This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Background Extraction Using Random Walk Image Fusion

Kai-Lung Hua, *Member, IEEE*, Hong-Cyuan Wang, Chih-Hsiang Yeh, Wen-Huang Cheng, *Senior Member, IEEE*, and Yu-Chi Lai, *Member, IEEE*

***Abstract*—It is important to extract a clear background for computer vision and augmented reality. Generally, background extraction assumes the existence of a clean background shot through the input sequence, but realistically, situations may violate this assumption such as highway traffic videos. Therefore, our probabilistic model-based method formulates fusion of candidate background patches of the input sequence as a random walk problem and seeks a globally optimal solution based on their temporal and spatial relationship. Furthermore, we also design two quality measures to consider spatial and temporal coherence and contrast distinctness among pixels as background selection basis. A static background should have high temporal coherence among frames, and thus, we improve our fusion precision with a temporal contrast filter and an optical-flow-based motionless patch extractor. Experiments demonstrate that our algorithm can successfully extract artifact-free background images with low computational cost while comparing to state-of-the-art algorithms.**

***Index Terms*—Background estimation, background extraction, background initialization, image fusion, random walk (RW).**

I. Introduction

OBJECT identification and tracking [1]–[7] often require to separate foreground objects from a background image before any further processing. Although there are research approaches such as those from the computer vision and pattern recognition Change Detection workshop [8]–[10] and from IEEE TRANSACTIONS ON CYBERNETICS [11], [12] focusing on utilizing a few frames to model the background and foreground information without extracting a clear background, the success of a clear background extraction can surely further improve the performance of foreground patch detection. In addition to differentiating the background and foreground information, some research also desires a clear background image in a complex scene with various moving objects for generation of special effects. For example, when visiting a popular tourist attraction point, people would like to have a photo without other tourists or moving objects [13]. Based on the above-mentioned reasons, there is a need to extract background information from a complex scene. Similar to [14] and [15], we develop a novel background extraction algorithm under the following assumptions.

1) The video sequence is taken by a static camera, i.e., all camera parameters including focal length, field of view, location, orientation, etc. are fixed.
2) Only foreground objects could be moving during the shooting period and static objects are part of the background.
3) Those background pixels must show up in at least one frame during the shooting sequence.

Furthermore, most algorithms make an extra assumption that there exist frames containing no foreground object or there are clean background shots in the beginning of the video. However, this assumption is nearly impossible for certain scenarios. For example, surveillance cameras shoot at a busy high way or the path way of a train station. Therefore, this paper focuses on solving these issues for acquisition of a clean and artifact-free background image from a video containing complex foreground occlusion and movement.

Traditional background extraction can be categorized into pixel-based, patch-based, and hybrid methods. Pixel-based methods [18] ignore spatial coherence to cause blurring and misdetection artifacts as shown in the first column of Fig. 1. Although patch-based methods [14], [19]–[21] improve pixel-based methods, shooting conditions could still affect correct background detection and result in blocking artifacts and seams as shown in the second column of Fig. 1. Chen and Aggarwal [16] combined advantages from both categories, but the algorithm may blur the results due to false selection of possible candidate pixels as shown in the third column of Fig. 1. Our system overcomes these issues by formulating background extraction as image fusion using general random walk (RW) with "harmonic weighting" to label each pixel from different frames, i.e., estimating the probability of a site (e.g., a pixel) being assigned a label based on known information. The proposed fusion algorithm has low computational complexity and produces a final fused background image with fine details and an optimal balance

Manuscript received July 6, 2016; revised October 15, 2016; accepted December 4, 2016. This work was supported by the Ministry of Science and Technology of Taiwan under Grant MOST 104-2221-E-011-091-MY2, Grant MOST 105-2622-E-011-012-CC3, Grant MOST 105-2628-E-001-003-MY3, Grant MOST 104-2221-E-011-029-MY3, Grant MOST 105-2218-E-011-005, and Grant MOST 105-2221-E-011-120-MY2. This paper was recommended by Associate Editor L. Shao.

K.-L. Hua, H.-C. Wang, C.-H. Yeh, and Y.-C. Lai are with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: hua@mail.ntust.edu.tw; d10115016@mail.ntust.edu.tw; m10315001@mail.ntust.edu.tw; yu-chi@mail.ntust.edu.tw).

W.-H. Cheng is with Academia Sinica, Taipei 115, Taiwan (e-mail: whcheng@citi.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2640288

2168-2267 © 2016 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 inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Fig. 1. Two exemplar results generated by a pixel-based median filter, a patch-based PBI method [14], and a hybrid method proposed by Chen and Aggarwal [16], and an image inpainting tool provided by Adobe Photoshop [17], respectively. The median filter method generally leaves blurring artifacts, the patch-based method generally has obvious seams along the block boundaries, the hybrid method still has blurring artifacts, and the inpainting method leaves blurring and incorrectly filling artifacts in the gap.

between two quality measures. In order to avoid blocking and blurring artifacts, we design two quality measures to consider coherence and distinctness among pixels. Furthermore, motion provides important hints about whether pixels belonging to a foreground object or not and thus, we design two different methods, temporal contrast estimation and optical-flow-based motionless patch extraction, to take the temporal coherence into consideration. When comparing the performance of each motion measure, we find that motionless patch extraction can almost remove all ghosting artifacts. Finally, when comparing to results generated by state-of-the-art algorithms, our approach can successfully extract clear and high-quality background images at low computational cost.

This paper makes the following contributions. First, we improve the fundamental tool for other computer vision research: precise background extraction from a video sequence with complex and occluded foreground object movements. Second, the core of our method is to formulate the problem as a probabilistic RW model to merge patches into a clean background image by considering contrast sensitivity, intensity consistency, and temporal coherence. Furthermore, we propose two methods to consider the temporal coherence among different frames including temporal contrast and motionless patch extraction. Finally, we also provide a mechanic to compute ground truth backgrounds from shooting frames for performance comparison among different algorithms. As demonstrated in the results, our system can extract a artifact-free background image from a complexly occluded video sequence while comparing to other algorithms.

The rest of this paper is organized as follows. Section II reviews those previous research done related to this paper. Sections III and IV present the details of our RW background extraction method. Section V shows the results of our algorithm and comparisons to other state-of-the-art methods. Section VI concludes with a discussion of limitations and future works.

II. Related Work

Background extraction is a research topic for computer vision and object detection and identification, and there are a large number of different research. In the following, we discuss those research directly related to ours.

A. Background Extraction

There are a large number of background extraction methods. For example, there are research in the Scene Background Modeling and Initialization (SBMI) workshop [22] aiming at background initialization, and Maddalena and Petrosino [23] provided a survey on background extraction with a static camera. For clarity, Toyama *et al.* [24] categorized background extraction into pixel-based and patch-based methods. Pixel-based methods process each pixel individually and independently. For example, a simple median filter can be applied for each pixel for background detection from values of different frames. Long and Yang [18] used the length of a pixel's consecutive period of similar intensity to determine its background pixel value. Yang *et al.* [25] proposed the Pixel-to-Model (P2M) paradigm and developed the P2M distance to indicate background pixels. Sobral *et al.* [26] provided a comparison of several matrix completion algorithms on the SBMI dataset [22]. Although pixel-based methods are computationally efficient, they ignored the spatial coherence and affection among other pixels. As a result, there are serious artifacts such as blurring and misdetection existing in the extracted background image when the scene contains complex movement and occlusion. Additionally, when foreground objects first stop and then move or move and stop, the detection may be problematic. Our algorithm uses color consistency measure to consider the relationship among neighboring pixels and find a balance using the global optimization of the RW framework.

Patch-based methods determine the background color of a pixel based on the information collected from a patch of pixels. Colombari and Fusiello [14] used sum of square distance (SSD) to group pixels in the temporal domain and apply visual grouping to spatially select proper object patches to composite the background. Reddy *et al.* [19], [20] also used SSD to group pixels in temporal domain but they select background colors based on features of frequency response instead of visual grouping. Ortego *et al.* [21] used motion filtering and dimensionality reduction trained by a set of labeled data to obtains a set of candidate blocks and spatially reconstruct the background image with these candidates. Although patch-based methods can improve the precision of background selection by using its neighboring information, shadows, reflections, fluctuation in lighting, and complex movement could still affect correct background detection. Additionally, when working on patches, all information is determined as a patch. Blocking artifacts and seams easily show up between two consecutive blocks. Our algorithm overcome these limitations by using temporal coherence and contrast consistency as measures.

Furthermore, Chen and Aggarwal [16] combined both patch-based and pixel-based concepts to average a set of selected pixels from neighboring regions and consecutive frames for the pixel's background color. However, the algorithm may still blur the extracted background due to false selection of possible candidate pixels. Therefore, we use "harmonic weighting" based on contrast sensitivity and color consistency to consider pixel coherence for avoidance of blurring and blocking artifacts.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

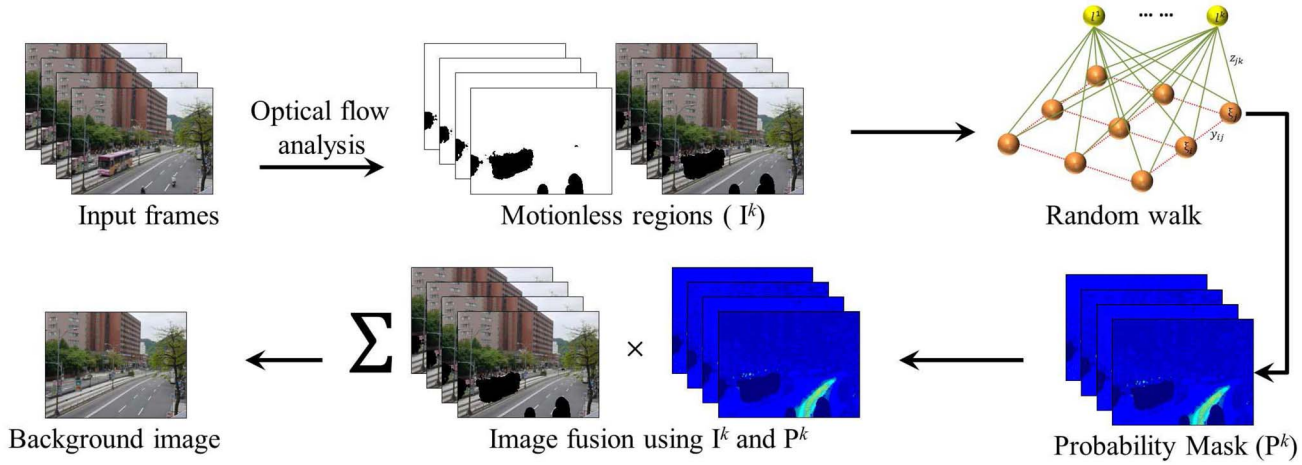


Fig. 2. Our system first analyzes a sequence of frames based on pixel-based motion analysis such as optical flow analysis. We incorporate these motion information into RW for the construction of probability maps. Our algorithm uses the probability maps to fuse the frames into a clean background image.

B. Background Extraction Datasets and Algorithmic Analysis

In order to examine the effectiveness of different background extraction algorithms, the SBMI workshop [22] provides the scene background initialization (SBI) dataset [27] for quantitative evaluation of the results. Furthermore, Maddalena and Petrosino [28] used the SBI dataset and proposed a common set of metrics to compare six background initialization methods. They give detailed analysis of the video contents and experimental results with these different methods. We analyze our algorithm with the SBI dataset and use Maddalenas six metrics to compare against other state-of-the-art algorithms.

Image inpainting such as [29] and [30] which is the process of reconstructing lost or deteriorated parts of images and videos can be used to repair those foreground-covered regions. Although the inpainted result might look real, the algorithm synthesizes the missing pixels from neighboring information which may deviate from the correct background information. The fourth column of Fig. 1 shows two examples. For each pixel, our algorithm select correct information from different frames to faithfully reconstruct the background image.

Image fusion merges pixels from different images to form a complete one and is employed in various applications. Different types of sensors such as computed tomography and magnetic resonance imaging or visible-band and infrared have different advantages themselves. Kumar and Dass [31] and Zheng *et al.* [32] combined information from multimodality sensors to preserve all relevant information. Because the optics of lenses have limited depth of field, it is often too difficult to make all relevant objects be in focus in an image. Therefore, Li *et al.* [33] and Zhao *et al.* [34] combined information from multifocus images to create a single image having objects in the scene be in focus. There are also research [35]–[38] merging details of the same scene revealed under different exposures to create images of a high dynamic range (HDR). Qin *et al.* [37] proposed patch-based exposure fusion to preserve the moving objects of dynamic scenes. Shen *et al.* [38] used the hybrid weight to produce a high quality image from different exposure images. Although researchers propose general fusion approaches [39], [40], none of them focuses on generating a clean background image. RW is widely applied in computer vision, such as image segmentation [41]–[43], video segmentation [44], visual saliency [45], visual tracking [46], and image fusion [47], [48]. Shen *et al.* [42] proposed lazy RW to obtain global probability maps and used a commute time strategy to segment weak boundaries and complicated texture regions. Dong *et al.* [43] proposed sub-Markov RW with a label prior to solve the segmentation problem of objects with thin and elongated parts. Liang *et al.* [44] proposed a novel video supervoxel generation algorithm using partially absorbing RWs and a built Laplacian optimization structure. This paper reformulates the generalized RW (GRW) framework proposed for image segmentation [41] and multiexposure merging for HDR images [48] to calculate an optimal set of probabilities subject to two quality measures, local and temporal contrast and color consistency. The local and temporal contrast measure preserves details in the extraction and considers the temporal relationship among frames; the color consistency measure considers consistency in a neighborhood and consistency with the natural scene. The proposed method can produce a final fused background image with fine details and an optimal balance between the two quality measures with low computational complexity.

III. Method

Although a video sequence of complex foreground occlusion and movement may not contain a clear shot of the background, there are still clean patches of the background. As shown in Fig. 2, we stitch these candidate patches together with probabilistic composition image fusion. For a static camera, we assume a static background, and thus, pixel motion analysis and estimation assist the elimination of foreground pixels. Therefore, we design two pixel motion estimation methods including temporal contrast analysis and motionless patch extraction derived from optical flow analysis.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Motion analysis takes the temporal coherence into consideration for the compatibility measure along with contrast consistency for distinct and static background analysis. Along with the spatial color consistency measure, our algorithm applies RW analysis to balance these two measures among pixels and regions for the construction of a probability map for all frames. We obtain the final probabilities by seeking a global optimization solution of the proposed GRWs. These probability maps serve as weights in the linear fusion process to produce the final fused image. The following sections explain the details of each step.

A. Problem Formulation

In this paper, we formulate background reconstruction as an image fusion problem of a probabilistic composition process. We denote a video sequence as $\mathcal{I} = \{\mathbf{I}^1, \ldots, \mathbf{I}^K\}$ where each frame has N pixels. We denote a set of corresponding labels as $\mathcal{L} = \{l^1, \ldots, l^K\}$ where a label, $l^k \in \mathcal{L}$, indicates the selection of $\mathbf{I}^k \in \mathcal{I}$. Our algorithm reconstruct a background image, $\mathbf{I}^*$, using probabilistic composition of input frames as

$$\mathbf{I}_i^* = \sum_{k=1}^{K} \mathbf{P}_i^k \mathbf{I}_i^k \tag{1}$$

where i is the pixel index with its image coordinate of $\overline{p}_i = (x_i, y_i)$, and $\mathbf{P}_i^k \triangleq \mathbf{P}(\xi_i = l^k|\mathcal{I})$, in which ξ_i is the ith variable coming from a defined corresponding label map Ξ, denotes the probability of a pixel being assigned the label of l^k given $\mathcal{I}$ with $\sum_{k=1}^{K} \mathbf{P}_i^k = 1$. This helps us transform background fusion to compute marginal probabilities when the given input video is subject to our designed quality measures for achieving an optimal balance among different considerations. Previously, patch-based algorithms [14], [19] choose a fixed probability for each patch of all frames to form the background based on local criteria, the blocking artifacts and seams are serious on the formed background image. Therefore, we propose a global optimal fusion algorithm based on GRW to relieve these artifacts. Experiments (Section V) show that the results generated by the proposed probabilistic fusion technique are better than those by other background extraction algorithms.

B. Random Walks for Background Extraction

RW is a statistical model and uses probabilistic random walking to form a series of tracks for determining spectacular connections among unseed and seed nodes. RW is first used for texture discrimination [49], algebraic potential determination [50], image segmentation [41]–[43], and multiexposure HDR fusion [48]. This paper follows the same concept to formulate the selection of pixels as an RW probability problem. According to past research [51], [52], to determine the probability of nodes achieved by RW is equivalent to find the solution of the corresponding Dirichlet question by setting those labeled seed nodes to unity and others to zero [53], [54], and we can estimate a probability map for each pixel from each frame by solving the Dirichlet problem. The following gives the details of the structure, relation, and solution.

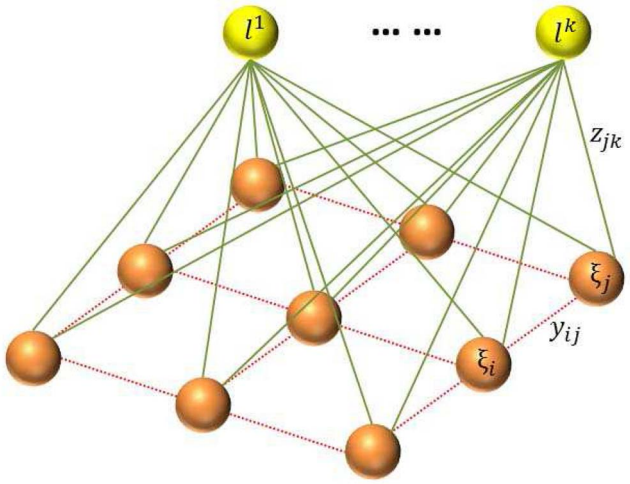


Fig. 3. Graph structure of the RW model used in our algorithm. Orange dots are variable nodes denoted as ξ_i, yellow nodes are label nodes denoted as l_i, red lines are variable edges denoted as y_{ij}, and green lines are label edges denoted as z_{jk}.

1) Frame Graph Representation: The RW model is built based on a graph structure, and thus, we first give the detailed definition of the frame graph used in this paper. As shown in Fig. 3, we present the relationship among all variables in the variable set Ξ and the label set $\mathcal{L}$ as a weighted undirected graph similar to the graph used by Shen *et al.* [48]. We construct the graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertices, $\mathcal{V} = \mathcal{L} \cup \Xi$, and edges, $\mathcal{E} = \mathcal{V} \times \mathcal{V}$. Our algorithm marks variable vertices with orange dots and label vertices with yellow dots. There are variable edges, $\mathcal{E}_\Xi$, connecting a variable vertex to its immediate four-connected variable vertices marked as red lines and label edges, $\mathcal{E}_\mathcal{L}$, connecting a variable vertex to all labels marked as green lines. We then define two measures: 1) $y_{ij} \triangleq y(\xi_i, \xi_j)$ for evaluating the spatial similarity of a variable edge connecting ξ_i and ξ_j in $\mathcal{E}_\Xi$ and 2) $z_{ik} \triangleq z(\xi_i, l^k)$ for evaluating the compatibility of a variable edge connecting ξ_i and l^k in $\mathcal{E}_\mathcal{L}$. The detailed definitions of spatial similarity y_{ij} and compatibility z_{ik} are shown in Sections III-C1 and III-C2.

2) Probability Computation From Dirichlet Solution: Let $\mathcal{V}$ be organized as follows: the first K nodes are the label nodes, i.e., $\mathcal{L} = \{v_1, \ldots, v_K\}$, and the last N nodes are variable nodes, i.e., $\Xi = \{v_{K+1}, \ldots, v_{K+N}\}$. Based on RW definition [53], we define the total energy for the system in Fig. 3 as

$$E = \frac{1}{2} \sum_{(v_i, v_j) \in \mathcal{E}} r_{ij} \big(u(v_i) - u\big(v_j\big)\big)^2 \tag{2}$$

where $u(\cdot)$ is the unknown potential defined in Ξ to minimize E, and r_{ij} is the compatibility function defined as

$$r_{ij} = \begin{cases} \tau_1 z_{i-K,j}, & v_i, v_j \in \mathcal{E}_\mathcal{L} \wedge v_j \in \mathcal{L} \\ \tau_2 y_{i-K,j-K}, & v_i, v_j \in \mathcal{E}_\Xi \end{cases} \tag{3}$$

where τ_1 and τ_2 are two user defined constants to balance the energy, and we choose 1.0 and 1.0, respectively, for the entire study. After selecting τ_1 and τ_2 and defining y_{ij} and z_{ik}, our algorithm solves u's by finding a harmonic function, $\nabla^2 u = 0$, in a similar manner as described in previous

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

research [41], [48]. We define a Laplacian matrix, $\mathbf{L}$, as

$$\mathbf{L}_{ij} = \begin{cases} d_i, & \text{if } i = j \\ -r_{ij}, & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $d_i = \sum r_{ij}$ is the degree of the ith vertex over all incident edges. We then rewrite the total energy as

$$E = \begin{bmatrix} u_{\mathcal{L}} \\ u_{\Xi} \end{bmatrix} \mathbf{L} \begin{bmatrix} u_{\mathcal{L}} \\ u_{\Xi} \end{bmatrix} \tag{5}$$

where $u_{\mathcal{L}} = [u(l_1), \ldots, u(l_K)^T]$ and $u_{\Xi} = [u(\xi_1), \ldots, u(\xi_N)^T]$. We solve the minimum energy solution by having $\nabla E = 0$ with respect to u_{Ξ}. After determining u's, we can determine the probability of each node as

$$\mathbf{P}^k(v_i) = \frac{u^k(v_i)}{\tau_3} \tag{6}$$

where $\tau_3 = \sum_{k=1}^{K} u(\xi_i)$ and $\xi_i \in \Xi$.

C. Quality Measure Function

Generally, the probability of a pixel reflects its affection in fusion, and it is determined by balancing different considerations formulated as quality measure functions, r_{ij}, which measure the selection quality of all frame pixels. This paper considers two aspects of data: spatial similarity and compatibility. The following gives the details of these measures.

1) Spatial Similarity: Since neighboring pixels have a highly chance from the same object with similar colors, we design spatial similarity or color consistency to maintain selection continuity and consistency among neighboring pixels using all three channels in the RGB color space and formulate it as

$$y_{ij} = e^{-\frac{\left\|\tilde{\mathbf{I}}_i - \tilde{\mathbf{I}}_j\right\|}{\overline{\sigma}_y}} \tag{7}$$

where i and j are two neighboring pixel indices, $\tilde{\mathbf{I}}_i = (1/K)\sum_{k=1}^{K} \mathbf{I}_i^k$ denotes the average pixel, $\mathbf{I}_i^k$ contains a color value, which is a triplet $(\mathrm{R}_i^k, \mathrm{G}_i^k, \mathrm{B}_i^k)$, $\|\cdot\|$ is Euclidean distance, and $\overline{\sigma}_y = \sigma_y/K$ is a user specific parameter and set as $\overline{\sigma}_y = 0.1$ in this paper. This term considers the spatial coherence of pixels and intends to choose similar labels for two neighboring pixels.

2) Compatibility: The input is a video sequence, object motion may induce blurring in addition to noises and artifacts. It is important to reconstruct distinct background, and distinctness can be indicated by local contrast which is usually related to variations in image luminance. The luminance value of each pixel is computed as $0.298R_i^k + 0.587G_i^k + 0.115B_i^k$.

Since contrast is usually related to variations in image luminance, and pixels shot at the same object should have similar colors, the local contrast measure should be biased toward pixels that provide less local luminance variations. In addition, since the background is motionless, the temporal variation in a pixel of corresponding frames should be small. As a result, in order to have high spatial contrast and distinctness and temporal coherence, this paper designs a compatibility term to select pixel labels as

$$z_{ik} = z_{ik}^S z_{ik}^T \tag{8}$$

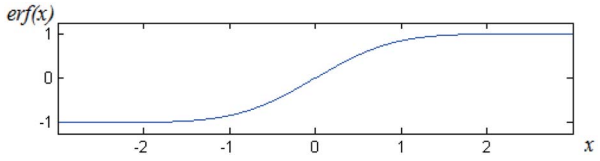


Fig. 4. Error function, Erf, which is a sigmoid function of $S(t) = [1/(1 + e^{-t})]$. It has an "S" shape to put more emphasis on the central part and less emphasis on both ends.

where i and k are the pixel and frame indices, respectively, the z^S term takes spatial distinctness into consideration, and the z^T term considers the temporal coherence. When considering the z^S term, we first denote the spatial contrast as c_i^k which is computed as the second-order spatial partial derivative of the luminance for $\mathbf{I}^k$ using the Marr–Hildreth [55] operator. The larger c_i^k is, the more contrast the pixel has. Furthermore, contrast may be perceived nonlinearly by humans, and thus, the estimated contrast is modulated by an error function defined in Fig. 4 to avoid overemphasis on high local variations. Furthermore, we also observe that when the intensity of a pixel is close to two extremities, i.e., zero and one, the information is less trustable, and the trustability can be emphasized with a Gaussian weighting. When taking these two factors into consideration, the spatial compatibility term is

$$z_{ik}^S = G\left(\mathbf{I}_i^k, \sigma\right)\left[\mathrm{Erf}\left(\frac{|c_i^k|}{\sigma_z}\right)\right] \tag{9}$$

where G is a Gaussian weighting function defined as $G(\mathbf{I}_i^k) = e^{-[(\mathbf{I}_i^k - 0.5)^2/2\sigma^2]}$, where σ is 0.2 in our experiments, Erf is a sigmoid function defined in Fig. 4, and σ_z is a weighting constant, and we set it to 1.0 in this paper.

Since we assume a static background, its corresponding temporal variation should be small. We denote the temporal variation as t_i^k computed as the forward difference of luminance in consecutive two frames

$$t_i^k = \left|\mathbf{I}_i^k - \mathbf{I}_i^{k+1}\right|. \tag{10}$$

The corresponding temporal compatibility term is designed as

$$z_{ik}^T = G\left(\mathbf{I}_i^k, \sigma\right)\left[\mathrm{Erf}\left(\frac{\sigma_z}{|t_i^k|}\right)\right]. \tag{11}$$

As shown in Section V, this temporal term can improve the blurring and ghosting artifacts. However, this temporal estimation does not really consider the motion relationship among pixels and thus, the results still contain ghosting artifacts. In order to further explore temporal coherence and motion relationship among pixels and objects, we design motionless patch extraction for better results whose details are given in Section IV.

D. Algorithm Summarization

Algorithm 1 shows the construction procedure of the probability maps. We first estimate the motion in each pixel based on temporal variation or optical flow analysis for taking the motionless background assumption into consideration. We then plug the estimated results into the compatibility estimation for labeling pixels according to spatial distinctness

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1 Background Extraction With RW Image Fusion

```
for each frame do
    Do temporal consistency analysis.
end for
for each pixel i in the frame do
    Compute color consistency, y_ij. (Eq. 7)
    Estimate spatial contrast sensitivity, z^S_ik. (Eq. 9)
    Estimate temporal variation, z^T_ik. (Eq. 11 or 18)
    Compute compatibility, z_ik. (Eq. 8)
    Compute compatibility function, r_ij. (Eq. 3)
    Construct the Laplacian matrix L. (Eq. 4)
end for
Solve u(L), u(Ξ), and P^k_i of all pixels in all frames.
Reconstruct the background image. (Eq. 1)
```

and temporal coherence. Furthermore, we consider the spatial coherence along with the compatibility for starting RWs for construction of probability maps. Finally, we use the probability maps to fuse the frames for the final result according to (1).

IV. Optical Flow Motionless Acceleration

Since we assume a background consisting of static and motionless objects through the whole video sequence, the corresponding background pixels should have no motion, and those pixels who contain obvious motion should be discarded. In the past, vision research identify moving objects using optical flow, and the results are generally efficient and robust. As a result, in order to reduce the computational cost and artifacts, we first use optical flow analysis [56], [57] to detect moving pixels and extract motionless patches from a video sequence by classifying pixels into two categories, motion and motionless. We determine the motion of a pixel by estimating temporal and spatial pixel intensity derivative from the consecutive frames in the neighboring region of the original pixel which matches the intensity of the original pixel as

$$\mathbf{I}(\overline{p}, t) = \mathbf{I}(\overline{p} + \Delta\overline{p}, t + \Delta t) \tag{12}$$

where $(\Delta\overline{p}, \Delta t)$ denotes the movement in the image plane and frame sequence, respectively. Furthermore, $\mathbf{I}(\overline{p} + \Delta\overline{p}, t + \Delta t)$ is approximated by its first degree of the Taylor series as

$$\begin{aligned} \mathbf{I}(\overline{p}, t) &= \mathbf{I}(\overline{p}, t) + \frac{\partial \mathbf{I}(\overline{p}, t)}{\partial \overline{p}} d\overline{p} + \frac{\partial \mathbf{I}(\overline{p}, t)}{\partial t} dt \\ 0 &= \frac{\partial \mathbf{I}(\overline{p}, t)}{\partial \overline{p}} d\overline{p} + \frac{\partial \mathbf{I}(\overline{p}, t)}{\partial t} dt \\ \frac{\partial \mathbf{I}(\overline{p}, t)}{\partial \overline{p}} \frac{\Delta\overline{p}}{\Delta t} &= -\frac{\partial \mathbf{I}(\overline{p}, t)}{\partial t} \end{aligned} \tag{13}$$

where $\vec{v} = (\Delta\overline{p}/\Delta t)$ denotes the motion vector at the pixel. We express (13) in the form of $\mathbf{A}\vec{v} = \vec{b}$ whose solution is estimated by finding the smallest motion vector, $\vec{v}$, of $\min(|\mathbf{A}\vec{v} - \vec{b}|^2)$, i.e., the optical flow vector.

Then, we use these optical flow vectors to determine the amount of information flowing in/out of a region as shown in Fig. 5. We estimate the amount of information flowing in/out

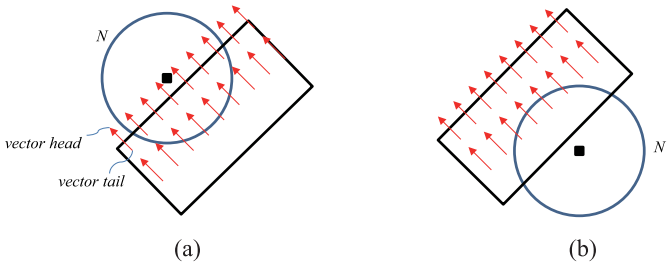


Fig. 5. Condition of information flowing (a) in when the optical flow vectors around a pixel point toward this pixel, and this should be equivalent to objects moving toward the pixel and (b) out when the optical flow vectors around a pixel point away from this pixel, and this should be equivalent to objects moving away from the pixel.

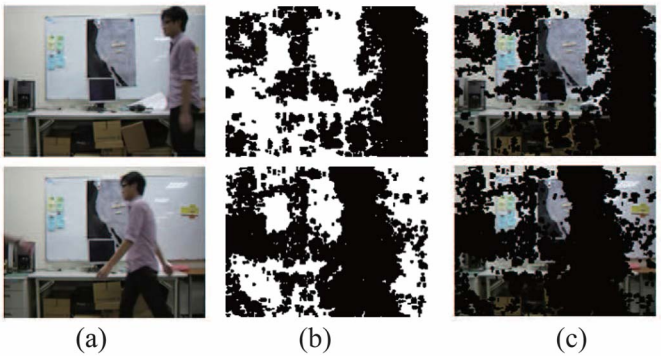


Fig. 6. Exemplar process of motionless patch extraction. (a) Two frames from a video sequence. (b) Corresponding masks constructed using the optical flow analysis described by (16). (c) Results of motionless patch extraction described by (17).

of a pixel by integrating flow vectors intersected with the destined circle, $\mathbf{N}$, centered at $\overline{p}_i$ of a radius of R [15]. We can then estimate the net flow at $\overline{p}_i$ for the kth frame using the Gaussian weighting scheme as

$$\mathbf{F}^k_{\text{in/out}}(\overline{p}_i) = \frac{1}{2\pi\sigma^2} \sum_{\overline{p} \in \mathbf{N}_{\text{in/out}}} e^{-\frac{1}{2}\frac{|\overline{p}-\overline{p}_i|^2}{\sigma^2}} \tag{14}$$

$$\mathbf{F}^k_{\text{net}}(\overline{p}_i) = \mathbf{F}^k_{\text{in}}(\overline{p}_i) - \mathbf{F}^k_{\text{out}}(\overline{p}_i) \tag{15}$$

where $\mathbf{N}_{\text{in/out}}$ denotes the neighboring pixels of the ith pixel whose motion vector crossing the boundary into/out of the virtual circle, respectively.

Fig. 5 shows two simple examples. If $\mathbf{F}^k_{\text{net}}(\overline{p}_i)$ is positive with a large magnitude, there is an object moving close to the point. If $\mathbf{F}^k_{\text{net}}(\overline{p}_i)$ is negative with a large magnitude, there is an object moving away from the point. Both situations indicate object movement, and the pixel should be classified as a moving pixel. When $\mathbf{F}^k_{\text{net}}(\overline{p}_i)$ is close to zero, the object movement at the pixel is small, and the pixel can be identified as a candidate background pixel. Therefore, if the absolute value of $\mathbf{F}^k_{\text{net}}(x, y)$ is lower than a user specified threshold T, we mark the pixel as motionless and assign the corresponding compatibility term to one (white). Otherwise, we assign it to zero (black). Our algorithm uses the mask to extract motionless patches as shown in Fig. 6

$$\mathbf{M}^k_i = \begin{cases} 1, & \text{if } |\mathbf{F}^k_{\text{net}}(\overline{p}_i)| < T \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$$\mathbf{I}_m^k = \mathbf{I}^k \mathbf{M}^k \tag{17}$$

$$z_{ik}^T = \mathbf{M}_i^k. \tag{18}$$

After setting the temporal compatibility function, our algorithm can achieve the image fusion more robustly and efficiently. In this paper, we select several parameters for all experiments. First, we set the radius of a circle to search nearby pixels to 10, and select the threshold to 250.

V. Experimental Results and Discussion

In order to evaluate the performance of our algorithm, we have constructed and collected several video sequences. All our test video sequences are available on the our demonstration website.[1] The following gives the details of the collection and construction process and comparisons to other background extraction algorithms.

A. Video Sequences and Ground Truth

There are two sources of our test videos. The first set of video sequences come from past research including a white board scene, a corridor scene, a room scene, a parking lot scene, a highway scene, another a corridor scene, another parking lot, and an eye checker scene collected from the website provided by Colombari and Fusiello [14] and the SBI dataset [27]. Because those video sequences used for evaluation in background extraction research are generally in form of low resolution, but currently, all surveillance cameras are in the full high-definition format, i.e., 1920×1080. In order to match the current video technology, we create and collect a set of video sequences in full HD format for the second set of video sequences. There are two categories of sequences, realistic and virtual. Computer graphics can easily create a 3-D virtual world, design moving patterns of foreground objects, and place the camera to construct 3-D animations. We design a movement pattern along with virtual background static objects or realistic skyboxes for testing different aspects. Since the camera in the virtual world uses the perfect pinhole model, a ground truth background can be created by shooting the virtual world after removing all foreground objects. Additionally, our algorithm mainly aims at extracting background from realistic videos shot at the world. Therefore, we examine the motion patterns in existing background extraction testing videos and mimic these patterns to create HD video sequences. Due to the length limitation, Fig. 7 only shows selected examples of both video sources, and the complete results can be found in the following Section V.

The SBI dataset [27] creates their own ground truths by manually selected pixels from video sequences and also provides six metrics to evaluate the accuracy of the result including average gray-level error (AGE), percentage of error pixels (pEPs), percentage of clustered error pixels (pCEPs), peak-signal-to-noise-ratio (PSNR), multiscale structural similarity index (MS-SSIM), and color image quality measure (CQM). The following sections use these metrics to compare our reconstructed results against others on the SBI sequences.

[1] http://graphics.csie.ntust.edu.tw/pub/BackgroundRW/main.html

Additionally, other research do not provide a ground truth background for each test case. We follow the same procedure as the SBI dataset to create the ground truth of these videos for using the same six error metrics for algorithm performance comparison. Because manual selection is strenuous, we would like to go one step further to have an automatic ground truth construction method and a performance comparison metric based on statistics. We propose a way to estimate the ground truth when collecting video sequences. Before having foreground objects moving inside each video sequence, we shot the background for a few seconds. Because the shooting period is short, it is reasonable to assume that the lighting variation is negligible. Since the CMOS response function, the valid shooting period, and other factors can be generally modeled with a Gaussian model, we assume that the shot color should have a Gaussian distribution. As a result, we take all color samples for a pixel from different frames to derive its Gaussian distribution to estimate its mean, μ_i, and standard deviation, $\sigma_{\mathbf{GT},i}$. We use these means of all pixels to form the ground truth and their deviation to compute the detection error rate. When two pixels come from the same Gaussian distribution (i.e., the pixel is correctly classified as background) with confidence α, they should be classified as success. Therefore, we adapt the metric proposed by Colombari and Fusiello [14] to estimate the number of miss-classified pixels as

$$\mathrm{Err}(\mathbf{GT}_i, \mathbf{B}_i) = \begin{cases} 1, & \text{if } \frac{1}{\sigma_{\mathbf{B},i}^2 + \sigma_{\mathbf{GT},i}^2} \|\mathbf{B}_i - \mathbf{GT}_i\|^2 < \chi_3^{-1}(\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where **GT** and **B** are the ground truth and reconstructed background, respectively.

B. Analysis of Free Parameters

Since different parameters may affect the result of reconstruction, we must choose a proper set for image fusion. This section discusses the influences of different parameters on the image fusion accuracy based on the reconstructed results of sequences from the SBI dataset. The accuracy is evaluated based on PSNR. There are five parameters, σ, $\overline{\sigma}_y$, σ_z, τ_1, and τ_2 in our algorithm. The default values of these parameters are set as $(\sigma, \overline{\sigma}_y, \sigma_z, \tau_1, \tau_2) = (0.2, 0.1, 1.0, 1.0, 1.0)$. When evaluating one parameter, the values of the other four parameters are set to the default values. For example, when analyzing the influence of σ, others parameters are fixed as $\overline{\sigma}_y = 0.1$, $\sigma_z = 1.0$, $\tau_1 = 1.0$, and $\tau_2 = 1.0$. We summarize the evaluation in Fig. 8. Note that as shown in Fig. 8(c), when σ_z is more than 0.2, the performance can be better. Additionally, σ_z would cause a considerable performance degradation (1.5 dB) when its value is equal to 0.1, but maintains stable good performance for other values. We can have a high PSNR value while the σ is between the 0.1 and 0.4. The values of $\overline{\sigma}_y$, τ_1, and τ_2 are similar to the ones used by Shen *et al.* [48]. Fig. 8(b), (d), and (e) shows that $\overline{\sigma}_y$, τ_1, and τ_2 have little influences on the performance of image fusion. Based on Fig. 8, we know that all the resultant PSNRs achieved at least 36 through various parameter settings, these experiments verified

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

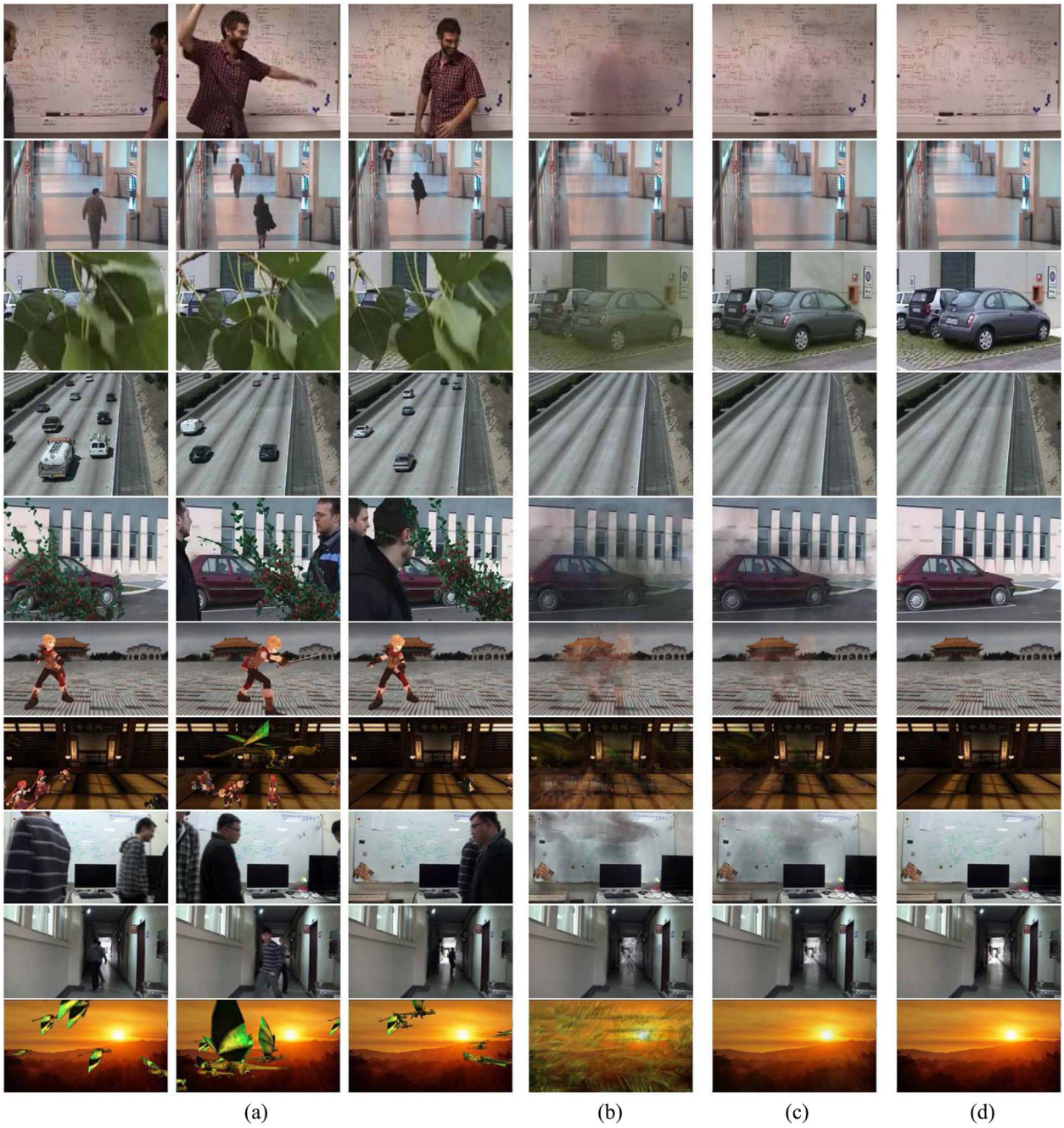


Fig. 7. Selected results of extracted background from video sequences used in the past research and our collected HD video sequences when using our algorithm without temporal consideration, with temporal contrast filtering, and with optical flow motion estimation. From top to bottom, these results are of a white board scene, a corridor scene, a parking lot scene, a highway scene, another parking lot, Shuan Kai-Sheng Memorial Hall, a virtual hall scene, a white board scene, a hall scene, and a virtual sunset. (a) Input. (b) Ours Navie. (c) Ours temporal. (d) Ours optical.

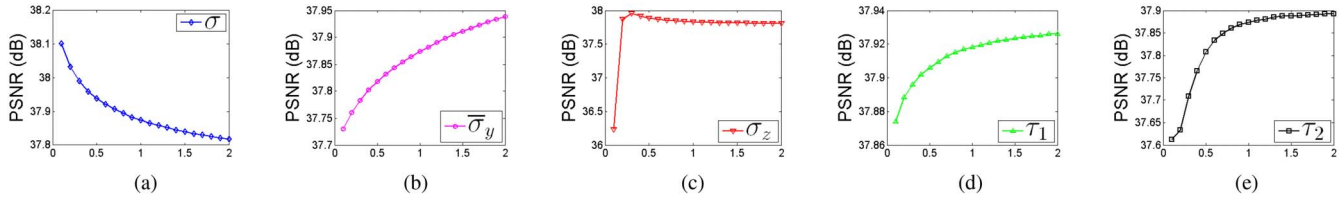


Fig. 8. Influences of different parameters, including (a) σ, (b) $\overline{\sigma}_y$, (c) σ_z, (d) τ_1, and (e) τ_2, on the image fusion based on the indication of PSNR.

the robustness of these parameters. Therefore, this paper sets them as $\sigma = 0.2$, $\overline{\sigma}_y = 0.1$, $\sigma_z = 1.0$, $\tau_1 = 1.0$, and $\tau_2 = 1.0$ for good reconstructed background results.

C. Results and Comparisons

After collecting the test video sequences, we first analyze the performance of our proposed temporal

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

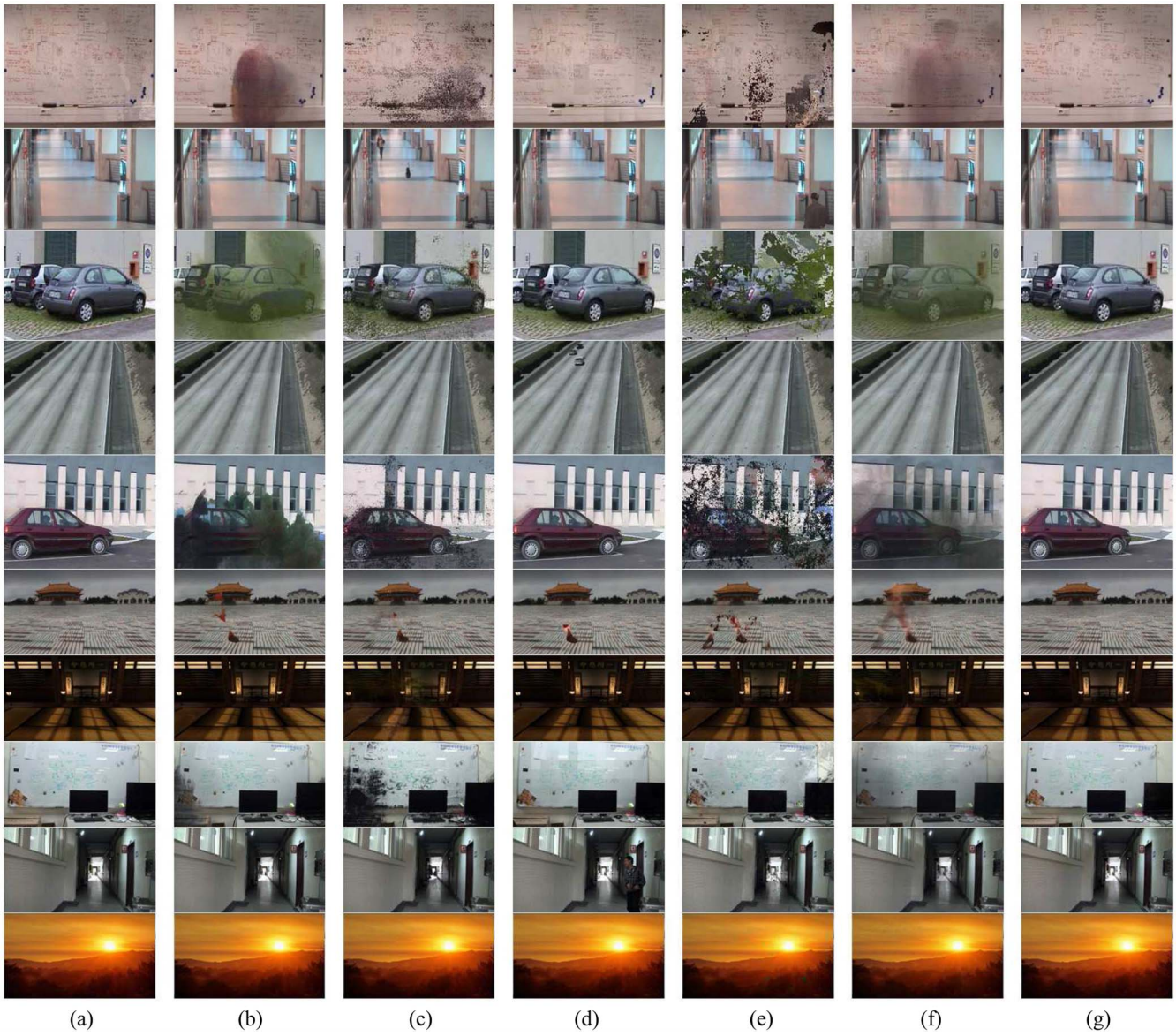


Fig. 9. Results of extracted background from video sequences used in the past research and our collected HD video sequences when using the pixel-based median filter, P2M [25], PBI [14], hybrid method [16], GMM [58], and our algorithm with optical flow motion estimation. From top to bottom, these results are of a white board scene, a corridor scene, a parking lot scene, a highway scene, another parking lot, Shuan Kai-Sheng Memorial Hall, a virtual hall scene, a white board scene, a hall scene, and a virtual sunset. (a) Ground truth. (b) Medium. (c) P2M. (d) PBI. (e) Hybrid. (f) GMM. (g) Ours optical.

coherence estimation. In this paper, we develop a temporal contrast filter and optical motion estimator to consider the temporal relationship among corresponding pixels for better results. As shown in Fig. 7, when applying RW fusion without considering the temporal relationship, ghosting artifacts are serious because all information are equally taken into consideration. Because the temporal contrast filter can consider the temporal coherence among corresponding pixels, the ghosting artifacts are relieved but cannot be totally removed. Optical flow motion estimation conservatively removes those possible moving pixels and can totally remove all ghosting artifacts. Furthermore, we have compared the extracted results with those generated by the pixel-based Median method, the P2M method [25], the hybrid method proposed by Chen and Aggarwal [16], the patch-based PBI method [14] and the Gaussian modeling method (GMM) [58] to further understand the effectiveness and performance of our RW background reconstructing method as shown in Fig. 9. Table I shows the comparison statistics while adapting the six metrics proposed in the SBI dataset. Our algorithm generally outperform the other algorithms in most cases based on the computed PSNR and MS-SSIM. Based on the numerically analyzed results of these metrics, our reconstructed backgrounds are very close to the respective ground truths. In addition to the above subjective and objective metrics, we also evaluate the computational cost by estimating the running time on a desktop computer equipped with an Intel Quad-Core i7 3.40 GHz CPU and 64 GB memory for all methods. As shown in Table I, the average running time of P2M [25], hybrid [16], PBI [14], GMM [58],

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
P2M DENOTES THE PIXEL-TO-MODEL METHOD [25]. PBI DENOTES THE PATCH-BASED METHOD [14]. HYBRID DENOTES THE HYBRID METHOD [16]. GMM STANDS FOR THE GMM METHOD [58]. OURS OPTICAL STANDS FOR OUR RW EXTRACTION METHOD WITH THE 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 MULTISCALE 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 Optical	**4.3391**	**0.8659**%	**0.2256**%	**0.9609**	**32.6254**	**50.6679**	173.6
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	**51.578**	**42.1**
			Ours Optical	2.5621	**0.2614**%	**0.1648**%	**0.9927**	**37.1089**	51.023	1462.6
CAVIAR2	384×288	460	P2M	0.9530	**0.0020**%	**0.0000**%	**0.9991**	**45.5059**	56.6271	2531.3
			PBI	1.4407	0.1923%	0.1180%	0.9968	39.7233	52.3532	40687.0
			Hybrid	1.4267	0.0407%	0.0142%	0.9970	41.4530	52.6182	1840.7
			GMM	**0.8052**	0.0356%	0.0020%	0.9988	44.3408	**59.7584**	**31.1**
			Ours Optical	0.8102	0.0142%	**0.0000**%	**0.9991**	45.1811	59.6862	1098.3
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**	**44.309**	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 Optical	2.0275	0.0347%	**0.0000**%	**0.9972**	39.3791	43.4583	267.8
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 Optical	**2.0158**	**0.2865**%	**0.0000**%	**0.9950**	**37.3437**	**47.0328**	924.8
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	40.479	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 Optical	**2.6030**	**0.0729**%	**0.0000**%	**0.9955**	**37.1408**	**49.4737**	162.5
People & Foliage	320×240	341	P2M	10.1301	14.9583%	2.4297%	0.8744	21.0926	28.4846	1503.7
			PBI	**1.738**	**0.0013**%	**0.0000**%	**0.9974**	**41.0196**	**47.7845**	6150.7
			Hybrid	33.4551	33.3893%	10.9766%	0.6015	12.3767	22.9357	15045.3
			GMM	32.069	57.4154%	49.4714%	0.7578	15.6736	31.6765	**18.9**
			Ours Optical	1.8231	0.0156%	**0.0000**%	0.9962	40.2077	46.5935	639.4
Snellen	146×150	321	P2M	32.1229	37.0756%	15.3887%	0.6164	13.8451	26.5883	409.7
			PBI	2.6997	**0.0000**%	**0.0000**%	0.9951	37.008	47.0115	1908.2
			Hybrid	4.4097	1.8663%	0.2894%	0.9381	25.3528	35.4596	6944.8
			GMM	40.3353	69.4252%	63.0546%	0.7546	14.5997	35.3227	**4.9**
			Ours Optical	**2.0174**	**0.0000**%	**0.0000**%	**0.9970**	**39.5654**	**47.9029**	163.1
Average			P2M	9.9508	11.6333%	3.6419%	0.8638	26.5313	39.0608	1426.9
			PBI	3.2889	1.0815 %	0.3095%	0.9774	35.0903	46.0899	38805.5
			Hybrid	11.4073	10.9187%	4.3413%	0.8375	26.0688	36.5068	7792.9
			GMM	16.5584	28.7072%	23.5545%	0.8418	26.7327	43.1504	**17.6**
			Ours Optical	**2.2748**	**0.1939**%	**0.0488**%	**0.9917**	**38.5690**	**49.4798**	611.5

TABLE II
P2M DENOTES THE PIXEL-TO-MODEL METHOD [25]. PBI DENOTES THE PATCH-BASED METHOD [14]. HYBRID DENOTES THE HYBRID METHOD [16]. GMM STANDS FOR THE GMM METHOD [58]. OURS TEMPORAL STANDS FOR OUR RW EXTRACTION METHOD WITH THE TEMPORAL VARIATION MODEL. OURS OPTICAL STANDS FOR OUR RW EXTRACTION METHOD WITH THE 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 Temporal		Ours Optical	
		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	30.7490	1.68e-1	**42.6540**	**3.80e-3**
HD2	196	35.7177	7.80e-3	32.0908	6.50e-3	23.6174	5.00e-2	32.3165	2.02e-1	21.3701	4.41e-1	**39.6056**	**1.00e-4**
HD3	87	30.3679	1.01e-1	49.5452	1.00e-3	35.2130	4.19e-2	33.2084	1.01e-1	29.3827	2.42e-1	**53.7193**	**1.00e-4**
HD4	135	24.2946	1.99e-1	13.6051	2.49e-1	20.7819	2.52e-1	24.7827	3.89e-1	22.2489	3.50e-1	**35.2984**	**5.12e-2**
HD5	154	23.1159	2.90e-1	27.9751	2.65e-1	27.0693	3.22e-1	27.3665	2.73e-1	25.9273	2.43e-1	**28.4660**	**1.11e-1**
HD6	152	14.8634	3.55e-1	28.2896	4.08e-1	23.1306	3.18e-1	17.2153	5.37e-1	19.0189	4.87e-1	**33.5139**	**2.12e-2**
HD7	155	26.2205	5.69e-2	23.8395	1.07e-1	25.3421	3.26e-1	28.6920	6.02e-2	29.8877	6.79e-2	**31.8903**	**3.53e-2**
HD8	155	41.5755	3.00e-3	42.1136	2.00e-3	39.9098	3.55e-2	**42.9385**	**8.00e-4**	29.7988	1.05e-2	42.7282	1.00e-3
HD9	130	43.9635	9.96e-4	46.0475	**2.31e-5**	42.0886	2.28e-2	41.5060	1.60e-3	36.1047	3.70e-3	**46.3560**	1.00e-4
Average		30.0759	1.17e-1	32.7026	1.17e-1	29.2387	1.57e-1	30.7169	1.86e-1	27.1653	2.24e-1	**39.3591**	**2.49e-2**

and the proposed method per test video is approximately 1426.9, 38805.5, 7792.9, 17.6, and 611.5 s, respectively. Our proposed method is quite efficient and can reconstruct pretty good results based on objective measures. Table II shows the error rate based on (19) when comparing to other algorithms. We can see the optical flow motion estimation better than

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) (b)

Fig. 10. Two exemplar reconstruction artifacts. (a) Ground truths of the scene. (b) Reconstruction results of our algorithm. Top: highway scene. Bottom: white board scene. Because the bushes have small local movements which can be identified by our optical-flow-based motionless patch extractor, the locally moving patches from different frames are fused to generate little blurring artifacts. Furthermore, although the shooting camera is static, there are uncertainties in the pixel capturing process and these uncertainties may result in blurring artifacts.

temporal contrast filter. In some cases, the temporal contrast filter can get the great result, i.e., sequence HD9 because the foreground objects are moving fast and temporal contrast filter can estimate motion pixels easy. But in most sequences, the temporal contrast filter cannot totally remove the artifacts and the results worse other algorithms. Therefore, our optical can remove all artifacts and generally perform better than the P2M, PBI, hybrid, and GMM methods.

While we carefully examine the reconstruction results, we find a few blurring artifacts which are hard to identify without zooming-in examination as shown in Fig. 10. Since we use the probability maps constructed by the RW method to fuse the final background image, the synthesized results can be disturbed to induce blurring artifacts while there are failures of the algorithmic assumptions. We find the following two common cases. First, we assume that the background objects are static, while there are background objects with small local movements, but movements are so little to be detected by our optical-flow-based motionless patch extract, the locally moving patches from different frames are fused to generate little blurring artifacts. Second, we assume a static camera, but there are uncertainties in the pixel capturing process, and these uncertainties would result the differences for different frames and cause blurring artifacts.

VI. Conclusion

Background extraction is very important for many research, and this paper proposes a background extraction method to formulate background extraction as a mathematical probability problem based on the RW theory. As shown in results, our system can successfully extract clean background images when inputs contained messy occlusions without a clean shot of the background in the entire sequence. When comparing to results generated by other algorithms, ours have better visual quality. In addition, we also propose a method to generate a ground truth background image for numerical comparison, and our results are still better. However, our system is not without limitations, and there are a few future research directions. First, we assume a static captured camera, and camera motion brings issues for our extraction. For example, our algorithm does not work on extracting from video sequences taken by a camera mounting on a plane. We would like to apply the structure from motion algorithm to track the camera motion and find the temporal relationship among frames to extract the static background. Second, although optical flow analysis can identify local motionless patches for better background extraction, it does not take all frames into consideration. We would like to design a global motionless patch extraction for better results.

References

[1] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.

[2] R. Kapoor and A. Dhamija, "Fast tracking algorithm using modified potential function," *IET Comput. Vis.*, vol. 6, no. 2, pp. 111–120, Mar. 2012.

[3] S.-C. Huang and B.-H. Do, "Radial basis function based neural network for motion detection in dynamic scenes," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 114–125, Jan. 2014.

[4] P. Chiranjeevi and S. Sengupta, "Detection of moving objects using multi-channel kernel fuzzy correlogram based background subtraction," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 870–881, Jun. 2014.

[5] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[6] N. Liu, H. Wu, and L. Lin, "Hierarchical ensemble of background models for PTZ-based video surveillance," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 89–102, Jan. 2015.

[7] L. Lin, Y. Lu, C. Li, H. Cheng, and W. Zuo, "Detection-free multiobject tracking by reconfigurable inference with bundle representations," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2447–2458, Nov. 2016.

[8] *Change Detection Workshop*. Accessed on Jun. 30, 2016. [Online]. Available: http://changedetection.net/

[9] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Providence, RI, USA, 2012, pp. 1–8.

[10] Y. Wang *et al.*, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Columbus, OH, USA, 2014, pp. 387–394.

[11] P. Chiranjeevi and S. Sengupta, "Interval-valued model level fuzzy aggregation-based background subtraction," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2585600.

[12] X. Cao, L. Yang, and X. Guo, "Total variation regularized RPCA for irregularly moving object detection under dynamic background," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 1014–1027, Apr. 2016.

[13] W.-H. Cheng, C.-W. Hsieh, S.-K. Lin, C.-W. Wang, and J.-L. Wu, "Robust algorithm for exemplar-based image inpainting," in *Proc. Int. Conf. Comput. Graph. Imag. Vis.*, Beijing, China, 2005, pp. 64–69.

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

[15] D. Gutchess, M. Trajkovics, E. Cohen-Solal, D. Lyons, and A. K. Jain, "A background model initialization algorithm for video surveillance," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1. Vancouver, BC, Canada, 2001, pp. 733–740.

[16] C.-C. Chen and J. K. Aggarwal, "An adaptive background model initialization algorithm with objects moving at different depths," in *Proc. IEEE Int. Conf. Image Process.*, San Diego, CA, USA, 2008, pp. 2664–2667.

[17] T. Georgiev, "Photoshop healing brush: A tool for seamless cloning," in *Proc. Workshop Appl. Comput. Vis. (ECCV)*, Nicosia, Cyprus, 2004, pp. 1–8.

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

[19] V. Reddy, C. Sanderson, and B. C. Lovell, "An efficient and robust sequential algorithm for background estimation in video surveillance," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, 2009, pp. 1109–1112.

[20] V. Reddy, C. Sanderson, and B. C. Lovell, "A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts," *J. Image Video Process.*, vol. 2011, no. 1, 2010, Art. no. 164956. [Online]. Available: http://jivp.eurasipjournals.springeropen.com/articles/10.1155/2011/164956

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[21] D. Ortego, J. C. SanMiguel, and J. M. Martínez, "Rejection based multipath reconstruction for background estimation in video sequences with stationary objects," *Comput. Vis. Image Understand.*, vol. 147, pp. 23–37, Jun. 2016.

[22] *Scene Background Modeling and Initialization*. Accessed on Jun. 30, 2016. [Online]. Available: http://sbmi2015.na.icar.cnr.it/

[23] L. Maddalena and A. Petrosino, "Background model initialization for static cameras," in *Background Modeling and Foreground Detection for Video Surveillance*. Boca Raton, FL, USA: Chapman and Hall, 2014, pp. 1–16.

[24] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1. 1999, pp. 255–261.

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

[26] A. Sobral, T. Bouwmans, and E. Zahzah, "Comparison of matrix completion algorithms for background initialization in videos," in *Proc. SBMI Workshop Conjunction ICIAP*, Genoa, Italy, 2015, pp. 510–518.

[27] *Scene Background Initialization (SBI) Dataset*. Accessed on Jun. 30, 2016. [Online]. Available: http://sbmi2015.na.icar.cnr.it/SBIdataset.html

[28] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *Proc. SBMI Workshop Conjunction ICIAP*, Genoa, Italy, 2015, pp. 469–476.

[29] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[30] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[31] M. Kumar and S. Dass, "A total variation-based algorithm for pixel-level image fusion," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2137–2143, Sep. 2009.

[32] S. Zheng, W.-Z. Shi, J. Liu, G.-X. Zhu, and J.-W. Tian, "Multisource image fusion method using support value transform," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1831–1839, Jul. 2007.

[33] S. Li, J. T.-Y. Kwok, I. W.-H. Tsang, and Y. Wang, "Fusing images with different focuses using support vector machines," *IEEE Trans. Neural Netw.*, vol. 15, no. 6, pp. 1555–1561, Nov. 2004.

[34] H. Zhao, Q. Li, and H. Feng, "Multi-focus color image fusion in the HSI space using the sum-modified-Laplacian and a coarse edge map," *Image Vis. Comput.*, vol. 26, no. 9, pp. 1285–1295, 2008.

[35] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Proc. Comput. Graph. Appl.*, 2007, pp. 382–390.

[36] L. Bogoni and M. Hansen, "Pattern-selective color image fusion," *Pattern Recognit.*, vol. 34, no. 8, pp. 1515–1526, 2001.

[37] X. Qin, J. Shen, X. Mao, X. Li, and Y. Jia, "Robust match fusion using optimization," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1549–1560, Aug. 2015.

[38] J. Shen, Y. Zhao, S. Yan, and X. Li, "Exposure fusion using boosting Laplacian pyramid," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1579–1590, Sep. 2014.

[39] G. Piella, "Image fusion for enhanced visualization: A variational approach," *Int. J. Comput. Vis.*, vol. 83, no. 1, pp. 1–11, 2009.

[40] V. S. Petrovic and C. S. Xydeas, "Gradient-based multiresolution image fusion," *IEEE Trans. Image Process.*, vol. 13, no. 2, pp. 228–237, Feb. 2004.

[41] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[42] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for super-pixel segmentation," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1451–1462, Apr. 2014.

[43] X. Dong, J. Shen, L. Shao, and L. V. Gool, "Sub-Markov random walk for image segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 516–527, Feb. 2016.

[44] Y. Liang, J. Shen, X. Dong, H. Sun, and X. Li, "Video supervoxels using partially absorbing random walks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 928–938, May 2016.

[45] J.-G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1661–1672, Sep. 2014.

[46] X. Li, Z. Han, L. Wang, and H. Lu, "Visual tracking via random walks on graph model," *IEEE Trans. Cybern.*, vol. 46, no. 9, pp. 2144–2155, Sep. 2016.

[47] K.-L. Hua, H.-C. Wang, A. H. Rusdi, and S.-Y. Jiang, "A novel multi-focus image fusion algorithm based on random walks," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 951–962, 2014.

[48] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3634–3646, Dec. 2011.

[49] H. Wechsler and M. Kidode, "A random walk procedure for texture discrimination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 3, pp. 272–280, Jul. 1979.

[50] N. Biggs, "Algebraic potential theory on graphs," *Bull. London Math. Soc.*, vol. 26, no. 6, pp. 641–682, 1997.

[51] S. Kakutani, "Markov processes and the Dirichlet problem," *Proc. Jpn. Acad.*, vol. 21, pp. 227–233, 1945.

[52] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*. Washington, DC, USA: Math. Assoc. America, 1984.

[53] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. New York, NY, USA: Wiley, 1989.

[54] R. Hersh and R. J. Griego, "Brownian motion and potential theory," *Sci. Amer.*, vol. 220, no. 3, pp. 67–74, 1969.

[55] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. Roy. Soc.*, vol. 207, no. 1167, pp. 187–217, 1980.

[56] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 2. Vancouver, BC, Canada, pp. 674–679, 1981.

[57] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

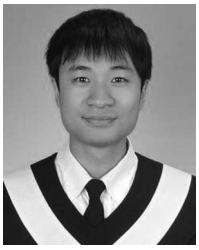
[58] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Fort Collins, CO, USA, 1999, pp. 246–252.



Kai-Lung Hua (M'12) received the B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2000, and the M.S. degree in communication engineering from National Chiao Tung University, Hsinchu, in 2002, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, in 2010.

He has been with the National Taiwan University of Science and Technology, Taipei, Taiwan, since 2010, where he is currently an Associate Professor with the Department of Computer Science and Information Engineering. His current research interests include digital image and video processing, computer vision, and machine learning.

Dr. Hua was a recipient of several research awards, including the Top 10% Paper Award of the 17th IEEE International Workshop on Multimedia Signal Processing, the Second Prize Award of the 2014 ACM Multimedia Grand Challenge, the Best Paper Award of the 2013 IEEE International Symposium on Consumer Electronics, the Best Poster Paper Award of the 2012 International Conference on 3D Systems and Applications, and the MediaTek Doctoral Fellowship. He is a member of Eta Kappa Nu and Phi Tau Phi.



Hong-Cyuan Wang received the M.S. degree from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2013, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering.

He joined the Multimedia Computing Laboratory, 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.



Chih-Hsiang Yeh received the B.S. degree from the Computer Science and Information Engineering Department, National Taiwan University of Science and Technology, Taipei, Taiwan, in 2014, where he is currently pursuing the M.S. degree.

His current research interests include graphics, vision, and mobile game, and application design.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Wen-Huang Cheng (SM'05) received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2002 and 2004, respectively, and the Ph.D. (Hons.) degree from the Graduate Institute of Networking and Multimedia, National Taiwan University, in 2008.

He is an Associate Research Fellow (Associate Professor) with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, where he is the Founding Leader with the Multimedia Computing Laboratory, and an Associate Research Fellow with a joint appointment in the Institute of Information Science. Before joining Academia Sinica, he was a Principal Researcher with MagicLabs, HTC Corporation, Taoyuan, Taiwan, from 2009 to 2010. His current research interests include multimedia content analysis, multimedia big data, deep learning, computer vision, mobile multimedia computing, social media, and human–computer interaction.

Dr. Cheng was a recipient of numerous research awards, including the 2016 Y. Z. Hsu Scientific Paper Award, the 2015–2016 Presidential Achievement Award of Rotary International, the Outstanding Youth Electrical Engineer Award from the Chinese Institute of Electrical Engineering in 2015, the Top 10% Paper Award from the 2015 IEEE International Workshop on Multimedia Signal Processing, the Outstanding Reviewer Award from the 2015 ACM International Conference on Internet Multimedia Computing and Service, the Prize Award of Multimedia Grand Challenge from the 2014 ACM Multimedia Conference, the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014, the Outstanding Young Scholar Awards from the Ministry of Science and Technology in 2012 and 2014, the Outstanding Social Youth of Taipei Municipal in 2014, the Best Reviewer Award from the 2013 Pacific-Rim Conference on Multimedia, and the Best Poster Paper Award from the 2012 International Conference on 3D Systems and Applications. He is an APSIPA Distinguished Lecturer.



Yu-Chi Lai (M'13) received the B.S. degree from the Electrical Engineering Department, National Taiwan University, Taipei, Taiwan, in 1996, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2003 and 2009, respectively.

He is currently an Associate Professor with the National Taiwan University of Science and Technology, Taipei, Taiwan. His current research interests include graphics, vision, and multimedia.