VIDEO INPAINTING ALGORITHM FOR A DYNAMIC SCENE

Sang-Heon Lee, Soon-Young Lee, Jun-Hee Heu, Sang-Uk Lee

School of Electrical Engineering and Computer Science
Seoul National University, INMC
Seoul, Korea
hunlee@ipl.snu.ac.kr sylee@ipl.snu.ac.kr hjun77@ipl.snu.ac.kr sanguk@ipl.snu.ac.kr

ABSTRACT

A new video inpainting algorithm is proposed for removing unwanted objects or error of sources from video data. In the first step, the block bundle is defined by the motion information of the video data to keep the temporal consistency. Next, the block bundles are arranged in the 3-dimensional graph that is constructed by the spatial and temporal correlation. Finally, we pose the inpainting problem in the form of a discrete global optimization and minimize the objective function to find the best temporal bundles for the grid points. Extensive simulation results demonstrate that the proposed algorithm yields visually pleasing video inpainting results even in a dynamic scene.

Keywords: inpainting, restoration, optimization, video editing, spatio-temporal inpainting.

1. INTRODUCTION

Recently, video data are widely used in such areas as movie, home video and user created contents (UCC). As demands for reproducing video data are getting increased, the advanced post video processing techniques are needed. The object of inpainting technique is to remove unwanted objects or fill in missing parts in a scene with a visually pleasing way. Video inpainting is a key technique in video post processing areas, such as video editing, film postproduction and old film restoration.

Early studies expanded image inpainting techniques straightforwardly to video inpainting by considering a video sEquence as a set of independent still images [1-4]. However, the results yield unpleasant artifacts due to temporal aliasing. More effective approaches for video inpainting are to exploit high spatio-temporal correlation in a video sEquence [5-7]. Patwardhan et al. proposed a block-based greedy video inpainting using motion segmentation technique that separates foreground and background images to provide temporally consistent inpainting results [5]. Wexler et al. proposed a space-time inpainting technique for masked areas in a video sEquence [6]. These algorithms treat masked areas as 3-dimensional space-time volumes and fill them with cubic patches to keep the spatio-temporal consistency of the masked areas. Yu et al. expanded the priority BP (belief propagation) algorithm for still images to video data [7]. They applied the global optimization method the inpainting problem and reduced spatial artifacts.

However, the algorithms which have been suggested so far are efficient only for static and parallel camera motions and periodic moving objects. The method based on motion segmentation has the defective motion analysis that causes poor undesired results in dynamic conditions [5]. In addition, the methods using static 3-dimmensional graph structure are not adaptable in the dynamic situation, so the results go against the ground truth [6, 7]. Therefore, a video inpainting technique, which keeps the spatio-temporal consistency without the constraints, is still challenging.

In this paper, we propose the visually optimized inpainting technique which uses the motion information based on the space-time continuity. First of all, we define a block bundle that consists of temporally correlated blocks in order to exploit the temporal correlation of the video sEquences. Then, we construct the graph structure of grid points to arrange temporal block bundles and edges that connect a grid point to other neighbor points in the masking area of frames. We adopt a discrete Markov random field (MRF) as the model for inpainting problem and optimize the objective function with the conditions that the structure and colors are kept spatio-temporal consistency. Finally, the ideal block bundles are chosen by the global optimization.

This paper is composed as followed. First of all, we propose spatio-temporal inpainting technique with temporal block bundles in section 2. In section 3, the proposed technique is compared with the existing techniques through extensive simulation. Finally, the conclusion and future works are introduced in section 4.

2. PROPOSED ALGORITHM

Human visual system is sensitive particularly for continuity of edges which is the boundary of objects. In addition, it is also very keen on both changes in edge areas and continuity of objects moving. Therefore, the spatio-temporal consistency should be satisfied for removing the specific video area and filling in visually

" This research was supported by the **MKE (Ministry of Knowledge Economy),** Korea under the ITRC (Information Technolgy Research Center) Support program supervised by the IITA (Institute of Information Technology Advancement) " (IITA-2008-C1090-0801- 0018)

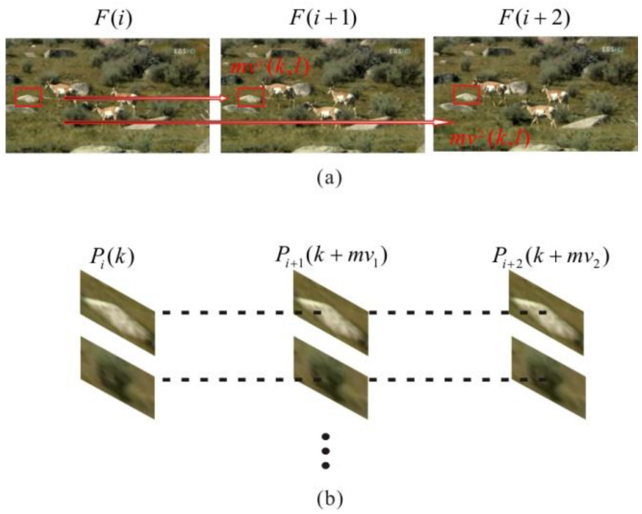

Fig. 1: The block bundle using motion vectors.

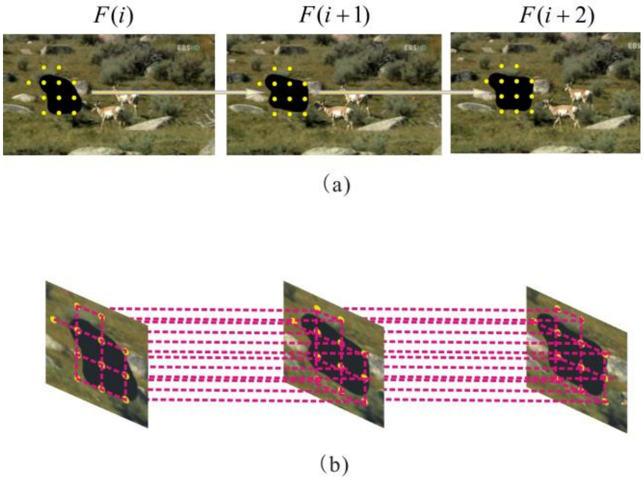

Fig. 2: The 3-dimmensional graph construction.

pleasing way. To keep the spatial and temporal consistency simultaneously, the best source block is selected from spatio-temporally correlated blocks, and then pasted to the target area covering over masked region.

The proposed algorithm defines the block bundle using motion information. In each frame, we set the block centered on each pixel and estimate the motion information. Then, the block bundles are defined as the combined matching blocks based on a temporal manner in the consecutive frames. Next, we place the grid points on the masking area in the frame with the half of block width spacing then restore the masking areas from correspondent candidate block bundles in the consecutive frames. To optimize the objective function globally, we use the model of the MRF, and so find the best block bundle for the grid points. The results of the iterative optimization method for the object function are improved in spatio-temporal consistency.

2.1 Temporal Block Bundles

To keep the spatio-temporal consistency of results, we have to pick out the best blocks in consecutive frames at once. But, as it is hard to satisfy the spatial and temporal consistency simultaneously, we try to keep only the temporal consistency of source blocks by combining a block with motion compensated blocks. By assuming that the motion of the local block is approximated to be a translation motion, we obtain the motion vectors by the block matching.

First of all, the motion information of frames is estimated to establish the motion block property. Note that $F(i)$ $(i = 0, \dots, N-1)$ the i th frame in video sEquence. We set the block $P_i(k)$ which has $w \times h$ size and centered at the pixel k in $F(i)$. Then, the motion vector for each corresponding block is given by

$$mv_1(k) = \arg\min_m \mathrm{SAD}(P_i(k), P_{i+1}(k+m)), \tag{1}$$

$$\mathrm{SAD}(P_i(k), P_{i+1}(k+m)) = \sum_{s \in [-\frac{w}{2}, \frac{w}{2}][-\frac{h}{2}, \frac{h}{2}]} |C_i(k+s) - C_{i+1}(k+m+s)|, \tag{2}$$

, where $C_i(k)$ means a 3-tuple R, G, B color vector at the pixel k in $F(i)$. In the same way, the mv_2 in $F(i+2)$ is obtained then $P_i(k)$, $P_{i+1}(k+mv_1)$ and $P_{i+2}(k+mv_2)$ are determined by mv_1, mv_2. After that, the temporal block bundle $B_{i,i+1,i+2}(k)$ is defined as the block set combined the motion compensated blocks $P_i(k)$, $P_{i+1}(k+mv_1)$ and $P_{i+2}(k+mv_2)$. The Fig. 1 (a) shows how to find motion vector and correspondent blocks of $P_i(k)$, and the Fig. 1 (b) illustrates the example of block bundles.

2.2 Spatio-temporal Graph Structure

We complete the mask area by filling with the best matching block bundles. The temporal block bundles are designed to keep only the temporal consistency of blocks in consecutive frames. Therefore, the 3-dimensional graph model is needed to optimize the spatial and the temporal consistency simultaneously. First of all, we construct the 2-dimensional graph which can estimate the similarity of neighbor blocks to keep the spatial consistency. The graph consists of grid points to place blocks in Fig. 2 (a). To achieve the spatio-temporal consistency in target area, we ensure enough overlapped region by placing the grid points to settle blocks with spacing of a half of the block size both horizontal and vertical.

Then, we construct the 3-dimensional spatio-temporal graph extended from the 2-dimensional spatial graph. Each grid point in $F(i)$ corresponds to the point in $F(i+1)$, $F(i+2)$ by motion vectors and combine the nodes in Fig. 2. But, it is hard to find adaptable motion vector for nodes in masking area where there are not available pixels. Thus, we estimate the motion information for nodes in masking

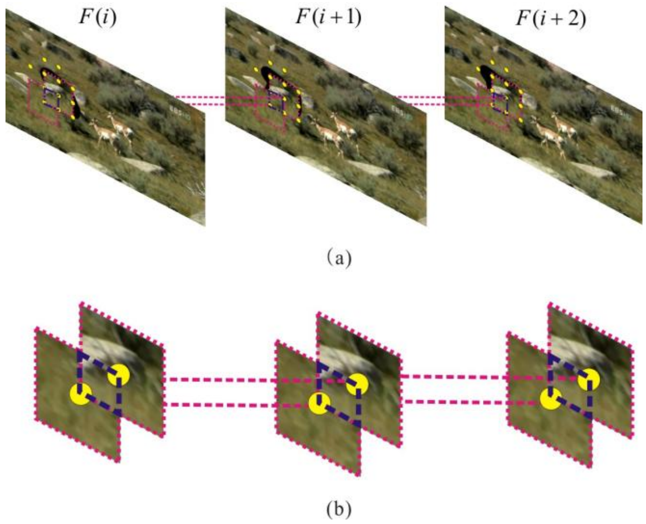

Fig. 3: The optimization with spatial consistency.

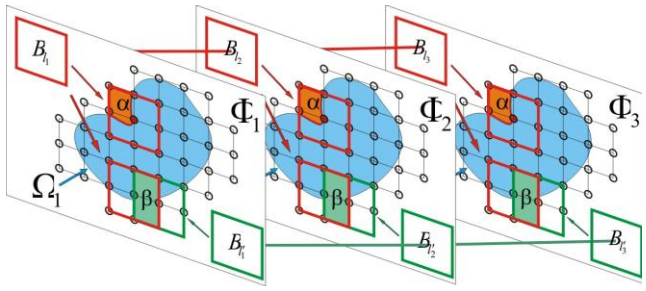

Fig. 4: The potential function of the MRF.

area with applying interpolation techniques to the neighbor motion vectors. The Fig. 2 (a) shows the matched grid points and estimated grids points in the mask area, and Fig. 2 (b) demonstrates the construction of 3-dimensional graph that consists of three temporal correlated 2-dimensional graphs. The spatio-temporal graphs improve the optimization of block bundles and help to make efficient use of overlapped area information for temporal consistency.

2.3 Spatio-temporal Optimization

We find the best block bundles among the candidate bundles for the grid points and paste the best bundles to the spatio-temporal graph in Fig. 3. To find the best temporal block bundles on the spatio-temporal graph, we model the graph to the MRF and solve it as well-known combinatorial problem. In Fig. 4, Ω is the masking region to restore and Φ is the source region. Grid points are nodes and edges are the lines that connect a point with the neighbor grid points. Our goal is to assign the best label l_i corresponding block bundle $B_{f,f+1,f+2}(l_i)$ to the i th node n_i in the graph so that the total energy $E(\{\hat{l}_i\})$ of the MRF is minimized.

$$E(\{\hat{l}_i\}) = \sum_{i=1}^{N} V_i(\hat{l}_i) + \sum_{(i,j)\in\varepsilon} V_{ij}(\hat{l}_i, \hat{l}_j), \qquad (3)$$

$$V_i(l) = \sum_{f=1}^{3} \sum_{s\in[-\frac{w}{2}\,\frac{w}{2}][-\frac{h}{2}\,\frac{h}{2}]} M_f(n_i+s)(C_f(n_i+s) - B_f^s(l))^2, \quad (4)$$

$$V_{ij}(l_i, l_j) = \sum_{s\in O} (B^s_{f,f+1,f+2}(l_i) - B^{\hat{s}}_{f,f+1,f+2}(l_j))^2, \qquad (5)$$

$$M_i(c) = \begin{cases} 0 & c \in \Omega_i \\ 1 & c \in \Phi_i \end{cases}. \qquad (6)$$

In the above Equations, Eq. (3) is the MRF energy Equation to optimize the spatio-temporal graph. The single node potential $V_i(l)$ for placing the block bundle with label l encodes how well that bundle agrees with the source region around n_i. At this time, $B_f^s(l)$ is defined as a 3-tupple R, G, B color vector at the pixel s in the block $B_f(l)$ corresponding to label l in the frame f. At this time, $M_i(c)$ in Eq. (5) denotes a binary mask, which is non zero only inside source region Φ. The α area in Fig. 4 shows single node potential $V_i(l)$.

The pair-wise potential $V_{ij}(l_i, l_j)$ in Eq. (3) for placing the block bundles corresponding to l_i, l_j over neighbors n_i, n_j, measures how well these bundles agree at the resulting region of overlap. The β area in Fig. 4 shows the potential $V_{ij}(l_i, l_j)$ that is given by the SSD over that region O overlapped by $B_{f,f+1,f+2}(l_i)$, $B_{f,f+1,f+2}(l_j)$ in Eq. (5). At this time, s and $\hat{s}$ are the overlapped pixels in O.

Finally, we use the graph-cut method to optimize the objective function of MRF. The graph-cut method guarantees considerably stable solution without delay.

3. EXPERIMENTAL RESULTS

The algorithm described in this paper is applicable to a various full color natural videos of complex dynamic scenes such as DVD movie or home DV etc. We compare the proposed method to existing algorithms [2]. There are two video data that are 10-frames of ‘Docuprime 5, the American western national park’ in EBS in Korea and ‘Brokeback Mountain’ for the experiment.

In Fig. 5, there is the result that the leading deer is removed in ‘Docuprime’ video. The original data is Fig. 5 (a), the result of Bertalmio et al. is 5 (b) and the last 5 (c) is the result of our proposed algorithm. In 5 (b), the result of Bertalmio et al. is over-blurred and is not able to keep the temporal consistency. On the contrary, the result in 5 (c) keeps the spatio-temporal consistency and is satisfied with the visuality. Also, the additional results about ‘Brokeback Mountain’ in Fig. 6 give support to prove the efficiency of the proposed algorithm.

4. CONCLUSION AND FUTURE WORKS

In this paper, we proposed the new inpainting algorithm which keeps the spatio-temporal consistency based on a global optimization using the notion of block bundle and

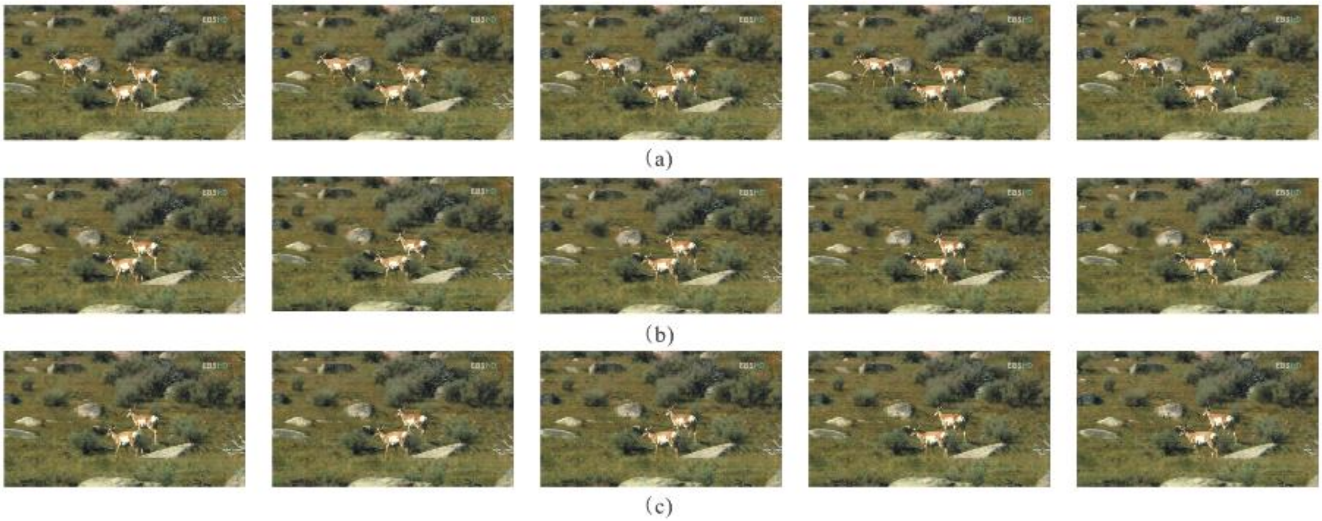

Fig. 5: The inpainting results of ‘Docuprime 5, the American western national park’.

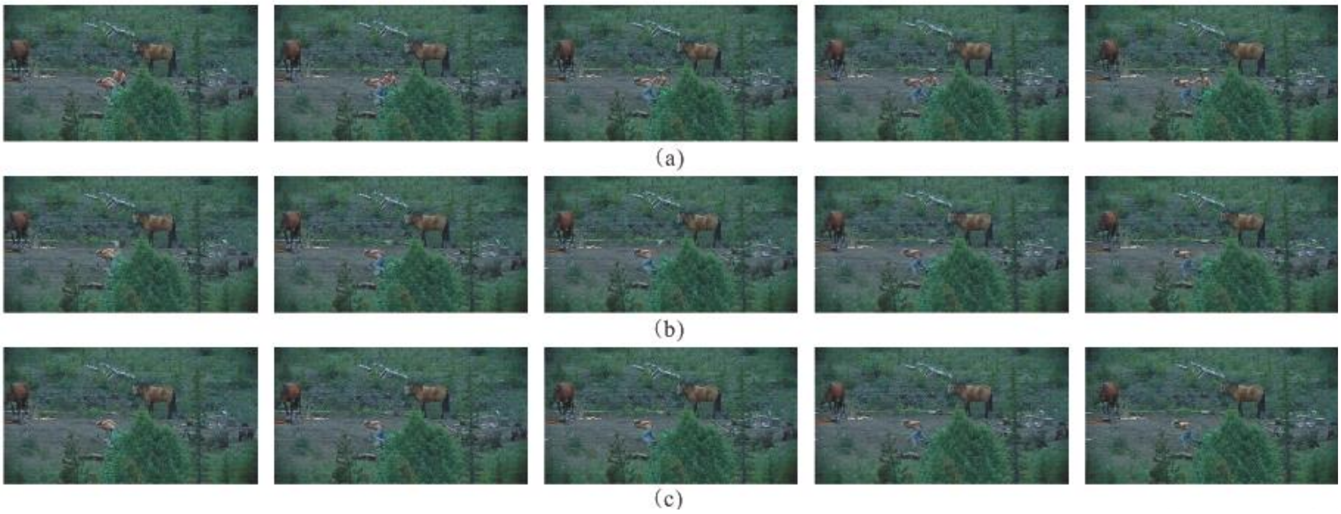

Fig. 6: The inpainting results of ‘Brokeback Mountain’.

spatio-temporal graph. First, we modeled a block bundle with motion vectors to maintain the time consistency. Then, using the 3-dimensional graph that is constructed by the spatial and temporal correlation, we optimized the spatial and temporal consistency simultaneously. Experimental results prove that our proposed algorithm can produce visually pleasant results.

The improvement of methods for film restoration, video editing and reproducing can be realizable by the proposed algorithm. The proposed method is expected to be the main technique for various researches about post video processing, such as video synthesis, video insertion and etc.

5. REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proc. ACM SIGGRAPH*, pp. 417-424, July 2000.

[2] M. Bertalmio, A. L. Bertozzi, and G. Sapiro,“Navier-stokes, fluid dynamics, and image and video inpainting,” in *Proc. CVPR*, vol. 1, pp. 355-362, June 2001.

[3] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Simultaneous structure and texture image inpainting,” *IEEE Trans. on Image Processing*, vol 12, no. 8, pp. 882-889, Aug. 2003.

[4] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Image Filling-In in a decomposition space,” in *Proc. ICIP*, vol 1, pp.853-855, Sep. 2003.

[5] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, “Video inpainting under con-strained camera motion,” *IEEE Trans. on Image Processing*, vol. 16, no. 2, pp. 545-553, Feb. 2007.

[6] Y. Wexler, E. Shechtman, and M. Irani,“Space-time completion of video,” *IEEE Trans. on PAMI*, vol. 29, no. 3, pp. 463-476, Mar. 2007.

[7] Y. Yu, D. Xu, C. Chen, and L. Zhao, “Video Completion Based on Improved Belief Propagation,” in *Proc. WSEAS*, vol 1, pp. 53-58, Sep. 2006.