AUTOMATIC INPAINTING OF LINEARLY RELATED VIDEO FRAMES

Yudong Xiao[1*], *Jinli Suo*[2], *Liheng Bian*[2], *Lei Zhang*[1] *and Qionghai Dai*[1,2]

Tsinghua National Laboratory for Information Science and Technology(TNList)
[1]Graduate School at Shenzhen, Tsinghua University [2]Department of Automation, Tsinghua University

ABSTRACT

This paper addresses automatic inpainting of a specific but common kind of videos captured by imaging a far or planar scene with a moving camera. The projective model tells that the frames of such videos can be approximately aligned by linear mappings except for some to-be-inpainted small regions. Mathematically, we treat inpainting as a global optimization with a linear system incorporating both the temporal consistency and the priors of the inpainting regions: (i) temporally registered frames form a low rank matrix; (ii) the pixels in the given inpainting regions destroy the low rankness with gross sparse errors. Besides, we also use a soft mask to ensure consistent global brightness before and after inpainting. Further, we propose a numerical solution to above optimization based on Augmented Lagrangian Method. The experiment results demonstrated our advantageous in both preserving thin scene structures and the details prone to be smoothed out by previous methods.

Index Terms— Video inpainting, low rank, sparse

1. INTRODUCTION

Video inpainting is of wide applications, e.g., excluding unwanted scratches or objects, background modeling, etc., and has been extensively explored. Past years have witnessed various approaches which generally fall into following types:

(i) Local block matching. Primary studies in video inpainting, such as [1], apply image inpainting methods frame by frame. Such direct image-to-video extension neglects the temporal continuity and the results suffer from fluctuations. Recent video inpainting methods fill holes using fragments in other frames. To handle complex videos, Shih et al.[2] and Patwardhan et al.[3][4] inpaint subregions separately with different methods or priorities. Some other researchers utilize the temporal consistency explicitly. For example, Mounira et al.[5] and Miguel et al.[6] apply motion compensation and conduct block matching across frames to estimate the unknown regions. Besides explicit motion compensation, Jia et al.[7] adopt additional techniques to eliminate spatial fusion artifacts and temporal fluctuations. Although the across-frame block matching helps obtaining reasonable results, the exploration of the temporal consistency is limited to several frames out of the whole sequence and artifacts exist widely.

(ii) Global optimization. Wexler et al.[8][9] extend the image inpainting approach by belief propagation along temporal dimension to build a spatiotemporal 3D graph model. Instead of optimizing the 3D graph directly, Liu et al.[10] complete the motion field before inpainting the regions. These approaches via global optimization are often time consuming and the large gap between simple energy definition and the diversity of nature videos tend to generate over smooth results.

In addition, either using local or global optimization, previous studies often attempt to propose a general inpainting method and trade off performance for the generality.

Motivation and our approach. We notice that in many cases, such as the planar or far scenes captured by a smoothly rotating or translating camera, the adjacent video frames are generally linearly related. These video frames can be registered by affine or homography transformations except for the to-be-inpainted regions, e.g., logo, subtitles. Frames of these scenes can all be aligned by either affine or homography transformation, and thus display a globally similar optical flow along the temporal dimension. Such observations inspire us to adopt a linear system to model such transformations and exploit temporal redundancy explicitly and automatically.

Mathematically, inpainting such video clips can be formulated as matrix completion robust to both transformation and gross error. In spite of the illposedness of the inpainting task, the low rank of the registered frame stack and the sparsity of the gross errors are both informative priors and would benefit generating reasonable inpainting result. In addition, the linearity of the latent transformation makes possible incorporating its estimation into a joint optimization framework. Inspired by the robust alignment method proposed by Peng et al.[11], we introduce two additional task specific priors for inpainting: (i) The position of the inpainting regions are given; (ii) The inpainted result should consist with the original video in global brightness. Fig. 1 gives one exemplar result, the region highlighted with dashed rectangle can be robustly inpainted and the clear background of the 'airplane' error term implies a good preservation of the global brightness.

This paper has some similarity with but largely differ-

[*]This work was supported by the Shenzhen upgrading project of key laboratories (No.CXB201104220042) and the Nanshan district project of core technology breakthrough (NO.AKC2013JSJS0009A).

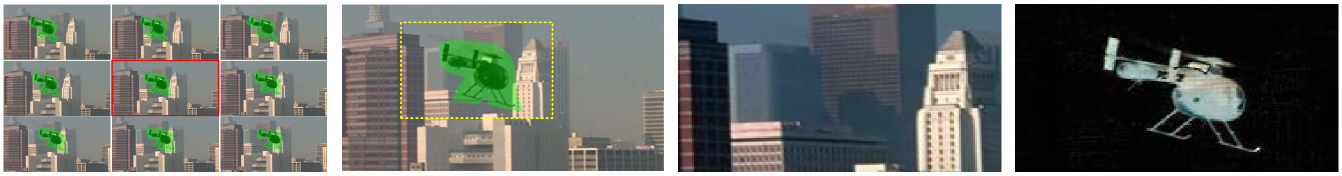


Fig. 1. Leftmost: a clip with labeled inpainting regions. Leftcenter and rightcenter: the highlighted frame in the clip and its inpainting result. Rightmost: the separated gross error.

ent from two previous works. Ding et al.[12] and Ling et al.[13][14] both adopt manifold learning techniques to explore the temporal redundancy of video frames for completion. These methods are applicable for objects that can be modeled by a template with some object specific labeling, while we focus on a different type of videos. Besides, our approach is under a new framework and can impose priors from the inpainting regions additionally and automatically.

2. FORMULATION AND ALGORITHMS

Let $\mathbf{V} = \{V_{1\cdots n}\}$ denote the n to-be-inpainted video frames, we can align them by applying a series of transformations $\mathbf{T} = \{T_{1\cdots n}\}$ to form a low rank matrix $\mathbf{L} = \{L_{1\cdots n}\}$, except for the entries in the binary inpainting region $\mathbf{R} = \{R_{1\cdots n}\}$. We also define an align operator $\circ$ as $\mathbf{V} \circ \mathbf{T} = [V_1 * T_1, V_2 * T_2, \cdots, V_n * T_n]$ and formulate video inpainting as following optimization:

$$\begin{aligned} \arg\min_{\mathbf{L},\mathbf{E},\mathbf{T}} \quad & ||\mathbf{L}||_* + \alpha||\mathbf{M} \odot \mathbf{E}||_1 \qquad (1) \\ s.t. \quad & \mathbf{V} \circ \mathbf{T} = \mathbf{L} + \mathbf{E}. \end{aligned}$$

Here we use nuclear norm $|| \cdot ||_*$ to model the low rank constraint; $\mathbf{E}$ is a sparse gross error in the pre-specified mask regions $\mathbf{M} = \mathbf{R} \circ \mathbf{T}$, we minimize its l_1 norm to force the sparsity and ensure color consistency out of the mask boundary; α is the coefficient balancing two energy terms.

To linearize the constraint in (2), we adopt the similar strategy in [11] to perform 1st order Taylor expansion and get

$$\begin{aligned} \arg\min_{\mathbf{L},\mathbf{E},\Delta\mathbf{T}} \quad & ||\mathbf{L}||_* + \alpha||\mathbf{M} \odot \mathbf{E}||_1 \qquad (2) \\ s.t. \quad & \mathbf{V} \circ \mathbf{T} + \sum_{i=1}^{n} J_i \Delta\mathbf{T} \epsilon_i \epsilon_i^T = \mathbf{L} + \mathbf{E}. \end{aligned}$$

Here $J_i = \frac{\partial}{\partial\zeta}\left(\frac{vec(V_i \circ \zeta)}{||vec(V_i \circ \zeta)||_2}\right)|_{\zeta = T_i}$ and ϵ_i is the standard basis of $\mathbb{R}^n$. To convexify above objective function, we introduce three auxiliary variables $\mathbf{S}$ and $\mathbf{h}_{1\cdots 2}$ and turn it into

$$\begin{aligned} \arg\min \quad & ||\mathbf{L}||_* + \alpha||\mathbf{S}||_1 \qquad (3) \\ s.t. \quad & \mathbf{h}_1 = \mathbf{S} - \mathbf{M} \odot \mathbf{E} = \mathbf{0} \\ & \mathbf{h}_2 = \mathbf{V} \circ \mathbf{T} + \sum_{i=1}^{n} J_i \Delta\mathbf{T} \epsilon_i \epsilon_i^T - \mathbf{L} - \mathbf{E} = \mathbf{0}, \end{aligned}$$

where $\odot$ is the component-wise product that for any two matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A} \odot \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$.

In this paper, we use ALM to optimize the above objective, with the Augmented Lagrangian equation defined as

$$\mathcal{L} = \|\mathbf{L}\|_* + \alpha\|\mathbf{S}\|_1 + \sum_{i=1}^{2}\left(< \mathbf{Y}_i, \mathbf{h}_i > + \frac{\mu}{2}\|\mathbf{h}_i\|_F^2\right), \quad (4)$$

in which $< \cdot, \cdot >$ denotes inner product, $\mathbf{Y}_i$ is the Lagrange multiplier matrix and μ is a positive scalar.

Update L. Removing the items irrelevant to $\mathbf{L}$ in (4) gets

$$\begin{aligned} \mathbf{L}^{(k+1)} &= \arg\min_{\mathbf{L}} \|\mathbf{L}\|_* \\ &+ \frac{\mu^{(k)}}{2}||\mathbf{L} - \mathbf{V} \circ \mathbf{T} - \sum_{i=1}^{n} J_i \Delta\mathbf{T}^{(k)} \epsilon_i \epsilon_i^T - (\mu^{(k)})^{-1}\mathbf{Y}_2^{(k)} + \mathbf{E}^{(k)}||_F^2, \end{aligned}$$

which is the typical nuclear norm optimization and thus $\mathbf{L}$'s update rule can be written as

$$\mathbf{L}^{(k+1)} = \mathbf{U}s_{(\mu^{(k)})^{-1}}(\Sigma)\mathbf{V}^T. \quad (5)$$

Here $\mathbf{U}\Sigma\mathbf{V}^T$ denotes the singular value decomposition of $(\mathbf{V} \circ \mathbf{T} + \sum_{i=1}^{n} J_i \Delta\mathbf{T}^{(k)} \epsilon_i \epsilon_i^T + (\mu^{(k)})^{-1}\mathbf{Y}_2^{(k)} - \mathbf{E}^{(k)})$ and the shrinkage operator $s_{\mu^{-1}}(\Sigma)$ keeps only the entries where $\Sigma_{ij} > \mu^{-1}$.

Update S. We keep only the items related to $\mathbf{S}$ and get a convex optimization function

$$\mathbf{S}^{(k+1)} = \arg\min_{\mathbf{S}} \alpha\|\mathbf{S}\|_1 + \frac{\mu^{(k)}}{2}\left\|\mathbf{S} - \mathbf{M} \odot \mathbf{E}^{(k)} + (\mu^{(k)})^{-1}\mathbf{Y}_1^{(k)}\right\|_F^2.$$

Following the ALM algorithm, $\mathbf{S}$'s updating rule is derived as

$$\mathbf{S}^{(k+1)} = s_{(\mu^{(k)})^{-1}\alpha}(\mathbf{M} \odot \mathbf{E}^{(k)} - (\mu^{(k)})^{-1}\mathbf{Y}_1^{(k)}). \quad (6)$$

Update E. Without a closed-form updating rule for $\mathbf{E}$, we adopt gradient decent method to update it iteratively:

$$\mathbf{E}^{(k+1)} = \mathbf{E}^{(k)} - \delta \times \frac{\partial\mathcal{L}}{\partial\mathbf{E}}|_{\mathbf{E}=\mathbf{E}^{(k)}}, \quad (7)$$

where δ is the gradient descent step parameter and

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathbf{E}} &= \mu^{(k)}[\mathbf{M} \odot \mathbf{E} - (\mathbf{S}^{(k)} + (\mu^{(k)})^{-1}\mathbf{Y}_1^{(k)})] \odot \mathbf{M} \qquad (8) \\ &+ \mu^{(k)}\left(\mathbf{E} - \mathbf{V} \circ \mathbf{T} - \sum_{i=1}^{n} J_i \Delta\mathbf{T}^{(k)} \epsilon_i \epsilon_i^T + \mathbf{L}^{(k)} - (\mu^{(k)})^{-1}\mathbf{Y}_2^{(k)}\right). \end{aligned}$$

Update ΔT. The energy terms with respect to $\mathbf{T}$ can be calculated in terms of MSE as

$$\Delta\mathbf{T}^{(k+1)} = \sum_{i=1}^{n} J_i^{\dagger}(\mathbf{L}^{(k)} + \mathbf{E}^{(k)} - \mathbf{V} \circ \mathbf{T} - (\mu^{(k)})^{-1}\mathbf{Y}_2^{(k)})\epsilon_i\epsilon_i^T. \quad (9)$$

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

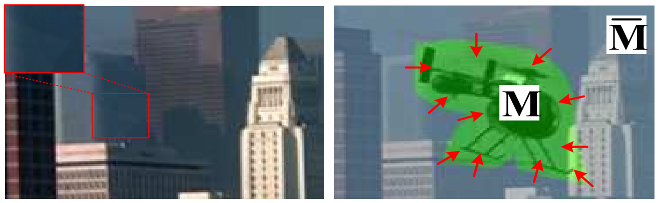


Fig. 2. The illustration of the fusion strategy.

Here $J_i^\dagger$ denotes the Moore-Penrose pseudoinverse of J_i.

Update $\mathbf{Y_1}, \mathbf{Y_2}$. According to the ALM algorithm, the Augmented Lagrangian Multiplier $\mathbf{Y_1}, \mathbf{Y_2}$ can be updated as

$$\mathbf{Y}_i^{(k+1)} = \mathbf{Y}_i^{(k)} + \mu^{(k)}\mathbf{h}_i^{(k)}. \tag{10}$$

Update μ. The parameter μ can be updated as

$$\mu^{(k+1)} = \min(\rho\mu^{(k)}, \mu_{max}), \tag{11}$$

The constant parameters are set as follows: α=3e-2, μ_{max}=10e6, $\rho = 1.1$ and δ=1e-3. For more clarity, the main steps are summarized in Alg. 1.

When the video is of time varying illuminations, directly warping back the recovered low rank stack according to the optimized **T** is unideal. As shown in Fig. 2, minimizing $\|\mathbf{M}\odot\mathbf{E}\|_1$ forces the separated error in $\overline{\mathbf{M}}$ close to zero and retains the illumination well, but in **M** the variation violates the low rank prior and thus causes artifacts. Therefore, we compute a scaling map for brightness adjustment: the scaling in $\overline{\mathbf{M}}$ is set as the ratio between original input and the recovered image, while that in **M** is calculated by spreading the factor in $\overline{\mathbf{M}}$ inward via heat diffusion. The simplicity of the global illumination pattern promises good result from such diffusion.

3. EXPERIMENT

We test the performance of our approach on a series of videos, with the binary masks labeling the to-be-inpainted regions provided. As for initial frame registration, we adopt SIFT algorithm to extract hundreds of marker points and select three that match best among all of the frames to fit a global transformation in terms of MSE. The transformations are defined over planar homography group.

We run our Matlab implementation on a workstation with Intel Xeon 2.27 GHz CPU and 4 GB memory. Without special optimization, current algorithm is able to handle 20 video frames of 300 × 300 pixels. It is worth noting that we can adopt a hierarchy strategy to handle high resolution videos. The algorithm usually needs 6-10 outer iterations to converge and the running time is around 200s for above data.

Fig. 1 and Fig. 3 respectively show inpainting results on three sequences: (1) excluding moving objects; (2) removing user specified logo or restoring destroyed regions; (3) recovering regions occluded by subtitles. The results demonstrate that the proposed approach is able to inpaint various occlusions successfully. Recall that our model imposes no constraint to the shape and position of the inpainting regions, such as its motion, shape, position, etc., so adjustment unnecessary for these different cases. The second video in Fig. 3 is of large illumination variance and our algorithm is able to preserve the original hue successfully. The closeup comparison with and advantages over two state-of-the-arts are shown in Fig. 4 and discussed in the figure caption.

```
Algorithm 1: Algorithm for the outer loop of optimization
  Input  : Original video $\mathbf{V}$ and corresponding mask $\mathbf{R}$,
           and initial transformation $\mathbf{T}^{(0)}$
  Output : Aligned inpainted video $\mathbf{L}$, sparse error $\mathbf{E}$ and
           transformation $\mathbf{T}$
  while not converged do
      • compute Jacobian matrices $\{J_{i\cdots n}\}$;
      • compute mask of aligned frames $\mathbf{M} \leftarrow \mathbf{R} \circ \mathbf{T}$;
      • solve the linearized convex optimization defined in
        Eq. 3 using Alg. 2;
      • update transformation $\mathbf{T} \leftarrow \mathbf{T} + \Delta\mathbf{T}^*$;
  end
```

```
Algorithm 2: Algorithm for the inner loop of optimization
  Input  : Original video $\mathbf{V}$, current mask $\mathbf{M}$, current
           transformation $\mathbf{T}$ and Jacobian matrices $\{J_i\}$
  Output : Aligned video $\mathbf{L}$, sparse error $\mathbf{E}$ and
           transformation increment $\Delta\mathbf{T}$
  $\mathbf{L}^{(0)} = \mathbf{V} \circ \mathbf{T}$, $\mathbf{E}^{(0)} = \mathbf{0}$, $\Delta\mathbf{T}^{(0)} = \mathbf{0}$, $\mathbf{Y}_{1,2}^{(0)} = \mathbf{0}$;
  while not converged do
      • update $\mathbf{L}^{(k+1)}$ according to (5);
      • update $\mathbf{S}^{(k+1)}$ according to (6);
      • update $\mathbf{E}^{(k+1)}$ according to (7);
      • update $\Delta\mathbf{T}^{(k+1)}$ according to (9);
      • update $\mathbf{Y}$, $\mu$ following (10) and (11), respectively;
      • $k := k + 1$;
  end
```

4. DISCUSSIONS AND FUTURE WORK

This paper proposes to use a convex optimization framework to inpaint a widely existing type of videos automatically and obtains promising results. The proposed approach makes extensive exploration of the temporal redundancy and is advantages in preserving details and global illumination. In addition, our algorithm is of much higher computational efficiency compare to the global optimization approaches.

The assumption that the video frames can be aligned by linear mappings is violated in some complex cases, e.g., non-rigid motion, multiple moving objects. Inpainting these videos needs combination with explicit alignment techniques robust to occlusion and will be studied in the near future.

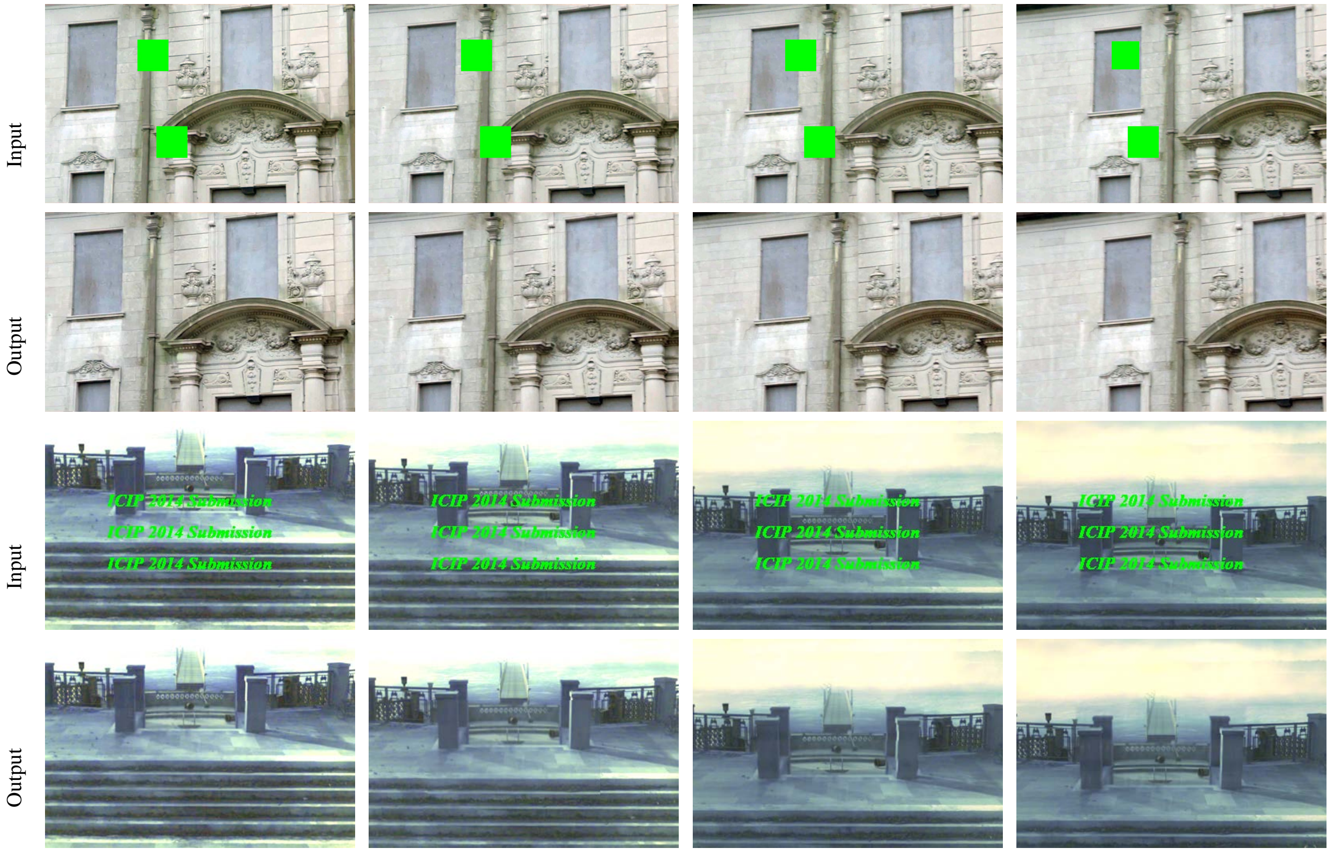
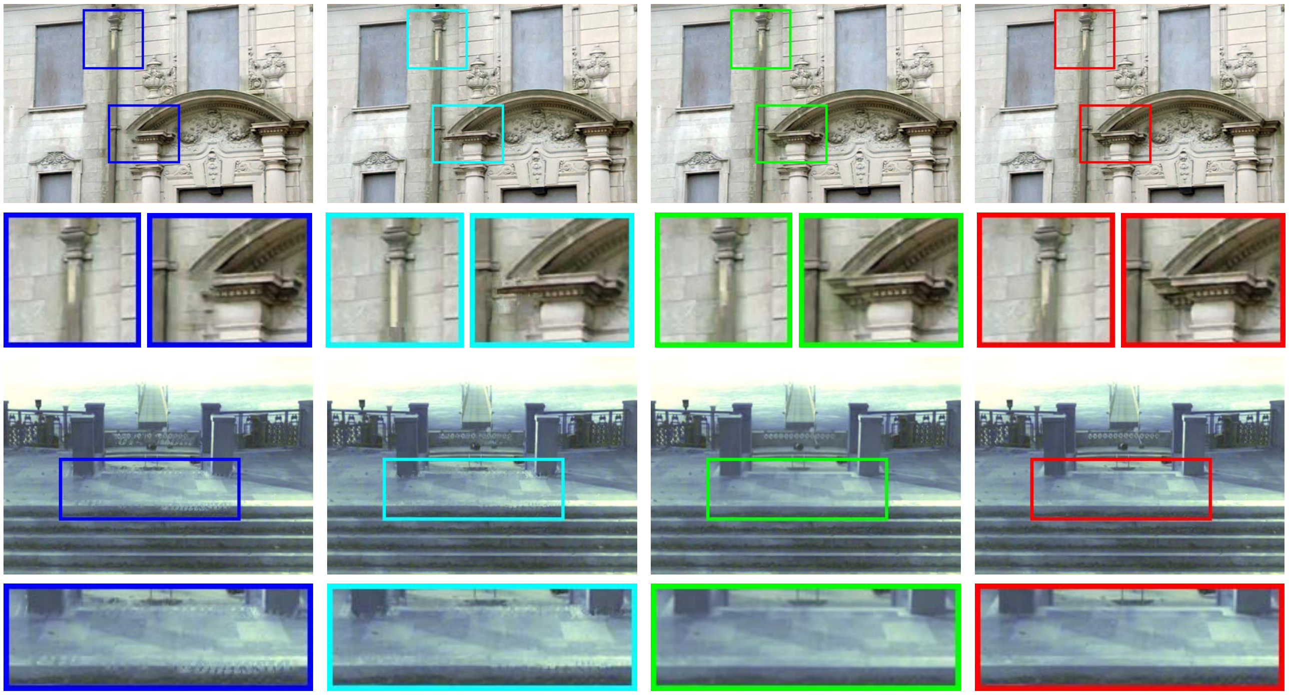


Fig. 3. Inpaint result of another two different cases. top: partial damage/occlusion; bottom: subtitle occlusion.



By Wexler et al.[9] By Patwardhan et al.[4] Our result Ground truth

Fig. 4. Comparison with state-of-the-arts on two clips in Fig. 3. The comparison demonstrates that our approach is closest to the ground truth, and is apparently advantages in removing unwanted occlusions neatly while preserving the details. This is mainly due to that the low rank prior makes use of the temporal redundancy of the whole sequence instead of one or a subset.

5. REFERENCES

[1] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *CVPR*. IEEE, 2001, vol. 1, pp. I–355–I–362.

[2] Timothy K Shih, Nick C Tang, and Jenq-Neng Hwang, "Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity," *IEEE Trans. CSVT*, vol. 19, no. 3, pp. 347–360, 2009.

[3] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmio, "Video inpainting of occluding and occluded objects," in *ICIP*. IEEE, 2005, vol. 2, pp. II–69–II–72.

[4] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmío, "Video inpainting under constrained camera motion," *IEEE Trans. Image Processing*, vol. 16, no. 2, pp. 545–553, 2007.

[5] Mounira Ebdelli, Christine Guillemot, and Olivier Le Meur, "Examplar-based video inpainting with motion-compensated neighbor embedding," in *ICIP*. IEEE, 2012, pp. 1737–1740.

[6] Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt, "Background inpainting for videos with dynamic objects and a free-moving camera," in *ECCV*, pp. 682–695. Springer, 2012.

[7] Yun-Tao Jia, Shi-Min Hu, and Ralph R Martin, "Video completion using tracking and fragment merging," *The Visual Computer*, vol. 21, no. 8-10, pp. 601–610, 2005.

[8] Yonatan Wexler, Eli Shechtman, and Michal Irani, "Space-time video completion," in *CVPR*. IEEE, 2004, vol. 1, pp. I–120.

[9] Yonatan Wexler, Eli Shechtman, and Michal Irani, "Space-time completion of video," *IEEE Trans. PAMI*, vol. 29, no. 3, pp. 463–476, 2007.

[10] Ming Liu, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang, "Video completion via motion guided spatial-temporal global optimization," in *International Conference on Multimedia*. ACM, 2009, pp. 537–540.

[11] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. PAMI*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[12] Tao Ding, Mario Sznaier, and Octavia I Camps, "A rank minimization approach to video inpainting," in *ICCV*. IEEE, 2007, pp. 1–8.

[13] Chih-Hung Ling, Yu-Ming Liang, Chia-Wen Lin, Yong-Sheng Chen, and H-YM Liao, "Human object inpainting using manifold learning-based posture sequence estimation," *IEEE Trans. Image Processing*, vol. 20, no. 11, pp. 3124–3135, 2011.

[14] Chih-Hung Ling, Chia-Wen Lin, Chih-Wen Su, Yong-Sheng Chen, and H-YM Liao, "Virtual contour guided video object inpainting using posture mapping and retrieval," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 292–302, 2011.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE