian flow formulation

The standard optical flow problem can be formulated as the maximization of the posterior probability

$$\arg\max_{\mathbf{U}_t} \; P(\mathbf{U}_t | \bar{\mathbf{I}}_t, \mathbf{U}_{t-1}) \tag{1}$$

where $\mathbf{U}_t$ represents the horizontal and vertical components of the optical flow field at time t and $\bar{\mathbf{I}}_t$ are the image observations for time $0, ..., t$, $\bar{\mathbf{I}}_t = [I_0, I_1, ..., I_{t-1}, I_t]$.

Using standard Markov assumptions and Bayes' rule, we rewrite the posterior probability as

$$\begin{aligned} P(\mathbf{u}_t | \bar{\mathbf{I}}_t, \mathbf{u}_{t-1}) \quad \propto \quad & P(I_t | I_{t-1}, \mathbf{u}_t) \\ & P(\mathbf{u}_t | \mathbf{u}_{t-1}) \\ & P(\mathbf{u}_t | \mathbf{u}_t(\mathcal{G}_\mathbf{x})) \end{aligned}$$

where, now, $\mathbf{u}_t = (u_t(\mathbf{x}), v_t(\mathbf{x}))$ is the horizontal and vertical flow at a pixel $\mathbf{x}$ and $\mathcal{G}_\mathbf{x}$ is the set of four neighbors for pixel $\mathbf{x}$. The global posterior in equation (1) is the product of this local posterior over all image locations (assuming conditional independence of neighbouring pixels). Here $\mathbf{U}_t$ is the optical flow field over the whole image and $\mathbf{u}_t$ is the optical flow field at a particular pixel.

Note that above posterior probability holds at every pixel $\mathbf{x}$ unless otherwise specified in the rest of the text. We omit $\mathbf{x}$ in our notation for the sake of simplicity.

Here $P(I_t | I_{t-1}, \mathbf{u}_t)$ is the observation likelihood that associates successive images with the motion that is being sought. It corresponds to the image brightness constancy assumption. The temporal and spatial coherence of motion are represented with the prior probabilities $P(\mathbf{u}_t | \mathbf{u}_{t-1})$ and $P(\mathbf{u}_t | \mathbf{u}_t(\mathcal{G}_\mathbf{x}))$ respectively. The temporal term simply enforces that the flow at the current instant is similar to the flow at the previous instant. The spatial term is a standard one based on the difference between neighboring horizontal and vertical flow values. All these terms are represented with a robust likelihood function [3]. For optimization we minimize the negative log of the posterior and these terms become robust error terms. Details are provided below.

3.2. Introducing Appearance Model

Let A_t be an *appearance model* (intensity-based model) at time t. This is introduced as a "memory" of what is being tracked and is lacking from standard optical flow formulations. It is introduced into the posterior as follows:

$$\begin{array}{rcl} P(A_t, \mathbf{u}_t | A_{t-1}, \bar{\mathbf{I}}_t, \mathbf{u}_{t-1}) & \propto & P(I_t | I_{t-1}, \mathbf{u}_t, A_t) \\ & & P(A_t | A_{t-1}) \\ & & P(\mathbf{u}_t | \mathbf{u}_{t-1}) \\ & & P(\mathbf{u}_t | \mathbf{u}_t(\mathcal{G}_\mathbf{x})). \end{array}$$

Here $P(I_t|I_{t-1}, \mathbf{u}_t, A_t)$ is the likelihood term and $P(A_t|A_{t-1})$ represents the temporal appearance prior. The goal here is to incrementally estimate the appearance model A_t by taking into account the observed image, the past appearance and the motion. The details will be described below.

3.3. Introducing Layers

To model the complexity of natural images where objects move and occlude each other, we introduce the notion of layers into the dense flow formulation. In particular, we introduce layers with both appearance and motion and estimate these from an image sequence. Here we focus on a simple case of two layers which can be thought of (roughly) as "foreground" and "background."

The posterior is now written as

$$P(\mathbf{A}_t, \mathbf{M}_t | \bar{\mathbf{I}}_t, \mathbf{A}_{t-1}, \mathbf{M}_{t-1}) \qquad (2)$$

where $\mathbf{A}_t = (A_t^b(\mathbf{x}), A_t^f(\mathbf{x}))$ are the appearance (intensity-based) models for foreground and background at every pixel location $\mathbf{x}$ and $\mathbf{M}_t = (\mathbf{u}_t^b, \mathbf{u}_t^f)$ are corresponding motion flow fields. Here $\mathbf{u}_t^b = (u_t^b(\mathbf{x}), v_t^b(\mathbf{x}))$ and $\mathbf{u}_t^f = (u_t^f(\mathbf{x}), v_t^f(\mathbf{x}))$. The superscripts b and f stand for background and foreground respectively.

Once again, the posterior probability can be simplified as

$$\begin{array}{rcl} P(\mathbf{A}_t, \mathbf{M}_t | \bar{\mathbf{I}}_t, \mathbf{A}_{t-1}, \mathbf{M}_{t-1}) & \propto & P(I_t | \mathbf{A}_t, \mathbf{M}_t, I_{t-1}) \\ & & P(\mathbf{A}_t | \mathbf{A}_{t-1}, \mathbf{M}_t) \\ & & P(\mathbf{M}_t | \mathbf{M}_{t-1}) \\ & & P(\mathbf{M}_t | \mathbf{M}_t(\mathcal{G}_\mathbf{x})). \end{array}$$

The appearance of the layers is assumed to change gradually and this temporal coherence is modeled by

$$P(\mathbf{A}_t | \mathbf{A}_{t-1}, \mathbf{M}_t) = \prod_{i=b,f} P(A_t^i(\mathbf{x}) | A_{t-1}^i(\mathbf{x}), \mathbf{u}_t^i) \quad (3)$$

where appearance model at the current time instant is associated with the appearance model at previous time instant via the motion of the corresponding layer.

The temporal and spatial coherence of motion are represented with

$$P(\mathbf{M}_t | \mathbf{M}_{t-1}) = \prod_{i=b,f} P(\mathbf{u}_t^i(\mathbf{x}) | \mathbf{u}_{t-1}^i(\mathbf{x})) \qquad (4)$$

and

$$P(\mathbf{M}_t | \mathbf{M}_t(\mathcal{G}_x)) = \prod_{i=b,f} P(\mathbf{u}_t^i(\mathbf{x}) | \mathbf{u}_t^i(\mathcal{G}_\mathbf{x})) \qquad (5)$$

respectively.

Assuming that each image in the sequence can be separated into foreground and background layers, the likelihood of observing image I_t can be represented as a mixture model

$$\begin{array}{rcl} P(I_t | \mathbf{A}_t, \mathbf{M}_t, I_{t-1}) & = & m_b\, p(I_t | A_t^b, \mathbf{u}_t^b, I_{t-1}) \\ & & +m_f\, p(I_t | A_t^f, \mathbf{u}_t^f, I_{t-1}) \\ & & +m_o p_o(I_t). \end{array} \qquad (6)$$

The probability of each pixel belonging to different layers is given by the mixture probabilities m_b, m_f and m_o; these mixing probabilities sum to 1, where m_o is the fixed outlier probability. In our experiments, we set $m_o = 0$.

For any pixel in the current image, the likelihood for each layer is

$$P(I_t | A_t^i, \mathbf{u}_t^i, I_{t-1}) = P(I_t | I_{t-1}, \mathbf{u}_t^i) \cdot P(I_t | A_t^i). \quad (7)$$

This likelihood simply enforces that the successive images in the sequence look similar when aligned using the motion of the corresponding layer and that each appearance model be similar to the current image in regions with high mixing probability (since the likelihood for each layer is multiplied by the corresponding mixing probability as can be seen in Eq. 6). The implementation details of these terms are provided below.

3.4. Optimization

Given images in a sequence, $\bar{\mathbf{I}}_t$, as well as flow fields and appearance models at the previous time instant, we seek the appearance models A_t^b and A_t^f and their corresponding flow fields $\mathbf{u}_t^b$ and $\mathbf{u}_t^f$ and the mixture probabilities m_b and m_f which provide a maximum likelihood fit to the data set.

The foreground/background separation problem can be considered as the maximization of the posterior probability. At every new time instant, we need to estimate the appearance models and their corresponding motion. We use the Expectation-Maximization (EM) algorithm [5] to solve for the $(A_t^i, \mathbf{u}_t^i)$ pairs.

According to the generalized EM algorithm, a local optimal solution can be achieved by iteratively optimizing the following function with respect to A_t^i and $\mathbf{u}_t^i$

$$\begin{array}{rcl} L(\mathbf{A}_t, \mathbf{M}_t) & = & log\, P(I_t | \mathbf{A}_t, \mathbf{M}_t, I_{t-1}) \\ & & +log\, P(\mathbf{A}_t | \mathbf{A}_{t-1}, \mathbf{M}_t) \\ & & +log\, P(\mathbf{M}_t | \mathbf{M}_{t-1}) \\ & & +log\, P(\mathbf{M}_t | \mathbf{M}_t(\mathcal{G}_\mathbf{x})) \\ & & +\lambda(1 - m_o - m_b - m_f) \end{array} \quad (8)$$

Note that the constraint that the mixing probabilities sum to one is added with Lagrange multiplier.

At a local extremum it can be shown that the parameters m_i and $\mathbf{A}_t, \mathbf{M}_t$ must satisfy

$$\begin{aligned} & q_i \cdot \frac{\partial}{\partial A_t^i}\left(log\ P(I_t|A_t^i)\right) \\ & + \frac{\partial}{\partial A_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right) = 0 \end{aligned} \qquad (9)$$

and

$$\begin{aligned} & q_i \cdot \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(I_t|I_{t-1}, \mathbf{u}_t^i)\right) \\ & + \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right) \\ & + \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_{t-1}^i)\right) \\ & + \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_t^i(\mathcal{G}_\mathbf{x}))\right) = 0. \end{aligned} \qquad (10)$$

Here q_i represents the *ownership probability*, that is the probability that the observed image value I_t belongs to the $i^t h$ layer. The ownership weights are defined by

$$q_i = \frac{m_i \cdot P(I_t|A_t^i, \mathbf{u}_t^i, \bar{\mathbf{I}}_{t-1})}{\sum_{j=b,f,o} m_j \cdot P(I_t|A_t^j, \mathbf{u}_t^j, \bar{\mathbf{I}}_{t-1})}. \qquad (11)$$

These equations for a maximum likelihood fit have been previously derived

simply by requiring that the partial derivative of $L(\mathbf{A}_t, \mathbf{M}_t)$ with respect to the parameters m_i and $\mathbf{A}_t, \mathbf{M}_t$ must vanish [10, 14]. For the details of the derivations see Appendix.

Given an initial guess for motion and appearance models, we first estimate the ownership probabilities q_i for each layer. This is the expectation step. Given these ownership probabilities, we compute the appearance models and motions that optimize the Eq. 9 and 10 in the maximization step.

The likelihoods and priors are modeled by a *t*- distribution of degree 3. The robust error function is given by the negative log.

$$\rho(x, \sigma, \alpha) = -log\left[\left(\frac{2\sigma^3}{\pi(\sigma^2 + x^2)^2}\right)^{\alpha}\right] \qquad (12)$$

where α is a parameter that determines the relative importance of each of the likelihood and prior terms. We define the derivative of this function as $\psi(x, \sigma, \alpha)$

$$\psi(x, \sigma, \alpha) = \frac{d}{dx}\rho(x, \sigma, \alpha) = \alpha\ \frac{-4x}{\sigma^2 + x^2}.$$

After the derivations, the actual equations in the M-step are as follows

$$\begin{aligned} & u^i(x)^{n+1} = u^i(x)^n \\ & - q_i(x) \cdot \psi(I_t(x) - I_{t-1}(x - u_t^i), \sigma_{IIi}, \alpha_{IIi}) \\ & - \psi(A_t^i(x) - A_{t-1}^i(x - u_t^i), \sigma_{AAi}, \alpha_{AAi}) \\ & - \psi(u_t^i(x) - u_{t-1}^i(x), \sigma_{temp_i}, \alpha_{temp_i}) \\ & - \sum_{\mu \epsilon \mathcal{G}_x} \psi(u_t^i(x) - u_t^i(\mu), \sigma_{sp_i}, \alpha_{sp_i}) \end{aligned}$$

and

$$\begin{aligned} & A^i(x)^{n+1} = A^i(x)^n \\ & - q_i(x) \cdot \psi(I_t(x) - A_t^i(x), \sigma_{IAi}, \alpha_{IAi}) \\ & - \psi(A_t^i(x) - A_{t-1}^i(x - u_t^i), \sigma_{AAi}, \alpha_{AAi}) \end{aligned}$$

where α_{IIi}, α_{AAi}, α_{temp_i}, α_{sp_i} are the α parameters for the image likelihood, appearance prior, and temporal and spatial motion priors respectively.

Intuitively, the above equations (derived from Eq. 9 and 10) can be interpreted as follows: there are two terms that contribute to the appearance models in the M-step for appearance optimization. The first term indicates that the appearance model should adapt to the new information in the current image, change appearance if necessary, and regions with high ownership weights are more likely to be adapted to the current image since the whole term is multiplied by q_i. For regions of low ownership weight, the second term dominates and associates successive appearances using the corresponding motion. Simply, this appearance prior term suggests that the appearance from the previous time instant be maintained after being warped by the layer motion. These two terms compete with each other and pull the optimal solution towards their extrema. The parameters used in modeling these terms become crucial in determining which term pulls the solution towards its extremum. Since we are using Eq. 12, we have variance parameters σ for each term and in addition to that we have α values for each term that determine the relative importance of these terms.

The M-step for motion optimization has four terms. The first term aligns successive frames in the sequence using the motion of the appropriate layer, but only in regions of high ownership weight. That is, background (foreground) motion should explain the correspondence between two successive images only in regions where background (foreground) ownership is high. The second term aligns successive appearance models using the motion of the appropriate layer. The third and the fourth terms suggest that the motion at a pixel should be similar to that of neighboring pixels in space and time.

3.5. Updating mixing probabilities

In our formulation, the mixing probabilities are simply the ownership weights. Yet, we expect these mixing probabilities which represent the assignment of the pixels to layers to be stable over time. To model this, we gradually update them as the ownership probabilities change. Moreover, we expect the background motion to be slower than those of the foreground and adding a prior that models this assumption helps separate foreground and background layers.

We initially set $m_b = m_f = 0.5$ and then the mixing probabilities for next time instant are updated by a linear combination of ownership weights and motion priors as follows

$$(m_i)_{t+1} = (1-\alpha_1-\alpha_2)(m_i)_t+\alpha_1 q_i+\alpha_2 p(u_t^i) \quad (13)$$

where $p(u_t^b) = exp(||u_t^b||, \sigma_{motion\_prior\_b})$ and $p(u_t^f) = 1 - exp(||u_t^f||, \sigma_{motion\_prior\_f})$. The mixing probabilities of each layer act as a prior on every pixel representing the probability of each pixel belonging to that layer. In our experiments, $\alpha_1 = \alpha_2 = 0.3$. The remaining parameters are specified in the Appendix.

4. Experimental Results

The flow diagram of our approach is illustrated in Figure 3. To cope with large motions and accelerate the convergence, a hierarchical process is employed. A P-level image pyramid is created and the estimation starts from the coarsest level. At each level, current appearance estimates are warped by the flow field estimates and projected onto a lower level as an initial estimate. We alternate the optimization of motion and appearance models and computation of ownership weights with optimized parameters at each level. Since it is difficult to optimize $(A_t^i, \mathbf{u}_t^i)$ pairs simultaneously, we adopt the strategy of improving each of them in turn with the other one fixed. This is a generalized EM algorithm and it can be proven that it converges to a local minimum. The optimization process is summarized in Figure 3. The computational cost is approximately the number of layers times the duration that standard optical flow [3] takes for every frame.

We obtained the background appearance model, for $t = 0$, by watching the scene long enough with a static camera and taking the median of those observations. The initial foreground appearance is currently chosen manually by determining the bounding box of object of interest. Once appearance models and corresponding flow fields are computed at time t, we warp these appearance models forward by the flow field estimates and use them as initial estimates at time $t + 1$.

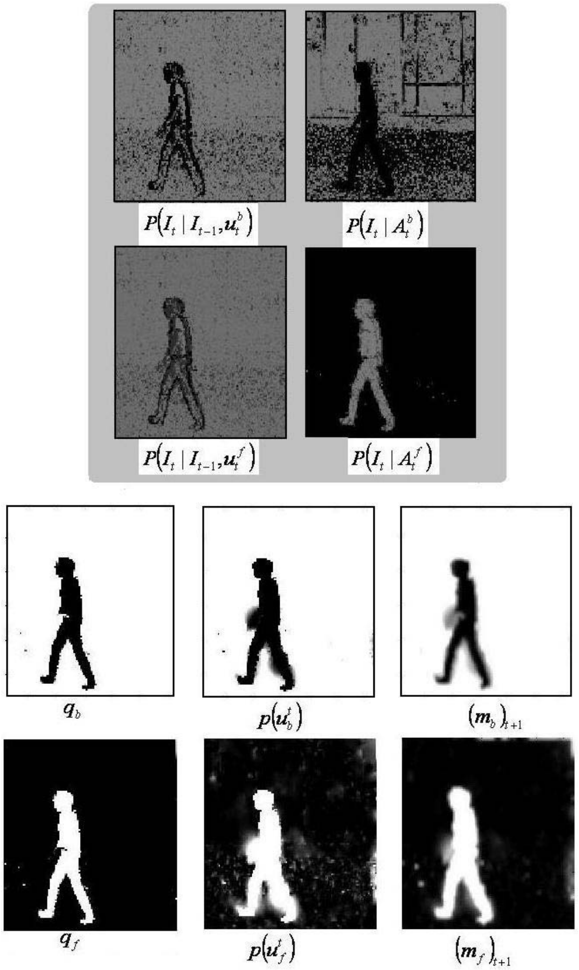
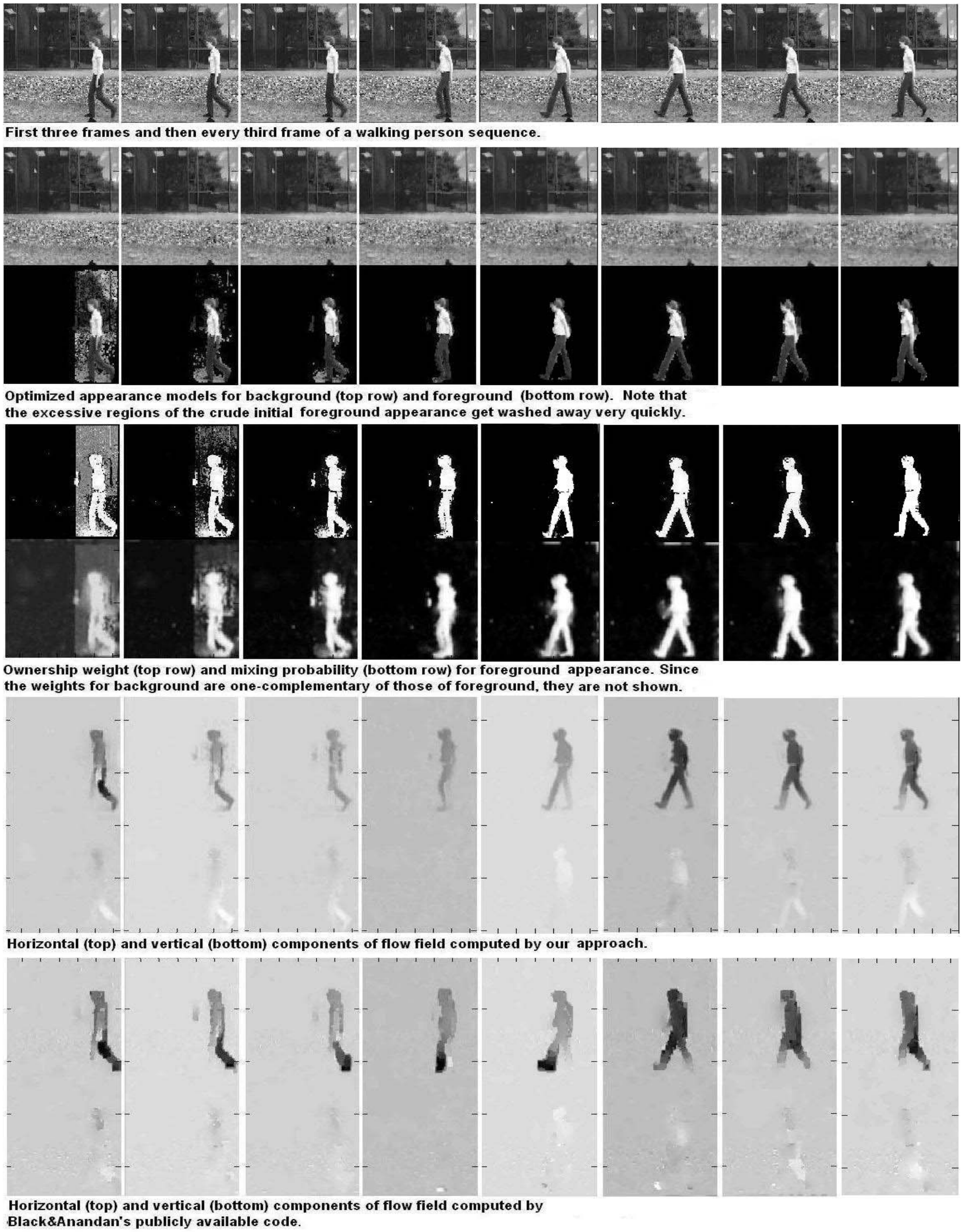


Figure 2. Intermediate results for the frames in Figure 1.

We experimented with our approach on a video clip of a walking person. Figures 4 - 5 illustrate the appearance models, flow fields, ownership weights and mixing probabilities for the background and foreground layers. We also computed the optical flow field by the approach in [3] for comparison. Inclusion of appearance models and layers in our approach visually improve the optical flow results as can be seen in Figures 4 - 5. In this paper, we focused on experiments with a static camera. Even in this case, our method helps to deal with the challenging problem of adapting the background appearance model and makes it more robust to illumination variations. In this simple case of static background, traditional background subtraction techniques could be employed. These techniques typically exhibit false positive and false negative detections which are treated with



First three frames and then every third frame of a walking person sequence.

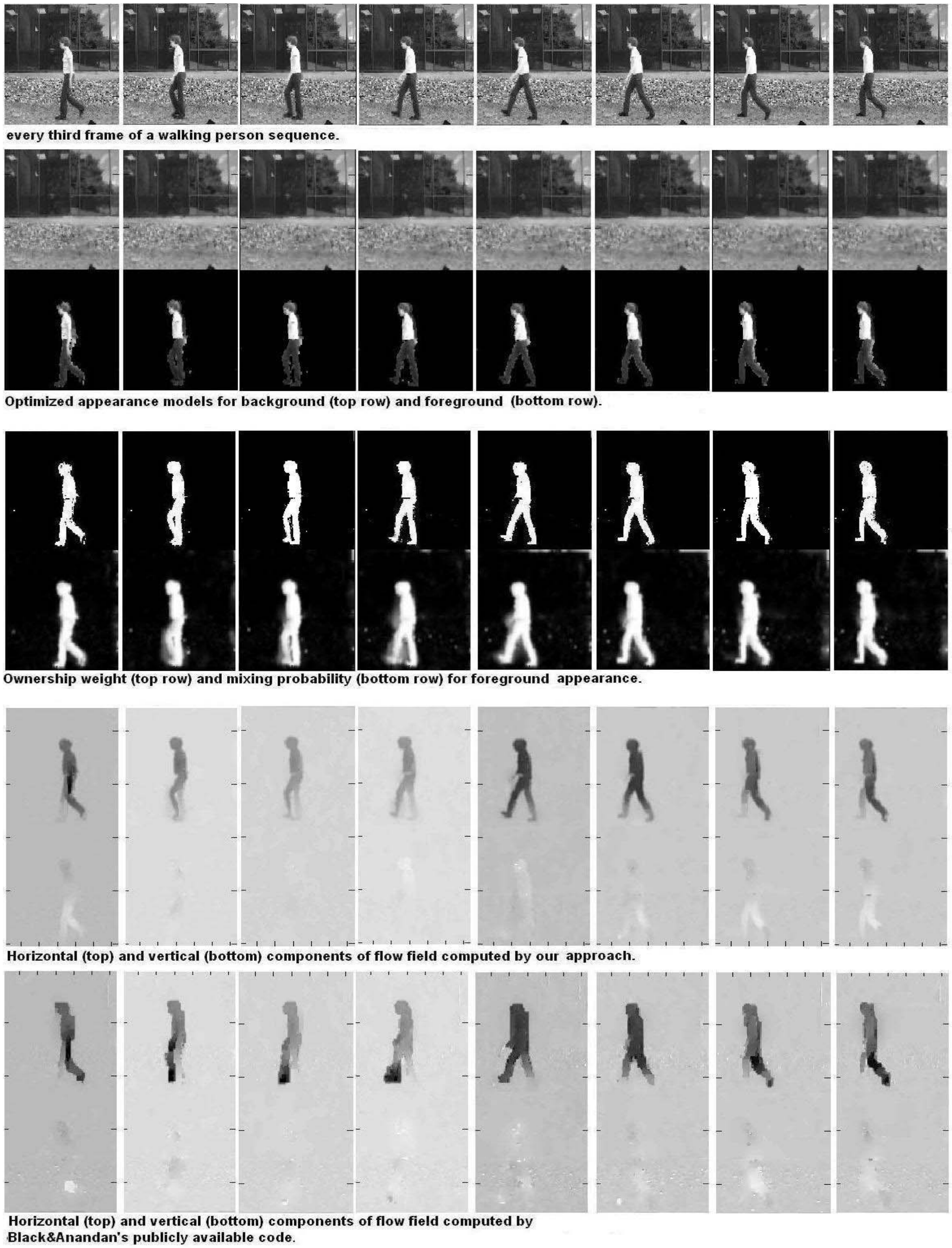
Optimized appearance models for background (top row) and foreground (bottom row). Note that the excessive regions of the crude initial foreground appearance get washed away very quickly.

Ownership weight (top row) and mixing probability (bottom row) for foreground appearance. Since the weights for background are one-complementary of those of foreground, they are not shown.

Horizontal (top) and vertical (bottom) components of flow field computed by our approach.

Horizontal (top) and vertical (bottom) components of flow field computed by Black&Anandan's publicly available code.

Figure 4. Results of our approach for a walking person sequence (first 3 frames and then every 3^{rd} frame) displayed. Note that the crude initial foreground appearance improves quickly after first few frames. The optical flow fields obtained by our approach are compared to those computed by Black and Anandan's publicly available code.



every third frame of a walking person sequence.

Optimized appearance models for background (top row) and foreground (bottom row).

Ownership weight (top row) and mixing probability (bottom row) for foreground appearance.

Horizontal (top) and vertical (bottom) components of flow field computed by our approach.

Horizontal (top) and vertical (bottom) components of flow field computed by Black&Anandan's publicly available code.

Figure 5. Results for every third frame following from Figure 4.

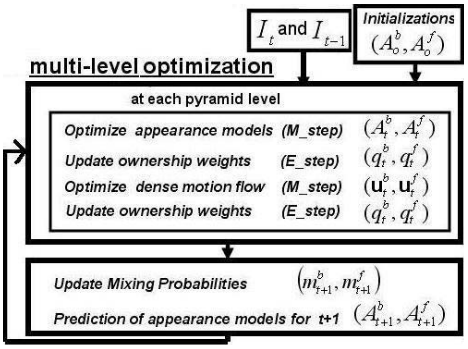

I_t and I_{t-1}
Initializations (A_o^b, A_o^f)
multi-level optimization
at each pyramid level
Optimize appearance models (M_step) (A_t^b, A_t^f)
Update ownership weights (E_step) (q_t^b, q_t^f)
Optimize dense motion flow (M_step) $(\mathbf{u}_t^b, \mathbf{u}_t^f)$
Update ownership weights (E_step) (q_t^b, q_t^f)
Update Mixing Probabilities (m_{t+1}^b, m_{t+1}^f)
Prediction of appearance models for t+1 (A_{t+1}^b, A_{t+1}^f)

Figure 3. Flow diagram of the approach

post-hoc filtering. In contrast, our approach addresses these problems in a Bayesian framework that enforces spatial and temporal continuity. Moreover, the approach provides a dense estimate of the motion. Future work includes experimentation with a moving camera or dynamic background.

The likelihoods that contribute to the ownership weight (Eq. 11) and adaptation of the mixing probabilities (Eq. 13) for frame 67 are illustrated in Figure 2. The ownership weights inherently act as a mask when optimizing the appearance models: regions with high ownership weight are quickly adapted to the current image whereas in regions of low ownership weight, this adapting occurs gradually. For those regions, only the appearance model prior term that associates successive appearance via the corresponding motion of the layer is active. So, as the person moves, the regions with high foreground ownership weight, converge to the appearance of the walking person whereas the appearance of the regions of low ownership weight are associated with the appearance model at the previous time instant (warped by the motion layer which is being simultaneously computed). Since the mixing probabilities are adapted over time, the ownership weights do not diminish immediately after disocclusion and the appearance of disoccluded regions maintains the values assigned to them previously. Modeling and integrating some appearance prior to deal with this disocclusion problem is an immediate extension for future work.

5. Conclusions

In this paper, we presented a Bayesian framework for estimating dense optical flow over time that explicitly estimates and exploits a persistent model of image appearance. We also exploited prior models that express how motion and appearance may change over time. We extended mixture model methods to the case of dense (rather than parametric) flow estimation and derived a mixture model formulation that includes explicit spatial and temporal priors.

The method is an extension of standard robust optical flow methods and it estimates dense, subpixel-accurate, flow fields and foreground and background appearance.

Future work involves estimating the number of layers and integrating the ordering of the layers into our framework.

References

[1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum likelihood estimation of mixture models and mdl encoding. In *ICCV95*, pages 777–784, 1995.

[2] M. Black and P. Anandan. Robust dynamic motion estimation over time. In *CVPR91*, pages 296–302, 1991.

[3] M. Black and P. Anandan. A framework for the robust estimation of optical flow. In *ICCV93*, pages 231–236, 1993.

[4] M. J. Black and A. D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *PAMI*, 18(10):972–986, 1996.

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

[6] B. Frey, N. Jojic, and A. Kannan. Learning appearance and transparency manifolds of occluded objects in layers. *CVPR03*, I:45–52, 2003.

[7] E. Hayman and H. Tao. A backgorund layer model for object tracking through occlusion. In *ICCV03*, pages 1079–1085, 2003.

[8] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using markov random fields. *PAMI*, 15(2):1217–1232, 1993.

[9] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[10] A. Jepson and M. Black. Mixture models for optical flow computation. In *CVPR93*, pages 760–761, 1993.

[11] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. In *CVPR01*, pages 415–422, 2001.

[12] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *CVPR01*, 2001.

[13] N. Jojic, B. Frey, and A. Kannan. A generative model of dense optical flow in layers. *University of Toronto TR PSI-2001-11*, August 2001.

[14] G. McLachlan and K. Basford. Mixture models: inference and applications to clustering. *Marcel Dekker Inc.*, 1988.

[15] E. Memin and P. Perez. A multigrid approach for hierarchial motion estimation. In *ICCV98*, pages 933–938, 1998.

[16] D. W. Murray and B. F. Buxton. Scene segmentation from visual motion using global optimization. *PAMI*, 9(2):220–228, 1987.

[17] N. Paragios and R. Deriche. Geodesic active regions for motion estimation and tracking. In *ICCV*, pages 688–694, 1999.

[18] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *PAMI*, 24(1):75–89, 2002.

[19] J. Weickert and C. Schnorr. Variational optic flow computation with a spatio-temporal smoothness constraint. *Journal of Mathematical Imaging and Vision*, 14(3):245–255, 2001.

[20] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *CVPR97*, pages 520–526, 1997.

[21] M. Ye, R. Haralick, and L. Shapiro. Estimating optical flow using a global matching formulation and graduated optimization. *PAMI*, 25(12):1625–1630, 2003.

6. Appendix

According to the generalized EM algorithm, a locally optimal solution can be achieved by iteratively optimizing Eq. 8 wrt to parameters A_t^i and $\mathbf{u}_t^i$.

6.1. Derivation of M-step and E-steps for Appearance Optimization

Taking derivative of $L(\mathbf{A}_t, \mathbf{M}_t)$ (Eq. 8) wrt the appearance models, we get

$$\begin{aligned}\frac{\partial L(\mathbf{A}_t, \mathbf{M}_t)}{\partial A_t^i} &= \frac{m_i \cdot \partial P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})/\partial A_t^i}{\sum_{j=b,f,o} m_j \cdot P(I_t|A_t^j, \mathbf{u}_t^j, I_{t-1})} \\ &+\frac{\partial}{\partial A_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right)\end{aligned}$$

Replacing $\partial P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})/\partial A_t^i$ by

$$P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1}) \cdot \frac{\partial}{\partial A_t^i}\left(log\ P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})\right)$$

and rewriting the equation with this replacement, we get

$$\begin{aligned}&\frac{\partial L(\mathbf{A}_t, \mathbf{M}_t)}{\partial A_t^i} \\ &= \frac{m_i P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})\frac{\partial(log\ P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1}))}{\partial A_t^i}}{\sum_{j=b,f,o} m_j \cdot P(I_t|A_t^j, \mathbf{u}_t^j, I_{t-1})} \\ &+\frac{\partial}{\partial A_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right).\end{aligned}$$

The replacement trick above lets us define

$$q_i = \frac{m_i \cdot P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})}{\sum_{j=b,f,o} m_j \cdot P(I_t|A_t^j, \mathbf{u}_t^j, I_{t-1})}. \tag{14}$$

Here q_i represents the *ownership probability*, that is the probability that the observed image I_t belongs to the i^{th} layer. Given some initial values for the appearance models and motion, these ownership weights are computed as the expectation, or E-step.

Then, the M-step is formulated in compact form as

$$\begin{aligned}\frac{\partial L(\mathbf{A}_t, \mathbf{M}_t)}{\partial A_t^i} &= q_i \cdot \frac{\partial}{\partial A_t^i}\left(log\ P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})\right) \\ &+\frac{\partial}{\partial A_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right).\end{aligned}$$

At a local extremum, the right hand side of the above equation will be equal to zero. Since the likelihood is defined as in Eq. 7, the above equation reduces to

$$\begin{aligned}&q_i \cdot \frac{\partial}{\partial A_t^i}\left(log\ P(I_t|A_t^i)\right) \\ &+\frac{\partial}{\partial A_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right) = 0.\end{aligned}$$

6.2. Derivation of M-step and E-steps for Motion Optimization

Similarly if we take derivative of $L(\mathbf{A}_t, \mathbf{M}_t)$ wrt $\mathbf{u}_t^i$, the M-step for motion optimization will be

$$\begin{aligned}\frac{\partial L(\mathbf{A}_t, \mathbf{M}_t)}{\partial u_t^i} &= q_i \cdot \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(I_t|A_t^i, \mathbf{u}_t^i, I_{t-1})\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_{t-1}^i)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_t^i(\mathcal{G}_\mathbf{x}))\right).\end{aligned}$$

Since the likelihood is defined as in Eq. 7, at a local extremum, the above equation reduces to

$$\begin{aligned}&q_i \cdot \frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(I_t|I_{t-1}, \mathbf{u}_t^i)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(A_t^i|A_{t-1}^i, \mathbf{u}_t^i)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_{t-1}^i)\right) \\ &+\frac{\partial}{\partial \mathbf{u}_t^i}\left(log\ P(\mathbf{u}_t^i|\mathbf{u}_t^i(\mathcal{G}_\mathbf{x}))\right) = 0.\end{aligned}$$

6.3. Parameters of Our Approach

The following parameters were used:

$\alpha_{IIb} = 3, \alpha_{IAb} = 3, \alpha_{AAb} = 4,$
$\alpha_{IIf} = 6, \alpha_{IAf} = 3, \alpha_{AAf} = 1,$
$\alpha_{sp_i} = 2.5, \alpha_{temp_i} = 2, \alpha_{motion\_prior\_i} = 1,$

$\sigma_{IIi} = 10, \sigma_{IAi} = 10, \sigma_{AAb} = 6, \sigma_{AAf} = 13,$
$\sigma_{sp_i} = 1, \sigma_{temp_i} = 1.5, \sigma_{motion\_prior\_i} = 0.1.$