Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 981724, 10 pages
http://dx.doi.org/10.1155/2014/981724

Research Article

Dynamic Scene Stitching Driven by Visual Cognition Model

Li-hui Zou,[1,2] Dezheng Zhang,[1,2] and Aziguli Wulamu[1,2]

[1] *School of Computer and Communication Engineering, University of Science and Technology, Beijing 100083, China*
[2] *Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China*

Correspondence should be addressed to Aziguli Wulamu; ali@bsw.gov.cn

Received 29 August 2013; Accepted 2 December 2013; Published 3 February 2014

Academic Editors: J. Shu and F. Yu

Dynamic scene stitching still has a great challenge in maintaining the global key information without missing or deforming if multiple motion interferences exist in the image acquisition system. Object clips, motion blurs, or other synthetic defects easily occur in the final stitching image. In our research work, we proceed from human visual cognitive mechanism and construct a hybrid-saliency-based cognitive model to automatically guide the video volume stitching. The model consists of three elements of different visual stimuli, that is, intensity, edge contour, and scene depth saliencies. Combined with the manifold-based mosaicing framework, dynamic scene stitching is formulated as a cut path optimization problem in a constructed space-time graph. The cutting energy function for column width selections is defined according to the proposed visual cognition model. The optimum cut path can minimize the cognitive saliency difference throughout the whole video volume. The experimental results show that it can effectively avoid synthetic defects caused by different motion interferences and summarize the key contents of the scene without loss. The proposed method gives full play to the role of human visual cognitive mechanism for the stitching. It is of high practical value to environmental surveillance and other applications.

1. Introduction

Wide field of view (FOV) is demanded in many application domains, such as intelligent transportation, military defense, and civil security. A larger scope of image information is beneficial for improving the reliability and the safety of the system. However, the FOV of an ordinary camera is usually much smaller than that of humans due to the limitations of the fabrication process of enlarging the sensor size. Image stitching technology supplies an effective solution for breaking the limitation of the camera FOV, which is getting more and more attentions of researches. It is to align a sequence of overlapping images and blend the overlapping regions to form a seamless wide FOV image. The techniques nowadays can be summarized into two mainstreams: one is represented by Szeliski who proposed the classical stitching model based on geometric relationships of camera motions [1], and the other is represented by Peleg et al. who proposed an improved adaptive manifold mosaicing model [2]. The former one is to extract the geometric transform between partly overlapped adjacent images for image registration and fusion [3–5]. This model is deemed as the foundation of image alignment and stitching research, handling many camera motions, that is, translation, rotation, affine, projective motion, and so forth. The latter one is to cut narrow strips which are perpendicular to optical flow from high-overlapping images and paste their warped strips whose optical flows become parallel to the camera motion direction to form the output manifold mosaics adaptively. Such stitching model breaks through the restriction of camera motions and promotes the development of image stitching, becoming a new research focus [6–9].

Both of the above two categories of image stitching algorithms address the registration and the blending processes on pixel levels, ignoring the visual perception mechanism of humans and the relations among image contents. Sometimes the algorithms cannot guarantee the integrity of interesting contents, especially when the camera capturing platform moves and the scanned scene contains multidimensional moving objects. Some potential stitching defects, for example, object clipping, motion blurring, or ghosting, caused by moving objects and scene movement as well as parallax, easily appear in the final mosaic image. There is still a practical challenge in such dynamic scene stitching.

Image is a kind of nonstructural perception information. Comparing to direct pixel operations, how to cooperate with human cognitive mechanism and relative mathematic models to construct new computational models and methods for image processing is a meaningful and necessary work. It is helpful for improving the process efficiency and further comprehensive understanding if considering the guide role of visual cognitive mechanism as much as possible. In this paper, we address the dynamic scene stitching from the visual cognition point of view, and an effective stitching approach driven by hybrid-saliency-based visual cognition model for dynamic video sequence is proposed to avoid synthetic defects caused by different movements of the scene. It considers human visual perception mechanism first. And a cognition model is proposed, consisting of multiple visual stimuli, that is, intensity, edge contour, and scene depth saliencies of the input frames. Moreover, under the manifold mosaicing framework, the stitching process is formulated as a cut path optimization problem in a constructed space-time graph from the original input video volume. The proposed cognitive model constrains the cutting energy function for column width selections during the manifold synthesis. The effectiveness of the idea, introducing visual cognitive mechanism into the image stitching process, is verified by the experiments. The key salient contents of the wide FOV scene can be summarized without missing or deforming. The algorithm is conducive to support further analysis of global situations within the wide FOV, and it could provide a concrete reference and some inspiration to other problems in image processing driven by visual cognition as well.

The paper is organized as follows. Section 2 formulates our dynamic scene stitching problem in mathematical descriptions. Section 3 addresses the hybrid-saliency-based visual cognition model and its calculation details. Section 4 gives the solutions of the output manifold via graph construction and cut path optimization at a minimum cognitive cutting cost. Section 5 shows the comparisons and experimental results in support of the effectiveness of the proposed method. Finally, we conclude this paper in Section 6.

2. Problem Formulation

We assume that the original dynamic scenes are captured by a camera settled on a horizontal stabled pan unit which can move in a smooth path, and scanning the scene with a certain semirotation, as shown in Figure 1. The video frames of the dynamic scenes are high-overlapped in the major scanning direction and seldom vertical movements. Since manifold mosaicing algorithm is an effective solution for breaking through the restriction of camera motions, we stitch the dynamic scene in this framework. The spirit of manifold-based mosaicing technique is to cut and paste proper strips, similar to the "scanning line" in 1D linear camera imaging, into an adaptive manifold of the output mosaics. The strips from each input frame are required to be perpendicular to the optical flow and proportional to the camera motion [2].

Under the scheme of manifold mosaicing and inspired by [7, 8], an idea of avoiding cutting moving objects or other regions for dynamic scene stitching is to select different widths of columns from every frame to form strips and align them smoothly into an adaptive nonlinear manifold. The process can be summarized as in Figure 2. The aligned neighboring strips must look locally like the real scene without any visual artifacts.

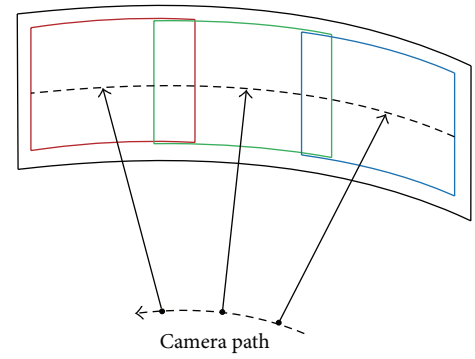


Figure 1: The camera path and its motion mode.

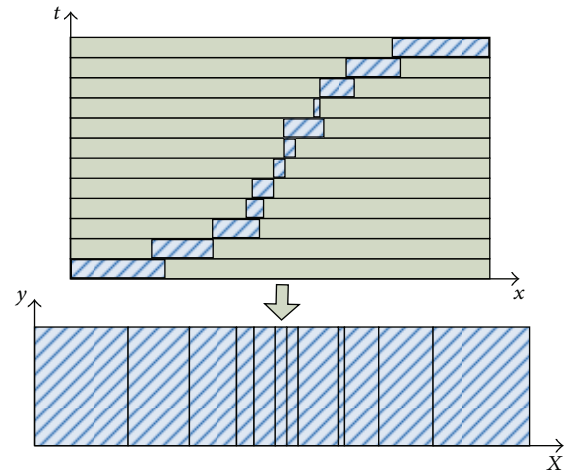


Figure 2: The principle for dynamic scene stitching.

Definition 1. Given a set of space-time volumes $V(x, y, t)$, where $V(x, t)$ represents the xth column of the tth frame, the output stitching image has L columns in all, that is, $\{\theta_i \mid i = [1, L]\}$. The mapping $\Gamma(\theta_i)$ between the output column θ_i and the input image column $V(x, \Delta y, t)$ is the vector of $(x, \Delta y, t)$, in which Δy is the vertical motion offset; if the major motion direction of the camera is horizontal, then $\Delta y \approx 0$. The set $P(x, t) = \{\Gamma(\theta_i)\}_{i=1}^{L}$ is defined as a *cut path* of the column strips along time t.

On the above hypothesis, each cut path corresponds to an output stitching image; therefore the process of dynamic scene stitching becomes to search and optimize the mapping relationships between the output columns and the input columns. We can then formulate the problem as the selection of the cut path at a minimum cutting cost throughout the space-time volumes.

Definition 2. If $\Gamma(\theta_i) = V(j, 0, k)$ and $\Gamma(\theta_{i+1}) = V(g, 0, h)$, that is, the θ_ith and the θ_{i+1}th columns of the output manifold are $V(j, k)$ and $V(g, h)$, that is, the jth column of the kth input frame and the gth column of the hth input frame, respectively, then the *cutting cost* is defined as follows:

$$C(\Gamma, \theta_i) = \min\left\{\left\|V(j,k) - V(g-1,h)\right\|, \left\|V(j+1,k) - V(g,h)\right\|\right\}. \quad (1)$$

The cost indicates the smoothness of the transition between consecutive column strips. If C is sufficiently small, it implies that the appearance of these two neighboring columns, $V(g, h)$ and $V(j, k)$, of the output manifold is as similar as that of the columns, $V(g-1, h)$ and $V(g, h)$, of the hth input frame, or as that of the columns, $V(j, k)$ and $V(j+1, k)$, of the kth input frame, keeping the local consistency of the original input frames. In this paper, a hybrid-saliency-based visual cognition model is proposed and a cognitive cutting cost is designed, specifically according to the model. More computation details are introduced in the following section.

3. Hybrid-Saliency-Based Computational Model of Visual Cognition

Visual attention mechanism plays an important role in visual cognition [10]. How to utilize this mechanism and relative mathematic representations to establish a computational model for visual cognition and guide for image processing to improve the performance of the algorithms is a meaningful research work. In this paper, we propose a visual cognition model, considering the potential directive effects as much as possible, and apply it to dynamic scene stitching to enhance the quality of the output mosaics.

The recent research on visual psychology shows that human's attention can be caused by the visual stimulus directly or by the observation task to find specific regions which are matched to the task. Based on these two kinds of causes, the visual attention patterns can be summarized into two categories: the bottom-up pattern driven by stimulus and the top-down pattern driven by task [11]. The most general dynamic scene stitching often encounters many different kinds of image contents, such as moving cars, people or animals, artificial buildings, and nature landscape. It is hard to define a uniform task to guide the stitching. Thus, the proposed visual cognition model adopts the bottom-up way to establish the computing model. Since interesting objects usually lie in the salient regions in regard to human visual perception, the model is established based on multiple visual stimuli by forming hybrid saliency maps of the moving targets and other interesting regions. The visual cognition model mainly involves three elements of different stimuli, defined as follows:

$$\mathrm{VCM}(I) = \alpha C_I(I) + \beta C_E(I) + \gamma C_D(I), \quad (2)$$

where $C_I(I)$, $C_E(I)$, and $C_D(I)$ are the intensity, the edge contour, and the scene depth information of the image, respectively, and α, β, γ are the weight coefficients of different stimuli. These elements reflect image saliency in different aspects. The intensity is the basic representation of an image. The edge contour is another important stimulus of image contents for analysis. And using depth information to distinguish the background and the objects is the fundamental function of biological vision [12]. The composing weights can be estimated by the content-based global amplification method [13]. Driven by the visual cognition model, the overall cutting cost of the optimum output manifold along the cut path P becomes

$$\begin{aligned}\min \mathrm{Cost}(P) &= \min \sum_{i=1}^{L} \left\|C_{\mathrm{VCM}}(\Gamma, \theta_i)\right\| \\ &= \min \sum_{i=1}^{L-1} \left\|\Gamma_{\mathrm{VCM}}(\theta_i) - \Gamma_{\mathrm{VCM}}(\theta_{i+1})\right\|,\end{aligned} \quad (3)$$

where $C_{\mathrm{VCM}}(\Gamma, \theta_i) = \alpha C_I(\Gamma, \theta_i) + \beta C_E(\Gamma, \theta_i) + \gamma C_D(\Gamma, \theta_i)$ indicates the saliency difference between the neighboring columns, $\Gamma_{\mathrm{VCM}}(\theta_i)$ and $\Gamma_{\mathrm{VCM}}(\theta_{i+1})$, of the hybrid visual cognitive map volumes. The cost reveals the smoothness of the transition between consecutive column strips in intensity, contour, and salient region. The hybrid saliency differences in different visual cognition aspects are calculated as follows.

3.1. Intensity Cognitive Saliency Difference. Intensity is deemed as the primitive features in psychological and biological visual cognition [14] and relatively easy to compute. We describe the intensity of input images by their gray or color values directly. The intensity difference between consecutive columns, $\Gamma(\theta_i) = V(x_i, t_i)$ and $\Gamma(\theta_{i+1}) = V(x_j, t_j)$, of the input volumes, is computed as follows:

$$\begin{aligned}C_I(\Gamma, \theta_i) = \min\Big\{&\left\|V_{\mathrm{gray}}(x_i, t_i) - V_{\mathrm{gray}}(x_j - 1, t_j)\right\|, \\ &\left\|V_{\mathrm{gray}}(x_i + 1, t_i) - V_{\mathrm{gray}}(x_j, t_j)\right\|\Big\},\end{aligned} \quad (4)$$

where $V_{\mathrm{gray}}(x_i, t_i)$ and $V_{\mathrm{gray}}(x_j, t_j)$ are the gray values of the x_ith and the x_jth columns in the t_ith and the t_jth frames, respectively. Minimizing this difference will maintain the basic visual appearance information among the neighboring strips. It is the primary premise to keep a smooth transition.

3.2. Edge Contour Cognitive Saliency Difference. Besides the salient intensity feature, the geometric contour structure is also a significant factor, impacting the continuity of the mosaic strips. The saliency of contour structures can be extracted by edge detectors, for example, Sobel, Canny, and so forth, which are easy to calculate and of definite physical meanings. Nevertheless, the detection results usually depend on the extent of luminance variance and contrast changes. We suggest extracting phase congruency (PC) which reflects the behavior in the frequency domain to express the saliency of contour structure of the image. Based on many physiological

and psychophysical evidences [15, 16], it is demonstrated that PC theory can provide a biologically plausible model for how human visual systems detect and identify features in an image. Compared with gradient-based edge detectors, it is not only invariant to illumination and contrast, but also superior in detecting and identifying multiple edge saliencies, including ramp edge, step edge, roof edge, and line edge. It can be considered as a dimensionless measure for the significance of a local structure. This property ensures that the PC-based contour saliency difference reflects the structural continuity cost among consecutive strips conforming to visual cognition behaviors. Therefore, the edge contour cognitive saliency difference between neighboring columns, $\Gamma(\theta_i) = V(x_i, t_i)$ and $\Gamma(\theta_{i+1}) = V(x_j, t_j)$, of the input volumes, is defined as follows:

$$C_E(\Gamma, \theta_i) = \min\left\{\left\|V_{\mathrm{PC}}(x_i, t_i) - V_{\mathrm{PC}}(x_j - 1, t_j)\right\|, \left\|V_{\mathrm{PC}}(x_i + 1, t_i) - V_{\mathrm{PC}}(x_j, t_j)\right\|\right\}, \tag{5}$$

where $V_{\mathrm{PC}}(x_i, t_i)$ and $V_{\mathrm{PC}}(x_j, t_j)$ are the phase congruency values of the x_ith and the x_jth columns in the t_ith and the t_jth PC maps, respectively.

The PC map volume $V_{\mathrm{PC}}(x, y, t)$ can be calculated from the input space-time volume $V(x, y, t)$. Rather than defining the saliency of edge features directly at points with sharp changes in intensity, the PC model postulates that features are perceived at points where the Fourier components are maximal in phase according to the psychophysical effects on human visual perception. It is derived from the local energy model [17], a salient feature measurement in frequency domain, and initially expressed as follows:

$$\mathrm{PC}(x) = \frac{|E(x)|}{\sum_n A_n(x)}, \tag{6}$$

where $A_n(x)$ is the amplitude of Fourier components at the location x in the signal and $|E(x)|$ is the local energy. The essence of the PC is to measure the phase similarity among all Fourier components. It is valued from 1 to 0, representing the saliency of features from significant down to none. However, this measure of PC does not provide good localization and it is also sensitive to noise. We adopt the improved PC based on banks of Log-Gabor wavelets and quadrature pairs of filters, which is developed by Kovesi [18] and widely used in the literature, to calculate $V_{\mathrm{PC}}(x, y, t)$. Since the local phase obtained by Log-Gabor wavelets lacks rotational invariance, orientation samplings are required to guarantee that the salient features are treated equally at all the possible orientations. The phase congruency at position (x, y) becomes

$$\mathrm{PC}_2(x, y) = \frac{\sum_o \sum_n W_o(x, y) \lfloor A_{no}(x, y) \Delta\Phi_{no}(x, y) - T_o \rfloor}{\sum_o \sum_n A_{no}(x, y) + \varepsilon}. \tag{7}$$

The symbols $\lfloor \cdot \rfloor$ denote that the enclosed quantity is equal to itself when its value is positive and zero otherwise. ε is a small constant to avoid division by zero. $W_o(x, y)$ is a factor weight for frequency spread in orientation o. T_o is the noise threshold, the estimated noise influence. Only energy values that exceed T_o are counted in the result. $A_{no}(x)$ is the local amplitude of frequency component on scale n and in orientation o. $\Delta\Phi_{no}(x, y)$ is the phase derivation function which can be expanded as follows:

$$\Delta\Phi_{no}(x, y) = \cos\left(\phi_{no}(x, y) - \overline{\phi}_o(x, y)\right) - \left|\sin\left(\phi_{no}(x, y) - \overline{\phi}_o(x, y)\right)\right|, \tag{8}$$

where $\overline{\phi}_o(x, y)$ is the local mean phase angle in orientation o. The product of $A_{no}(x)$ and $\Delta\Phi_{no}(x, y)$ can be calculated as follows:

$$A_{no}(x, y)\Delta\Phi_{no}(x, y) = e_{no}(x, y)\phi_e(x, y) + o_{no}(x, y)\phi_o(x, y) - \left|e_{no}(x, y)\phi_o(x, y) + o_{no}(x, y)\phi_e(x, y)\right|, \tag{9}$$

where $\phi_e(x, y) = \sum_n e_{no}(x, y)/E(x, y)$, $\phi_o(x, y) = \sum_n o_{no}(x, y)/E(x, y)$. The local energy $E(x, y)$ is defined as follows:

$$E(x, y) = \sqrt{\left(\sum_n e_{no}(x, y)\right)^2 + \left(\sum_n o_{no}(x, y)\right)^2}, \tag{10}$$

where $e_{no}(x, y) = V(x, y) * M^e_{no}$ and $o_{no}(x, y) = V(x, y) * M^o_{no}$ are the convolution results of the input image signal $V(x, y)$ with even- and odd-symmetric Log-Gabor filters, M^o_{no} and M^e_{no}, on scale n and in orientation o.

3.3. Scene-Depth Cognitive Saliency Difference. Depth is an important component channel in biological vision organisms. It assists in focusing attention on important locations and objects of the viewed scene. Since the human visual system has evolved predominantly in natural 3D environments, it is inspired to utilize depth information to accomplish visual task by instinct. There have been several efforts to include the depth channel in computational attention models to make the artificial visual attention biologically plausible [12, 19, 20]. In this paper, we take advantage of the characteristic that depth information has prominent effect on highlighting regional objects to define the scene-depth-based cognitive saliency difference between neighboring output columns, $\Gamma(\theta_i) = V(x_i, t_i)$ and $\Gamma(\theta_{i+1}) = V(x_j, t_j)$, as follows:

$$C_D(\Gamma, \theta_i) = \min\left\{\left\|V_{\mathrm{depth}}(x_i, t_i) - V_{\mathrm{depth}}(x_j - 1, t_j)\right\|, \left\|V_{\mathrm{depth}}(x_i + 1, t_i) - V_{\mathrm{depth}}(x_j, t_j)\right\|\right\}, \tag{11}$$

where $V_{\mathrm{depth}}(x_i, t_i)$ and $V_{\mathrm{depth}}(x_j, t_j)$ are the estimated depth values of the x_ith and the x_jth columns in the t_ith and the t_jth depth label maps, respectively. The depth saliency difference reflects the regional homogeneity in visual cognition.

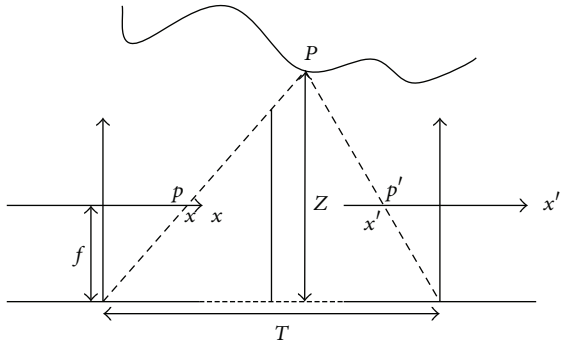


Figure 3: Relationship between depth and disparity.

Computing depth for an attention system is usually solved in stereo vision problems. In general, sensing the same scene from different view points, the depth information can be obtained by computing the disparity, that is, the parallax, between corresponding pixel pairs based on the triangulation principle [21]. The relationship between depth and disparity can be explained briefly as shown in Figure 3. Suppose that two corresponding projected pixels of the scene point P, whose depth is Z in the neighboring frames, are $p(x, y)$ and $p'(x, y)$, lying in the equal scanning line, that is, $\Delta y \approx 0$; then the disparity becomes $d(x, y) = x - x'$. If given the focal length f, according to the similar triangle principle, we have

$$\frac{Z}{T} = \frac{f}{d}. \tag{12}$$

It shows that if given the depth Z of a fixed point, then the disparity between its corresponding projected pixels is determined. Conversely, the depth can be also calculated by the disparity. Based on this fundamental correspondence, they are easy to interconvert with each other. With the increase of depth, the disparity goes down to 0 at the infinite points, whereas the nearest point with maximum disparity is denoted as $D_{\max}$. Thus, the disparity range of arbitrary points in the scene is $D_s = [0, D_{\max}]$, usually discretized as $D_s = \{0 = d_0 < d_1 < \cdots < d_n = D_{\max}\}$ in pixels. According to the above correspondence relationship, the depth map volume V_{depth} can be computed between each two neighboring frames from the disparity field, $d_p \in D_s$, by matching one to a reference one and mapping to the discrete disparity space to obtain the disparity of every pixel in the reference frame. In this paper, we simplify the disparity estimation algorithm of [22] and calculate the depth cognitive saliency difference based on mean-shift disparity filter, assuming that disparity values vary smoothly in homogeneous regions and depth discontinuities only occur on region boundaries. The specific steps for depth map calculation are as follows.

Step 1 (segment the homogeneous regions by mean-shift). We adopt mean-shift algorithm to decompose the reference frame into regions of homogeneous color or grayscale. It is easy to oversegment a whole region into multiple regions, which is preferred here to satisfy the disparity variance assumption in practice.

Step 2 (compute the local match cost in a bidirectional way). Taking each pair of neighboring frames as the reference image and the matched image, the match cost of pixel (x, y) and disparity d between the reference frame and the matched frame in a local window $N(x, y)$ are calculated in a bidirectional way. Consider

$$\begin{aligned} C_D(x, y, d) &= (1 - \omega) * C_{\text{SAD}}(x, y, d) \\ &\quad + \omega * C_{\text{GRAD}}(x, y, d), \\ C_{\text{SAD}}(x, y, d) &= \sum_{(i,j)\in N(x,y)} \left| I_1(i, j) - I_2(i + d, j) \right|, \\ C_{\text{GRAD}}(x, y, d) &= \sum_{(i,j)\in N_x(x,y)} \left| \nabla_x I_1(i, j) - \nabla_x I_2(i + d, j) \right| \\ &\quad + \sum_{(i,j)\in N_y(x,y)} \left| \nabla_y I_1(i, j) - \nabla_y I_2(i + d, j) \right|. \end{aligned} \tag{13}$$

The matching criterion C_D combines sum of absolute differences (SAD) and gradient absolute differences (GRAD). It is adaptive to the scene changes and would provide better accuracy, especially on the surface with textures.

Step 3 (estimate initial disparity map via cross-checking and WTA). In order to detect unreliable matches, a cross-checking procedure to the bidirectional matching cost is employed in conjunction with the winner-take-all (WTA) optimization strategy (choosing the disparity with the lowest matching cost). If given the range of disparities, $R_d = [d_{\min}, d_{\max}]$, in which the number of discrete disparities becomes $N_d = d_{\max} - d_{\min} + 1$, then the initial matched disparity of the reference frame is

$$D_{\text{int}}(x, y) = \arg\min_{d\in R_d} C_D(x, y, d). \tag{14}$$

Step 4 (simplify the computing by filtering the disparity map based on mean-shift segments). On the assumption that disparity values vary smoothly in homogeneous regions and depth discontinuities only occur on region boundaries, a single depth value is computed for each homogeneous region. The initial disparity map is filtered by taking the median disparity value of each mean-shift segment as its whole parallax, that is,

$$D_{s_i} = \text{median}\left(D_{\text{int}}(x, y)\right), \quad (x, y) \in \text{Seg}_i. \tag{15}$$

After the above disparity calculation steps, the depth information is obtained indirectly. We can transform the disparity map volume into its depth map volume at last.

4. Cut Path Optimization via Graph Construction

Since the columns of input frames are deemed as the basic elements for forming the output manifold, every column of

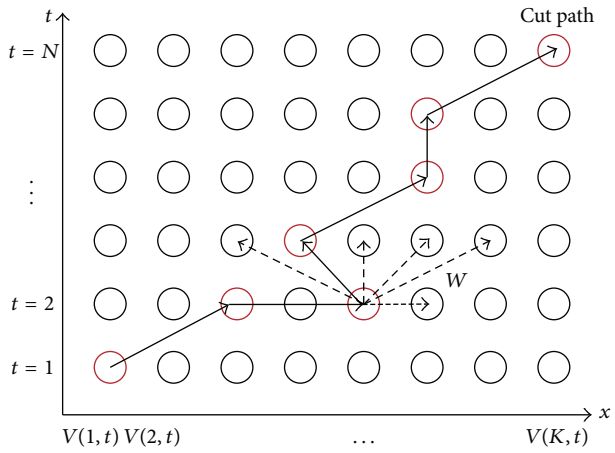


FIGURE 4: Graph construction in x-t space and its cut path for dynamic scene stitching.

the input images is regarded as a node so that the video volume can be abstracted as a graph. Let $G(V, E, W)$ denote the graph, as shown in Figure 4. The nodes $V = \{V(k, t)\}$ are the $K \times N$ image columns. The edges E encode the possible transitions among the columns. And each edge has an associated transition cost $W = C_{\mathrm{VCM}}(\Gamma, \theta_i)$, that is, the cutting cost, defined as the above cognitive saliency difference from the hybrid visual cognitive map volumes.

During the cost computation, due to the instability of point-to-point comparison of columns, we compute the hybrid cognitive saliency difference between $\Gamma(\theta_i)$ and $\Gamma(\theta_{i+1})$ in their centered rectangle windows. Moreover, there is no need to add all possible edges to the graph since the frames come from high-overlapping video sequences. Only those edges among nearby patches, as those dashed edges in Figure 4, are computed depending on the expected maximal motion velocity.

The goal of cut path optimization is to find a shortest path from V_{start} to V_{end}. It minimizes the cutting cost along the cut path, in which the salient regions are kept with minimum deformation as much as possible. Due to the efficiency of Dijkstra algorithm [23] for solving the shortest path problem between given nodes in a graph with nonnegative edge costs, we adopt it to search the cut path from the starting frame to the ending frame. Suppose $u_0 = V_{\mathrm{start}}$ is the source node and $v_0 = V_{\mathrm{end}}$ is the destination node. The basic idea of the algorithm is to calculate the shortest path and distance from u_0 to all the possible transition nodes of G, in the order of their distance to u_0. It stops until v_0 or covering all the possible transition nodes of G. In the meantime, labels are used to avoid repeating and keep the computing information of every step. The algorithm steps are as follows.

Step 1. Set $l(u_0) = 0$, and $l(v) = \infty, S_0 = \{u_0\}$, $i = 0$ for $v \neq u_0$. If $|V| = 1$, then stop; otherwise go to Step 2.

Step 2. For each v in $V \setminus S_i$, that is, $v \in \overline{S}_i$ ($\overline{S}_i = V \setminus S_i$), replace $l(v)$ by $\min_{u \in S_i}\{l(v), l(u) + w(uv)\}$. When $v \neq u$, $w(uv) = \infty$. If $l(v)$ is replaced, put a label $(l(v), u_i)$ on v.

Step 3. Compute $\min_{u \in S_i}\{l(v)\}$, $v \in \overline{S}_i$ ($\overline{S}_i = V \setminus S_i$), and denote the corresponding node as u_{i+1}; then let $S_{i+1} = S_i \cup \{u_{i+1}\}$.

Step 4. If $i = |V| - 1$, then stop. If $i < |V| - 1$, then replace i by $i + 1$ and go to Step 2.

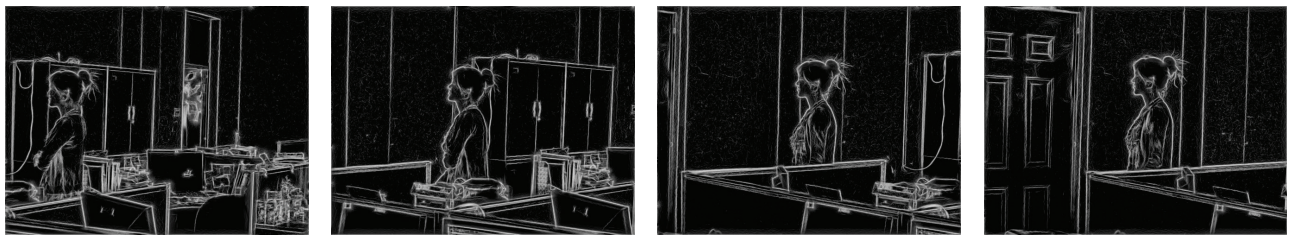
After optimizing the cut path of the dynamic volume, select the column strips in every frame according to the path and paste them together into a large adaptive manifold. The output stitching scene is then composed without deformations or other artificial defects.

5. Experiments and Comparisons

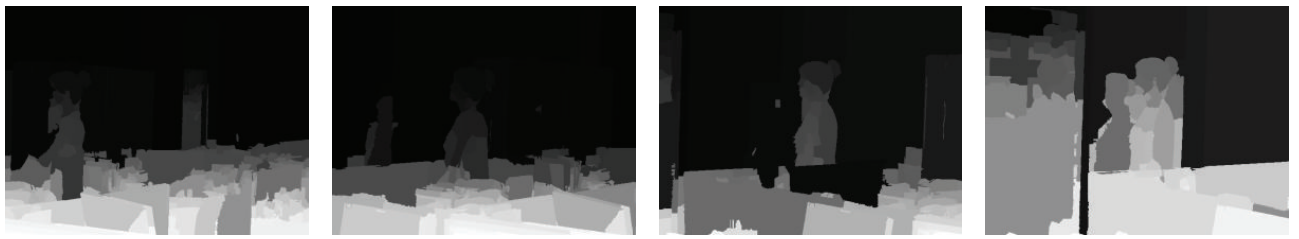
In order to testify the performance of the proposed algorithm in dealing with moving objects, we captured a series of video sequences under the previously described camera motion mode. Different complex human movements were involved in the videos. And the proposed method was also compared to another two manifold mosaicing algorithms [6, 8].

Two typical examples of dynamic scene stitching are shown in Figures 5 and 6. The visual cognitive maps of different salient stimuli are seen in Figures 5 and 6, in which (a)s are original inputs in gray intensities, (b)s are the PC-based contour saliency maps, and (c)s are the depth saliency maps. The stitching results of [6, 8] and our proposed algorithm are shown in Figures 7–9, respectively. Reference [6] estimates the global motion parameters between neighbor frames at first by iterating Lucas-Kanade optical flow under Gaussian pyramid strategy. And then the strips are selected from the middle of video frames according to the classical manifold mosaicing technique. The final output mosaics are composed without any a priori perception or optimization. The algorithm of [6] can stitch the background of the scene entirely, but it is poor in dealing with nonrigid moving objects, as seen in Figure 7. Instead of treating scene stitching as geometrical alignment, [8] poses it as a minimal appearance distortion in pure pixel processing level. The algorithm of [8] shows some effectiveness on maintaining moderate moving objects during the stitching, as the walking person in scene 1 is only elongated a little bit; see Figure 8(a). Nevertheless, when the scene contains more complex movements, such as the movements of the person, cleaning the blackboard, in scene 2, the moving object would be easily clipped, as seen in Figure 8(b). The performance of the algorithm in [8] needs to be improved. The stitching results of the proposed method are shown in Figure 9. And the corresponding cut paths are shown in Figure 10. It can be seen that, driven by the visual cognition model, the cut paths successfully avoid cutting the salient movement regions and the backgrounds as well, whereas the other two algorithms cannot guarantee the integrity of the moving objects in different degree, especially when the object moves in high mobility, since their manifold

(a) Original inputs in gray intensities



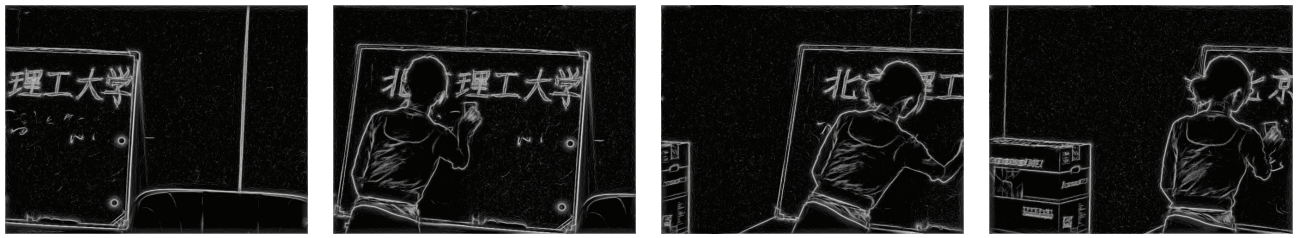
(b) PC-based contour saliency maps



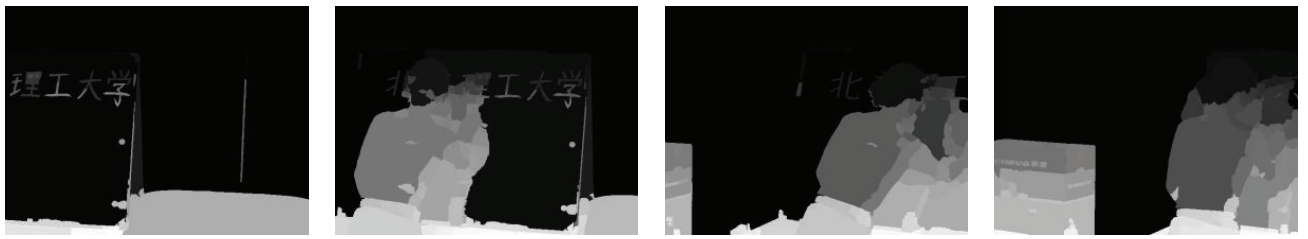
(c) Depth saliency maps

Figure 5: Multiple visual stimuli of dynamic scene 1.

(a) Original inputs in gray intensities



(b) PC-based contour saliency maps



(c) Depth saliency maps

Figure 6: Multiple visual stimuli of dynamic scene 2.

(a) Dynamic scene 1 (b) Dynamic scene 2

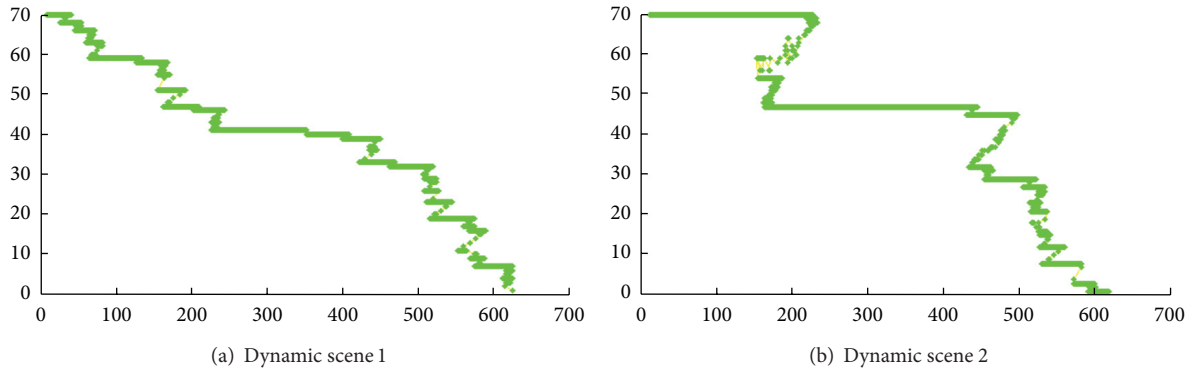
Figure 7: Stitching results of [6].

(a) Dynamic scene 1 (b) Dynamic scene 2

Figure 8: Stitching results of [8].

(a) Dynamic scene 1 (b) Dynamic scene 2

Figure 9: Stitching results of the proposed method.



(a) Dynamic scene 1 (b) Dynamic scene 2

Figure 10: Cut paths for forming the optimum output manifolds.

synthesis processes neglect the visual cognitive mechanism stimulated by multichannel saliencies.

After a sequence of experimental tests, it shows that the proposed method is robust to different dynamic movements. The hybrid-saliency-based cognitive model guarantees the stitching effect nicely. The proposed method can solve the dynamic scene stitching problem effectively.

6. Conclusions

This paper investigates dynamic video sequence stitching, especially under the situation that the scene, captured on a movable platform, contains moving objects or other important interesting regions. There is a great challenge to preserve the moving objects and the salient regions in the final stitching image as their original looks without any missing or deformation. In our research work, we proceed from human visual cognitive mechanism and analyze multiple visual stimuli to construct a hybrid-saliency-based cognitive model. Constrained by this model and combined with the manifold mosaicing framework, we proposed an effective dynamic scene stitching algorithm without any camera calibration and motion estimation. It can give full play to the role of visual cognitive mechanism of human in image synthesis for global scenes and reasonably avoid synthetic defects, such as motion blur and object clipping. The experimental results show that the proposed method performed quite well. It can be applied to wide-field monitoring system for supporting global situation judgments and decision-making, or other security investigations. The next goal is to study the sensitivity towards the selection of parameters in the cognition model, cooperating with quantitative stitching assessment.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by China Postdoctoral Science Foundation Funded Project (no. 2013M540863) and the National Key Technology R&D Program in the 12th Five-Year Plan of China (no. 2013BAI13B06).

References

[1] R. Szeliski, "Video mosaics for virtual environments," *IEEE Computer Graphics and Applications*, vol. 16, no. 2, pp. 22–30, 1996.

[2] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet, "Mosaicing on adaptive manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1144–1154, 2000.

[3] L. Zou, J. Chen, J. Zhang, and J. Lu, "Image mosaicing algorithm for dynamic scenes using multi-scaled PHOG feature and optimal seam," *Pattern Recognition and Artificial Intelligence*, vol. 25, no. 4, pp. 624–631, 2012.

[4] J. Jia and C.-K. Tang, "Image stitching using structure deformation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 617–631, 2008.

[5] Z. S. Zhao, X. Feng, S. H. Teng et al., "Multiscale point correspondence using feature distribution and frequency domain alignment," *Mathematical Problems in Engineering*, vol. 2012, Article ID 382369, 14 pages, 2012.

[6] L.-H. Zou, J. Chen, J. Zhang, and L.-H. Dou, "An image mosaicing approach for video sequences based on space-time manifolds," in *Proceedings of the 29th Chinese Control Conference (CCC '10)*, pp. 3003–3006, Beijing, China, July 2010.

[7] A. Rav-Acha, G. Engel, and S. Peleg, "Minimal Aspect Distortion (MAD) mosaicing of long scenes," *International Journal of Computer Vision*, vol. 78, no. 2-3, pp. 187–206, 2008.

[8] Y. Wexler and D. Simakov, "Space-time scene manifolds," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 858–863, Beijing, China, October 2005.

[9] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg, "Dynamosaicing: mosaicing of dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1789–1801, 2007.

[10] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: a survey," *ACM Transactions on Applied Perception*, vol. 7, no. 1, article 6, 2010.

[11] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," , *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–206, 2013.

[12] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Proceedings of 15th International Conference on Pattern Recognition (ICPR '00)*, vol. 1, pp. 375–378, Barcelona, Spain, 2000.

[13] L. Itti and C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems," in *Human Vision and Electronic Imaging IV*, vol. 3644 of *Proceedings of the SPIE*, pp. 473–482, San Jose, Calif, USA, January 1999.

[14] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[15] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[16] L. Henriksson, A. Hyvärinen, and S. Vanni, "Representation of cross-frequency spatial phase relationships in human visual cortex," *Journal of Neuroscience*, vol. 29, no. 45, pp. 14342–14351, 2009.

[17] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, no. 5, pp. 303–313, 1987.

[18] P. Kovesi, "Phase congruency: a low-level image invariant," *Psychological Research*, vol. 64, no. 2, pp. 136–148, 2000.

[19] C. I. Penaloza, Y. Mae, K. Ohara et al., "Using depth to increase robot visual attention accuracy during tutoring," in *Proceedings of IEEE International Conference on Humanoid Robots-Workshop of Developmental Robotics*, pp. 14–19, Osaka, Japan, 2012.

[20] C. Lang, T. V. Nguyen, H. Katti et al., "Depth matters: influence of depth cues on visual saliency," in *Proceedings of the 12th European Conference on Computer Vision*, pp. 101–115, Springer, Berlin, Germany, 2012.

[21] I. P. Howard and B. J. Rogers, *Perceiving in Depth, Volume 2: Stereoscopic Vision*, no. 29, Oxford University Press, 2012.

[22] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, pp. 15–18, Hongkong, August 2006.

[23] R. Bauer, D. Delling, P. Sanders et al., "Combining hierarchical and goal-directed speed-up techniques for dijkstra's algorithm," *Journal of Experimental Algorithmics*, vol. 15, Article 2.3, 2010.