A Multi-transformational Model for Background Subtraction with Moving Cameras*

Daniya Zamalieva, Alper Yilmaz, and James W. Davis

The Ohio State University, Columbus OH, USA

Abstract. We introduce a new approach to perform background subtraction in moving camera scenarios. Unlike previous treatments of the problem, we do not restrict the camera motion or the scene geometry. The proposed approach relies on Bayesian selection of the transformation that best describes the geometric relation between consecutive frames. Based on the selected transformation, we propagate a set of learned background and foreground appearance models using a single or a series of homography transforms. The propagated models are subjected to MAP-MRF optimization framework that combines motion, appearance, spatial, and temporal cues; the optimization process provides the final background/foreground labels. Extensive experimental evaluation with challenging videos shows that the proposed method outperforms the baseline and state-of-the-art methods in most cases.

Keywords: Background subtraction, moving camera, moving object detection.

1 Introduction

Background subtraction is essential for many high level tasks in computer vision, including but not limited to object detection, object recognition, tracking, 3D scene recovery, and action recognition. Considering its precursory nature in the computer vision pipeline, the performance of background subtraction directly affects the quality of each task it precedes as well as the final results in the pipeline. For over two decades, a significant number of background subtraction methods have been published under the assumption that the camera capturing the scene is stationary. Needless to say, none of these algorithms are applicable in the case when the camera is moving. The ever increasing use of mobile phones and handheld cameras introduces a need for new background subtraction methods that alleviate a stationary camera requirement.

When the camera moves during acquisition, the pixels corresponding to background no longer maintain their positions in consecutive frames. This observation severely complicates the traditional background subtraction process and

* Electronic supplementary material -Supplementary material is available in the online version of this chapter at `http://dx.doi.org/10.1007/978-3-319-10590-52`. Videos can also be accessed at `http://www.springerimages.com/videos/978-3-319-10589-5`

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 803–817, 2014.

requires compensation of the camera motion. The printed literature contains only a handful of studies on background subtraction for freely moving cameras [15,10,7,22,8,9,4,5]. A typical first step in all these methods is background motion estimation, which can be broadly classified into two categories: model-based estimation [10,8,9,7] and trajectory-based estimation [15,4,5]. The model-based methods assume that the majority of the visible scene is the background, and they estimate either the homography transform [10,7] or fundamental matrix [8,9] between frames. The homography transform, however, is valid only when the scene is planar or when the camera does not translate. On the other hand, the fundamental matrix is only valid for nonplanar scenes and can be computed when the camera translates creating parallax. Consequently, homography methods are prone to parallax, while fundamental matrix methods are susceptible to small camera motion. Since homography and fundamental matrices are complementary, neither can be used alone to model unknown camera motion. Alternatively to model-based methods, trajectory-based methods rely on dense long-term pixel trajectories to infer background motion [15,4,5]. These methods, however, are sensitive to tracking errors and short or incomplete trajectories especially when the camera motion is fast.

Methods in both categories remedy their drawbacks by employing appearance modeling and spatial smoothing. The appearance modeling is achieved by generating and transforming the background and foreground models for spectrally consistent results. Similarly, spatial smoothing ensures similar labeling results for proximal pixels. Both of these constraints work reasonably well in the case when the motion estimation prior to their application is acceptable; however, they are ineffectual when the motion estimation is incorrect.

In order to overcome the aforementioned problems related to implicit background motion estimation, we propose to use both the homography transform and the fundamental matrix (see Figure 1 for algorithmic flow). At each frame, we first estimate a dense motion field, then use it to compute the geometric transformations which are later used to propagate appearance models from the previous frame to the current frame. From among the two geometric transformations, the appropriate one is selected by adopting the Geometric Robust Information Criterion (GRIC) [18,6]. The application of the GRIC improves the motion estimation by choosing the appropriate geometric model for the short baseline case videos; hence, it makes the proposed background subtraction scheme immune to geometric degeneracies. In the case when the GRIC score favors the homography transform, the appearance propagation becomes a 1-1 mapping. On the other hand, if the fundamental matrix is chosen, we propagate the appearance models by estimating a series of homography transforms. The propagated appearance models provide the likelihood of each pixel used for background/foreground labeling. The appearance models we implement are similar to most background subtraction methods for stationary cameras, where the background appearances are modeled using a mixture of Gaussians per pixel [17]. Finally, we combine motion, appearance, spatial, and temporal cues in a MAP-MRF optimization framework to obtain the final background/foreground labels.

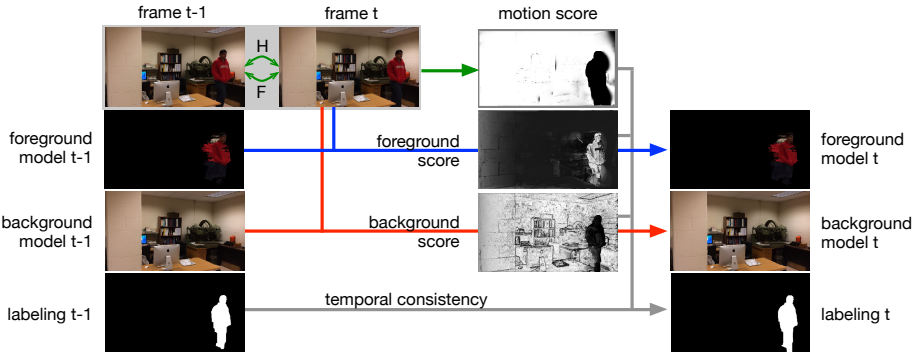

Fig. 1. Overview of the proposed method. First, the geometric transformation between frames $t-1$ and t is selected using the GRIC score, and the motion scores for each pixel are computed. Then, the appearance models from previous frame are propagated and compared with the current frame to obtain background/foreground scores. The motion and appearance scores are combined with the labeling of the previous frame to obtain the labeling for the current frame. Finally, the appearance models are updated accordingly.

The contributions of this paper can be summarized as follows. We provide novel methods in context of background subtraction for: 1) determining the best fitting geometric model that relates consecutive frames; 2) propagating learned background/foreground appearance via a multi-transformational model; 3) labeling pixels that is robust to occlusions and optical flow errors; 4) incorporating motion, appearance, spatial, and temporal cues for the final labeling.

2 Related Work

A common assumption for background subtraction methods is having a stationary camera for modeling the background appearance. The stationary camera assumption can be formulated as an identity transformation between the incoming frame and the background model. There are many papers in the literature that assume a stationary camera setup, and they have been discussed in comprehensive surveys [13,2]. There are also studies that perform background subtraction for stationary cameras but with dynamic backgrounds that exhibit non-stationary properties in time [12,16,11].

When the camera motion is not constrained, the background subtraction problem becomes complicated. Among the few papers published on this topic, a common treatment is to estimate the view geometric transformations such that the regions that do not fit the estimated transformations are labeled as foreground. Following this scheme, [9] uses the fundamental matrix for initial background/foreground labeling, which is iteratively refined by imposing temporal and spatial smoothness. In their method, the image is divided into blocks and the temporal models of each block are propagated using optical flow. This method, however,

is prone to small moving objects and degeneracies in estimating the fundamental matrix. In [8], the authors improved their method by refining the initial labeling using belief propagation. While providing a better postprocessing procedure, the performance still suffers from view geometric degeneracies in the fundamental matrix estimation, such as when the camera does not translate, frame-to-frame motion is very small, or the scene is planar. In contrast, [10] uses the homography transform to model and transfer the background model. View geometric degeneracies of homography estimation, such as nonplanar scenes, however, degrade the performance of their method. In order to avoid the problems caused by complex background scenes, [7] estimates separate homography transforms for a number of planes by applying a cascade of RANSAC steps. Since the points used for estimation include foreground objects, estimated homography transforms may be for non-existing planes which cause incorrect transformations. In addition, their method cannot tolerate when the first frame of the sequence contains moving objects or when the object enters the camera view together with a previously unseen part of the background. Alternative to implicit view geometry based methods, [22] performs full 3D recovery using a combination of structure from motion and bundle adjustment. Their approach requires a set of computationally expensive steps which is not suitable for background subtraction, which is usually considered an initial step in high-level computer vision tasks.

In contrast to the use of geometric transformations, some researchers analyze long-term trajectories to find moving objects in the sequence. In [15], the authors assume that the trajectories of background features form a 3D subspace, which can be estimated with factorization based shape-from-motion. In [4], a similar method is introduced, where the factorization is guided by group sparsity constraints defined for foreground. Both methods, however, strongly rely on long-term feature tracking which inhibits their real time application. A recent method [5] represents trajectories in a low-dimensional space and groups them by relearning the Gaussian Mixture Model at each frame. The decision of which trajectory groups belong to background or foreground is given by a set of heuristics such as compactness, surroundedness, and spatial closeness. These heuristics may fail for complex background scenes and non-rigid foreground objects.

In this paper, we model background motion by choosing appropriate geometric transformation for each frame instead of committing to a single model. Our method accommodates multiple transformational models for appearance model propagation, which are applied according to the selected geometric transformation.

3 Choosing Frame-to-Frame Transformations

Assuming the camera reference frame coincides with the world reference frame, the projection of a 3D point can be written as $\mathbf{x} = \mathtt{P}\mathbf{X} = \mathtt{K}[\mathtt{I}|\mathbf{0}]\mathbf{X}$, where $\mathtt{P}$ is a 3×4 projection matrix, $\mathtt{K}$ is the camera calibration matrix. When the camera rotates and translates, point $\mathbf{X}$ projects to the new image by $\mathbf{x}' = \mathtt{K}'[\mathtt{R}|\mathbf{t}]\mathbf{X}$. In this case, there is no one-to-one mapping between the image points $\mathbf{x}$ and $\mathbf{x}'$. These points, however, satisfy the fundamental matrix $\mathtt{F}$: $\mathbf{x}'^{\top}\mathtt{F}\mathbf{x} = 0$. The fundamental matrix is a geometrically valid transformation except for the

following degenerate cases: 1) the camera does not translate and only rotates, 2) all matching points are coplanar. In addition, for small camera baseline the equation system for estimating fundamental matrix becomes ill-conditioned.

For the degenerate cases stated above the transformation becomes a 1-1 mapping. This can be shown by dropping the last column of the projection matrix $\mathtt{P}$, in the case when the camera does not translate: $\mathbf{x} = \mathtt{K}\mathbf{X}$ and $\mathbf{x}' = \mathtt{K}'\mathtt{R}\mathbf{X}$, such that 1-1 mapping between points becomes $\mathbf{x}' = (\mathtt{K}'\mathtt{R}\mathtt{K}^{-1})\mathbf{x} = \mathtt{H}_\mathtt{R}\mathbf{x}$, where $\mathtt{H}_\mathtt{R}$ is referred to as the rotational homography. For the second case in which all points are coplanar, without loss of generality, we can assume that the points lie on $Z = 0$ plane. In this case, the third column $\mathbf{p}_3$ of $\mathtt{P}$, which gets multiplied with the point's Z coordinate, is not relevant in projection and can be dropped. The resulting projections in case become $\mathbf{x} = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_4]\mathbf{X} = \mathtt{H}\mathbf{X}$, and $\mathbf{x}' = [\mathbf{p}'_1, \mathbf{p}'_2, \mathbf{p}'_4]\mathbf{X} = \mathtt{H}'\mathbf{X}'$, such that $\mathbf{x}' = (\mathtt{H}'\mathtt{H}^{-1})\mathbf{x} = (\mathtt{H}_\pi)\mathbf{x}$, where $\mathtt{H}_\pi$ is the homography transform with respect to plane π.

The homography transform and the fundamental matrix constitute all possible frame-to-frame geometric transformations for a static scene. In order words, when they are used interchangeably, they can model all camera motions and scene geometries. In order to realize this observation and allow a freely moving camera in arbitrary background, we use both geometric transformations instead of committing to only one of them (opposed to the published literature summarized in Section 2).

A straightforward selection of the appropriate geometric transformation for consecutive frames is to first estimate both transformations and compare the sum of fitting errors for each one individually. This approach, however, is not a well posed due to the fact that the homography transform is a bijective 2D map and results in a two-dimensional error; while the fundamental matrix is a many-to-one mapping and provides a one-dimensional error. In order to define a 1D geometric distance, we follow the convention described in [20,19], which computes the approximation of the squared geometric distance $e^2_{i,\mathtt{H}}$ and $e^2_{i,\mathtt{F}}$ from the 4D joint-space point $[\mathbf{x}_i; \mathbf{x}'_i]$ to the homography $\mathtt{H}$ and fundamental matrix $\mathtt{F}$ manifold, respectively.

Using these distance measures, [18] introduces the Geometric Robust Information Criterion (GRIC), which is a Bayesian model selection scheme for the two geometric transformations. In order to offset measurement errors in model estimation, a search region S is defined in which the distances are assumed acceptable. This model is later modified by [6] for a 3D scene recovery problem, where the authors suggest to add another search criteria R which defines the range of disparity along which the feature match is expected to occur. Since the original GRIC [18] is biased towards selection of homography transform as addressed in [6], we adopt the modified GRIC score for $m = \{\mathtt{H}, \mathtt{F}\}$ given as:

$$GRIC_m = \sum_i \rho_2 \left(\frac{e^2_{i,m}}{\sigma^2} \right) + n \left((D - d_m) \log 2\pi\sigma^2 + 2 \log \frac{c_m}{\gamma} \right) + k_m \log n, \quad (1)$$

where D is the dimensionality of an observation ($D = 4$ for a pair of 2D points), d_m is the dimensionality of the underlying model manifold ($d_H = 2$, $d_F = 3$), σ is

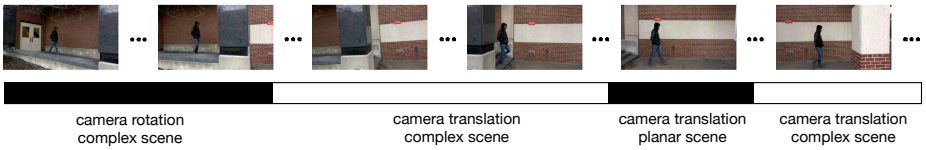

Fig. 2. GRIC score based selection for the *outdoor* sequence. The bar below exemplar frames indicates which geometric transformation is selected. Selection of `H` and `F` is respectively denoted by black and white regions.

the standard deviation of the measurement error, γ is the prior expectation that a correspondence is an inlier, k_m is the number of model parameters ($k_H = 8$, $k_F = 7$), n is the number of observations, and $\rho_2(x) = \min\{x, T_m\}$ with

$$T_m = 2\log\left(\frac{\gamma}{1-\gamma}\cdot\frac{\nu}{c_m}\right) - (D - d_m)\log 2\pi\sigma^2. \tag{2}$$

For an $L \times L$ image, while an arbitrary correspondence may occur in the volume $\nu = L \times L \times S \times S$, the GRIC score assumes an inlier correspondence is only distributed in the volume c_m, where $c_H = L \times L$ and $c_F = L \times L \times R$. In this framework, the lower GRIC score indicates the better geometric model.

In our implementation of the GRIC score, for each new frame I^t, we first compute point correspondences between I^t and I^{t-1} using optical flow. The matching features are then used to estimate both the fundamental matrix `F` and the homography transform `H` using RANSAC. Each geometric model is then subjected to Eqn. (1). The best fitting geometric model is selected based on the lowest GRIC score (see Figure 2). Once the model is chosen, the corresponding measurement error $e^2_{i,m}$ is used to assign each pixel $\mathbf{x}_i$ a motion based score $m(\mathbf{x}_i)$ which indicates the likelihood of $\mathbf{x}_i$ being a background pixel:

$$m(\mathbf{x}_i) = \exp\left(-\frac{e^2_{i,m}}{2\sigma^2_m}\right), \tag{3}$$

where σ_m controls the normalization of the motion score $m(\mathbf{x}_i)$.

4 Appearance Modeling

The motion score of a pixel $m(\mathbf{x})$ which is computed from the estimated motion model can be used to tentatively label a pixel as background or foreground. The resulting labeling based purely on motion is often noisy and prone to errors in optical flow; hence one can conjecture that the motion information alone is insufficient for the labeling. Besides motion, the appearance provides a strong clue indicating the presence or absence of a foreground object. To leverage the information provided by appearance changes, we maintain a background model $\mathcal{B}(\mathbf{x})$ and a foreground model $\mathcal{F}(\mathbf{x})$ for each pixel $\mathbf{x}$.

The challenge that makes background subtraction for a moving camera a hard problem is the requirement of registering the current frame with the background and foreground models. If such alignment is computed, any method proposed for a stationary camera can be applied to perform the background subtraction. In this paper, we adopt a commonly used Gaussian mixture model [17] to represent the background and foreground models.

The proposed method accommodates both the homography and a fundamental matrix based transformations by employing different mapping strategies. In the following text, we focus on how the selected geometric model can be used to compute an appearance based background and foreground scores from the generated appearance models.

4.1 Background Score for Homography Transform

When the GRIC metric results in selection of the homography transformation from the current frame, I^t, and the background model, $\mathcal{B}$, it can be directly used to map point $\mathbf{x}$ in I^t to background model $\mathcal{B}$:

$$\mathbf{x}' = \mathtt{H}_0\mathbf{x}. \tag{4}$$

The appearance based background score of $\mathbf{x}$ given the model $\mathcal{B}(\mathbf{x}')$ is then computed from:

$$s(\mathbf{x}|\mathcal{B}(\mathbf{x}')) = \sum_{j=1}^{g} w_j^b \ \exp\left(-\frac{1}{2}(I^t(\mathbf{x}) - \mu_j^b)^\top (\Sigma_j^b)^{-1}(I^t(\mathbf{x}) - \mu_j^b)\right) + w_0^b c, \tag{5}$$

where g is the number of mixture components at $\mathcal{B}(\mathbf{x}')$, $w_j^b, \mu_j^b, \Sigma_j^b$ are respectively the weight, mean, and covariance of jth component, c is a constant, and w_0 is the weight of a constant component which prevents from setting $w_1 = 1$ when the model is first initialized. Assuming independence of color channels, the covariance matrix is set to a diagonal matrix of the form $\Sigma = \sigma^2\mathbf{I}$.

4.2 Background Score for Fundamental Matrix

In the case when the GRIC metric chooses the fundamental matrix as the geometric transformation, the mapping between the image and the model becomes one-to-many, such that a point $\mathbf{x}$ in I^t maps to an epipolar line $\mathtt{F}\mathbf{x}$ in $\mathcal{B}$. This one-to-many mapping, however, does not register the image with the background model, and inhibits proper updating of the model. Geometrically, the choice of fundamental matrix suggests that the background scene contains more than a single physical plane, each of which can be transformed by a different homography transform that can be computed from a cascade of RANSAC steps.

In order to find such one-to-one mappings, correspondences with low motion score $m(\mathbf{x})$ given in Eqn. (3) are removed from the background pixel set due

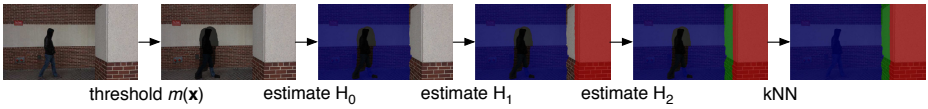

Fig. 3. The estimation of multiple homography transforms. First, the pixels with low motion scores are excluded (shaded with black in the second image). At each step, a homography transform $\mathtt{H}_i$ is estimated, and the inliers (shaded with navy, red, green) are excluded from subsequent homography estimations.

to the fact that they most likely correspond to foreground regions as discussed earlier. The remaining points are used to estimate the homography $\mathtt{H}_1$ using RANSAC. The inlier point set satisfying $\mathtt{H}_1$ are excluded from the set and the procedure is repeated for the remaining points to estimate $\mathtt{H}_2, \mathtt{H}_3, \ldots$ until the number of correspondences left is small (see Figure 3). In this scheme, each estimated homography transform $\mathtt{H}_i$ corresponds to a different plane π_i within the background scene and can be used to perform one-to-one mapping of pixels on respective planes to the background model by $\mathbf{x}'_i = \mathtt{H}_i\mathbf{x}$ for $i \geq 1$. While the inlier sets provide a list of pixels for each plane π_i, they by no means provide a complete set of pixels due to observation noise; hence, we transform each pixel in the image using all computed homography transforms and select the plane that satisfies:

$$\mathbf{x}' = \underset{\mathbf{x}'_i}{\operatorname{argmin}} \, ||I^t(\mathbf{x}) - I^{t-1}(\mathbf{x}'_i)||^2. \tag{6}$$

This process provides the transformation that best satisfy appearance similarity and the transformed pixel $\mathbf{x}'$ is used to compute the background score in Eqn. (5).

While this approach works well for visible background pixels, occluded background pixels that become visible after the foreground object moves require special treatment. This observations also holds for pixels with noisy optical flow. This is due to the fact the appearance constraint in Eqn. (6) is not satisfied. For such a pixel, we consider its k-nearest unoccluded pixels that are associated with one of the planes π_i and perform a majority voting to associate it to a plane. The corresponding homography is then used to compute the background score $s(\mathbf{x}|\mathcal{B}(\mathbf{x}'))$ using Eqn. (5).

Note that, if the homography transforms are estimated directly, one cannot avoid the estimation of a homography that maps the foreground objects between consecutive frames. If such a transformation is included, the foreground object will be incorporated into the background model after a number of frames. Moreover, moving objects present in the first frame would be directly included in the background model, and with the corresponding homography estimated, they cannot be distinguished from the background as the object moves. We avoid estimating the foreground homography by excluding the pixels with low scores $m(\mathbf{x})$ that indicate the presence of a moving object.

4.3 Foreground Score

Considering that the foreground objects can be non-rigid and nonplanar, we estimate the mapping of a pixel based on its optical flow:

$$\mathbf{x}' = \mathbf{x} + \boldsymbol{u}(\mathbf{x}), \tag{7}$$

where $\boldsymbol{u}(\mathbf{x})$ is optical flow of $\mathbf{x}$, and $\mathbf{x}'$ is the projected location we will use to estimate the foreground score. We should note that, for foreground objects, aside from having low background scores from Eqn. (5), the projections by optical flow and the estimated background homography transforms are different, which is encoded in the motion score $m(\mathbf{x})$. The foreground score $\mathcal{F}$ after the projection can be written similar to the background model of Eqn. (5):

$$s(\mathbf{x}|\mathcal{F}(\mathbf{x}')) = \sum_{j=1}^{g} w_j^f \exp\left(-\frac{1}{2}(I^t(\mathbf{x}) - \mu_j^f)^\top (\Sigma_j^f)^{-1}(I^t(\mathbf{x}) - \mu_j^f)\right) + w_0^f c, \tag{8}$$

where the subscript f indicates foreground.

5 Background/Foreground Labeling

Given the projection model, and the background and foreground scores for each pixel, our objective is to estimate a binary label $\mathcal{L}^t$ at time t, which denotes if the pixel belongs to background, $\mathcal{L}^t(\mathbf{x}) = 0$, or foreground, $\mathcal{L}^t(\mathbf{x}) = 1$. The cues introduced in the Section 3-4 provide necessary constraints for the labeling problem. In particular, the motion of pixel $\mathbf{x}$ and how well it satisfies the background motion based on Eqn. (3) can be used to reflect the cost of a background or foreground label:

$$\mathcal{M}(\mathbf{x}) = \begin{cases} 1 - m(\mathbf{x}) & \text{if } \mathcal{L}^t(\mathbf{x}) = 0 \\ m(\mathbf{x}) & \text{otherwise} \end{cases}. \tag{9}$$

In similar fashion, how well the appearance of the pixel fits to the background or the foreground model can be competed using Eqns. (5) and (8) by:

$$\mathcal{A}(\mathbf{x}) = \begin{cases} s(\mathbf{x}|\mathcal{F}(\mathbf{x} + \boldsymbol{u})) & \text{if } \mathcal{L}^t(\mathbf{x}) = 0 \\ s(\mathbf{x}|\mathcal{B}(\mathtt{H}_i\mathbf{x})) & \text{otherwise} \end{cases}, \tag{10}$$

where $i \geq 0$. Aside from the motion and appearance based terms, one can expect that the label of a pixel should be both temporally and spatially consistent. These constraints are typically introduced to the labeling cost function as smoothness terms that penalize the assignment of different labels to pixel's spatial or temporal neighborhood. Let a pixel $\mathbf{x}$ in frame t corresponds to pixel $\mathbf{x}'$ in the previous frame $t-1$. The temporal smoothness $T(\mathbf{x})$ is defined based on the neighborhood $G(\mathbf{x})$ of $\mathbf{x}$ and can be computed as:

$$T(\mathbf{x}) = (1 - \delta(\mathcal{L}^t(\mathbf{x}) - \mathcal{L}^{t-1}(\mathbf{x}'))) \exp\left(\frac{-\|I^t(\mathbf{x}) - I^{t-1}(\mathbf{x}')\|^2}{2\sigma_T^2}\right), \tag{11}$$

where $\delta(\cdot)$ is a Kronecker delta function.

The spatial smoothness $V(\mathbf{x}_i, \mathbf{x}_j)$ enforces the adjacent pixels $\mathbf{x}_i$ and $\mathbf{x}_j$ in frame I^t to have the same label and can be formulated as:

$$V(\mathbf{x}_i, \mathbf{x}_j) = (1 - \delta(\mathcal{L}^t(\mathbf{x}_i) - \mathcal{L}^t(\mathbf{x}_j))) \exp\left(\frac{-\|I^t(\mathbf{x}_i) - I^t(\mathbf{x}_j)\|^2}{2\beta}\right), \quad (12)$$

where the constant β is defined [9] as:

$$\beta = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{x}_j \in G(\mathbf{x}_i)} \|I(\mathbf{x}_i) - I(\mathbf{x}_j)\|^2 \quad (13)$$

and $G(\mathbf{x}_i)$ is a set of neighboring pixels around $\mathbf{x}_i$. Given the motion, appearance, and smoothness constraints, the pixels can be labeled as foreground or background by minimizing the labeling cost function E is given by:

$$E(\mathcal{L}^t, \mathcal{X}) = \sum_{\mathbf{x}_i \in I^t} \mathcal{M}(\mathbf{x}_i) + \lambda_A \sum_{\mathbf{x}_i \in I^t} \mathcal{A}(\mathbf{x}_i) + \lambda_T \sum_{\mathbf{x}_i \in I^t} T(\mathbf{x}_i) + \lambda_S \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{N}} V(\mathbf{x}_i, \mathbf{x}_j), \quad (14)$$

where the appearance, temporal, and spatial terms are weighted by λ_A, λ_T, λ_S, and $\mathcal{N}$ is the neighboring system on pixels. The solution of energy minimization can be efficiently computed using the graph-cut algorithm [3].

6 Appearance Model Update

Once the labels for all the pixels are assigned, the new background and foreground observations can be used to update the background and foreground models by mapping the pixels based on the associated transformations. Let $\mathbf{x}'$ be the model location of pixel $\mathbf{x}$ in frame I^t. For a pixel with background label $\mathbf{x}' = \mathtt{H}_i\mathbf{x}$ for $i \geq 0$, while for a pixel with foreground label $\mathbf{x}' = \mathbf{x} + \boldsymbol{u}$. In order to update the appropriate component of the Gaussian mixture in $\mathcal{B}(\mathbf{x})$ or $\mathcal{F}(\mathbf{x})$, the color at $I(\mathbf{x})$ is checked against each component until a match is found. The parameters of a distribution that matches the current pixel are updated as:

$$\mu_i^m(\mathbf{x}) \leftarrow (1 - \alpha)\mu_i^m(\mathbf{x}') + \alpha I^t(\mathbf{x}), \quad (15)$$

$$\sigma_i^m(\mathbf{x}) \leftarrow (1 - \alpha)\sigma_i^m(\mathbf{x}') + \alpha(I^t(\mathbf{x}) - \mu_i^m(\mathbf{x}'))^\top (I^t(\mathbf{x}) - \mu_i^m(\mathbf{x}')), \quad (16)$$

where i indicates the selected component of the mixture model, $m = \{f, b\}$, and α is the learning rate. In this process, the mean and standard deviation of the unmatched distributions remain unchanged. If the components are updated, the weight of the matching component in the new mixture distributions are computed as follows:

$$w_i \leftarrow (1 - \alpha)w_i + \alpha, \quad (17)$$

while the weight of the remaining mixture components are updated by:

$$w_j \leftarrow (1 - \alpha)w_j. \quad (18)$$

which are renormalized to satisfy $\sum_{i=0}^{g} w_i = 1$. If none of the g components of the mixture model match $I^t(\mathbf{x})$, the component with lowest weight is replaced by normal distribution $N(I^t(\mathbf{x}), \sigma_{init})$, where σ_{init} is set to a high value.

For the foreground model, the above procedure is applied to $\mathcal{F}^{t-1}(\mathbf{x}^f)$ to construct $\mathcal{F}^t(\mathbf{x})$ if $\mathcal{L}^t(\mathbf{x}) = 1$ or $m(\mathbf{x}) < 0.5$. Otherwise, we update only $w_0 = w_0 + \alpha$, while other parameters remain unchanged. This formulation prevents the foreground model from learning the background.

7 Experiments

In contrast to the stationary camera case, there is no benchmark dataset for evaluating performances of background subtraction methods for moving cameras. Due to this unavailability, some studies do not provide quantitative comparisons [22,10]. In this paper, we use a set of sequences from the Hopkins dataset [21] (*cars1-8, people1-2*) and from [14] (*cars, person*) which have been used by recent quantitative papers on the topic [9,8,15,5]. The sequences in Hopkins dataset, however, typically contain 20 to 50 frames and does not contain the challenges posed in realistic scenarios. Hence, we additionally include two very challenging sequences (*indoor, outdoor*) that reflect a real-world setting acquired with a smartphone camera. For quantitative evaluation, we generated the ground truth by manually extracting all moving objects in all frames.

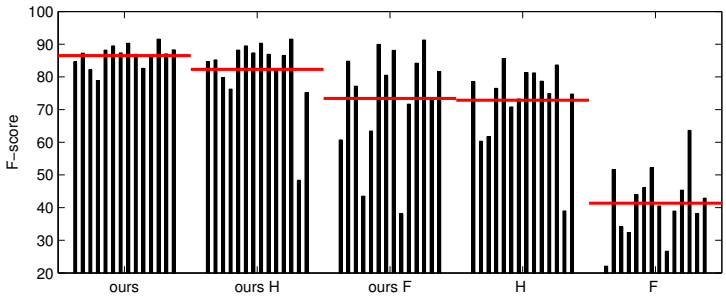

Fig. 4. The F-score computed from all sequences for our method and its variations. The bars in a group correspond to the sequences in the following order: *cars*, *person*, *cars1-8*, *people1-2*, *outdoor*, *indoor*. The red lines indicate the average F-score across the sequences for each approach.

Given a video sequence, our implementation generates dense point correspondences between consecutive frames from optical flow per pixel estimated using [1]. These correspondences are used to estimate both the fundamental matrix and the homograhy transform, which is followed by computing the respective GRIC score using Eqn. (1). The search region and the range of disparity in the GRIC are respectively set to $S = 30$, $R = 2$, $\sigma = 0.3$, and $\gamma = 0.6$. The geometric transformation providing the lowest GRIC score is selected to compute the

Table 1. Average precision (P), recall (R), and F-score (F) values for our and state-of-the-art methods. The best scores are denoted in bold.

	ours			Sheikh *et al.* [15]			Lim *et al.* [9]			Kwak *et al.* [8]		
	P	R	F	P	R	F	P	R	F	P	R	F
cars	**83.9**	**85.6**	**84.6**	65.1	84.6	72.7	79.4	64.4	71.0	59.5	62.6	60.7
person	82.3	93.6	**87.3**	69.6	**95.1**	80.0	**83.5**	83.1	82.7	53.9	62.8	56.8
cars1	72.9	**94.5**	**82.2**	68.5	74.3	67.9	63.0	87.2	72.6	**84.3**	73.8	78.5
cars2	69.8	**90.8**	78.9	54.7	81.7	63.6	**95.2**	77.2	**85.0**	67.9	74.1	70.5
cars3	**82.0**	95.6	**88.2**	62.8	**97.4**	76.1	70.6	87.7	77.9	80.4	80.2	80.2
cars4	**87.7**	**91.7**	**89.5**	68.5	88.3	76.2	80.8	73.1	75.1	57.5	67.9	62.1
cars5	**89.2**	**85.7**	**87.4**	62.7	79.7	66.2	69.4	82.5	75.3	62.3	68.0	64.5
cars6	**86.8**	94.2	**90.3**	68.8	**96.9**	79.8	64.4	73.1	68.4	62.4	89.0	73.1
cars7	80.2	**95.0**	86.9	81.3	94.4	**87.0**	**88.9**	84.2	86.2	66.2	72.9	69.1
cars8	73.7	**94.4**	**82.6**	**81.6**	85.4	82.2	73.7	76.2	74.9	77.5	76.6	76.7
people1	**92.5**	**81.6**	**86.6**	40.5	80.9	51.7	38.5	80.9	49.7	49.2	69.3	56.3
people2	**93.9**	89.5	**91.6**	72.6	88.0	78.2	71.6	**93.8**	80.5	85.0	77.4	80.8
outdoor	**91.3**	85.0	**87.0**	22.1	79.8	28.0	9.3	26.6	10.4	45.9	**86.4**	54.5
indoor	**91.6**	**87.4**	**88.3**	35.3	63.6	38.8	15.0	41.5	19.1	11.5	23.1	14.2

background model transformation (Section 4). The transformed model is used in the MAP-MRF framework with the following parameters $\lambda_A = 2$, $\lambda_T = 0.5$, $\lambda_S = 10$, and $\sigma_T = 10$. The final labels are then used to update the background and foreground models. During this step, a 3×3 window around projected pixel $\mathbf{x}'$ is evaluated, and the pixel with the highest probability is updated to avoid rounding and errors during the projection. In order to adapt to changes in appearance, we set the learning rate used for the model update to $\alpha = 0.05$.

We provide extensive comparison of the proposed method with its variations and the state-of-the-art. For different variations of our approach, we use 1) the complete method (*ours*), 2) our method with `H` only (*ours* `H`), 3) our method with `F` only (*ours* `F`), and two baseline methods that are obtained by theshold-ing of the motion score $m(\mathbf{x})$ computed with 4) homography only (`H`) and 5) fundamental matrix only (`F`). The competitive approaches are four state-of-the-art methods [15,9,8,7]. The implementation of [8] is provided by the authors[1], and we implemented the remaining methods. We selected the best parametric settings for all comparison after numerous experimental trials for quantitative evaluation. The overlap between the detected regions and the ground truth is analyzed by precision, recall, and their harmonic mean F-score.

In Figure 4, we plot the F-scores for different variations of our approach and two baseline methods. As expected, application of the appearance, spatial, and temporal constraints significantly improves the labeling compared to using only the motion scores (`H` and `F`). We observe that, for the Hopkins dataset and *cars*/*person* sequences, mostly the homography transform is chosen by GRIC. It can be attributed to the fact that the camera capturing these sequence moves very slowly, resulting in a very small baseline. As a result, for the aforementioned

[1] `http://cv.postech.ac.kr/research/gbs/`

Fig. 5. Qualitative results for (row 1) the proposed method, (row 2) Sheikh *et al.* [15], (row 3) Lim *et al.* [9], and (row 4) Kwak *et al.* [8] for *cars4*, *people2*, *indoor* and *outdoor* sequences. Ten more sequences are included in supplemental material.

sequences, the performance of our method is comparable to the case where the homography transform alone is used (*ours* `H`). Note that, however, in many cases, our method results in a considerably higher performance compared to always choosing the fundamental matrix (*ours* `F`). The strength of our method can be realized in more complex sequences, that contain both camera rotation only and camera translation with complex scenes, such as the *indoor* and *outdoor* sequences. For these sequences, the alternated usage of `H` and `F` results in a higher performance than that of using `H` or `F` alone. Detailed results for this figure are tabulated in supplemental material.

In Table 1, we present quantitative comparisons of our method with the state-of-the-art methods. Note that our approach mostly outperforms the competitive methods, and it results in a significantly higher accuracy for long and realistic sequences (*indoor* and *outdoor*). As presented in qualitative results in Figure 5 (more results are included in supplemental material), we observe that [15] is susceptible to the noise in trajectories around the moving objects and image

boundaries. This method also suffers from inconsistencies due to the lack of appearance models and temporal constraints. The methods introduced in [9] and [8] rely on the fundamental matrix for inferring camera and object motion and the propagation of appearance models. As a result, their models become corrupted when the fundamental matrix estimation is unsuccessful for a few consecutive frames. We also observed that, both of these methods are highly dependent on the correct background/foreground initialization in the first frame. Due to this requirement, these methods are initialized with the homography transform in the first frame in cases when the corresponding fundamental matrix estimation is observed to be incorrect. The results across the sequences for [7] are not provided since the algorithm is not appropriate when moving objects are present in the first frame, which is the case for all sequences except the *indoor*. For the *indoor* sequence, [7] results in 11.71 precision, 51.88 recall and 18.22 F-score values. The low performance can be attributed to the fact that [7] may estimate homography transforms for foreground objects as if they are part of the background.

While the proposed method outperforms the state-of-the-art, we observed the following limitations during our experiments. When the moving object is present in the scene in the first frame, the occluded parts that become visible as the object moves may be initially misdetected as foreground, especially for cluttered backgrounds. On the other hand, a moving object entering a previously unseen part of the scene revealed as the camera moves may be introduced as part of the background unless its motion is not significantly different from that of the camera. However, once the object continues to move, our algorithm correctly labels it as the foreground region.

8 Conclusions

We present a new method for background subtraction for moving cameras. Instead of committing to a single geometric transformation, we employ the Bayesian selection scheme to choose the model that best describes the transformation between the frames. As a result, the proposed method can adapt to various combinations of camera motions and scene structures. We maintain background and foreground models that are propagated using homography transform(s). The background/foreground labeling is obtained by combining the motion, appearance, spatial, and temporal cues in a MAP-MRF optimization framework. Extensive experimental results with challenging videos show that the proposed method outperforms the state-of-the-art in most cases.

References

1. Black, M.J., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. CVIU 63(1), 75–104 (1996)
2. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection: A systematic survey. Recent Patents on Computer Science 4, 147–176 (2011)

3. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. TPAMI 23(11), 1222–1239 (2001)
4. Cui, X., Huang, J., Zhang, S., Metaxas, D.N.: Background subtraction using low rank and group sparsity constraints. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 612–625. Springer, Heidelberg (2012)
5. Elqursh, A., Elgammal, A.: Online moving camera background subtraction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 228–241. Springer, Heidelberg (2012)
6. Gauglitz, S., Sweeney, C., Ventura, J., Turk, M., Höllerer, T.: Live tracking and mapping from both general and rotation-only camera motion. In: International Symposium on Mixed and Augmented Reality (2012)
7. Jin, Y., Tao, L., Di, H., Rao, N., Xu, G.: Background modeling from a free-moving camera by multi-layer homography algorithm. In: ICIP (2008)
8. Kwak, S., Lim, T., Nam, W., Han, B., Han, J.H.: Generalized background subtraction based on hybrid inference by belief propagation and Bayesian filtering. In: ICCV (2011)
9. Lim, T., Han, B., Han, J.H.: Modeling and segmentation of floating foreground and background in videos. PR 45(4), 1696–1706 (2012)
10. Liu, F., Gleicher, M.: Learning color and locality cues for moving object detection and segmentation. In: CVPR (2009)
11. Mahadevan, V., Vasconcelos, N.: Background subtraction in highly dynamic scenes. In: CVPR (2008)
12. Mittal, A., Paragios, N.: Motion-based background subtraction using adaptive kernel density estimation. In: CVPR, pp. 302–309 (2004)
13. Piccardi, M.: Background subtraction techniques: a review. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3099–3104 (2004)
14. Sand, P., Teller, S.: Particle video: long-range motion estimation using point trajectories. In: CVPR, pp. 2195–2202 (2006)
15. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV (2009)
16. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. TPAMI 27(11), 1778–1792 (2005)
17. Stauffer, C., Eric, W., Grimson, L.: Learning patterns of activity using real-time tracking. TPAMI 22, 747–757 (2000)
18. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. IJCV 50(1), 35–61 (2002)
19. Torr, P.H.S., Murray, D.W.: The development and comparison of robust methods for estimating the fundamental matrix. IJCV 24(3), 271–300 (1997)
20. Torr, P.H.S., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. CVIU 78(1), 138–156 (2000)
21. Tron, R., Vidal, R.: A benchmark for the comparison of 3-D motion segmentation algorithms. In: CVPR (2007)
22. Zhang, G., Jia, J., Hua, W., Bao, H.: Robust bilayer segmentation and motion/depth estimation with a handheld camera. TPAMI 33, 603–617 (2011)