

# A Multilayer-Based Framework for Online Background Subtraction with Freely Moving Cameras

Yizhe Zhu

Department of Computer Science, Rutgers University, Piscataway, NJ 08854  
yizhe.zhu@rutgers.edu

Ahmed Elgammal

elgammal@cs.rutgers.edu

## Abstract

*The exponentially increasing use of moving platforms for video capture introduces the urgent need to develop the general background subtraction algorithms with the capability to deal with the moving background. In this paper, we propose a multilayer-based framework for online background subtraction for videos captured by moving cameras. Unlike the previous treatments of the problem, the proposed method is not restricted to binary segmentation of background and foreground, but formulates it as a multi-label segmentation problem by modeling multiple foreground objects in different layers when they appear simultaneously in the scene. We assign an independent processing layer to each foreground object, as well as the background, where both motion and appearance models are estimated, and a probability map is inferred using a Bayesian filtering framework. Finally, Multi-label Graph-cut on Markov Random Field is employed to perform pixel-wise labeling. Extensive evaluation results show that the proposed method outperforms state-of-the-art methods on challenging video sequences.*

## 1. Introduction

The identification of regions of interest is typically the critical preprocessing step for various high-level computer vision applications, including event detection, video surveillance, human motion analysis, etc. Background subtraction is a widely-used technique to perform pixel-wise segmentation of foreground regions out of background scenes. Unlike foreground object detection algorithms, background subtraction methods typically produce much more accurate segmentation of foreground regions rather than merely detection bounding boxes, without the need to train individual object detectors. A great number of various traditional background subtraction methods and algorithms have been proposed [34, 8, 21, 32, 12, 28]. Most of these methods focused on modeling background under the assumption that the camera is stationary. However, more

and more videos are captured from moving platforms, such as camera phones, and cameras mounted on ground vehicles, robots, ariel drones, etc. Traditional background subtraction algorithms are no longer applicable for such videos captured from a non-stationary platform [7]. The exponentially increasing use of moving platforms for video capture introduces a high demand for the development of general background subtraction algorithms that are not only as effective as traditional background subtraction but also applicable to moving-camera videos.

Similar to most video segmentation methods, a few works [26, 5] resort to processing the whole video offline. Offline methods can typically produce good results on short sequences since the information in latter frames can significantly benefit the segmentation in the earlier frames. However, since it needs to store and process the information over the whole video, the memory and computational cost increase exponentially as the number of frames to process grows [7]. Additionally, in various cases, such as video surveillance and security monitoring, videos need to be analyzed as they being streamed in real time, where an efficient online background subtraction method is greatly in demand.

The key to handling long sequences in an online way is to learn and maintain models for the background and foreground layers. Such models accumulate and update the evidence over a large number of frames and also supply valuable knowledge foundation to high-level vision tasks. Recently, a few online background subtraction methods with moving cameras have been proposed [31, 20, 9, 22, 41]. Most methods formulate it as a binary segmentation problem with the assumption of only one foreground object, naturally resulting in bad segmentation when multiple moving objects appear in the scene. Especially in the case where objects go across each other, motion estimation for objects suffers great confusion, which further degrades the performance of background subtraction.

To remedy this drawback, we propose a general multilayer-based framework with the capability of handling multiple foreground objects in the scene. The objects can be automatically detected based on motion inconsistency and

an independent processing layer is assigned to every foreground object as well as the background. In each layer, we would like the same process to be preformed concurrently inside the “processing block”, which takes the accumulated information and the new evidence of each layer as input, and outputs a probability map indicating the confidence of pixels belonging to each layer. In this paper, we elaborately design such a “processing block” with three steps as follows. (a) Motion model estimation is first performed based on Gaussian Belief Propagation [1] with motion vectors of corresponding trajectories as the evidence. (b) The appearance model and the prior probability are predicted by propagating the previous one with the estimated motion model. (c) Given the current frame as new evidence, Kernel Density Estimation [8] is employed to infer the probability map as the output. Finally, based on the collection of probability maps produced by each “processing block”, the pixel-wise segmentation for the current frame is generated by Multi-label Graph-cut.

Besides, since it is the first work to tackle multilabel background subtraction problem in moving camera scenarios, to our best knowledge, we also design a methodology to evaluate the performance and show our method outperforms the state-of-the-art methods.

## 2. Related Work

**Motion Estimation and Compensation:** The freely moving camera introduces a movement in the projected background scene, and thus complicates the background subtraction problem. An intuitive idea to tackle such a problem is compensating the camera motion. A few pioneering works resort to estimating a homography [13] that characterizes the geometric transformation of background scene between consecutive frames. Typically RANSAC [11] and its variants MLESAC [36] are employed to achieve robust estimation using many matches of feature points. Jin *et al.* [15] model the scene as a set of planer regions where each background pixel is assumed to belong to one of these regions. Homographies are used to rectify each region to its corresponding planer representation in the model. Zamalieva *et al.* [41] leverage geometric information to develop multiple transformation models, and choose one that best describes the relation between consecutive frames.

Recently, motion estimation has been widely employed to comprehensively specify the motion for every pixel [25, 2, 20, 22]. These works [20, 22] used optical flow as the evidence. Kwak *et al.* [20] divided the images into small blocks in a grid pattern, and employed nonparametric belief propagation to estimate the motion field based on average optical flow of each block. Its following work [22] improved the quality of motion estimation by replacing blocks with superpixels as the model unit. On the other hand, in [25, 2], optical flow orientations were claimed independent

of object depth in the scene, and used to clusters pixels that have similar real-world motion, irrespective of their depth in the scene. However, high dependency on the optical flow makes these methods susceptible to the noise in the estimation of optical flow. In contrast, our method improves motion model estimation by employing Gaussian Belief Propagation [1] with the motion vectors of sparse feature points as more robust evidence.

**Appearance Modeling:** Traditionally, statistical representations of the background scene have been proposed to estimate spatially extendable background models. Hayman *et al.* [14] built a mixture of Gaussian mosaic background model. Ren *et al.* [29] used motion compensation to predict the position of each pixel in a background map, and model the uncertainty of that prediction by a spacial Gaussian distribution. The construction of image mosaic associated with a traditional mixture Gaussian background model was also claimed to be effective in [23, 30]. However, the hyper-parameter required by this parametric model restricts its adaptability and application. On the contrary, we employ nonparametric Kernel Density Estimation method [8] to build models of the appearance of foreground and background regions, making our approach more stable and applicable.

**Layered Representation:** The layered representation, referring to approaches that model the scene as a set of moving layers, has been used for foreground detection [27, 16], motion segmentation [39, 37, 19]. In [27], the background was modeled as the union of nonparametric layer-models to facilitate detecting the foreground under static or dynamic background. Kim *et al.* [16] proposed a layered background model where a long-term background model is used besides several multiple short-term background models. Wang *et al.* [39] used an iterative method to achieve layered-motion segmentation. Torr *et al.* [37] modeled the layers as planes in 3D and integrating priors in a Bayesian framework. [19] models spatial continuity while representing each layer as composed of a set of segments. A common theme of these layered models is the assumption that the video is available beforehand [7]. Such an assumption prevents the use of such approaches for processing videos from streaming sources. Some dynamic textures methods [3, 4, 24] also employed the layered model to tackle the complex dynamic background, but with stationary cameras. To the best of our knowledge, the proposed method is the first layered model applied in the moving camera scenarios.

The organization of this paper is as follows. The overview of our proposed framework are presented in Section 3. Section 4 describes the trajectory labeling. The components inside “processing block” are described in Section 5&6 and the final pixel-wise multi-labeling are presented in section 7. Finally, Section 8 illustrates quantitatively and qualitatively experimental results based on two criteria.

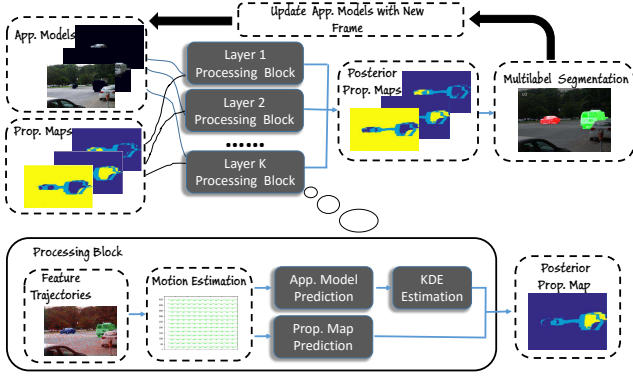


Figure 1: The framework of our proposed approach.

### 3. Framework Overview

The proposed framework is demonstrated in Figure 1. First we employ the feature point tracking method presented in [35] to generate feature trajectories. Then the generated trajectories are clustered into different layers based on motion inconsistency, and the labels of trajectories are continuously propagated over frames using the dynamic label propagation [42]. Each cluster of trajectories is assigned to the corresponding layer, and the number of layers is adapted according to that of foreground objects appearing in each frame.

Each layer possesses an independent “processing block” that produces a posterior probability map. Let  $k$  and  $t$  denote the index of the layer and the frame respectively. Inside the “processing block”, we have two sources of input: the appearance model  $\mathcal{A}_{t-1}^k$  and the prior probability map of labels  $P(\mathcal{L}_{t-1}^k)$  produced from the previous frame, and a group of corresponding sparse trajectories  $\mathcal{P}_t^k$ . The first step is the inference of the motion model in each layer using Gaussian Belief Propagation with the motion vectors of corresponding trajectories as the evidence. Then, the new appearance model  $\mathcal{A}_t^k$  is obtained by shifting the previous appearance model  $\mathcal{A}_{t-1}^k$  based on the estimated motion model  $\mathcal{M}_t^k$ , and the new prior probability map  $P_{prior}(\mathcal{L}_t^k | \mathcal{P}_t^k)$  can be inferred from the previous one in the same way.

Given the current frame  $\mathcal{I}_t$  as the new observation, the likelihood  $P(\mathcal{I}_t | \mathcal{L}_t^k)$  is estimated by Kernel density Estimation (KDE) [8] with the propagated appearance model in every layer. Then the posterior probability map for each layer is inferred as

$$P_{post}(\mathcal{L}_t^k | \mathcal{I}_t, \mathcal{P}_t^k) = \frac{1}{Z} P(\mathcal{I}_t | \mathcal{L}_t^k) P_{prior}(\mathcal{L}_t^k | \mathcal{P}_t^k), \quad (1)$$

where  $Z = \sum_k P_{post}(\mathcal{L}_t^k | \mathcal{I}_t, \mathcal{P}_t^k)$  is the partition function. With a collection of the posterior probability maps from each layer, we achieve the final pixel-wise labeling by optimizing a cut on a multilabel graph with the minimal cut energy. At the end of the whole process, appearance mod-

els are updated with the current frame and labels. In the following sections, we will describe the process steps in detail.

## 4. Trajectory Labeling and Propagation

The feature point tracking method [35] we employ has achieved a good performance in feature trajectories generation. To cluster trajectories, several motion segmentation methods [26, 40] have provided good solutions, but may fail when the video does not meet the assumption of the affine camera model. To get rid of such an assumption, an online method proposed by Elqursh *et al.* [10] considered sparse trajectory clustering as the problem of manifold separation and dynamically propagate labels over frames. Inspired by [10], we first cluster the trajectories in the initialization frames, and continuously propagates the label of trajectories over frames using the dynamic label propagation [42]. We briefly describe the algorithm here.

### 4.1. Trajectory Clustering

Given  $n$  trajectories, two distance matrices  $D_M^t$  and  $D_S^t$  are defined to represent the difference between trajectories in motion and spatial location. The entries  $D_{ij}^M = d_{1:t}^{1:t}(T_i, T_j)$  and  $D_{ij}^S = d_{1:t}^{1:t}(T_i, T_j)$  are the distances between  $i$ -th and  $j$ -th trajectories over frames up to  $t$ . For detailed definition, please refer to [26]. The affinity matrix over  $n$  trajectories is then formulated as

$$A = \exp(-(\lambda D_M^t + (1 - \lambda) D_S^t)), \quad (2)$$

where  $\lambda$  is the parameter to balance two distances.

Considering each trajectory as a node and the affinity matrix as the edge weights, we cast the trajectory clustering to graph cut problem. Starting from the initial cluster that contains all trajectories, normalized cuts [33] are employed to perform optimal binary cut on initial cluster and again on the generated clusters. This recursive process continues until the evaluated normalized cut cost on the cluster is above the threshold ( $10^{-4}$  in our work), which indicates this cluster of trajectories belongs to the same component (i.e. objects or the background), and needs no further splitting. All trajectories are assigned with labels according to which cluster they belong to.

### 4.2. Label Propagation

With the labels of trajectories in the initial frames, the label propagation, as a semi-supervised learning method, is adopted to infer the labels of trajectories in subsequent frames. We first construct a graph  $G$  with the trajectories in the previous and current frames as the labeled and unlabeled nodes respectively. The affinity matrix  $A$  involving the labeled and unlabeled trajectories are calculated and used as edge weights between corresponding nodes. Let  $Y_l$

and  $Y_u$  denote the labeling probability matrix corresponding to labeled and unlabeled node respectively. Each row of  $Y_l$  is a one-hot vector with one at the location corresponding to the label of node and zero otherwise. To estimate the labeling probability matrix  $Y_u$ , we apply Markov random walks on the graph [43]. A transition matrix  $P$  is defined as  $[p_{ij} = A_{ij} / \sum_k A_{ik}]$ , where the entry  $p_{ij}$  is the transition probabilities from  $i$ th node to  $j$ th node. Given the partition of nodes into labeled and unlabeled nodes, the matrix  $P$  can be splitted into 4 blocks:

$$P = \begin{bmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{bmatrix}. \quad (3)$$

The closed form solution of the labeling probability matrix of unlabeled node is  $Y_u = (I - P_{uu})^{-1} P_{ul} Y_l$ . The label for the node  $i$  can be obtained by

$$l_i = \arg \max_j y_{ij}, \quad (4)$$

where  $y_{ij}$  is the entry in the row corresponding to node  $i$ .

It is worth noting that the label propagation algorithm can only assign the trajectories with known labels. However, when new objects move into the scene, new labels should be introduced in time. To accomplish it, after the labels are predicted in each frame, a normalized cut cost in each cluster is evaluated. A small cost indicates a great intra-cluster variation inside the cluster. If the cost is below the threshold ( $10^{-4}$  in our work), the cluster should be further splitted and a new label is assigned to the cluster with more different appearance from the previous one. When an object moves out of the scene, few trajectories are assigned with the corresponding label and the corresponding cluster is removed. In this way, the number of clusters changes adaptively according to how many moving objects appear in the scene. After clustering process is done, all trajectories in each cluster are further assigned to each layer.

## 5. Motion Model Estimation

Although providing the motion information only at sparse pixels, trajectories of feature points are more accurate and less noisy compared with optical flow. With these accurate motion vectors of trajectories as evidence, we can estimate the motion field model for the whole frame (i.e. to estimate the motion vector of every pixel in each layer). To accomplish the pixel-wise motion estimation, we construct a pairwise MRF on a grid structure, where each vertex represents a pixel and the set of edges  $\varepsilon$  represents pairwise neighborhood relationship on this structure. Two sets of potentials are involved: edge potentials  $\Psi_{ij}$  measuring the similarity between neighborhood vertices, and self-potentials  $\Phi_i$  measuring the likelihood of evidence. If these potentials are defined as Guassian distribution, the task is thus formulated as a Gaussian Belief Propagation (GaBP)

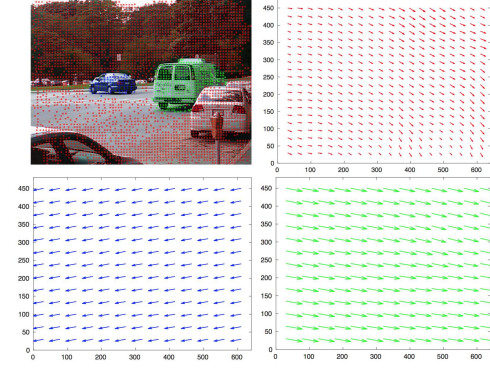


Figure 2: Visualization of labeled trajectories and motion models of three layers. The first figure is the original image covered with labeled trajectories: the trajectories are clustered into three layers: **background layer (Red)**, **car1 layer (Blue)** and **car2 layer (Green)**. The following figures show the estimated motion models for the three layers, where the arrow and length indicate the direction and magnitude of motion.

problem [1]. Given the motion vectors of trajectories in the  $k$ -th layer, the conditional joint probability distribution of motion model can be inferred as:

$$P(\mathcal{M}_t^k | \mathcal{P}_t^k) \propto \prod_{(i,j) \in \varepsilon} \Psi(m_{k,t}^i, m_{k,t}^j) \prod_{i \in S_{k,t}} \Phi(m_{k,t}^i), \quad (5)$$

where  $\mathcal{P}_{k,t}$  donates the set of feature points along with trajectories that are clustered to  $k$ -th layer. The edge and self potentials are defined as  $\Psi(m_{k,t}^i, m_{k,t}^j) = \mathcal{N}(m_{k,t}^i | m_{k,t}^j, \Sigma_m)$  and  $\Phi(m_{k,t}^i) = \mathcal{N}(m_{k,t}^i | m_{p,t}^i, \Sigma_p)$ , where  $m_{p,t}^i$  represents the motion vector of corresponding trajectories associated with  $i$ -th pixel. Our formulation encourages the similarity between motion vectors of neighborhood points and that between estimated motion vectors of feature points and the evidence (i.e. motion of trajectories). According to GaBP, this equation can be rewritten as

$$P(\mathcal{M}_t^k | \mathcal{P}_t^k) \propto \exp(-\frac{1}{2} m^T A m + m^T b), \quad (6)$$

where the inverse covariance matrix  $A$  is defined to show the connection of every pair of nodes, and the shift vector  $b$  is defined by the motion of trajectory. A closed form solution for marginal posterior probability is  $p(m_{k,t}^i | \mathcal{P}_t^k) = \mathcal{N}(\mu_{k,t}^i, \Sigma_{k,t}^i)$ , where  $\mu_{k,t}^i$  is the  $i$ -th entry of  $A^{-1}b$  and  $\Sigma_{k,t}^i$  is the entry  $\{A^{-1}\}_{ii}$ . The estimated motion field is demonstrated in Figure 2.

## 6. Bayesian Filtering

The inference of probability maps can be performed as the sequential Bayesian filtering on the Hidden Markov

Model. In this section, we first predict appearance model and prior probability for each layer given the motion models, followed by the inference of the posterior probability with new observations in the current frame. For simplicity of the expression, we remove the subscript  $k$  in the rest of paper. Note that the processes in each layer are identical.

### 6.1. Model Propagation

Given the probability distribution of the estimated motion model  $P(\mathcal{M}_t|\mathcal{P}_t)$ , we can estimate the appearance model and the probability map in the current frame by propagating the corresponding model and map from the previous frame. Specifically, the motion vector describes exactly how a pixel shifts between the consecutive frames. Therefore, armed with the motion information of a pixel in the current frame, we can easily obtain the appearance of the pixel by propagating that of the corresponding pixel in the previous frame. However, the motion vector of each pixel has Gaussian distribution over two-dimensional space. It is impractical to marginalize the appearance over the whole Gaussian distribution. If we discard the uncertainty in the motion vector and set the variance to 0, the gaussian distribution is reduced to a Dirac delta function  $P(m_t^i|\mathcal{P}_t) = \delta(m_t^i - \mu_{k,t}^i)$ . Then the marginalization over the whole appearance model is reduced to the involvement of the particular pixel. The appearance model of  $i$ -th pixel can be readily obtained by

$$a_t^i = a_{t-1}^{\phi(i, \mu_t^i)}, \quad (7)$$

where the function  $\phi(i, \mu_t^i)$  is to obtain the corresponding index in the previous frame by reversely shifting the  $i$ -th pixel according to the associated motion vector  $\mu_t^i$ . In practice, we found this approximation have little performance degradation while saving much computational cost.

Similarly, the prior probability of  $i$ -th pixel in each layer is obtained by

$$P_{prior}(l_t^i|\mathcal{P}_t) = P(l_{t-1}^{\phi(i, \mu_t^i)}). \quad (8)$$

### 6.2. Probability Map Update

Once the appearance model and probability map are propagated from the previous frame, the posterior probability map of each layer can be inferred with the current Frame  $\mathcal{I}_t$  as the new observation. With the assumption of independence between pixels, the posterior probability of  $i$ -th pixel is computed by Eq. (9).

$$P_{post}(l_t^i|I_t^i, \mathcal{P}_t) \propto P(I_t^i|l_t^i)P_{prior}(l_t^i|\mathcal{P}_t), \quad (9)$$

The likelihood of the observation in  $i$ -th pixel  $P(I_t^i|l_t^i)$  is essentially describing how well the appearance  $I_t^i$  fits the appearance model in each layer. Kernel Density Estimation(KDE), as a nonparametric method, can effectively

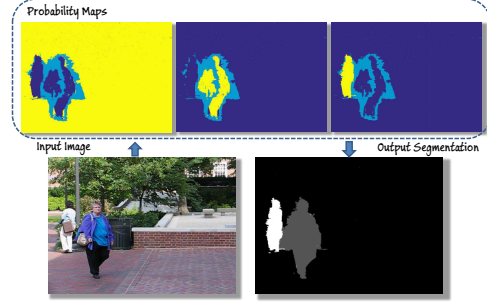


Figure 3: Inside the dotted box, three probability maps are demonstrated: the background layer, foreground1 layer(woman on the right), foreground2 layer(woman on the left). Note that the lighter color represents the high confidence. Two images in the bottom show the input image and the result of Multilabel Segmentation.

model the appearance for background or foreground without the restriction of hyper-parameter estimation. For its simplicity and effectiveness, we employ KDE technique for appearance modeling. The appearance model of  $i$ -th pixel  $a_t^i$  is involved with a pool of color features  $\{I_f\}_N^i$  accumulated from the previews frames. The likelihood of the observation  $P(I_t^i|l_t^i)$  estimated as:

$$P(I_t^i|l_t^i) = \frac{1}{N} \sum_{f=1}^N K_G(I_t^i - I_f^i), \quad (10)$$

where  $K_G(\cdot)$  is the Gaussian kernel function,  $I_f^i$  is the color feature in  $f$  frame stored for KDE modeling, and  $N = 20$  is the number of stored previous frames. The posterior probability map produced in each layer is normalized by partition function, and then used as the knowledge for multilabel segmentation.

## 7. Multilabel Segmentation

With the collection of normalized probability maps of all layers at hand, our final task for foreground objects segmentation is to perform pixel-wise labeling for the whole frame. This segmentation problem can be converted into an energy minimization problem on the pairwise MRF which can be polynomially solvable via Graph Cuts [17]. Due to ill-posed nature of the segmentation problem, regularizations are always required. For our problem, we designed two regularizers: a smoothing cost and a label cost in preference of smoothness of labeling and fewer unique labels assigned, respectively. The global energy function is formulated as:

$$E(\mathcal{L}) = \sum_{k \in S(\mathcal{L})} \sum_{i: l_i = k} E(I_i) + \lambda_1 \sum_{(i,j) \in \mathcal{N}} q_{ij} E(I_i, I_j) + \lambda_2 \sum_{k \in S(\mathcal{L})} h_k \delta_k(\mathcal{L}) \quad (11)$$

where  $S(\mathcal{L})$  denotes the set of unique labels of  $\mathcal{L}$ . The three terms in the right-side of equation are data cost, smoothness cost and label cost, respectively. The data cost  $E(I_i)$  is defined as the negative log probability of  $i$ th pixel belonging to a certain layer  $-\log P_{post}(l_t^i|I_t^i)$ . The smoothness cost is defined as the similarity of two neighboring pixels  $E(I_i, I_j) = -\log \mathcal{N}(I_i|I_j, \Sigma_a)$ , and  $q_{ij}$ , as a sign function, equals to +1 if  $l_i, l_j$  have the same label; otherwise -1. Moreover, in the term of label cost,  $h_k$  is the non-negative label cost of the label  $k$ . And  $\delta_k(f)$  is the corresponding indicator function with the definition  $\delta_k(\mathcal{L}) = 1$  if  $\exists i : l_i = k$ ; and 0 otherwise.  $\lambda_1, \lambda_2$  are non-negative parameters to balance data cost against such two regularizations (both are 1 in our work). The energy minimization function can be solved efficiently using the method presented in [6].

It is worth noting that when a new foreground object appears in the scene or initial frames are processed, the feature samples to build the appearance model usually are not sufficient for Kernel Density Estimation. In these cases that KDE is invalid, the probability maps is no longer available as the data cost for the multilabel segmentation. An alternative way to define the term of data cost in energy function is based on the labeled trajectories.  $E(I_i) = c$  if  $l_i = l_p$  and  $-c$  otherwise, where  $c$  is a negative constant, and  $l_p$  is the label of trajectory associated with  $i$ -th pixel. This definition simply utilizes the motion informance rather than the appearance model.

Finally, according to the labels of pixels, the appearance models  $\mathcal{A}_t$  are updated by adding the color feature  $I_t^i$  in the current frame to the corresponding appearance model.

## 8. Experiments

We evaluate our algorithm qualitatively and quantitatively based on two criteria: one is the normal two-label background subtraction and one is multilabel background segmentation. The former is evaluated using F-score, which is the common measurement. For the latter, since, to our best knowledge, no one has done such work before, we carefully design a reasonable measurement. The result shows our method outperforms the state-of-the-art approaches in both settings.

**Dataset:** Experiments were run on two public datasets. A set of sequences (cars1-8, people1-2) in the Hopkins dataset [38] is commonly used for quantitative evaluation on this topic, some of which contain multiple foreground objects. To quantitatively evaluate the performance, we produced the groundtruth mask manually for all frames, including the discrimination of foreground objects. Another one is Complex Background Dataset(CBD) [25], where the complex background and camera rotation introduce a great challenge.

	MLBS	GBSSP	FOF	OMCBS	GBS	BSFMC
cars1	<b>92.04</b>	87.14	50.84	91.77	80.30	73.13
cars2	<b>90.16</b>	82.17	56.60	69.13	68.45	55.68
cars3	<b>93.16</b>	72.94	73.57	41.27	79.22	60.91
cars4	<b>91.55</b>	88.24	47.96	73.65	66.63	54.81
cars5	<b>86.62</b>	81.66	70.94	60.44	74.56	51.97
cars6	<b>92.23</b>	81.44	84.34	90.31	73.34	37.56
cars7	<b>91.17</b>	90.86	42.92	89.87	69.10	40.35
cars8	85.93	<b>86.85</b>	87.61	83.84	80.29	62.42
people1	<b>81.38</b>	81.21	69.59	64.04	80.19	34.25
people2	<b>94.34</b>	84.74	88.40	89.32	81.40	64.92
drive	<b>65.95</b>	53.55	61.80	13.68	5.18	2.02
forest	72.20	<b>91.44</b>	31.44	42.99	23.19	16.76
parking	<b>83.66</b>	68.97	73.19	11.47	11.02	13.05
store	<b>86.28</b>	83.44	70.74	10.18	21.42	8.83
traffic	48.19	31.31	<b>71.24</b>	41.49	24.14	27.49
Overall	<b>83.66</b>	77.73	65.45	58.23	55.90	40.28

Table 1: Two-label background subtraction performance comparison with the F-score(%) on **Hopkins** and **CBD** Dataset. Best performance scores are highlighted in bold.

### 8.1. Two-label Background Subtraction

The performance of our framework, represented as MLBS (Multi-Layer Background Subtraction), is compared to five state-of-the-art algorithms: GBSSP [22], FOF [25], OMCBS [9], GBS [20], and BSFMC [31]. GBS requires the initialization labels of each frame as additional inputs, and GBSSP needs the groundtruth of the first frame as the initialization. For fair comparisons, we provide GBS with the labeling results of BSFMC, and offer GBSSP the labeling result of the first frame generated by our method. Note that instead of the requirement of these additional informations, our MLBS method completes self-initialization automatically. Moreover, we use the parameters provided by authors in external methods and fixed them for each algorithm in all experiments.

The quantitative comparisons on two dataset are shown in Table 1. It can be seen that the proposed method outperforms other methods in the literature on most test sequences. Especially in the case where the objects are occluded and then separated, such as the cars3 and people2, the multilayer strategy boosts the performance with a great jump on F-score. This is because our method can accurately estimate separate motion models for foreground objects, instead of only one ambiguous model for the whole foreground, and separate appearance models comprehensively provide the evidence of probabilistic inference. The overall F-score of our method is higher than the best score of the state-of-the-arts by a noticeable margin (83.66% vs 77.73%).

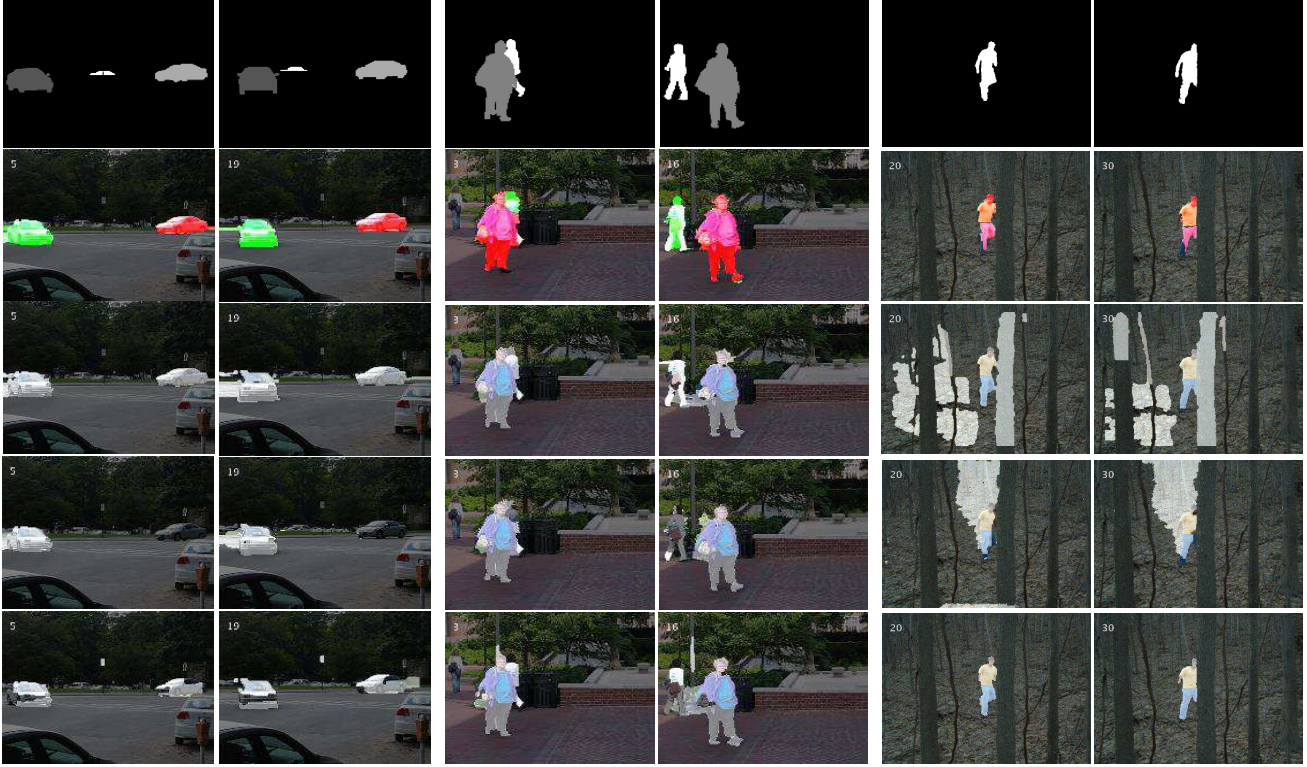


Figure 4: Comparison of our MLBS, and three state-of-the-art algorithms on the sequences cars5, people2 and forest, separated in three blocks. Row1: Groundtruth; Row2: our MLBS; Row3: GBS [20]; Row4: OMCBS [9]; Row5: GBSSP [22]. For clear demonstration, the background is darkened and the foreground objects are marked in different colors. (Better seen in color)

To evaluate the ability of the proposed method handling multiple objects, we categorize the sequences into three scenarios according to the number of moving objects in the scene, and report the average F-score on each scenario in Table 2. In the single object scenario, our method outperform the second best method by a small margin (84.05% vs 80.70%). But as the number of objects increases, the margin grows (3.4% vs 8.0% vs 8.7%). It clearly demonstrates the outstanding capability of our method in complex scenes.

	MLBS	GBSSP	FOF	OMCBS	GBS	BSFMC
1	<b>84.05</b>	80.70	59.20	54.22	47.82	31.20
2	<b>91.14</b>	81.51	83.19	71.48	80.30	62.75
$\geq 3$	<b>74.99</b>	65.05	66.26	57.02	55.72	45.05

Table 2: Two-label background subtraction performance comparison on the videos with different numbers of moving objects. Best performance scores are highlighted in bold.

Qualitative comparisons with three algorithms are illustrated in Figure 4. We pick two representative sequences (cars5 and people2), where at least two foreground objects appear in the scene, from Hopkins Dataset, and one se-

quence (forest) from CBD. It is notable that our MLBS algorithm separates moving objects and background accurately when the background scene is complex and the motions of objects are not simply in the same direction. For instance, in the sequence of people2, two women, who are walking in different directions, occluded and then separated, are separated precisely while other methods wrongly label the new appearing background region (e.g. the black rubbish bin) as foreground.

## 8.2. Multilabel Background Subtraction

We also evaluate the capability to separate different foreground objects. Since no one, to our best knowledge, has done such work before, we have designed a baseline to compare the performance. Details about baseline are shown in the Supplementary Material. Besides, [26] have proposed an offline approach (SLT) to tackle video segmentation problem based on long-term video analysis, and achieved a cutting edge performance. Since this method has no discrimination of foreground and background, we modify it by manually assigning the background label to the best fit segmented region, and make a comparison with it.

For quantitative comparison, we design a measurement

		Precision	Recall	F-Score
First 10 Frames	ours	89.79	<b>83.00</b>	<b>85.06</b>
	SLT	<b>90.70</b>	78.94	83.41
	Baseline	61.23	62.32	61.35
All Frames	ours	88.60	<b>85.05</b>	<b>86.16</b>
	SLT	<b>89.51</b>	78.24	82.00
	Baseline	63.63	66.22	64.35

Table 3: Performance comparison of Multi-foreground segmentation on **Hopkins** Dataset

by modifying the metric used in [26]. Let  $g_i$  denote the foreground region in the groundtruth,  $c_i$  the corresponding regions in the mask generated by algorithms, and  $|\cdot|$  the number of pixels inside the region. For each foreground region in the groundtruth, precision, recall and F-score are defined as:

$$P_i = \frac{|c_i \cap g_i|}{|c_i|} \quad R_i = \frac{|c_i \cap g_i|}{|g_i|} \quad F_i = \frac{2P_i R_i}{P_i + R_i} \quad (12)$$

Overall metric is obtained by averaging the measures of single regions. And the best one-to-one assignment of generated regions to groundtruth regions is found by the Hungarian method [18]. In the case where there exist generated regions without the assignment of groundtruth regions, we set the precision and recall of such regions to 1 and 0 respectively. It's worth mentioning that our revised metric is calculated over only foreground regions to keep consistency with the binary background subtraction, as the measurement values are the same as that of the binary one when there is only one foreground object in the scene.

The performance evaluation is shown in Table 3. We have compared the performance of the approaches evaluated on both all frames and the first ten frames. It can be seen that the proposed method outperforms SLT on both sets and has a great jump from the baseline. It's notable that the F-score of our method on all frames is higher than that on first ten frames due to the nature of online methods. Unlike the offline methods that hold the knowledge of the whole sequence at the beginning of the sequence process, our online approach has little prior knowledge and requires the self-initialization step during the first several frames, which surely leads to lower performance due to the insufficiency of prior knowledge. But it affects little as the sequence becomes longer.

Qualitative evaluations of our method and SLT are demonstrated in Figure 5. Our proposed method can separate the objects more accurately while SLT may falsely recognize two objects as only one when objects are very near and moving in the similar direction (see Block2). Furthermore, our method could detect new objects immediately when they enter the scene (see Block2). With the ability of accurate and robust foreground objects detection and seg-

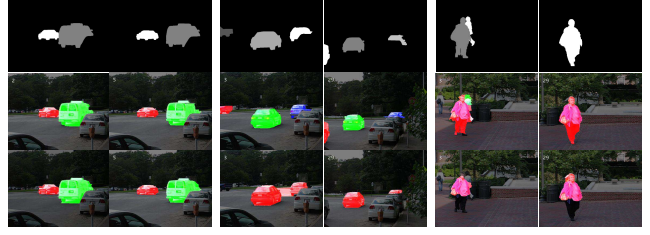


Figure 5: The Multilabel segmentation performance comparison with SLT on the sequence cars3, cars2 and people2, separated in three blocks. Row1: Groundtruth; Row2: MLBS; and Row3: SLT [26]. (Better seen in color)

mentation, our method produces a proper number of foreground object regions, which is reflected by the higher recall in Table 3. Additionally, equipped with appearance models for each layer, our MLBS method is capable of dealing with the articulated motion (see Block3) to a certain extent.

tclb	mtest	mdprop	grapcut	upmdl	Total
1109	3675	144	1369	559	6856

Table 4: Average computation time (msec) over different stages on a single frame from cars1. The stages are: trajectory clustering and label propagation (**tclb**), motion estimation(**mtest**), model propagation (**mdprop**), graphCut (**grapcut**) and update model(**upmdl**).

Table 4 shows the average computational time per frame. Our un-optimized Matlab implementation takes around 7 seconds per frame with an Intel Xeon-E5 CPU and 16GB memory. The computational time is dominated by motion estimation and graphcut. Motion estimation is done mainly by matrix inversion and multiplication. Since such operations can be readily parallelized on the GPU, we believe the real-time performance can be achieved by the optimization with the GPU and faster multi-thread implementation.

## 9. Conclusion

We propose a novel online multilayer-based framework for background subtraction with moving camera. In our framework, every foreground object and the background are assigned to an independent processing layer. A processing block is carefully designed to perform the posterior inference using Bayesian Filtering Framework, and Multi-label Graph-cut is employed to produce the pixel-wise segmentation for every video frame based on the normalized probability maps. Experiments show that our method performs favorably against other state-of-the-art methods, with outstanding ability to segment multiple foreground objects.

## References

- [1] D. Bickson. Gaussian belief propagation: Theory and application. *CoRR*, abs/0811.2518, 2008.
- [2] P. Bideau and E. Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. *arXiv preprint arXiv:1604.00136*, 2016.
- [3] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):909–926, 2008.
- [4] A. B. Chan and N. Vasconcelos. Layered dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1862–1879, 2009.
- [5] X. Cui, J. Huang, S. Zhang, and D. N. Metaxas. Background subtraction using low rank and group sparsity constraints. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, pages 612–625, 2012.
- [6] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *International journal of computer vision*, 96(1):1–27, 2012.
- [7] A. Elgammal. *Background Subtraction: Theory and Practice*. Morgan & Claypool Publishers, 2014.
- [8] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, Jul 2002.
- [9] A. Elqursh and A. Elgammal. Online moving camera background subtraction. In *European Conference on Computer Vision*, pages 228–241. Springer, 2012.
- [10] A. Elqursh and A. Elgammal. Online motion segmentation using dynamic label propagation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2008–2015, 2013.
- [11] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.
- [12] B. Han and L. S. Davis. Density-based multifeature background subtraction with support vector machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1017–1023, May 2012.
- [13] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [14] E. Hayman and J.-O. Eklundh. Statistical background subtraction for a mobile observer. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 67–74. IEEE, 2003.
- [15] Y. Jin, L. Tao, H. Di, N. I. Rao, and G. Xu. Background modeling from a free-moving camera by multi-layer homography algorithm. In *2008 15th IEEE International Conference on Image Processing*, pages 1572–1575. IEEE, 2008.
- [16] K. Kim, D. Harwood, and L. S. Davis. Background updating for visual surveillance. In *International Symposium on Visual Computing*, pages 337–346. Springer, 2005.
- [17] V. Kolmogorov and R. Zabini. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, Feb 2004.
- [18] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [19] M. P. Kumar, P. H. Torr, and A. Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76(3):301–319, 2008.
- [20] S. Kwak, T. Lim, W. Nam, B. Han, and J. H. Han. Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering. In *2011 International Conference on Computer Vision*, pages 2174–2181, 2011.
- [21] D.-S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, May 2005.
- [22] J. Lim and B. Han. Generalized background subtraction using superpixels with label integrated motion estimation. In *European Conference on Computer Vision*, pages 173–187. Springer, 2014.
- [23] A. Mittal and D. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 160–167 vol.2, 2000.
- [24] A. Mumtaz, W. Zhang, and A. B. Chan. Joint motion segmentation and background estimation in dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 368–375, 2014.
- [25] M. Narayana, A. Hanson, and E. Learned-Miller. Coherent motion segmentation in moving camera videos using optical flow orientations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1577–1584, 2013.
- [26] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, June 2014.
- [27] K. Patwardhan, G. Sapiro, and V. Morellas. Robust foreground detection in video using pixel layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):746–751, 2008.
- [28] X. Ren, T. X. Han, and Z. He. Ensemble video object cut in highly dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1947–1954, June 2013.
- [29] Y. Ren, C.-S. Chua, and Y.-K. Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 24(1):183–196, 2003.
- [30] S. Rowe and A. Blake. Statistical mosaics for tracking. *Image and Vision Computing*, 14(8):549–564, 1996.
- [31] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1219–1225, Sept 2009.
- [32] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1778–1792, Nov 2005.

- [33] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.
- [34] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.
- [35] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010.
- [36] P. Torr and A. Zisserman. Mlesac. *Comput. Vis. Image Understand.*, 78(1):138–156, Apr. 2000.
- [37] P. H. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):297–303, 2001.
- [38] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.
- [39] J. Y. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, 1994.
- [40] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *European conference on computer vision*, pages 94–106. Springer, 2006.
- [41] D. Zamaliev, A. Yilmaz, and J. W. Davis. A multi-transformational model for background subtraction with moving cameras. In *European Conference on Computer Vision*, pages 803–817. Springer, 2014.
- [42] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *IN ICML*, pages 912–919, 2003.
- [43] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions.