Generalized Background Subtraction Based on Hybrid Inference by Belief Propagation and Bayesian Filtering

Suha Kwak   Taegyu Lim   Woonhyun Nam   Bohyung Han   Joon Hee Han
Department of Computer Science and Engineering, POSTECH, Korea
{mercury3,limtk77,xgene,bhhan,joonhan}@postech.ac.kr

Abstract

We propose a novel background subtraction algorithm for the videos captured by a moving camera. In our technique, foreground and background appearance models in each frame are constructed and propagated sequentially by Bayesian filtering. We estimate the posterior of appearance, which is computed by the product of the image likelihood in the current frame and the prior appearance propagated from the previous frame. The motion, which transfers the previous appearance models to the current frame, is estimated by nonparametric belief propagation; the initial motion field is obtained by optical flow and noisy and incomplete motions are corrected effectively through the inference procedure. Our framework is represented by a graphical model, where the sequential inference of motion and appearance is performed by the combination of belief propagation and Bayesian filtering. We compare our algorithm with the existing state-of-the-art technique and evaluate its performance quantitatively and qualitatively in several challenging videos.

1. Introduction

Background subtraction typically refers to an algorithm to detect moving objects in the scene when the video is captured by a stationary camera. Various background subtraction algorithms have been proposed so far [1, 4, 6, 7, 10, 13, 15, 16, 19, 21, 23], where the main concern is adaptive background modeling for each pixel or region in the static monocular camera environment. The separation between foreground objects and background scene in videos captured by a moving camera is much more difficult; motion estimation frequently suffers from various challenges due to complex scene structures, motion blurs and inconsistent features, and modeling and updating foreground and background appearances are not straightforward because of the error accumulation in image registration and appearance modeling procedure. Since moving camera setup assumes

more general environment in background subtraction, we call the problem *generalized background subtraction.*

Generalized background subtraction problem is not completely new, and there are several studies related to foreground and background segmentation in a moving camera environment. Motion segmentation is one of the most popular approaches for this problem, where camera motion is canceled by estimating dominant background motion to identify foreground objects [5, 9]. However, these methods are based on a strong assumption that the background is able to be modeled effectively with a single plane, which is not generally valid. A more advanced approach is the combination of plane and parallax framework, where a homography is first computed to match the features in two consecutive frames and the residual pixels are further registered by parallax estimation [22]. This technique involves less restrictions than the homography-only based algorithms, but still assumes that there exists a dominant plane for matching by homography.

On the other hand, [14] combines image registration and appearance modeling for foreground/background segmentation. In this technique, the factorization method [18] is utilized for image registration since it can handle more general 3-D motions conceptually. However, this algorithm is based on a simple modeling and propagation of foreground and background appearances, and requires a reliable long-term feature tracking method to run the factorization method; it depends heavily on the performance of particle video [12], which is the technique used for the robust motion estimation in [14]. The appearance models by this method are susceptible to be corrupted and unreliable due to error accumulation by temporary failures in motion estimation and feature sparsity in particle video.

We propose a systematic probabilistic inference framework based on the combination of nonparametric belief propagation (BP) and sequential Bayesian filtering for generalized background subtraction. Nonparametric BP is employed for the robust motion estimation resistant to noisy and incomplete observations, and Bayesian filtering propagates the appearance models sequentially in a reliable man-

ner by integrating previous appearances and current image observations.

In our framework, motion is first obtained by optical flow and then estimated by nonparametric BP in Markov Random Field (MRF). The prior appearance models are predicted by integrating the appearance models of the previous frame subject to the currently estimated current motion. The foreground/background likelihood ratio of each pixel is computed based on both of the motion and the prior appearance information; the likelihood ratio determines the label of each pixel. The labels are used to update motion, and the predictions of foreground/background appearances are improved by the updated motion; the labels are re-estimated based on the updated motions and appearances. Such iterative procedure is repeated until convergence in each frame. After the labels converge, the prior appearance models are combined with the foreground/background observations in the current frame to obtain the posterior appearance models in the sequential Bayesian filtering framework.

The rest of the paper is organized as follows. We describe the overview of our algorithm in Section 2, and foreground/background appearance modeling technique with motion estimation is presented in Section 3. Section 4 illustrates the performance of our algorithm with challenging videos.

2. Algorithm overview

We estimate foreground/background appearance models in the videos captured by a moving camera, where a video frame is divided into a regular grid of blocks. Our algorithm is based on the following assumptions:

- At the beginning of a sequence, the major motion in the scene belongs to background and the outliers are considered as foreground.

- Spatially adjacent background (or foreground) areas have similar motions.

- The temporal variations of foreground and background appearances are smooth subject to the proper image registration by motion estimation.

The first assumption is to obtain reliable foreground and background models through motion segmentation in the first few frames. The second and third assumptions allow us to take spatial and temporal evidences for estimating blockwise motion and appearance models.

The graphical model reflecting the last two assumptions is illustrated in Figure 1; the second assumption is embedded by four-neighborhood pairwise MRF on the motion random variables, and the third assumption is reflected by the directed edges in the graphical model. There are two layers—motion and appearance—in the graphical model.
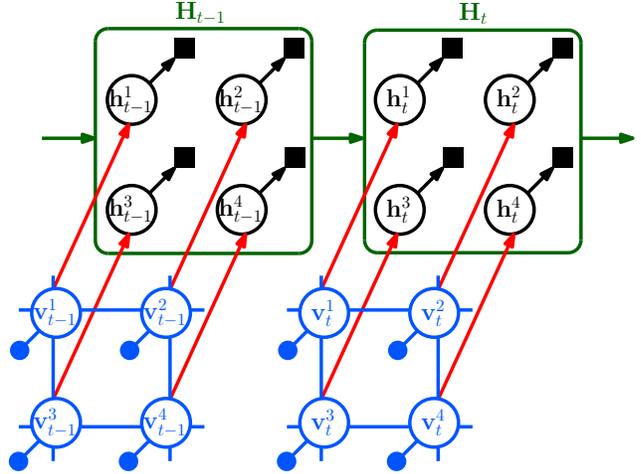

Figure 1. The proposed graphical model for blockwise model estimation. The solid discs and squares represents the optical flow observations for motion models and image observations for appearance models, respectively. The same graphical model is applied to foreground and background.

Let $\mathbf{h}_{\ell,t}^i$ be a random variable following a color distribution based on the pixels with foreground or background labels $\ell \in \{f, g\}$ in the ith block at time t, and $\mathbf{H}_{\ell,t} = \{\mathbf{h}_{\ell,t}^1, \ldots, \mathbf{h}_{\ell,t}^N\}$ be a set of the random variables, where N is the number of the blocks. Also, denote by $\mathbf{v}_{\ell,t}^i$ a random variable representing a motion distribution in the ith block with label ℓ, and by $\mathbf{V}_{\ell,t} = \{\mathbf{v}_{\ell,t}^1, \ldots, \mathbf{v}_{\ell,t}^N\}$ a set of the random variables. Both of $\mathbf{h}_{\ell,t}^i \in \mathbb{R}^3$ and $\mathbf{v}_{\ell,t}^i \in \mathbb{R}^2$ are continuous random variables. Note that $\mathbf{H}_t \in \{\mathbf{H}_{b,t}, \mathbf{H}_{f,t}\}$ and $\mathbf{V}_t \in \{\mathbf{V}_{b,t}, \mathbf{V}_{f,t}\}$.

We estimate all marginal posterior probabilities with respect to each of the random variables in the graphical model. At the beginning of each frame, initial foreground/background labels ($\mathcal{L}_t^{\text{init}}$) are given by a simple motion segmentation; the label at pixel $\mathbf{x}$ at time t is denoted by $\mathcal{L}_t(\mathbf{x}) \in \{f, b\}$. The models of $\mathbf{V}_t$ are estimated by a nonparametric BP based on $\mathcal{M}_t$, which is the pixelwise optical flow in an input image $\mathcal{I}_t$. After the inference of $\mathbf{V}_t$ is completed, the prediction of $\mathbf{H}_t$ denoted by $\mathbf{H}_{t|t-1}$ is estimated based on $\mathbf{V}_t$ and $\mathbf{H}_{t-1}$. Note that the models of $\mathbf{V}_t$ and $\mathbf{H}_{t|t-1}$ are then used to revise $\mathcal{L}_t$. Because different labels organize different foreground/background observations for $\mathbf{V}_t$, the algorithm repeats the model estimation for $\mathbf{V}_t$ and $\mathbf{H}_{t|t-1}$ until the $\mathcal{L}_t$ is converged. Finally, the model of $\mathbf{h}_t^i$ is computed by the product of the prediction (the model of $\mathbf{h}_{t|t-1}^i$) and the observation likelihood from the ith block in $\mathcal{I}_t$; it is a Bayesian filtering approach.

3. Model estimation and labeling by inference

Both of the motion and appearance random variables are in continuous state space; we model their distributions

by Gaussian kernel density estimation. As discussed in Section 2, the motion is estimated by nonparametric BP in MRF, and the appearance model is propagated in a sequential manner by Bayesian filtering. The inference for the graphical model is performed for foreground and background separately but identically, and we sometimes omit the label ℓ in the random variables for simplicity. The details of our hybrid inference algorithm is described below.

3.1. Motion estimation by nonparametric BP

We estimate the motion of the ith block at time t by the marginal posterior probability $p(\mathbf{v}_t^i|\mathcal{M}_t)$. BP is a well-known framework to find such marginal probabilities, which we call believes, for the entire latent variables simultaneously [11]. However, the conventional discrete BP is not available in our case because $\mathbf{v}_t^i$ is in a continuous state space. So, we adopt nonparametric BP whose believes and messages are Gaussian mixtures [17].

We construct a pairwise MRF; the set of vertices $\mathcal{V}$ contains the random variables in $\mathbf{V}_t$ and the set of edges $\mathcal{E}$ represents pairwise neighboring relationships in the regular grid structure of the blocks as in the motion layer in Figure 1. The joint posterior probability with respect to $\mathbf{V}_t$ in the MRF model is given by

$$p(\mathbf{V}_t|\mathcal{M}_t) \propto \prod_{i \in \mathcal{V}} \Phi(\mathbf{v}_t^i, \mathcal{M}_t) \prod_{(i,j) \in \mathcal{E}} \Psi(\mathbf{v}_t^i, \mathbf{v}_t^j). \quad (1)$$

The observation clique potentials Φ are modeled by Gaussian kernel density estimation whose kernel points are the associated image observations in the ith block as

$$\Phi(\mathbf{v}_{\ell,t}^i, \mathcal{M}_t) = \sum_{\{\hat{\mathbf{x}}|\hat{\mathbf{x}} \in R(i), \mathcal{L}_t(\hat{\mathbf{x}})=\ell\}} \alpha \, \mathcal{N}(\mathbf{v}_{\ell,t}^i; \mathcal{M}_t(\hat{\mathbf{x}}), \Sigma_\Phi), \quad (2)$$

where α is a normalized weight, $R(i)$ is a set of pixels in the ith block, and Σ_Φ is kernel bandwidth for motion observation. The compatibility clique potentials Ψ are based on a single Gaussian distribution given by

$$\Psi(\mathbf{v}_t^i, \mathbf{v}_t^j) = \mathcal{N}(\mathbf{v}_t^i - \mathbf{v}_t^j; 0, \Sigma_\Psi), \quad (3)$$

where Σ_Ψ is kernel bandwidth for motion compatibility. The compatibility clique potential between a pair of neighboring blocks encourages the blocks to have similar motions; it corresponds to our second assumption.

The marginal posteriors for each $\mathbf{v}_t^i$ are approximated by sum-product message-passing algorithm [11]. In loopy BP, the messages are synchronously updated by iteration. At the first iteration, the message sent from $\mathbf{v}_t^j$ to $\mathbf{v}_t^i$ is initialized as

$$m_{j \to i}^1(\mathbf{v}_t^i) = \int_{\mathbb{R}^2} \Psi(\mathbf{v}_t^i, \mathbf{v}_t^j) \Phi(\mathbf{v}_t^j, \mathcal{M}_t) d\mathbf{v}_t^j, \quad (4)$$

and the message at iteration $n > 1$ is updated to

$$m_{j \to i}^n(\mathbf{v}_t^i)$$
$$= \int_{\mathbb{R}^2} \Psi(\mathbf{v}_t^i, \mathbf{v}_t^j) \Phi(\mathbf{v}_t^j, \mathcal{M}_t) \prod_{k \in \eta(j) \setminus i} m_{k \to j}^{n-1}(\mathbf{v}_t^j) d\mathbf{v}_t^j, \quad (5)$$

where $\eta(j)$ denotes the set of the indices of the neighboring blocks of the jth block. The messages are iteratively updated until they converge or the predefined number of iterations is reached. Finally, the belief of $\mathbf{v}_t^i$, which means the motion model of ith block, is derived from the incoming messages and its own observation as

$$p(\mathbf{v}_t^i|\mathcal{M}_t) \propto \Phi(\mathbf{v}_t^i, \mathcal{M}_t) \prod_{j \in \eta(i)} m_{j \to i}^*(\mathbf{v}_t^i), \quad (6)$$

where $m_{j \to i}^*(\mathbf{v}_t^i)$ is the message in the last iteration. Note that all messages and believes are Gaussian mixtures. Because we define Ψ as a function of the difference between two latent variables, a message is simply derived from a Gaussian convolution of a product of Gaussian mixtures, which is also a Gaussian mixture (Eq. (4) and (5)). However, the number of mixture components increases exponentially by the product of Gaussian mixtures; our algorithm approximates the original density by a Gibbs sampling [3] and bounds the number of mixture components [17] during motion estimation.

The nonparametric BP for motion model estimation reduces the noise in optical flow and recovers the missing background motion by message passing as in Figure 2. So, our algorithm can predict appearances for unseen blocks because the motion models of such blocks are estimated through the inference by message passing.

3.2. Appearance prediction via motion

The appearance model of the ith block at the time step t is the posterior probability with respect to $\mathbf{h}_t^i$ given observations and is estimated by Bayesian filtering as

$$p(\mathbf{h}_t^i|\mathcal{M}_{1:t}, \mathcal{I}_{1:t}) \propto p(\mathcal{I}_t|\mathbf{h}_t^i) p(\mathbf{h}_t^i|\mathcal{M}_{1:t}, \mathcal{I}_{1:t-1}), \quad (7)$$

where $p(\mathcal{I}_t|\mathbf{h}_t^i)$ is the observation likelihood and $p(\mathbf{h}_t^i|\mathcal{M}_{1:t}, \mathcal{I}_{1:t-1})$ is the prediction with respect to $\mathbf{h}_t^i$. With consideration of the corresponding block motion, the prediction is given by

$$p(\mathbf{h}_t^i|\mathcal{M}_{1:t}, \mathcal{I}_{1:t-1})$$
$$= \int_{\mathbb{R}^2} p(\mathbf{h}_t^i|\mathbf{v}_t^i, \mathcal{M}_{1:t-1}, \mathcal{I}_{1:t-1}) p(\mathbf{v}_t^i|\mathcal{M}_t) d\mathbf{v}_t^i. \quad (8)$$

So, we first estimate the motion model $p(\mathbf{v}_t^i|\mathcal{M}_t)$ as described in Section 3.1, and then compute the prediction of the current block appearance via the motion model. Given a backward-motion $\hat{\mathbf{v}}_t^i$, we model $p(\mathbf{h}_t^i|\hat{\mathbf{v}}_t^i, \mathcal{M}_{1:t-1}, \mathcal{I}_{1:t-1})$

(a) Input image

(b) Pixelwise motions

(c) Blockwise BG mean motions
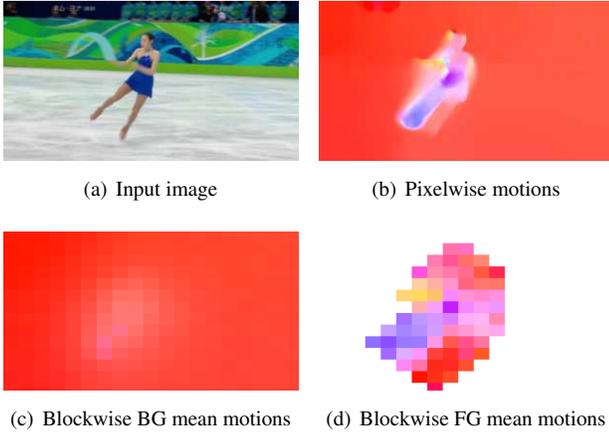
(d) Blockwise FG mean motions

Figure 2. Motion model estimation of background (BG) and foreground (FG) by nonparametric BP. Color and intensity denote the direction and size of motion, respectively. The missing background motion was recovered by message passing as shown in (c). In a similar manner, the blocks that do not contain foreground observations but are near foregrounds also have foreground motion models estimated by incoming messages as shown in (d).

in Eq. (8) as a mixture of predictions from the appearance models in the previous frame, which is given by

$$p(\mathbf{h}_t^i|\hat{\mathbf{v}}_t^i, \mathcal{M}_{1:t-1}, \mathcal{I}_{1:t-1})$$
$$= \sum_{k=1}^N \gamma_k^i(\hat{\mathbf{v}}_t^i) \int_{\mathbb{R}^3} p(\mathbf{h}_t^i|\mathbf{h}_{t-1}^k)p(\mathbf{h}_{t-1}^k|\mathcal{M}_{1:t-1}, \mathcal{I}_{1:t-1})d\mathbf{h}_{t-1}^k, \tag{9}$$

where $\gamma_k^i(\hat{\mathbf{v}}_t^i)$ is the responsibility of $\mathbf{h}_{t-1}^k$ with respect to $\mathbf{h}_t^i$ by considering the given backward motion $\hat{\mathbf{v}}_t^i$; it considers how much area of the kth block is overlapped with the ith block translated by $\hat{\mathbf{v}}_t^i$, and is defined by

$$\gamma_k^i(\hat{\mathbf{v}}_t^i) = \int_{R(k)} \mathcal{U}(\boldsymbol{\nu}; \mathbf{c}^i + \hat{\mathbf{v}}_t^i)d\boldsymbol{\nu}$$
$$\approx \int_{R(k)} \mathcal{N}(\boldsymbol{\nu}; \mathbf{c}^i + \hat{\mathbf{v}}_t^i, \Sigma_\gamma)d\boldsymbol{\nu} = \int_{R(k)} \mathcal{N}(\boldsymbol{\nu} - \hat{\mathbf{z}}_t^i; 0, \Sigma_\gamma)d\boldsymbol{\nu}, \tag{10}$$

where $\mathbf{c}^i$ and $\hat{\mathbf{z}}_t^i = \mathbf{c}^i + \hat{\mathbf{v}}_t^i$ are the current and previous center location of the ith block, respectively. The exact responsibility of the kth block is obtained by integrating the uniform distribution $\mathcal{U}(\boldsymbol{\nu}; \hat{\mathbf{z}}_t^i)$ on the kth block region $R(k)$, where $\mathcal{U}(\boldsymbol{\nu}; \hat{\mathbf{z}}_t^i)$ is nonzero in the region of the unit block centered at $\mathbf{z}_t^i$. For the ease of implementation, we approximate the uniform distribution to a Gaussian distribution whose mean and covariance are $\mathbf{z}_t^i$ and Σ_γ, respectively (Figure 3(a)). Finally, the state transition probability in Eq. (9), which involves our third assumption, is given by

$$p(\mathbf{h}_t^i|\mathbf{h}_{t-1}^k) = \mathcal{N}(\mathbf{h}_t^i - \mathbf{h}_{t-1}^k; 0, \Sigma_{tr}), \tag{11}$$

Figure 4. Sequential estimation of blockwise appearance models in the *skating* sequence. **(Row 1)** Input image. **(Row 2)** Mean colors of background block appearance models. **(Row 3)** Mean colors of foreground block appearance models.

where the covariance matrix Σ_{tr} allows smooth variations between appearances in time.

The block appearance prediction in Eq. (8) is simplified by Eq. (9) to (11), and is derived in Eq. (12). Note that π_k^i is the integration of the Gaussian convolution of the density function with respect to $\mathbf{z}_t^i$ (T1 in Eq. (12)) in the kth block region $R(k)$ (Figure 3(b)); the density function $p(\mathbf{z}_t^i|\mathcal{M}_t)$ is directly derived from the motion model $p(\mathbf{v}_t^i|\mathcal{M}_t)$. $p_k(\mathbf{h}_t^i|\mathcal{M}_{1:t-1}, \mathcal{I}_{1:t-1})$ (T2 in Eq. (12)) is obtained by the Gaussian convolution of the kth block appearance model in the previous time step.

In summary, the prediction of the block appearance model is a weighted sum of Gaussian-blurred block appearance models of the previous frame, where the weights are derived from the expected responsibilities considering the distribution of the block motion. Our Bayesian filtering successfully estimates the appearance models in time despite dynamic scene changes as illustrated in Figure 4. Note that the appearance models of the occluded background regions are estimated reasonably via prediction.

3.3. Pixelwise label estimation

The pixelwise label $\mathcal{L}_t$ is just the final output in an ordinary background subtraction algorithm. However, in our framework, the labels in a block determine the observation likelihood $p(\mathcal{I}_t|\mathbf{h}_t^i)$ and the prior density $p(\mathbf{h}_t^i|\mathcal{M}_{1:t}, \mathcal{I}_{1:t-1})$ for block appearance estimation; $\mathcal{L}_t$ is important to obtain accurate models by propagation in time.

We employ the standard loopy BP [2, 11] for label inference on the four-connected image grid. We denote two likelihood functions, based on the block motion models and the predictions of block appearances, by

$$\boldsymbol{\zeta}_t^i(\hat{\mathbf{v}}, \ell) = p(\mathbf{v}_{\ell,t}^i = \hat{\mathbf{v}}|\mathcal{M}_t), \tag{13}$$
$$\boldsymbol{\xi}_t^i(\hat{\mathbf{h}}, \ell) = p(\mathbf{h}_{\ell,t}^i = \hat{\mathbf{h}}|\mathcal{M}_{1:t}, \mathcal{I}_{1:t-1}). \tag{14}$$

The observation clique potential of the pixel $\mathbf{x}$ in the ith

$$p(\mathbf{h}_t^i|\mathcal{M}_{1:t},\mathcal{I}_{1:t-1})$$

$$= \sum_{k=1}^{N} \int_{\mathbb{R}^2} \gamma_k^i(\mathbf{v}_t^i)p(\mathbf{v}_t^i|\mathcal{M}_t)d\mathbf{v}_t^i \cdot \int_{\mathbb{R}^3} p(\mathbf{h}_t^i|\mathbf{h}_{t-1}^k)p(\mathbf{h}_{t-1}^k|\mathcal{M}_{1:t-1},\mathcal{I}_{1:t-1})d\mathbf{h}_{t-1}^k$$

$$\approx \sum_{k=1}^{N} \int_{R(k)} \underbrace{\int_{\mathbb{R}^2} \mathcal{N}(\boldsymbol{\nu}-\mathbf{z}_t^i;0,\Sigma_\gamma)p(\mathbf{z}_t^i|\mathcal{M}_t)d\mathbf{z}_t^i}_{\text{T1: Gaussian convolution of the density w.r.t. the previous location}} d\boldsymbol{\nu} \cdot \underbrace{\int_{\mathbb{R}^3} \mathcal{N}(\mathbf{h}_t^i-\mathbf{h}_{t-1}^k;0,\Sigma_{tr})p(\mathbf{h}_{t-1}^k|\mathcal{M}_{1:t-1},\mathcal{I}_{1:t-1})d\mathbf{h}_{t-1}^k}_{\text{T2: Gaussian convolution of the previous block appearance model}}$$

$$= \sum_{k=1}^{N} \pi_k^i\, p_k(\mathbf{h}_t^i|\mathcal{M}_{1:t-1},\mathcal{I}_{1:t-1}). \tag{12}$$

$\mathcal{U}(\nu;\hat{\mathbf{z}}_t^i=\mathbf{c}^i+\hat{\mathbf{v}}_t^i)$ $\mathcal{N}(\nu;\hat{\mathbf{z}}_t^i=\mathbf{c}^i+\hat{\mathbf{v}}_t^i,\Sigma_\gamma)$ $p(\mathbf{z}_t^i|\mathcal{M}_t)$ $\int \mathcal{N}(\boldsymbol{\nu}-\mathbf{z}_t^i;\mathbf{0},\Sigma_\gamma)p(\mathbf{z}_t^i|\mathcal{M}_t)d\mathbf{z}_t^i$ $\{\pi_k^i\}_{k=1}^N$

(a) Approximation of the responsibility distribution   (b) Computation of the mixture weights by blockwise integrations
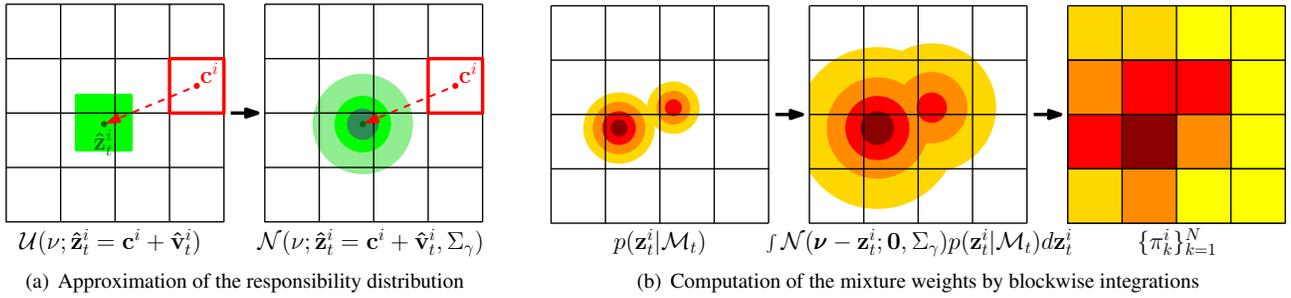
Figure 3. Visualization to compute responsibilities and their expectations for the mixture of predictions. (a) A uniform responsibility distribution $\mathcal{U}(\boldsymbol{\nu};\hat{\mathbf{z}}_t^i)$ is determined by a given backward motion $\hat{\mathbf{v}}_t^i$; the exact responsibilities for the blocks are obtained by integrating $\mathcal{U}(\boldsymbol{\nu};\hat{\mathbf{z}}_t^i)$ on each block region. Because we approximate $\mathcal{U}(\boldsymbol{\nu};\hat{\mathbf{z}}_t^i)$ to a single Gaussian distribution $\mathcal{N}(\boldsymbol{\nu}|\hat{\mathbf{z}}_t^i,\Sigma_\gamma)$, the responsibilities are also approximated (Eq. (10)). (b) Because a responsibility is a function of $\mathbf{v}_t^i$, which is a random variable following own distribution $p(\mathbf{v}_t^i|\mathcal{M}_t)$, we obtain the expectation of the responsibility by integrating it over $\mathbf{v}_t^i$ with $p(\mathbf{v}_t^i|\mathcal{M}_t)$. Consequently, the expected responsibility of the kth block, which is denoted by π_k^i, is derived by integrating the convolution of $p(\mathbf{z}_t^i|\mathcal{M}_t)$ and $\mathcal{N}(\boldsymbol{\nu}|\mathbf{z}_t^i,\Sigma_\gamma)$ on the region of the kth block. The expected responsibilities are the mixture weights for aggregating the predictions (Eq. 12).

block is given by

$$\varphi(\mathcal{L}_t(\mathbf{x})=b) = \frac{\zeta_t^i(\mathcal{M}_t(\mathbf{x}),b)\cdot\xi_t^i(\mathcal{I}_t(\mathbf{x}),b)}{\sum_{\ell\in\{b,f\}}\zeta_t^i(\mathcal{M}_t(\mathbf{x}),\ell)\cdot\xi_t^i(\mathcal{I}_t(\mathbf{x}),\ell)},$$

$$\varphi(\mathcal{L}_t(\mathbf{x})=f) = 1-\varphi(\mathcal{L}_t(\mathbf{x})=b). \tag{15}$$

The compatibility clique potential between a pair of neighboring pixels $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\psi(\mathcal{L}_t(\mathbf{x}),\mathcal{L}_t(\mathbf{y})) = \begin{cases} \lambda, & \text{if } \mathcal{L}_t(\mathbf{x})=\mathcal{L}_t(\mathbf{y}), \\ 1-\lambda, & \text{otherwise}, \end{cases} \tag{16}$$

where $0.5 < \lambda < 1$. The believes with respect to the pixelwise labels are computed by the sum-product message-passing algorithm [11] with the potentials we designed; the labels are determined by comparing the foreground and background believes at each pixel. For the details about pixelwise inference, refer to [2].

Because the label $\mathcal{L}_t$ changes the configuration of the motion observations too, the algorithm repeats the motion estimation (Section 3.1) and appearance prediction (Section 3.2) until $\mathcal{L}_t$ converges.

The evidences for pixelwise labeling—motion likelihood (Eq. (13)), appearance likelihood (Eq. (14)), and their combination—are illustrated in Figure 5. Moving objects similar in motion with camera or background may not be detected by motion likelihoods as in Figure 5(b). Appearance likelihoods are not reliable when foreground and background have similar colors as in Figure 5(c). By the combination of both likelihoods in Eq. (15), we obtain more stable evidences for pixelwise label estimation (Figure 5(d)).

3.4. Appearance estimation by Bayesian filtering

When $\mathcal{L}_t$ is converged, the observation likelihood in Eq. (7) is modeled by Gaussian kernel density estimation with the image observations from the corresponding block region and four neighboring block regions as

$$p(\mathcal{I}_t|\mathbf{h}_{\ell,t}^i) = \sum_{\{\hat{\mathbf{x}}|\hat{\mathbf{x}}\in R(i),\mathcal{L}_t(\hat{\mathbf{x}})=\ell\}} \beta_1\,\mathcal{N}\big(\mathbf{h}_{\ell,t}^i;\mathcal{I}_t(\hat{\mathbf{x}}),\Sigma_{\mathbf{h}}\big)$$

$$+ \sum_{j\in\eta(i)}\sum_{\{\hat{\mathbf{x}}|\hat{\mathbf{x}}\in R(j),\mathcal{L}_t(\hat{\mathbf{x}})=\ell\}} \beta_2\,\mathcal{N}\big(\mathbf{h}_{\ell,t}^i;\mathcal{I}_t(\hat{\mathbf{x}}),\Sigma_{\mathbf{h}}\big), \tag{17}$$

(a) Input image   (b) Motion likelihoods
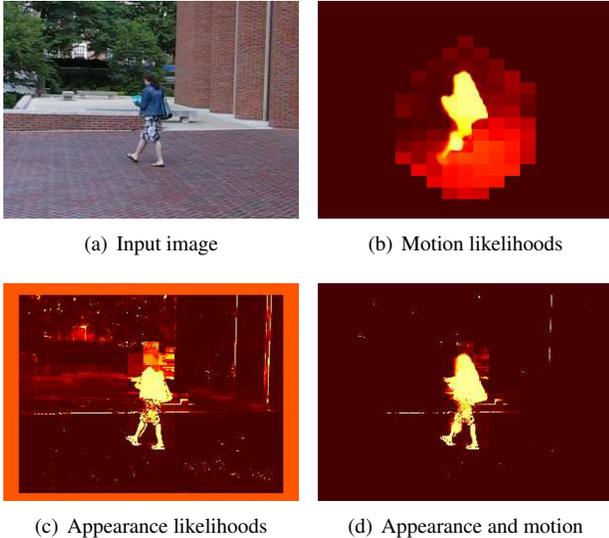
(c) Appearance likelihoods   (d) Appearance and motion

Figure 5. The pixelwise likelihood ratios. Both motion and appearance likelihoods have limitations—missing leg in (b) and noisy head in (c), but the combination of motion and appearance cues improves the result significantly.

where β_1 and β_2 are normalized weights ($\beta_1 > \beta_2$).

Then the block appearance model (Eq. (7)), which is equivalent to the posterior probability with respect to $\mathbf{h}_t^i$, is estimated by product of the prediction (Eq. (12)) and the likelihood (Eq. (17)). Because both of the prediction and the likelihood are Gaussian mixtures, the appearance model is again a Gaussian mixture; we employ a Gibbs sampling [3] to prevent the exponential increase of the number of mixture components over time.

4. Experiments

We evaluate our algorithm qualitatively and quantitatively in five challenging sequences. Our results are compared with other methods including [14], one of the state-of-the-art algorithms for generalized background subtraction.

4.1. Preprocessing

Optical flow and initial motion segmentation. We compute optical flows by [8] to obtain the observation of motion. At the first stage of each frame t, we perform motion segmentation over $\mathcal{M}_t$ by RANSAC with epipolar constraints and initialize the labels; $\mathcal{L}_t^{\text{init}}(\hat{\mathbf{x}}) = f$ if $\mathcal{M}_t(\hat{\mathbf{x}})$ is outlier, where $\hat{\mathbf{x}}$ is a pixel location in image. The motion segmentation based on epipolar constraint can handle more general motions than homography and produces reasonable results in practice. Note that the motion estimation and segmentation is not our objective. We attempt to estimate foreground and background models accurately through probabilistic inference given reasonable low-level motion analysis.

Initialization for filtering. To estimate the initial appearance model $p(\mathbf{h}_1^i|\mathcal{M}_1, \mathcal{I}_1)$, we first generate $\mathcal{M}_1$ by comparing two images $\mathcal{I}_0$ and $\mathcal{I}_1$, and then construct $\mathcal{L}_1^{\text{init}}$ by the motion segmentation of $\mathcal{M}_1$. Based on the motion segmentation result, the global, not blockwise, foreground and background appearance models are obtained from foreground and background region in the scene, respectively, by kernel density estimation. The global appearance models are useful when the motion segmentation misses parts of foreground objects that are similar in motions with camera. The foreground/background likelihood ratios for each pixel are computed based on the global models, and are used to revise $\mathcal{L}_1$ by BP on the four-connected pixelwise MRF [2]. Our algorithm alternates modeling and labeling procedure several times; after the labels by the global models are obtained, blockwise foreground and background appearance models are estimated by kernel density estimation based on the labels. Then, we start the normal procedure for the estimation of motions and appearances introduced in Section 3.

4.2. Experimental results

Our algorithm was tested in three videos—*Car1*, *People1*, and *People2*—in the Hopkins 155 dataset [20], and another two videos downloaded from *YouTube*—*skating* and *cycle*. In the *Hopkins-Car1* sequence, motion segmentation is errorneous; the motion outliers come from the road, not from the car (Figure 6). The motions in the *Hopkins-People1* and *Hopkins-People2* sequences are relatively gentle, but the accurate motion estimation is not straightforward since the foreground/background colors are similar in some areas (Figure 7). The *skating* involves articulated motions (Figure 9) and the *cycle* contains rapid motions and nonplanar background (Figure 10).

To demonstrate the effectiveness of our hybrid inference technique, we compared our algorithm with three other methods—motion segmentation, [14][1], and a reduced version of our algorithm, which excludes nonparametric BP in the motion layer inference.

The quantitative comparison results are illustrated in Figure 11, where the performances are computed based on the manually annotated groundtruth in every 5 frame. Our algorithm is the best in all videos except the *Hopkins-People2*, where we still have a decent performance. It is because the inference by nonparametric BP and Bayesian filtering improves the quality of motion and appearance models and our pixelwise labeling reduces the segmentation noises by the combination and motion and appearance likelihoods.

5. Conclusion

We proposed a novel algorithm for generalized background subtraction through foreground/background model

---

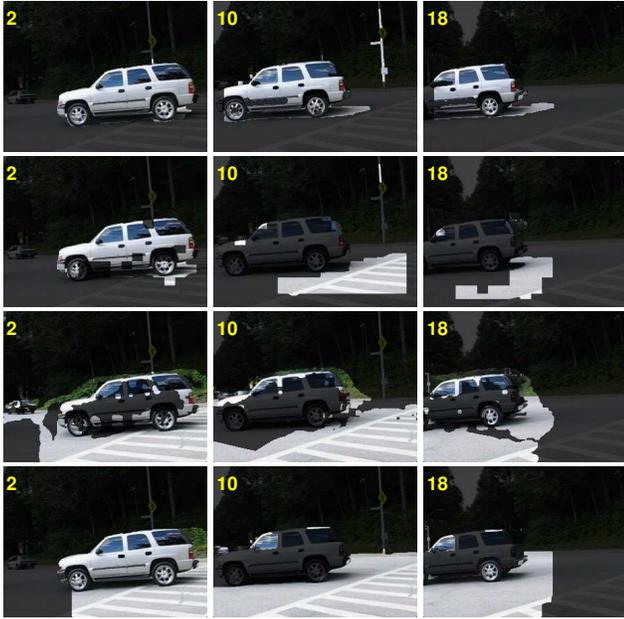[1]We used optical flow instead of particle video [12] for fair comparison.

Figure 6. The results of the *Hopkins-Car1* sequence. **(Row 1)** Ours. **(Row 2)** Ours except motion layer inference. **(Row 3)** Motion segmentation. **(Row 4)** [14].
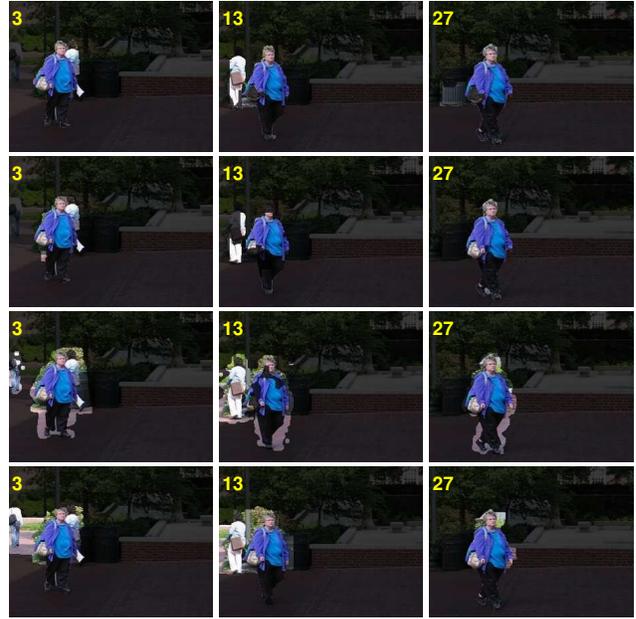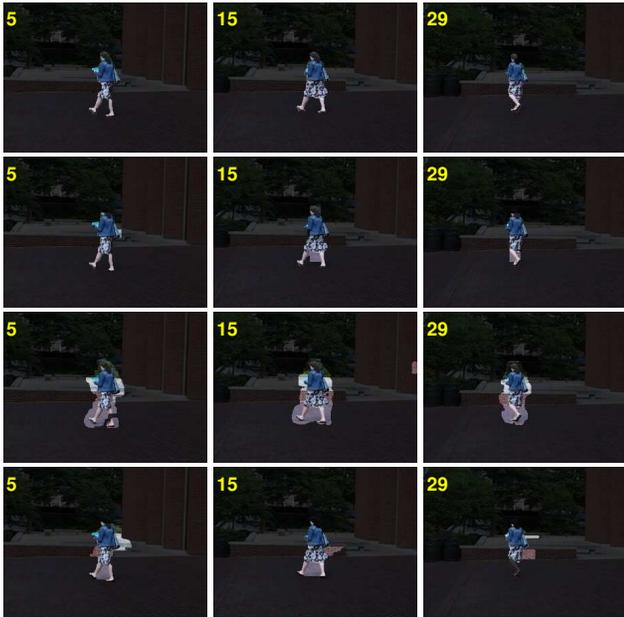

Figure 7. The results of the *Hopkins-People1* sequence. **(Row 1)** Ours. **(Row 2)** Ours except motion layer inference. **(Row 3)** Motion segmentation. **(Row 4)** [14].

Figure 8. The results of the *Hopkins-People2* sequence. **(Row 1)** Ours. **(Row 2)** Ours except motion layer inference. **(Row 3)** Motion segmentation. **(Row 4)** [14].

estimation by a hybrid inference based on nonparametric BP and Bayesian filtering. Motion models are estimated by nonparametric BP; it reduces noisy and incomplete motions. Bayesian filtering propagates appearance models by combining the prior appearance model and the current observation sequentially. We validated our algorithm qualitatively and quantitatively in various sequences, and showed successful background subtraction results.

Acknowledgement

References

[1] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of IEEE*, 90:1151–1163, 2002. 1

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006. 4, 5, 6

[3] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE TPAMI*, 6(6):721–741, 1984. 3, 6

[4] B. Han, D. Comaniciu, and L. Davis. Sequential kernel density approximation through mode propagation: Applications to background modeling. In *ACCV*, 2004. 1

[5] E. Hayman and J. O. Eklundh. Statistical background subtraction for a mobile observer. In *ICCV*, 2003. 1
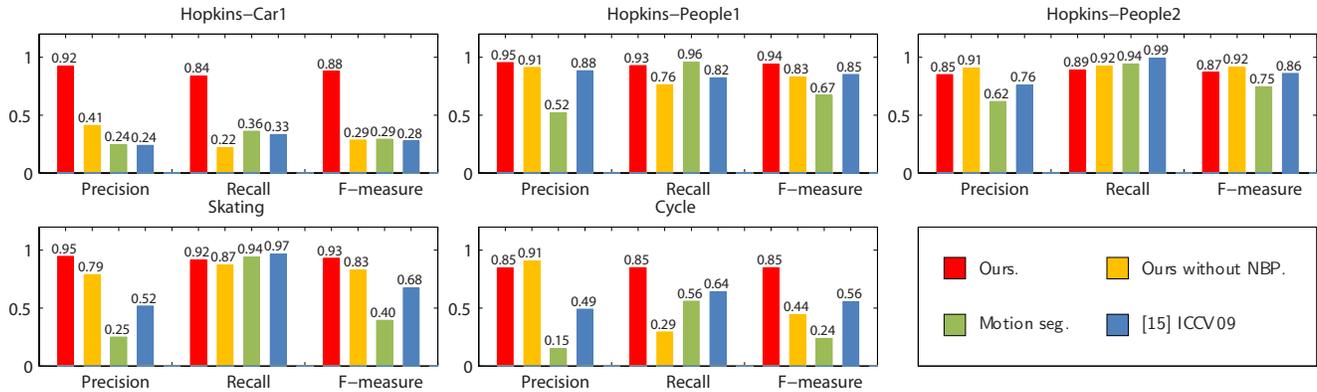
Figure 11. Performance comparisons with other methods

Figure 9. The results of the *skating* sequence. **(Row 1)** Ours. **(Row 2)** Ours except motion layer inference. **(Row 3)** Motion segmentation. **(Row 4)** [14].

Figure 10. The results of the *cycle* sequence. **(Row 1)** Ours. **(Row 2)** Ours except motion layer inference. **(Row 3)** Motion segmentation. **(Row 4)** [14].

[6] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11(3):172–185, 2005. 1

[7] D. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE TPAMI*, 27(5):827–832, 2005. 1

[8] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 6

[9] A. Mittal and D. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *CVPR*, 2000. 1

[10] A. Mittal and N. Paragios. Motion-based backgroudn subtraction using adaptive kernel density estimation. In *CVPR*, 2004. 1

[11] J. Pearl. *Probabilistic Reasoning in Inteligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988. 3, 4, 5

[12] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. In *CVPR*, pages 2195–2202, 2006. 1, 6

[13] M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background subtraction based on coocurence of image variations. In *CVPR*, 2003. 1

[14] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 1, 6, 7, 8

[15] Y. Sheikh and M. Shah. Bayesian object detection in dynamic scenes. In *CVPR*, 2005. 1

[16] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE TPAMI*, 22(8):747–757, 2000. 1

[17] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Non-parametric belief propagation. In *CVPR*, pages 605–612, 2003. 3

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992. 1

[19] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV*, pages 255–261, 1999. 1

[20] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007. 6

[21] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE TPAMI*, 19:780–785, 1997. 1

[22] C. Yuan, G. Medioni, J. Kang, and I. Cohen. Detecting motion regions in the presence of a string parallax from a moving camera by multiview geometric constraints. *IEEE TPAMI*, 20, 2007. 1

[23] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust Kalman filter. In *CVPR*, pages 44–50, 2003. 1