This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

Background Extraction Based on Joint Gaussian Conditional Random Fields

Hong-Cyuan Wang, Yu-Chi Lai, Wen-Huang Cheng, Chin-Yun Cheng, and Kai-Lung Hua*

***Abstract*—Background extraction is generally the first step in many computer vision and augmented reality applications. Most existing methods, which assume the existence of a clean background during the reconstruction period, are not suitable for video sequences such as highway traffic surveillance videos, whose complex foreground movements may not meet the assumption of a clean background. Therefore, we propose a novel joint Gaussian conditional random field (JGCRF) background extraction algorithm for estimating the optimal weights of frame composition for a fixed-view video sequence. A maximum a posteriori problem is formulated to describe the intra- and inter-frame relationships among all pixels of all frames based on their contrast distinctness and spatial and temporal coherence. Because all background objects and elements are assumed to be static, patches that are motionless are good candidates for the background. Therefore, in the algorithm method, a motionless extractor is designed by computing the pixel-wise differences between two consecutive frames and thresholding the accumulation of variation across the frames to remove possible moving patches. The proposed JGCRF framework can flexibly link extracted motionless patches with desired fusion weights as extra observable random variables to constrain the optimization process for more consistent and robust background extraction. The results of quantitative and qualitative experiments demonstrated the effectiveness and robustness of the proposed algorithm compared with several state-of-the-art algorithms; the proposed algorithm also produced fewer artifacts and had a lower computational cost.**

***Index Terms*—Gaussian conditional random fields, background extraction, background initialization, background estimation, image fusion**

I. Introduction

Background extraction is a key step in many computer vision and augmented reality applications such as object identification and tracking [3]–[8] and online model estimation [9] that generally require a clear background for successful foreground patch detection. Foreground extraction research [10]–[12] in the CVPR Change Detection workshop has applied a sequence of frames to model background and foreground information; a clear preprocessing background image can make the modeling process more effective. Furthermore, in some situations, a clear background image is required of a complex scene with various moving objects. For example, when taking a picture at a popular tourist attraction, people desire to remove other tourists or moving objects in the foreground [13]. As a result, there is a need to extract a clear background image from a complex scene. Previous studies [1], [2], [14]–[16] generally faced two problems, (Fig. 1). First, some of them assumed that frames contain no foreground object or that there are clean background shots at the beginning of the video. However, video sequences generally contain complex foreground movements such as busy highway traffic and human motion in city streets; therefore, the assumption is invalid. Consequently, the reconstructed background image may contain blurring artifacts, as shown in Fig. 1. Second, although the shooting duration is not long enough to assume static lighting in the environment, the camera automatically changes its aperture and shutter speed owing to varying evaluated meterings of moving objects. This results in different exposure settings for different frames, leading to varying lighting appearance across frames. Therefore, these two problems must be addressed for acquiring a clean and artifact-free background image from a complex foreground and varying-exposure video.

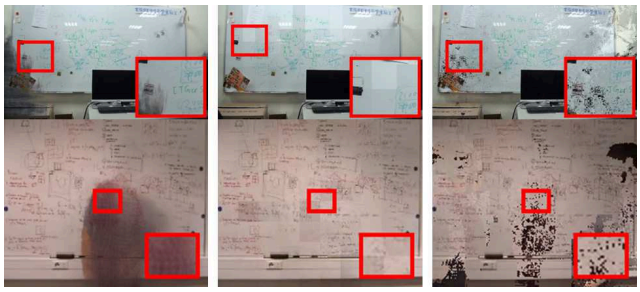


Fig. 1. Results obtained using several available background extraction algorithms with two white board scenes. The first column are background images extracted by the median filter, and they generally possess blurring artifacts due to foreground movement. The second column are background images extracted by the patch-based PBI method [1], and they generally possess blocking artifacts due to inconsistent selection from different frames. The third column are background images extracted by the hybrid method proposed by Chen *et al.* [2], and they generally possess blurring artifacts due to false selection of foreground patches. Furthermore, although the shooting duration is not long enough to assume static lighting in the environment, the camera automatically changes its aperture and shutter speed owing to varying evaluated meterings of moving objects. This results in different exposure settings for different frames, leading to varying lighting appearance across frames. As a result, this induces reconstructed variations across patches, especially on the top of the scene.

In general, background extraction techniques are classified into three categories: pixel-based, patch-based, and hybrid methods. Pixel-based methods, such as that in [14], consider only pixel-wise temporal coherence and ignore

Manuscript received at December 14, 2016.

H.-C. Wang, Y.-C. Lai, C.-Y. Cheng, and K.-L. Hua (corresponding) are with CSIE, NTUST, Taipei, Taiwan, ROC, corresponding e-mail: hua@mail.ntust.edu.tw.

W.-H. Cheng is with Academia Sinica, Taipei, Taiwan, ROC.

Wang and Lai contribute equally.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

spatial coherence, resulting in blurring and misdetection artifacts (Fig. 1, first column). Patch-based methods, such as those in [1], [15], and [16], consider localized spatial coherence to improve the selection process of pixel-based methods; however, the distinct blocks used for background extraction result in blocking artifacts and seams (Fig. 1, second column). This problem becomes even more obvious when varying exposure results in differences in lighting appearance in frames and causes serious blocking. Hybrid methods, such as that in [2], combine the advantages of the methods from the first two categories; however, false selection of possible candidate pixels blurs the background (Fig. 1, third column). We thus propose a system that overcomes these problems by formulating background extraction as an image fusion process in which pixels from different frames are fused; moreover, optimal frame weights are estimated using a joint Gaussian conditional random field (JGCRF) to model the distinctness and spatial and temporal coherence among pixels. Although JGCRF has been applied for image denoising [17], image labeling [18], and multi-exposure high dynamic range fusion [19], to the best of our knowledge, our proposed method is the first to apply JGCRF for background extraction. The proposed fusion algorithm has low computational complexity and produces a final fused background image with fine details by optimally balancing the two quality measures. Additionally, weighting merging can implicitly consider a pixel's exposure distribution of candidate background frames to statistically compute the represented background value to overcome artifacts due to varying exposure. However, when a naive implementation is used, there still exists some ambiguity among pixels owing to the false selection of foreground patches, and this induces artifacts. To address this problem, the observation that motion provides important hints about whether pixels belong to a foreground object helps in designing a motionless patch extraction method based on pixel-based temporal difference and frame-based temporal variation to better consider the temporal coherence. Motionless patches are constructed with moving foreground candidates having a value of zero and motionless background candidates having a value given by the temporal contrast weighted by the sum of the contrast of all other background candidates at the same location. They act as extra observable random variables linking with desired fusion weights to constrain the solving process of JGCRF for a cleaner background. A performance comparison showed that the proposed motionless patch extraction method can remove almost all ghosting artifacts. Moreover, compared with other state-of-the-art algorithms, our system can successfully obtain a clear and high-quality background image with low computational cost.

This paper makes the following contributions:

1) Because complex foreground occlusion and temporal exposure variation cause blurring and blocking artifacts in various extraction algorithms, our background extraction algorithm is aimed at removing blurring and blocking artifacts by using the proposed JGCRF image fusion approach that considers contrast sensitivity, intensity consistency, and temporal coherence. The fusion weights are then formulated as a maximum a posteriori (MAP) problem, and a power-law normalization is used to further address the blurring problem.
2) Pixel movement provides important hints for separating the foreground and the background. However, previous studies have not considered it; therefore, they suffered from false foreground detection problems that degraded the resultant fused background (Fig. 1). By contrast, our system extracts conservative motionless patches by discarding pixels that likely belong to foreground moving objects. Potential motion pixels are identified by using the temporal difference between two consecutive frames and accumulating the temporal variation in the pixel from different frames to successfully avoid false background labeling. In addition, our system computes the naive weight of background candidates as the ratio of the corresponding contrast to the sum of all valid contrasts at the same location; the computed weight acts as the initial weight to constrain the JGCRF optimization process.

The results showed that our algorithm can more effectively and robustly estimate a clear background image from a complexly occluded video sequence, compared with other state-of-the-art algorithms. The remainder of this paper is organized as follows. Section II reviews previous related studies. Section III discusses the details of our proposed JGCRF background extraction method. Section IV explains the concept of our motionless patch extraction method. Section V presents the results of our algorithm as well as comparisons with other state-of-the-art methods. Section VI concludes with a discussion of the limitations of this study and suggestions for future works.

II. Related Work

Background extraction has been extensively studied in the fields of computer vision and augmented reality. Owing to space limitations, only studies that are directly related to the present study are discussed in this paper.

Background extraction: Numerous background extraction methods have been proposed thus far. For example, several researchers at the 2015 Scene Background Modeling and Initialization workshop [20] aimed to achieve background initialization. Maddalena *et al.* [21] surveyed fixed-view background extraction methods and presented a thorough overview of this field. Toyama *et al.* [22] categorized background extraction methods into pixel-based, patch-based, and hybrid methods. Pixel-based methods extract a background image based on information about individual and independent pixels. In the simplest such approach, the background is reconstructed by applying a median filter to each pixel across all frames. Long and Yang [14] labeled the intensity appearing at the longest intervals in a pixel in video sequences as the intensity of the background pixel.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

Stauffer and Grimson [23] used a Gaussian mixture model (GMM) to estimate detailed shapes of foreground objects in video sequences for compositing a background image. Yang *et al.* [24] developed the Pixel-to-Model (P2M) distance to indicate the background possibility of a pixel for background reconstruction. Li *et al.* [25] separated the foreground and the background by using their proposed Bayesian framework. Zhou *et al.* [26] combined object detection and background learning into a single optimization process for efficiency and robustness. Sobral *et al.* [27] and Liu *et al.* [6] have applied matrix completion to an incomplete joint observation in motion detection and frame selection to conduct extra motion analysis during extraction. The main advantage of their approach is efficiency; however, their results could be easily affected by complex foreground movements and surrounding illumination variations because they ignored the spatial coherence and influence among other pixels. Additionally, when foreground objects first stop and then move or first move and then stop, detection may be problematic. The color consistency measure is used to consider the relationship among neighboring pixels and to find a balance with optimal weights for pixels from different frames based on the JGCRF framework. Our motionless extraction approach further improves the composition through a global consistency analysis of temporal motions.

Patch-based methods extract background pixels from a video sequence using information computed from image patches rather than from individual pixels. Russell and Gong [28] composed the background by using patch-based block matching. Colombari and Fusiello [1] grouped pixels in the temporal domain by using the Sum of Squared Differences (SSD) criterion and selected background patches based on the consistency analysis of corresponding patches. Similarly, Reddy *et al.* [15], [16] have grouped pixels using SSD but selected patches based on the frequency response of corresponding patches. Ortego *et al.* [29] used motion filtering and dimensionality reduction trained by a set of labeled data to obtain a set of candidate blocks, and then they spatially reconstructed the background image using these candidates. Patch-based methods use localized block information to improve the precision of background selection; however, shadows, reflections, varying exposure lighting appearance, and complex movements could still degrade the accuracy of background detection. Patch-based methods do not take into consideration interpatch spatial coherence and interframe exposure statistics, and thus, some blocking artifacts and seams exist between two consecutive blocks. These limitations can be overcome by designing temporal coherence and contrast consistency as measures for the optimal weight searching process using JGCRF, in addition to merging weights across multiple frames to consider the statistics of the shooting conditions.

Hybrid methods integrate the concepts of both pixel- and patch-based methods to achieve better results. Chen and Aggarwal [2] selected neighboring regions from different frames to compose background images based on optical flow analysis. Huang *et al.* [30] also used optical flow analysis to estimate the pixel motion for a combination of RGB colors. However, their algorithm may still blur the extracted background owing to the false selection of possible candidate pixels; furthermore, its computational cost is high. Our proposed approach computes fusion frame weights based on the JGCRF model that jointly considers the contrast sensitivity, color consistency, and temporal coherence to handle blurring and blocking artifacts.

Background extraction datasets and algorithmic analysis: The Scene Background Initialization (SBI) dataset [31] published in the Scene Background Modeling and Initialization workshop [20] offers a few test videos along with the corresponding ground truth results for both quantitative and qualitative evaluations. Moreover, Maddalena *et al.* [32] proposed six metrics for evaluating commonly available background initialization methods based on the SBI dataset. The algorithm proposed in the current study was analyzed in detail using the SBI dataset and Maddalenas six metrics and then compared with other state-of-the-art algorithms.

III. Mathematical Background and Our Background Reconstruction Method

A random field is a generalization of a stochastic process of multidimensional vectors or points on some manifold. When used in the natural sciences, values in a random field are often spatially correlated. Conditional random fields (CRFs) are used for structured predictions by taking context into account. They are generally modeled as a discriminative undirected probabilistic graphical model to encode known relationships between observations and to construct consistent interpretations. Previous studies [17]–[19], [33] have used CRFs to select a set of optimal weights to combine different candidates of solutions; background extraction can be viewed as the selection of pixels from different frames to form a complete and clear background image. The algorithm proposed in this study represents each pixel as a random variable whose space is the probable range of a pixel value; then, a clear background image is reconstructed via the formulation of CRFs. The following sections explain our formulation of CRFs in detail.

A. Joint Gaussian Conditional Random Field Formulation for Background Extraction

A video sequence is denoted as $\mathcal{F} = \{\mathbf{F}^1, \cdots, \mathbf{F}^K\}$ with the same size of N pixels in each frame, where K represents the total number of frames, $\mathbf{F}_i^k = 0.298R_i^k + 0.587G_i^k + 0.114B_i^k$ represents the illuminance of the i-th pixel at the k-th frame, and (R_i^k, G_i^k, B_i^k) represents the color triple of the pixel. All pixels of all frames where $N \gg K$ are registered for background reconstruction. We formulate the reconstruction of a background image, $\mathbf{O}$, as an image fusion problem with a set of optimal weights selected based on the JGCRF model expressed as the average weighting of all frames:

$$\mathbf{O}_i = \sum_{k=1}^{K} \mathbf{W}_i^k \mathbf{F}_i^k, \tag{1}$$

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

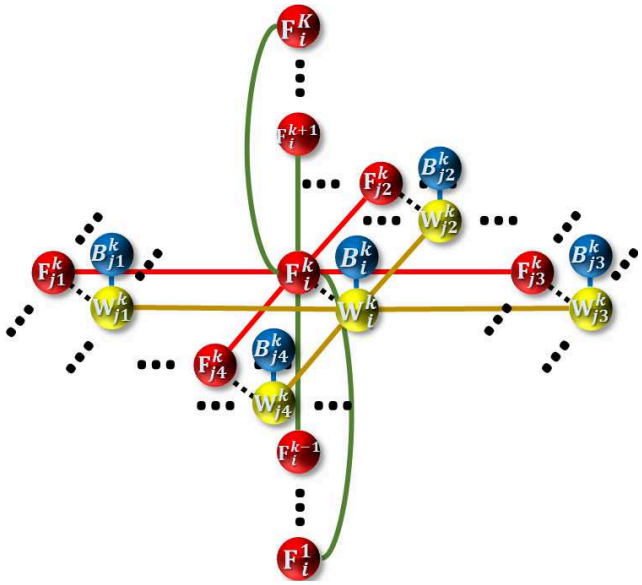


Fig. 2. Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ used in our joint Gaussian conditional random field model. Nodes consist of observable random variables illuminance, $\mathcal{F}$, marked as red circles, observable random variables naive weight, $\mathcal{B}$, marked as blue circles, and desired hidden random variables weight, $\mathcal{W}$, marked as yellow circles. The linking edges are the interactions between aforementioned random variables including those among $\mathcal{F}$ and $\mathcal{W}$, marked with dot lines, those among $\mathcal{B}$ and $\mathcal{W}$ marked with blue lines, those among two spatially neighboring $\mathcal{F}$ marked with red lines, those among two $\mathcal{F}$ which are at the same pixel location from two different frames marked with green lines, and those between two spatially neighboring $\mathcal{W}$ marked with brown lines.

where i and k are the pixel and frame index, respectively, and $\mathcal{W} = \{\mathbf{W}^1, \cdots, \mathbf{W}^K\}$ are frame weights recorded as maps for the contribution of the corresponding pixel in each frame.

Before defining and formulating JGCRF, we must define two quantities, contrast sensitivity and pixel similarity, for clarity.

- **Contrast sensitivity, $\mathbf{C}_i^k$**: Generally, the reconstructed result should be distinct and clear. Local contrast can reveal the degree of distinctness. Furthermore, contrast may be perceived nonlinearly by humans in the following way: When the pixel intensity is close to two extremes, (i.e., zero and one), the information is less trustable, and the trustability can be emphasized with a Gaussian weighting. Therefore, we define the value contrast sensitivity as

$$\mathbf{C}_i^k(\mathcal{A}) = \exp\left(-\frac{(\mathbf{A}_i^k - 0.5)^2}{2\sigma^2}\right), \tag{2}$$

  where $\mathcal{A}$ is the sequence of maps (i.e., $\mathcal{A}$ is either the illuminance $\mathcal{F}$ or naive weight $\mathcal{B}$ maps), $\mathbf{A}_i^k$ is the illuminance $\mathbf{F}_i^k$ or naive weight $\mathbf{B}_i^k$ of the i-th pixel at the k-th frame, $\exp(\cdot)$ is the exponential function, i is pixel index, k is the frame index, and σ is a user-defined parameter for controlling the response of illuminance to contrast.
- **Pixel similarity, $\mathbf{S}_{i,j}^k$**: Because spatially and temporally neighboring pixels have a high chance of being from the same object, they have a high chance of having similar colors and similar weights. Therefore, the random field should consider the smoothness and similarity among neighboring frame pixels, $\mathbf{F}_i^k$, and desired weights, $\mathbf{W}_i^k$, by using the difference between two neighboring pixels.

$$\mathbf{S}_{i,j}^k(\mathcal{A}) = \begin{cases} \exp\left(-\frac{\|\mathbf{A}_i^k - \mathbf{A}_j^k\|}{\sigma_2}\right), & if\ j \in \mathcal{N}_r(i), \\ 0, & otherwise. \end{cases} \tag{3}$$

  where i and j are two pixel indices; k is frame index; $\|\cdot\|$ is the Euclidean distance; $\mathcal{N}_r(i)$ is a neighboring region centered at pixel i with a radius of r, where $r = 1$ implies the four adjacency neighbors of i; and σ_2 is a user-defined parameter for controlling the response between difference and similarity. When the difference is small, the similarity is high. Additionally, when the value is large, the value is directly set to zero to avoid overemphasizing the difference. Pixel similarity is designed to maintain selection continuity and consistency among neighboring pixels and to reduce the influence of noise and the variation in the lighting appearance induced by varying exposure.

As shown in Fig. 2, our JGCRF is presented with an undirected random field graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ [34] whose vertices are a set of observable random variables, $\mathbf{X}$, and hidden random variables, $\mathbf{Y}$; the edges are the interactions between vertices. The observable random variables, $\mathbf{X}$, consist of all pixels of the observed frames denoted as observable random variable illuminance, $\mathcal{F}$, and all pixels of the predicted naive weighting maps denoted as observable random variable naive weight, $\mathcal{B}$; this allows the flexibility to provide extra information to direct the optimization direction. In other words, $\mathcal{B}$ provides reference information for solving JGCRF, and this work preserves the relationship among the observable random variable illuminance levels, $\mathcal{F}$, by setting $\mathbf{B}_i^k = \frac{\mathbf{C}_i^k(\mathcal{F})}{\sum_{k=1}^K \mathbf{C}_i^k(\mathcal{F})}$, which is the ratio of its contrast to the sum of all contrasts at the same pixel location. The hidden random variables, $\mathbf{Y}$, are all pixels of the intended weighting maps denoted as hidden random variable fusion weight, $\mathcal{W}$. Correspondingly, the linking edges between vertices indicate the interactions and compatibility among these random variables, as shown in Fig. 2. An observable random variable illuminance value, $\mathbf{F}_i^k$, is connected to its four connected observable random variable illuminance levels in the same frame for spatial analysis. The link energy is expressed by the pixel similarity, $\mathbf{S}_{i,j}^k(\mathcal{F})$, between $\mathbf{F}_i^k$ and $\mathbf{F}_j^k$. When $\mathbf{F}_i^k$ and its neighboring pixels have strong linking edges, the similarity level is high. An observable random variable illuminance, $\mathbf{F}_i^k$, is also linked to all other observable random variable illuminance levels, $\mathbf{F}_i^l$, at the same pixel location in different frames for temporal analysis. If two pixel frames have high temporal coherence, the linking edge is strong; that is, all pixels come from the same background objects. This link energy is set to the accumulated pixel similarities in its neighborhood as $\prod_{j \in \mathcal{N}_r(i)} \mathbf{S}_{i,j}^k(\mathcal{F}) \mathbf{S}_{i,j}^l(\mathcal{F})$, where k and l

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

are two frame indices. These two types of edges are used to depict the intra- and inter-frame relationship.

An edge also exists between an observable random variable illuminance, $\mathbf{F}_i^k$, and its corresponding hidden random variable weight, $\mathbf{W}_i^k$, whose energy is expressed as $\sum_{j\in\mathcal{N}_r(i)} \mathbf{S}_{i,j}^k(\mathcal{F}) + \mathbf{C}_i^k(\mathcal{F})$. If a frame pixel has high contrast and has a similar value to the neighboring pixels, the linking edge is strong. A hidden random variable weight, $\mathbf{W}_i^k$, is connected to its four connected hidden random variable pixel weights, $\mathbf{W}_j^k$, in the same frame for spatial coherence in order to choose similar labels for two neighboring pixels whose energy is expressed as $\mathbf{S}_{i,j}^k(\mathcal{F})$. Here, the temporal coherence among $\mathcal{W}$ is implicitly encoded with the edge to its corresponding observable random variable illuminance, $\mathbf{F}_i^k$. Finally, observable random variable naive weights, $\mathbf{B}_i^k$, are connected to their corresponding hidden random variable weights, $\mathbf{W}_i^k$, for constraining and initiating the optimization process of weight estimation. If the observable random variable naive weight, $\mathbf{B}_i^k$, has a high value, its corresponding hidden random variable weight, $\mathbf{W}_i^k$, should also have a high value. The edge energy is set to $\mathbf{C}_i^k(\mathcal{B})$.

$(\mathbf{X}, \mathbf{Y})$ is a conditional random field in which the hidden random variables, $\mathbf{Y}$, conditioned on $\mathbf{X}$, obey the Markov property with respect to the graph. Finally, we can determine the optimal weighting maps by solving a MAP problem on this undirected graph, as discussed in the next section. Our system uses the estimated weighting maps to fuse the input image to form a clean background according to (1). Experiments (Section V) showed that the results obtained using the proposed random field fusion technique were better than those obtained using other background extraction algorithms.

B. Optimization with MAP

The previous section first derives the relation of observable and hidden random variables in the form of an undirected graph, as shown in Fig. 2. Then, background extraction is formulated into a joint Gaussian CRF statistical model, and the pixel fusion weights are computed based on the MAP estimation. To use the JGCRF model to describe their relationship, the following assumptions are made:

1) The difference between the corresponding desired pixel weights and naive pixel weights falls in a Gaussian distribution, and this relationship is denoted as $(\mathbf{W}_i^k - \mathbf{B}_i^k) \sim N(0, \alpha^{-1})$ where α^{-1} is the standard deviation of the Gaussian distribution.
2) The desired weight variation should be smooth among neighboring pixels, and their relationship should follow the Gaussian distribution. We denote their relationship as $(\mathbf{W}_i^k - \mathbf{W}_j^k) \sim N(0, \beta^{-1})$, where β^{-1} is the standard deviation of the Gaussian distribution.

Accordingly, the joint Gaussian form can be used to express the related probability density function [19] as

$$p(\mathcal{W}|\mathcal{B}, \Lambda, \Sigma) \propto \exp\left(\mathrm{Tr}\left(-\mathcal{W}^T\Lambda\mathcal{B} - \frac{1}{2}\mathcal{W}^T\Sigma\mathcal{W}\right)\right), \tag{4}$$

where $\mathrm{Tr}(\cdot)$ is the trace operation, and Λ and Σ are its two precision matrices that are designed to describe the inter- and intra-frame relationships, respectively. Λ is used to model the interframe relationship between the observable random variable naive weight, $\mathcal{B}$, and hidden random variable weight, $\mathcal{W}$. Σ is used to describe the intraframe relationship within the hidden random variable weight, $\mathcal{W}$. In the graph definition, $\mathcal{W}$ and $\mathcal{B}$ are denoted as

$$\mathcal{W} = \begin{pmatrix} \mathbf{W}_1^1 & \cdots & \mathbf{W}_1^K \\ \vdots & \ddots & \vdots \\ \mathbf{W}_N^1 & \cdots & \mathbf{W}_N^K \end{pmatrix}, \mathcal{B} = \begin{pmatrix} \mathbf{B}_1^1 & \cdots & \mathbf{B}_1^K \\ \vdots & \ddots & \vdots \\ \mathbf{B}_N^1 & \cdots & \mathbf{B}_N^K \end{pmatrix}, \tag{5}$$

where N is the total number of pixels in each frame, and K is the total number of video sequences. Because the goal of this JGCRF model is to reconstruct a clear background, the precision matrices are formulated based on a comprehensive and integrated survey of all information at the same pixel location from different frames. Additionally, the $N \times N$ precision matrix, Λ, can be further decomposed to indicate the interframe and interweight relationships of $\mathcal{W}$ and $\mathcal{B}$; our framework decomposes it into $\mathcal{U}$ and $\mathcal{V}$ to depict the interframe contribution and correlation between $\mathcal{W}$ and $\mathcal{B}$ as

$$\Lambda = \mathcal{U} + \mathcal{V}, \tag{6}$$

where $\mathcal{U}$ is a diagonal matrix and $\mathcal{V}$ is a symmetric matrix. Because $\mathcal{U}$ describes the contribution at the corresponding pixel location based on interframe features, the clearness and distinctness at the pixel location that are contributed by different frames (i.e., the sum of the contrasts at that location) constitute a good indicator. Accordingly, $\mathcal{U}$ is defined as

$$\mathcal{U}_{i,j} = \begin{cases} \sum_{k=1}^K \mathbf{C}_i^k, & if\ i = j, \\ 0, & otherwise. \end{cases} \tag{7}$$

Because $\mathcal{V}$ depicts the weighting correlation between two distinct neighboring pixel locations based on the interframe features, the dissimilarity between the accumulated contrast at two neighboring pixel locations from different frames is a good indicator. Accordingly, $\mathcal{V}$ is defined as

$$\mathcal{V}_{i,j} = \begin{cases} -\prod_{k=1}^K \exp\left(-\frac{\|\mathbf{C}_i^k - \mathbf{C}_j^k\|}{\sigma_1}\right), & if\ j \in \mathcal{N}_r(i), \\ 0, & otherwise. \end{cases} \tag{8}$$

where σ_1 is a user-defined parameter for controlling the response of contrast difference to the correlation. The other $N \times N$ precision matrix, Σ, can also be decomposed to take the intra-frame relationship into consideration to maintain the smoothness of the solution; our framework decomposes it into $\mathcal{P}$ and $\mathcal{Q}$ to depict the intraframe contribution and

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

correlation of $\mathcal{W}$ as

$$\Sigma = \mathcal{P} + \mathcal{Q}, \tag{9}$$

where $\mathcal{P}$ is a diagonal matrix and $\mathcal{Q}$ is a symmetric matrix. Generally, a clear composed background has pixels with similar color values for the same object and distinct values for different objects. In other words, both the accumulated similarity and contrast at the location are taken into consideration. Accordingly, $\mathcal{P}$, which indicates the contribution at the pixel location based on the intraframe features, is defined as

$$\mathcal{P}_{i,j} = \begin{cases} \sum_{j' \in \mathcal{N}_r(i)} \prod_{k=1}^{K} \mathbf{S}_{i,j'}^{k} + \sum_{k=1}^{K} \mathbf{C}_{i}^{k}, & if\ i = j, \\ 0, & otherwise. \end{cases} \tag{10}$$

Because $\mathcal{Q}$ describes the correlation between two distinct neighboring pixel locations based on the intraframe features, the similarity between the accumulated contrast at two neighboring pixel locations from different frames is a good indicator. $\mathcal{Q}$ is defined as

$$\mathcal{Q}_{i,j} = \begin{cases} -\prod_{k=1}^{K} \mathbf{S}_{i,j}^{k}, & if\ j \in \mathcal{N}_r(i), \\ 0, \ \text{otherwise.} \end{cases} \tag{11}$$

After setting up $\mathcal{U}$, $\mathcal{V}$, $\mathcal{P}$, and $\mathcal{Q}$, we can derive the final fusion weights by transforming (4) into a MAP problem as follows:

$$\begin{aligned} \mathcal{W}^{*} &= \arg\max_{\mathcal{W}} \exp\left(\mathrm{Tr}\left(-\mathcal{W}^{T}\Lambda\mathcal{B} \ - \ \frac{1}{2}\mathcal{W}^{T}\Sigma\mathcal{W}\right)\right) \\ &= \arg\min_{\mathcal{W}} \left(\mathrm{Tr}\left(\mathcal{W}^{T}\Lambda\mathcal{B} \ + \ \frac{1}{2}\mathcal{W}^{T}\Sigma\mathcal{W}\right)\right). \end{aligned} \tag{12}$$

Then, the optimized fusion weights, $\mathcal{W}$, can be found by solving the preceding MAP problem. This is equivalent to solving the linear system of $\Lambda\mathcal{B} = -\Sigma\mathcal{W}$ as

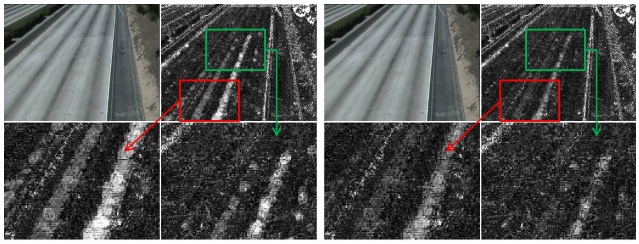
$$\begin{aligned} \mathcal{W}^{T}\Lambda\mathcal{B} &= -\frac{1}{2}\mathcal{W}^{T}\Sigma\mathcal{W} \\ \Lambda\mathcal{B} &= -\Sigma\mathcal{W} \\ \mathcal{W} &= -\Sigma^{-1}\Lambda\mathcal{B} \\ \mathcal{W} &= -(\mathcal{P}+\mathcal{Q})^{-1}(\mathcal{U}+\mathcal{V})\mathcal{B}. \end{aligned} \tag{13}$$

C. Weight Adjustment

After the JGCRF model is solved, a set of weights are determined for image fusion. However, when this set of weights is used directly, they remove too much distinctness, and the results are therefore generally blurred. To address this problem, the power-law normalization method is applied to adjust the weights; specifically, pixels with higher weights are emphasized to remove blurring artifacts. The weight renormalization can be expressed as

$$\mathbf{W}_{i}^{k\prime} = \frac{(\mathbf{W}_{i}^{k})^{n}}{\sum_{k=1}^{K}(\mathbf{W}_{i}^{k})^{n}}, \tag{14}$$

where n is a user-defined parameter that lies in the range $[1, +\infty)$. Fig. 3 presents the effect of weight adjustment. Note that this study used $n = 2$ in all comparisons.



(a) (b)

Fig. 3. Extracted background images using proposed JGCRF framework (a) without and (b) with weight adjustment. For each sub-figure, the upper left and the upper right show the original extracted background image and the normalized (enhanced) difference between the extracted result and the ground truth, respectively. Furthermore, the second row shows two zoomed in versions of the red and green boxes at upper-right.

IV. Motionless Patch Extraction

Although a video sequence may contain complex foreground movements that prevent a clear shot of the background, clean partial views of the background should exist. Previous studies [6], [27] have considered this observation to analyze the amount of pixel movement for generating a motion mask to assist background initialization. These concepts are extended to estimate the probability of a pixel belonging to the foreground objects by analyzing the pixels motion; pixels that contain obvious movement are discarded to prevent the incorporation of foreground information into the final result. Because a CRF allows initial conditions to be set so as to constrain the final solution, and because we assume a static background and camera, motionless patches are good background candidates for forming a composite and serve as the initial conditions for a clear background. Therefore, as shown in Fig. 4, these possible static patches are first labeled as motionless masks by thresholding the squared distance between corresponding pixels of two consecutive frames to consider the temporal coherence. Subsequently, the frame-based variation is used to aggregate the masks along with a morphological operator to generate smoother patch boundaries. Additionally, its temporal contrast weighted by the sum of the contrasts of all other background candidates at the same location is used to compute the initial naive weight of background candidates. Generally, CRFs are defined as an undirected graph model [34]. Thus, we first build an undirected graph model by having all frame pixels as nodes of observable random variable illuminance levels and connecting these vertices to spatial and temporal neighbors as edges for constructing CRFs; the dependency of the CRFs follows a joint normal distribution. Then, the weight of all frame pixels is set as a hidden random variable over the corresponding weight to link with the corresponding pixel node, and the input frames are analyzed to extract intra- and inter-frame features expressed as spatial color consistency and contrast measures. Furthermore, motionless patches act as observable random variable naive weights, and they are connected to their corresponding hidden random variable weights. The optimized fusion weight estimation can be solved as a MAP problem with the naive weights of

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

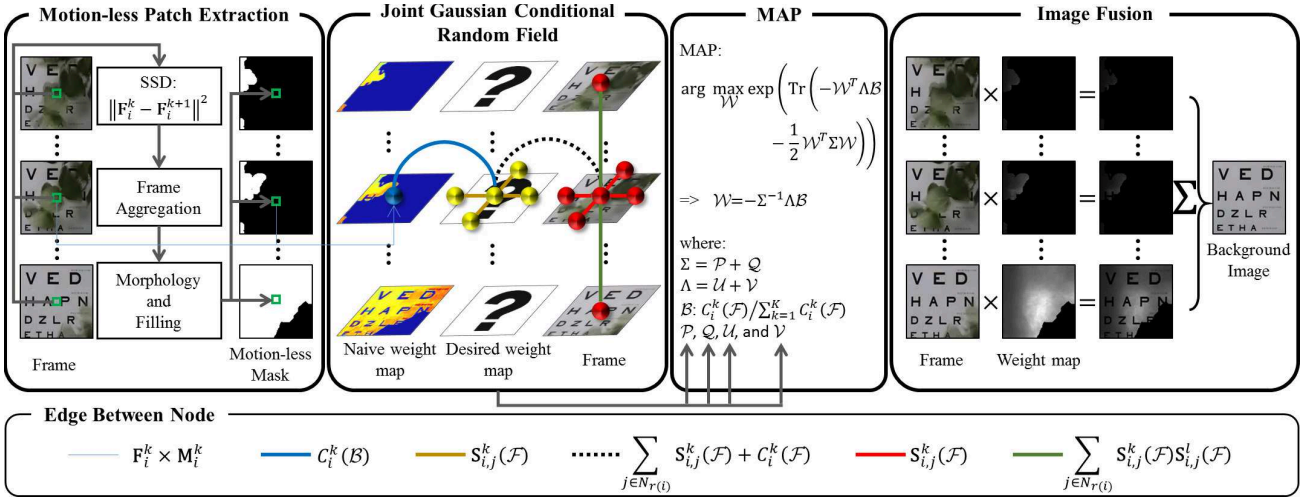


Fig. 4. Motionless patches are first estimated by using the temporal difference between the corresponding pixels of two consecutive frames and aggregating the frame-based variations to set foreground candidates to zero and background candidates to the value of the temporal contrast weighted by the sum of the contrasts of all other background candidates at the same location. Then, the JGCRF graph is constructed by setting all pixels of the input sequence as nodes of observable random variable illuminance levels and linking each node to its neighbors in the frame for the intraframe relationship and to its temporal neighbors in consecutive frames for the interframe relationship. Additionally, the weight of all pixels is set as hidden random variable weight and linked with the corresponding observable random variable illuminance. All hidden random variables are also linked to their neighbors for spatial coherence. Motionless patches are input as observable random variable naive weights to link with the corresponding hidden random variable weights to constrain the solving procedure. The fusion weight map of each frame is estimated as a MAP problem formulated using the JGCRF graph. Finally, frames are merged into a clear background image with the optimal frame weights.

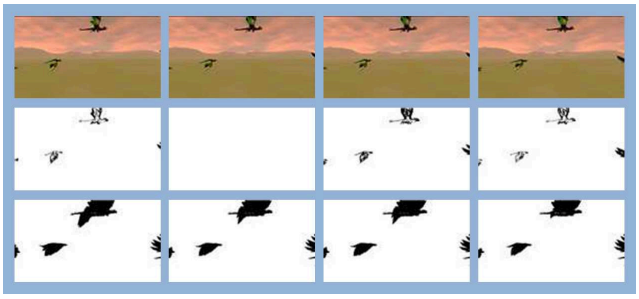


Fig. 5. This shows exemplar masks, $\mathbf{M}$, that indicate the candidate background pixels from our collected HD2 sequence. The first row shows the input frames (8, 9, 10, and 11) respectively. The second row shows the extracted masks using (16). The third row shows the aggregation results derived using frame variation.

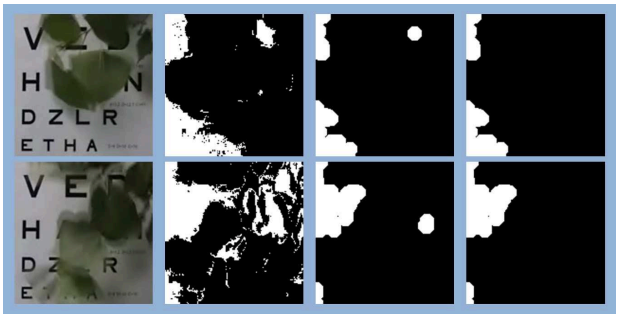


Fig. 6. This shows two intermediate results in the motionless mask construction process. The first column shows two exemplar input frames (1 and 133) respectively from an eye checker video sequence [35]. The second column shows the extracted results after (17). The third column shows the results after applying closing operations. The fourth column shows the results after applying filling operations.

motionless patches. Finally, the estimated weight of each pixel in each frame is used to fuse the final background result. To consider the temporal coherence and variation among pixels from consecutive frames, JGCRF uses temporal edges linking among observable random variable illuminance levels at the same pixel location from different frames. However, because this temporal analysis is only localized, and because merging weights for these varied and possible moving pixels are set to be small instead of zero, the reconstructed results would generally incorporate some foreground information, and foreground objects occupy a large portion of the view and appear for a long time with a complex and staggering motion pattern.

Because a background is generally assumed to have only static objects throughout the shooting period, the corresponding background pixels should have no motion. When foreground objects move, the difference between corresponding pixels from consecutive frames is large, and thus, the square difference is used to represent the movement possibility.

$$\mathrm{SD}_i^k = \|\mathbf{F}_i^k - \mathbf{F}_i^{k+1}\|^2, \tag{15}$$

After the difference is computed, a threshold, T_p, is chosen to determine whether a pixel is moving or motionless. The mask is computed as

$$\mathrm{g}_i^k = \begin{cases} 1, & \mathrm{SD}_i^k \le T_p \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

As shown in the second row of Fig. 5, when the temporal difference is used, misclassifications may occur when foreground objects are motionless for a while during the shooting. When a large portion of the frame contains

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

moving objects, the similarity between two consecutive frames is low; thus, frame similarity is used to aggregate motion information to address this problem. Therefore, our mask is given as

$$\mathbf{M}_i^k = \prod_{l:\mathrm{PSNR}(\mathbf{F}^k,\mathbf{F}^l)\geq T_q} \mathrm{g}_i^l, \tag{17}$$

where i is the pixel index, l and k are frame indices, T_q is a user-defined parameter, and $\mathrm{PSNR}(\cdot)$ is the peak signal-to-noise ratio (PSNR) between two frames, and our algorithm uses it to estimate the frame similarity.

$$\mathrm{PSNR}(\mathbf{F}^k,\mathbf{F}^l) = 10\times\log\left(\frac{255^2}{\mathrm{MSE}(\mathbf{F}^k,\mathbf{F}^l)}\right), \tag{18}$$

where $\mathrm{MSE}(\cdot)$ is the mean squared error of the frame difference, and it is defined as

$$\mathrm{MSE}(\mathbf{F}^k,\mathbf{F}^l) = \frac{1}{N}\sum_i \|\mathbf{F}_i^k - \mathbf{F}_i^l\|^2. \tag{19}$$

The third row in Fig. 5 shows the aggregation results derived using frame variation. Because moving pixels are assumed to be foreground objects, the detected foreground regions should be smooth and connected. However, the estimated $\mathbf{M}^k$ is shown to contain numerous rough and disconnected regions; the closing morphological operation could be used to connect these disconnected and rough regions, as shown in the third column in Fig. 6. In addition, there still exist holes that cannot be connected through the morphological operation; hence, the filling hole transform [36] is used to connect and remove these holes conservatively, as shown in the fourth column in Fig. 6.

After computing the masks, our algorithm sets the weight of foreground candidates to zero and that of background candidates to the value of the temporal contrast weighted by the sum of the contrasts of all other background candidates at the same location. The initial naive weight maps are used to form the random field graph for constraining the final weight optimization process for background fusion. Our extracted motionless patches can be used to conservatively remove all possible moving pixels, and they can also assist other background reconstruction methods in removing those pixels that do not belong to the background before evaluation; that is, they can discard these moving pixels from the evaluation to avoid the incorporation of foreground information.

V. Experimental Results and Discussion

After designing and implementing our algorithm, we applied it to reconstruct background images from a set of test video sequences. The collection and construction processes for these video sequences are described as follows, and comparisons against several state-of-the-art algorithms are presented. Owing to length limitations, this section only shows selected reconstruction results, and the complete test sequences, and their corresponding ground truth and reconstructed results, are provided on the supplemental website[1].

[1]web site: 140.118.155.223/BackgroundJGCRF/main.html

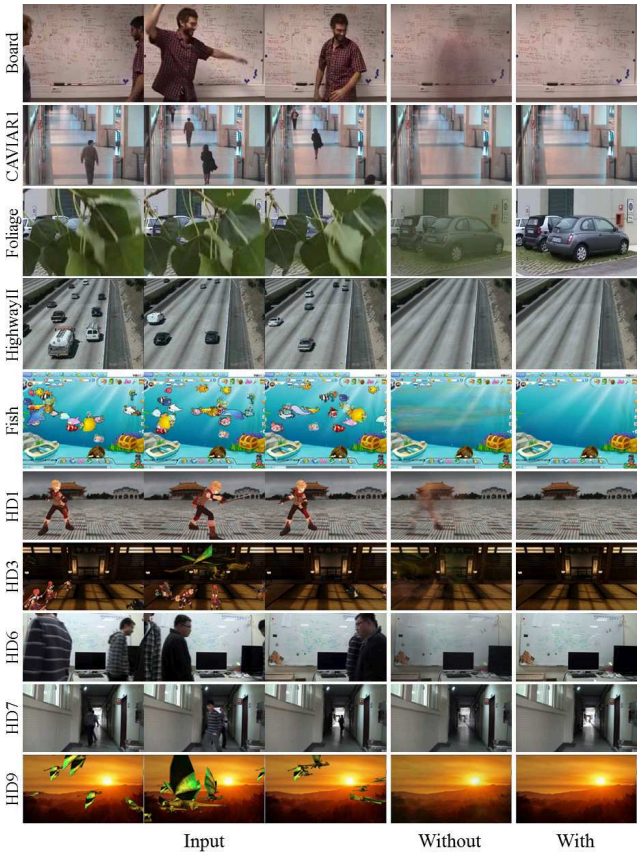


Fig. 7. Extracted background images from selected video sequences using our algorithm without/with motionless patch extraction.

A. Video Sequences and Ground Truth

Our test videos were obtained from two sources. First, video sequences including a whiteboard scene, a corridor scene, a person, a parking lot scene, a highway scene, another corridor scene, and another parking lot were collected from the websites provided by Colombari *et al.* [35] and the SBI dataset [31]. Second, sequences collected in previous studies generally have low resolution; however, all commercially available surveillance cameras are high-definition (HD) cameras; that is, they have 1920×1080 pixel resolution. Therefore, to evaluate the performance of our algorithm under current technologies, a set of video sequences was created in HD format. Our collected dataset comprised two categories of videos: virtual and realistic. Currently available game engines can build a virtual world with designed object motions and a virtual camera for animation generation. To test different aspects of our algorithm, complex movement patterns were designed for several foreground objects along with a static background. Additionally, we developed our algorithm with the main objective of extracting a background image from realistic videos of the real world. Therefore, we examined the motion patterns in existing testing videos used for background extraction and imitated these patterns to shoot similar HD video sequences. Notably, previous studies have collected their videos carefully to avoid exposure variation; however,

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

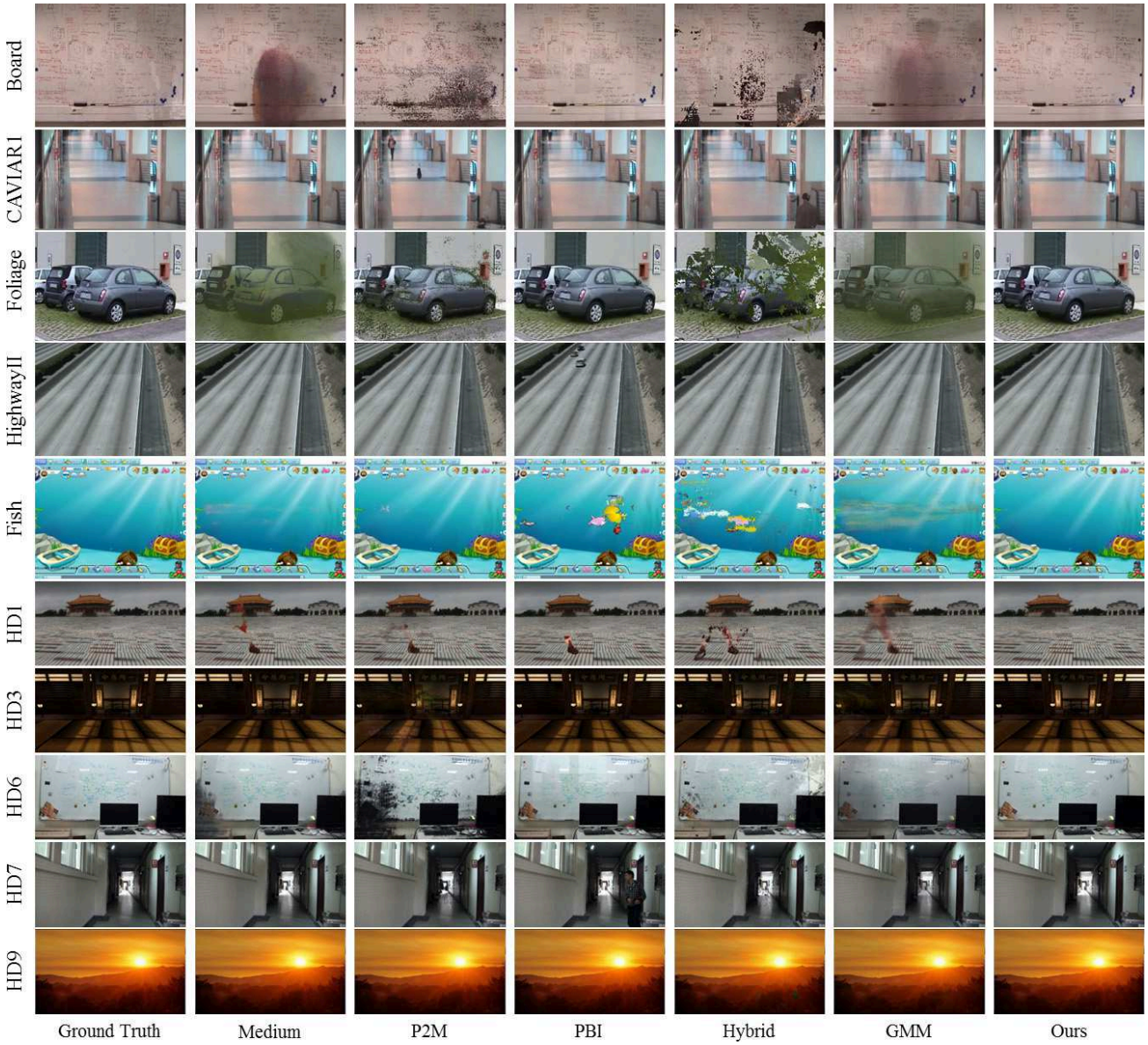


Fig. 8. Extracted background images from selected video sequences using the median filter, P2M [24], PBI [1], hybrid method [2], GMM [23], and our algorithm with motionless patch extraction.

during our collection, variation occurred occasionally. This is because in autoexposure mode, various aperture and shutter speed settings are selected based on the metering of objects in the scene, and the metered objects can change during shooting, thereby inducing illumination variations in different frames.

All test video sequences in the SBI dataset [31] have their own corresponding ground truth created by manually selecting pixels from video sequences. Furthermore, Maddalena *et al.* [32] proposed six metrics for evaluating the accuracy of results: average gray-level error (AGE), percentage of error pixels (pEPs), percentage of clustered error pixels (pCEPs), PSNR, multi-scale structural similarity index (MS-SSIM), and color image quality measure (CQM). We used these metrics to compare our reconstructed results against other results on the SBI sequences.

However, sequences from other studies do not have their corresponding ground truth data, and creating ground truth data for these video sequences entails using the same manual procedure applied to the SBI dataset. Nevertheless, manual selection is strenuous; hence, we implemented an automatic ground truth construction method and a performance comparison metric based on statistics. Because our generation fixes the camera parameters, all foreground objects can be simply removed to render the view to generate the ground truth. To produce video sequences, a standard procedure was developed to automatically generate the ground truth. This study first captured the background without any foreground objects for a few seconds. Because of the availability of only a few frames, the lighting variation can be considered negligible. The process was performed randomly, and all disturbances and factors, such

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

as shooting period and CMOS response, could be modeled randomly by using a Gaussian model. In other words, the color in each pixel from different frames should follow a Gaussian distribution. Therefore, our system can directly estimate the mean, μ_i, and standard deviation, $\sigma_{\mathbf{GT},i}$, of each pixel from the video sequence. Then, when two pixels come from the same Gaussian distribution (i.e., the pixel is correctly classified as background) with confidence α, we can classify them as a success. Therefore, the metric proposed by Colombari *et al.* [1] was adapted to estimate the number of misclassified pixels as

$$Err(\mathbf{GT}_i, \mathbf{B}_i) = \begin{cases} 1, & \text{if } \frac{1}{\sigma^2_{\mathbf{B}_i}+\sigma^2_{\mathbf{GT}_i}} \parallel \mathbf{B}_i - \mathbf{GT}_i \parallel^2 < \chi_3^{-1}(\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where $\mathbf{GT}$ and $\mathbf{B}$ are the ground truth and reconstructed background respectively.

B. Analysis of Free Parameters and Motionless Support

After obtaining the test video sequences and their corresponding ground truth data, we evaluated the influence of our parameters and the performance while conducting motionless analysis. Our method uses six user-defined parameters: σ, σ_1, σ_2, n, T_p, and T_q. In this study, we set $\sigma = 0.5$, $\sigma_1 = 0.05$, and $\sigma_2 = 0.02$ for all experiments. Generally, T_p and T_q are related to the length and resolution of the sequence. T_p and T_q were set as $(1500, 35)$ for sequences from the SBI dataset and as $(2000, 25)$ for our created HD video sequences. Furthermore, previous studies have generally revealed that n has essential effects on the reconstructed results. Hence, we visually examined the difference in reconstruction performance between different n values; we set n to 1, 2, and 3, and we observed that the performance difference between $n = 2$ and $n = 3$ was negligible. Consequently, we compared only the performance difference between $n = 1$ and $n = 2$, and the results are denoted in this paper as JGCRF-1 and JGCRF-2, respectively. Regarding JGCRF-1, the original weight determination method by our JGCRF model was used to construct the final results. The results were generally blurred and lacked detail. To address this problem, we applied (14) to adjust the weights for the initial weights used for JGCRF-2. This could prioritize important pixels to remove blurring artifacts. Therefore, the performance metrics derived for JGCRF-2 were superior to those derived for JGCRF-1, as shown in TABLES I and II. The JGCRF-2 result was close to the ground truth data. For example, the average values of AGE, pEPs, and pCEPs were reduced compared with those derived for JGCRF-1, and the average of misclassified pixels was also reduced, as presented in TABLE II. The effectiveness of our motionless patch extraction was also evaluated by executing the proposed JGCRF background extraction algorithm on these test sequences. As illustrated in Fig. 7, we observed severe ghosting artifacts when we did not apply the motionless patches in our JGCRF background extraction algorithm, and this is because all information is considered equally in the method. Because our motionless patch extraction algorithm analyzes pixel motions across frames temporally to eliminate possible moving pixels, using it with JGCRF enables our method to focus on smaller and better candidates to completely remove all ghosting artifacts.

Because different parameters may affect the reconstruction results, we examined the effects of different parameters on the image fusion accuracy during the reconstruction of the SBI sequence, as determined in terms of the PSNR. The six parameters (σ, σ_1, σ_2, n, T_p and T_q) in our algorithm were set as $(\sigma, \sigma_1, \sigma_2, n, T_p, T_q) = (0.5, 0.05, 0.02, 2.0, 1500, 25)$. When evaluating one parameter, our system keeps the values of the other parameters as defaults. For example, during the analysis of the influence of σ, the other parameters would be set as $\sigma_1 = 0.05$, $\sigma_2 = 0.02$, $n = 1.0$, $T_p = 1500$, and $T_q = 25$. The evaluation results are summarized in Fig. 9. When σ was more than 0.3, the reconstruction results generally had higher PSNR values. In addition, the reconstruction results were generally good for $n = 2$ and 3. σ_1 and σ_2 had little influence on the image fusion performance. For good results, T_p and T_q should be set higher than 1200 and 25, respectively. Therefore, these parameters were set as $\sigma = 0.5$, $\sigma_1 = 0.05$, $\sigma_2 = 0.02$, $n = 2.0$, $T_p = 1500$, and $T_q = 25$ for obtaining good background reconstruction results. Concurrently, our reconstruction results could achieve a high PSNR of at least 38 dB with these parameter settings; this verifies the robustness of these parameters.

C. Results and Comparisons

To better analyze and compare our algorithm against other state-of-the-art algorithms including the pixel-based Median method, P2M method [24], hybrid method [2], patch-based PBI method [1] and Gaussian mixture method (GMM), the properties of the collected sequences were examined in terms of the appearance period of foreground objects (long vs. short), motion speed (fast vs. slow), motion complexity (complex vs. simple), motion pattern (consistent vs. staggering), and exposure variation (steady vs. varied), as presented in TABLE III. As shown in Fig. 8, a clear background could be successfully reconstructed by our algorithm, despite complex foreground movements, a long foreground occupation period, different movement speeds, various motion patterns, and varied exposure settings; by contrast, different artifacts existed in the reconstruction results obtained using other methods. The performance of the other algorithms is summarized as follows.

Generally, if foreground objects appear only for a short period, such as those in CAVIAR1, HighwayI, HighwayII, IBMTest2, HD2, HD5, HD8, and HD9, the median filter can operate adequately. By contrast, when foreground objects appear in the view for a long period, such as those in Board, Foliage, Fish, Lab, HD1, HD3, HD4, HD6, and HD7, the foreground information is included in the reconstruction results, thus engendering artifacts; for example, the middle section of the backgrounds reconstructed from Board and HD1 is shown to contain serious artifacts. The

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

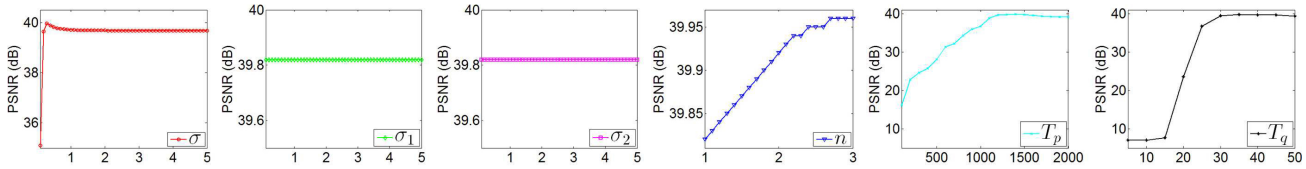


Fig. 9. Influences of the parameters σ, σ_1, σ_2, n, T_p, and T_q of our JGCRF model while using PSNR.

median filter also removes exposure variations that can be modeled with a normal distribution, as depicted in Board, HD5, and HD6.

P2M [24] uses the last 50 frames to compute pixel features for background reconstruction based on the appearance portion, and while foreground objects appear and move slowly during the reconstruction period, they are incorporated into the reconstructed background, resulting in blurring artifacts, as presented in Board, CAVIAR1, Foliage, Fish, IBMTest2, Lab, HD1, HD2, HD3, HD4, HD5, HD6, and HD7. While foreground objects move quickly in HD8, HD9, HighwayI, and HighwayII, P2M [24] can successfully discriminate them to reconstruct the background. Similarly, P2M [24] may select different pixels from various frames with different exposure settings, resulting in spatial lighting appearance variations.

Because PBI [1] selects a patch by analyzing the local consistency among patches, it can select the correct background object for reconstruction in Board, CAVIAR1, Foliage, HighwayI, IBMTest2, HD3, HD5, HD6, HD8, and HD9, and this is because foreground objects in these images move consistently and can thus be accurately detected; by contrast, objects in Fish, Lab, HighwayII, HD1, HD2, HD4, and HD7 may remain in a region for a specific period, thus resulting in artifacts. However, PBI [1] selects only a good candidate patch from the sequence; therefore, it may chose patches shot at different exposure settings, thus inducing blocking artifacts and illumination variations (as shown in Board, HD5, and HD6).

The hybrid method [2] depends on optical-flow-based motion analysis for a clear reconstruction. Thus, because foreground objects move obviously and consistently in HighwayI, HighwayII, and HD5, the algorithm can successfully remove them from the results. However, when the movement is relatively small, it fails to identify foreground objects, thus resulting in blurring artifacts, as shown in the reconstruction in CAVIAR1, IBMTest2, HD8, and HD9. Furthermore, when the movement is back-and-forth as in Board, Foliage, Fish, Lab, HD1, HD2, HD3, HD4, HD6, and HD7, motion analysis fails markedly, inducing obvious artifacts in the results. Similarly, it may choose pixels with different lighting appearances from different frames, as in Board, HD5, and HD6, leading to exposure artifacts in the reconstruction.

GMM [23] uses local appearance statistics to analyze consistency; thus, when foreground objects move consistently and rapidly, such as those in HighwayI, HighwayII, IBMTest2, HD8, it can reconstruct the background successfully. However, when objects move slowly or remain stagnant, as in CAVIAR1 and Fish, it fails to discriminate foreground objects from the background, resulting in blurring artifacts. Moreover, when the movement pattern is complex and objects move across a large portion of the view, as in Board, Foliage, Lab, HD1, HD2, HD3, HD4, HD5, HD6, HD7, and HD9, it fails to discriminate foreground objects from the background, resulting in blurring artifacts. Similar to the median filter, GMM can model exposure variations for the removal of lighting appearance variations.

In addition to the visual examination of the extracted results, we used the six metrics proposed in the SBI dataset to compare the extracted results with those generated by the P2M [24] method, patch-based PBI method [1], hybrid method [2], and the GMM [23]; the comparison results are shown in TABLE I. As indicated by the results obtained using these numerical metrics, our reconstructed backgrounds were very close to the respective ground truth data. Our reconstructed results were generally superior to those obtained using previous methods, as determined by the corresponding PSNR and MS-SSIM. Furthermore, our algorithm had the best results in terms of the PSNR and misclassified pixels in the HD dataset (TABLE II).

After evaluating and comparing the performance of our algorithm against other state-of-the-art algorithms, we analyzed the computational cost of the different algorithms. TABLE I lists the running times of all test algorithms when used to process different video sequences on a desktop computer equipped with an Intel Quad-Core i7 3.40 GHz CPU and 64 GB memory. The average running times of P2M [24], PBI [1], hybrid [2], GMM [23] and the proposed method per test video were approximately 1056.9, 333393.8, 6731.8, 11.9, and 48.4s, respectively. Although our method does not have the lowest computational complexity, it is robust for extracting a clear background image from video sequences containing complex foreground movements and temporal illuminance variations.

VI. Conclusion

Extracting a clear background is an important task in many applications. This study proposes an algorithm that successfully extracts clear background images from fixed-view video sequences with varying exposure lighting appearances, complex foreground movements, and occlusions without a clean shot of the background. The extraction process in this algorithm is formulated as image fusion with optimal frame weights estimated using a JGCRF model while considering distinctness and spatiotemporal coherence. For the further removal of blurring artifacts, motionless patches are extracted based on the temporal

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I

PERFORMANCE COMPARISONS WITH OTHER ALGORITHMS USING SIX MADDALENA'S METRICS [32] WHILE APPLYING TO SEQUENCES COLLECTED FROM PAST RESEARCH. P2M DENOTES THE PIXEL-TO-MODEL METHOD [24]. PBI DENOTES THE PATCH-BASED METHOD [1]. Hybrid DENOTES THE HYBRID METHOD [2]. GMM STANDS FOR THE GMM METHOD [23]. Ours JGCRF-1 STANDS FOR OUR JGCRF METHOD OF $n = 1$ ALONG WITH OUR MOTIONLESS PATCH EXTRACTION. Ours JGCRF-2 STANDS FOR OUR JGCRF METHOD OF $n = 2$ ALONG WITH OUR MOTIONLESS PATCH EXTRACTION. AGE DENOTES THE AVERAGE GRAY-LEVEL ERROR. pEPs DENOTES THE PERCENTAGE OF ERROR PIXELS. pCEPs DENOTES THE PERCENTAGE OF CLUSTERED ERROR PIXELS. MS-SSIM DENOTES THE MULTI-SCALE STRUCTURAL SIMILARITY INDEX. PSNR DENOTES THE PEAK SIGNAL-TO-NOISE RATIO. CQM DENOTES THE COLOR IMAGE QUALITY MEASURE. Time DENOTES THE COMPUTATIONAL TIME IN SECONDS.

Video Sequence	Resolution	Frames	Method	AGE	pEPs	pCEPs	MS-SSIM	PSNR	CQM	Time
Board	200 × 164	228	P2M	14.6891	17.9939%	4.3018%	0.6396	19.3845	36.2423	422.8
			PBI	8.3397	6.0366%	1.4085%	0.8878	27.2436	49.4015	1300.0
			Hybrid	19.4046	18.4421%	7.2866%	0.5900	17.5456	28.5480	2490.5
			GMM	21.9175	40.061%	33.814%	0.5721	18.5178	43.9199	**4.9**
			Ours JGCRF-1	4.3621	0.5213%	0.0274%	0.9645	32.7551	**51.3203**	15.2
			Ours JGCRF-2	**4.3007**	**0.4421**%	**0.0122**%	**0.9646**	**32.8820**	51.2372	16.1
CAVIAR1	384 × 288	600	P2M	4.2338	1.7721%	1.0478%	0.9391	27.4782	41.6122	3351.6
			PBI	**2.2186**	0.4293%	0.2777%	0.9899	34.7024	48.4701	46492.2
			Hybrid	3.3438	3.2888%	2.6713%	0.9678	28.8630	41.3713	4616.8
			GMM	3.7511	1.9653%	1.3194%	0.9536	32.0858	51.5780	**42.1**
			Ours JGCRF-1	2.4950	0.2218%	0.1170%	**0.9942**	37.6238	**51.5631**	140.2
			Ours JGCRF-2	2.4888	**0.2126**%	**0.1068**%	**0.9942**	**37.6939**	51.5485	143.3
Foliage	200 × 148	394	P2M	8.4035	11.8229%	0.4931%	0.9284	23.4306	35.7070	664.1
			PBI	**2.0303**	**0.0000**%	**0.0000**%	**0.9970**	**39.4755**	44.3090	912.6
			Hybrid	22.9960	29.4201%	13.3958%	0.6361	15.6689	30.3775	12787.9
			GMM	28.1972	60.2917%	40.7639%	0.7117	17.7199	27.6801	**7.6**
			Ours JGCRF-1	2.3081	**0.0000**%	**0.0000**%	0.9968	38.5990	**44.3938**	30.3
			Ours JGCRF-2	2.3484	**0.0000**%	**0.0000**%	0.9967	38.3189	44.0384	30.4
HighwayI	320 × 240	440	P2M	2.6527	0.9297%	0.0664%	0.9797	34.3766	47.4902	1003.1
			PBI	2.4808	0.5104%	0.0430%	0.9807	36.0805	57.2108	1950.7
			Hybrid	3.1211	0.6836%	**0.0143**%	0.9680	34.6029	56.4894	11429.4
			GMM	4.9577	0.3984%	0.0286%	0.9475	31.1735	**59.2909**	**5.8**
			Ours JGCRF-1	**1.6668**	**0.2161**%	0.0221%	**0.9915**	**38.9436**	58.7516	73.0
			Ours JGCRF-2	1.6747	0.2240%	0.0234%	0.9914	38.8793	58.5481	72.0
HighwayII	320 × 240	500	P2M	2.3065	0.4232%	0.0026%	0.9935	35.7604	43.4384	2147.0
			PBI	3.4889	1.4922%	0.5299%	0.9737	30.3793	38.9104	212328.3
			Hybrid	3.1326	0.5000%	0.0612%	0.9863	34.5069	41.0814	17704.0
			GMM	2.3605	0.3698%	0.0013%	0.9928	35.6755	46.2098	**26.6**
			Ours JGCRF-1	2.0043	0.3477%	**0.0000**%	**0.9951**	37.3915	46.7019	89.4
			Ours JGCRF-2	**1.9838**	**0.2760**%	**0.0000**%	**0.9951**	**37.8582**	**46.9710**	86.0
IBMtest2	320 × 240	90	P2M	6.7676	9.0182%	5.4714%	0.9197	25.7532	43.7863	384.7
			PBI	4.3554	0.5000%	0.1419%	0.9813	31.1707	40.4790	665.0
			Hybrid	3.0895	0.4023%	0.0352%	0.9828	32.7834	39.6624	913.1
			GMM	3.0315	0.0938%	0.0091%	0.9928	35.2485	49.0578	**5.1**
			Ours JGCRF-1	**2.5081**	**0.0195**%	**0.0000**%	**0.9961**	**37.7096**	49.6714	12.0
			Ours JGCRF-2	2.5272	**0.0195**%	**0.0000**%	**0.9961**	37.6611	**49.7184**	12.4
Fish	296 × 236	128	P2M	0.5103	0.1432%	0.0043%	0.9943	42.1387	64.5712	263.2
			PBI	2.1535	2.8144%	1.4287%	0.9044	28.9439	48.5460	2489.7
			Hybrid	4.2341	3.8680%	1.5274%	0.8299	24.7867	47.0206	1577.0
			GMM	2.4931	2.2546%	1.2197%	0.9351	32.7221	74.0855	**1.6**
			Ours JGCRF-1	0.1824	0.0215%	**0.0000**%	0.9993	51.3988	**75.6748**	14.1
			Ours JGCRF-2	**0.1809**	**0.0129**%	**0.0000**%	**0.9994**	**51.9074**	75.1320	14.3
Lab	240 × 160	192	P2M	21.0497	28.3047%	10.6302%	0.6311	16.1505	25.1252	218.9
			PBI	4.4358	3.6745%	2.5182%	0.9113	26.1499	34.8628	1011.9
			Hybrid	24.7799	32.1875%	18.0833%	0.4451	15.1593	23.7052	2335.7
			GMM	10.3409	21.5156%	15.9115%	0.8905	24.2246	32.3346	**1.4**
			Ours JGCRF-1	1.1553	**0.0000**%	**0.0000**%	**0.9991**	44.1249	**52.8020**	12.5
			Ours JGCRF-2	**1.1447**	**0.0000**%	**0.0000**%	**0.9991**	**44.1576**	52.5263	12.5
Average			P2M	7.5767	8.8010%	2.7522%	0.8782	28.0591	42.2466	1056.9
			PBI	3.6879	1.9322%	0.7935%	0.9533	31.7682	45.2737	33393.8
			Hybrid	10.5127	11.0991%	5.3844%	0.8008	25.4896	38.5320	6731.8
			GMM	9.6312	15.8688%	11.6334%	0.8745	28.4210	48.0196	**11.9**
			Ours JGCRF-1	2.0853	0.1685%	0.0208%	**0.9921**	39.8183	**53.8599**	48.3
			Ours JGCRF-2	**2.0812**	**0.1484**%	**0.0178**%	**0.9921**	**39.9198**	53.7150	48.4

TABLE II

THIS SHOWS THE COMPARISON RESULTS AGAINST OTHER ALGORITHMS WITH OUR OWN METRIC USING OUR COLLECTION VIDEO SEQUENCES. P2M DENOTES THE PIXEL-TO-MODEL METHOD [24]. PBI DENOTES THE PATCH-BASED METHOD [1]. Hybrid DENOTES THE HYBRID METHOD [2]. GMM STANDS FOR THE GMM METHOD [23]. Ours JGCRF-1 STANDS FOR OUR JGCRF METHOD OF $n = 1$ ALONG WITH OUR MOTIONLESS PATCH EXTRACTION. Ours JGCRF-2 STANDS FOR OUR JGCRF METHOD OF $n = 2$ ALONG WITH OUR MOTIONLESS PATCH EXTRACTION. Err. DENOTES THE DIFFERENCE BETWEEN THE GROUND TRUTH AND THE EXTRACTED BACKGROUND. PSNR DENOTES THE PEAK SIGNAL-TO-NOISE RATIO.

Video Sequence	Frames	P2M		PBI		Hybrid		GMM		Ours JGCRF-1		Ours JGCRF-2	
		PSNR	Err.	PSNR	Err.	PSNR	Err.	PSNR	Err.	PSNR	Err.	PSNR	Err.
HD1	152	30.5637	3.80e-2	30.8166	1.35e-2	25.9956	4.10e-2	28.4263	1.09e-1	42.8286	3.00e-3	**42.8290**	**2.97e-3**
HD2	196	35.7177	7.80e-3	32.0908	6.50e-3	23.6174	5.00e-2	32.3165	2.02e-1	**39.5558**	**1.16e-4**	39.4974	1.17e-4
HD3	87	30.3679	1.01e-1	49.5452	1.00e-3	35.2130	4.19e-2	33.2084	1.01e-1	**52.9384**	**4.60e-5**	52.9303	4.80e-5
HD4	135	24.2946	1.99e-1	13.6051	2.49e-1	20.7819	2.52e-1	24.7827	3.89e-1	**35.0191**	6.43e-2	34.9919	**6.30e-2**
HD5	154	23.1159	2.90e-1	**27.9751**	**2.65e-1**	27.0693	3.22e-1	27.3665	2.73e-1	26.5606	4.07e-1	26.7628	3.83e-1
HD6	152	14.8634	3.55e-1	28.2896	4.08e-1	23.1306	3.18e-1	17.2153	5.37e-1	**33.9547**	**3.24e-1**	33.3948	3.25e-1
HD7	155	26.2205	5.69e-2	23.8395	1.07e-1	25.3421	3.26e-1	28.6920	6.02e-2	31.7435	**3.53e-2**	**31.7696**	3.54e-2
HD8	155	41.5755	3.00e-3	42.1136	2.00e-3	39.9098	3.55e-2	42.9385	8.00e-4	**42.6759**	7.68e-4	42.6683	**7.64e-4**
HD9	130	43.9635	9.96e-4	46.0475	2.31e-5	42.0886	2.28e-2	41.5060	1.60e-3	**46.0483**	5.80e-5	46.0431	**5.70e-5**
Average		30.0759	1.17e-1	32.7026	1.17e-1	29.2387	1.57e-1	30.7169	1.86e-1	**39.0361**	9.27e-2	38.9875	**8.99e-2**

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III
THIS LISTS THE CHARACTERISTICS OF OUR TEST SEQUENCE. F.P. DENOTES THE APPEARING PERIOD OF FOREGROUND OBJECTS; L. AND SH. DENOTE A LONG AND A SHORT PERIOD, RESPECTIVELY. M.S. DENOTES THE MOTION SPEED; F. AND SL. DENOTE FAST AND SLOW MOTION, RESPECTIVELY. M.C. DENOTES THE MOTION COMPLEXITY; C. AND SI. DENOTE COMPLEX AND SIMPLE MOTION, RESPECTIVELY. M.P. DENOTES THE MOTION PATTERN; CT. AND STA. DENOTE "CONSISTENT AND STEADY MOVEMENT" AND "INTERTWINED MOVEMENT OF MOVING AND STOPPING", RESPECTIVELY. E.V. DENOTES THE EXPOSURE VARIATION; ST. AND V. DENOTE STEADY AND VARIED EXPOSURE SELECTION, RESPECTIVELY.

	F.P.	M.S.	M.C.	M.P.	E.V.
Board	L.	F.	C.	CT.	V.
CAVIAR1	Sh.	Sl.	Si.	Sta.	St.
HighwayI	Sh.	F.	Si.	CT.	St.
HighwayII	Sh.	F.	Si.	CT.	St.
Foliage	L.	F.	C.	CT.	St.
Fish	L.	Sl.	C.	Sta.	St.
IBMTest2	Sh.	F.	Si.	CT.	St.
Lab	L.	Sl.	C.	Sta.	V.
HD1	L.	F.	C.	Sta.	St.
HD2	Sh.	Sl.	C.	Sta.	St.
HD3	Sh.	Sl.	C.	CT.	St.
HD4	L.	F.	C.	Sta.	St.
HD5	Sh.	F.	Si.	CT.	V.
HD6	L.	F.	Si.	CT.	V.
HD7	L.	F.	Si.	CT.	St.
HD8	Sh.	F.	Si.	CT.	St.
HD9	L.	F.	Si.	CT.	St.

differences and variations and are input to the JGCRF model as extra observable constraints for determining the merging weights. Our experimental results demonstrate our algorithm to be more effective and robust with better output quality compared with other state-of-the-art algorithms. However, our system has some limitations, and a few future research directions are possible. In the proposed, background images of fixed-view video sequences can be extracted successfully; however, camera motion causes problems in the system. For example, our system may fail to compute a clear background from sequences taken by a camera that is mounted on top of a car. Therefore, it is desirable to incorporate camera tracking into our framework for broader applications. Second, although our motionless patch extraction conservatively selects candidates with high possibility, it does not globally examine the motion consistency of these candidates for the removal of patches that are locally static but globally dynamic. Thus, another appealing future research direction would be to design a global motionless extraction algorithm for achieving better results.

ACKNOWLEDGEMENT

This work was supported by Ministry of Science and Technology of Taiwan under Grants MOST106-3114-E-011-004, MOST106-2218-E-011-009-MY2, MOST106-2221-E-011-154-MY2, MOST104-2221-E-011-091-MY2, and MOST105-2228-E-011-005.

REFERENCES

[1] A. Colombari and A. Fusiello, "Patch-based background initialization in heavily cluttered video," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 926–933, Apr. 2010.

[2] C.-C. Chen and J. K. Aggarwal, "An adaptive background model initialization algorithm with objects moving at different depths," *IEEE International Conference on Image Processing*, 2008.

[3] J.-W. Seo and S. D. Kim, "Recursive on-line $(2D)^2$PCA and its application to long-term background subtraction," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2333–2344, Dec. 2014.

[4] J. Wen, Y. Xu, J. Tang, Y. Zhan, Z. Lai, and X. Guo, "Joint video frame set division and low-rank decomposition for background subtraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 12, pp. 2034–2048, Dec. 2014.

[5] A. Staglian, N. Noceti, A. Verri, and F. Odone, "Online space-variant background modeling with sparse coding," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2415–2428, Aug. 2015.

[6] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.

[7] D. K. Panda and S. Meher, "Detection of moving objects using fuzzy color difference histogram based background subtraction," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 45–49, Jan. 2016.

[8] X. Zhang, C. Zhu, S. Wang, Y. Liu, and M. Ye, "A bayesian approach for camouflaged moving object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, Apr. 2016.

[9] J. Ventura and T. Hllerer, "Online environment model estimation for augmented reality," *IEEE International Symposium on Mixed and Augmented Reality*, pp. 103–106, 2009.

[10] Change detection workshop. [Online]. Available: http://changedetection.net/

[11] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–8, Jun. 2012.

[12] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 393–400, Jun. 2014.

[13] W.-H. Cheng, C.-W. Hsieh, S.-K. Lin, C.-W. Wang, and J.-L. Wu, "Robust algorithm for exemplar-based image inpainting," *The International Conference on Computer Graphics, Imaging and Vision*, pp. 64–69, Jul. 2005.

[14] W. Long and Y.-H. Yang, "Stationary background generation: an alternative to the difference of two images," *Pattern Recognition*, vol. 23, no. 12, pp. 1351–1359, 1990.

[15] V. Reddy, C. Sanderson, and B. C. Lovell, "An efficient and robust sequential algorithm for background estimation in video surveillance," *IEEE International Conference on Image Processing*, pp. 1109–1112, 2009.

[16] ——, "A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts," *EURASIP Journal on Image and Video Processing*, Apr. 2011.

[17] R. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freema, "Learning Gaussian conditional random fields for low-level vision," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[18] X. Wang, X.-P. Zhang, I. Clarke, and Y. Yakubovich, "A new Gaussian mixture conditional random field model for indoor image labeling," *In Proceedings of ACM Multimedia Workshop on Interactive Multimedia for Consumer Electronics*, pp. 51–56, 2009.

[19] R. Shen, I. Cheng, and A. Basu, "QoE-based multi-exposure fusion in hierarchical multivariate Gaussian CRF," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2469–2478, Jun. 2013.

[20] Scene Background Initialization (SBI) dataset. [Online]. Available: http://sbmi2015.na.icar.cnr.it/SBIdataset.html

[21] L. Maddalena and A. Petrosino, *Background Modeling and Foreground Detection for Video Surveillance*. Chapman and Hall/CRC, 2014, ch. Background Model Initialization for Static Cameras, pp. 1–16.

[22] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *IEEE International Conference on Computer Vision*, vol. 1, pp. 255–261, 1999.

[23] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, Jun. 1999.

[24] L. Yang, H. Cheng, J. Su, and X. Li, "Pixel-to-model distance for robust background reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 5, pp. 903–916, May 2016.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2733623, IEEE Transactions on Circuits and Systems for Video Technology

[25] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.
[26] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.
[27] A. Sobral, T. Bouwmans, and E. Zahzah, "Comparison of matrix completion algorithms for background initialization in videos," *SBMI 2015 Workshop in conjunction with ICIAP 2015*, pp. 510–518, 2015.
[28] D. Russell and S. Gong, "A highly efficient block-based dynamic background model," *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 417–422, Sep. 2005.
[29] D. Ortego, J. C. SanMiguel, and J. M. Martínez, "Rejection based multipath reconstruction for background estimation in video sequences with stationary objects," *Computer Vision and Image Understanding*, vol. 147, no. C, pp. 23–37, 2016.
[30] S.-S. Huang, L.-C. Fu, and P.-Y. Hsiao, "Region-level motion-based foreground segmentation under a bayesian network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 4, pp. 522–532, Apr. 2009.
[31] "Scene Background Initialization (SBI) dataset." [Online]. Available: http://sbmi2015.na.icar.cnr.it/SBIdataset.html
[32] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," *SBMI 2015 Workshop in conjunction with ICIAP 2015*, Sep. 2015.
[33] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, Boston, MA, 2005.
[34] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 282–289.
[35] "Background initialization website." [Online]. Available: http://www.diegm.uniud.it/fusiello/demo/bkg/
[36] P. Soille, *Morphological Image Analysis: Principles and Applications*, 2nd ed. Springer-Verlag New York, Inc., 2003.

Hong-Cyuan Wang received the M.S. degree from National Taiwan University of Science and Technology, Taipei, Taiwan, in 2013. He is pursuing the Ph.D degree in Department of Computer Science and Information Engineering from National Taiwan University of Science and Technology, Taipei, Taiwan. He joined the Multimedia Computing Laboratory (MCLab) in Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, in 2013, as a research assistant. His current research interests include image fusion, background modeling, video streaming, and image and video processing.

Yu-Chi Lai received the B.S. from National Taiwan University, Taiwan, R.O.C., in 1996 in Electrical Engineering Department. He received his M.S. and Ph.D. degrees from University of Wisconsin-Madison in 2003 and 2009 respectively in Electrical and Computer Engineering and his M.S. and Ph.D. degrees in 2004 and 2010 respectively in Computer Science. He is currently an associate professor in NTUST and his Research interests are in the area of graphics, vision, and multimedia.

Weh-Huang Cheng received the Ph.D. (Hons.) degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2008. He is an Associate Research Fellow (Associate Professor) with the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, where he is the Founding Leader with the Multimedia Computing Laboratory (MCLab). His current research interests include multimedia content analysis, multimedia big data, deep learning, computer vision, mobile multimedia computing, social media, and human computer interaction.

Chin-Yun Cheng received the B.S. degree in computer science and information engineering from National Taiwan University of Science and Technology, Taiwan in 2014. Currently, he is a M.S. student in Department of Computer Science and Information Engineering at National Taiwan University of Science and Technology. His research interests include digital image and video processing and computer vision.

Kai-Lung Hua received the B.S. degree in electrical engineering from National Tsing Hua University in 2000, and the M.S. degree in communication engineering from National Chiao Tung University in 2002, both in Hsinchu, Taiwan. He received the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, in 2010. Since 2010, Dr. Hua has been with National Taiwan University of Science and Technology, where he is currently an associate professor in the Department of Computer Science and Information Engineering. He is a member of Eta Kappa Nu and Phi Tau Phi, as well as a recipient of MediaTek Doctoral Fellowship. His current research interests include digital image and video processing, computer vision, and multimedia networking.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.