Video Stabilization and Completion Using Two Cameras

Jie Zhou, *Senior Member, IEEE,* Han Hu, and Dingrui Wan

***Abstract*—Video stabilization is important in many application fields, such as visual surveillance. Video stabilization and completion based on a single camera have been well studied in recent years, but it remains a very challenging problem. In this paper, we propose a novel framework to produce a stable high-resolution video for visual surveillance by using two cameras, in which one static camera serves to capture low-resolution wide-view-angle images, and the other is a pan-tilt-zoom camera to capture high-resolution images. Different with using a single camera, the interesting target can be detected and tracked more effectively and much more high-resolution information can be utilized for the stabilization and completion by using two videos from two cameras. A three-step stabilization approach is designed to deal with the resolution's discrepancy between two synchro videos and a four-stage completion strategy is taken to utilize more high-resolution information. Experimental results show that the proposed algorithm has a satisfying performance.**

***Index Terms*—High-zoom video, video completion, video stabilization, visual surveillance.**

I. Introduction

HIGH-RESOLUTION videos are useful and important for visual surveillance. Compared with low-resolution ones, high-resolution videos can provide more detailed information which can be used for object identification, behavior and activity analysis, as well as security evidence collection. However, due to the movement of targets and camera, many original high-resolution videos are unstable and the interesting objects might be incomplete in the view. Thus, it is needed to reproduce stable high-resolution videos from original unstable ones.

Single static camera is unsuitable for capturing high-resolution video with moving targets, because image resolution conflicts with the scope of field of view (FOV). Camera should be kept in a low-zoom level to maintain the target staying in FOV. Using a single active camera, such as a pan-tilt-zoom (PTZ) camera, could solve the above conflict by changing its view angle [1]–[5]. However, high resolution video captured by an active camera with a high zoom level is usually unstable and incomplete, because: 1) a same angular speed of camera movement will cause faster pixel movement for high-zoom capturing than that for low-zoom capturing, and 2) both automatical and manual camera controls are prone to cause overshoot and undershoot due to the time delay in mechanical movement.

In this paper, we propose a system to produce a stable high-resolution video by using two cameras, in which one static camera captures a wide-view-angle video with a low resolution (low-zoom) but a large FOV (i.e., wide view angle), and the other one is a PTZ camera to capture high-resolution images at a high zoom value. The active PTZ camera is controlled by either the wide-view-angle camera or manual operation. Since the discrepancy in resolution between two synchro videos might increase the registration difficulty, we propose a three-step stabilization approach to deal with it. In order to make full use of the high-resolution information, we propose four types of image completion strategies: current high-resolution image inpainting; high-resolution background model inpainting; sample patch with motion field based foreground inpainting and current scaled low-resolution image inpainting. Compared with the systems of using a single PTZ camera, this configuration has the following advantages for high-resolution video stabilization and completion.

1) The interesting target can be easily segmented, detected, and tracked in the static low-resolution wide-angle views. Then by registering the high-resolution image to the low-resolution image, the task of stabilization and completion is much easier even when the correspondences among successive high-zoom images are failed to calculate.
2) By using the low-zoom views as a bridge, much more high-spatial-resolution information can be found for the inpainting, which is difficult or impossible in high-zoom views directly; furthermore, the low-resolution image information from the static camera can serve as the safeguard to guarantee the integrity of the output video, when there is no available high-resolution information.

This paper is organized as follows. Section II describes an overall framework of the proposed system. In Section III, the details of video stabilization are discussed. From Sections IV to VI, the steps of completing are described. The experimental results are provided in Section VII. In Section VIII, we summarize this paper with some conclusions.

Manuscript received June 4, 2010; revised March 4, 2011; accepted April 18, 2011. Date of publication May 12, 2011; date of current version December 7, 2011. This work was supported by the Natural Science Foundation of China, under Grants 61020106004, 61021063, and 60721003, and by the Research Fund from the Ministry of Communication of China. This paper was recommended by Associate Editor T. Fujii.

The authors are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: jzhou@tsinghua.edu.cn; huh04@mails.tsinghua.edu.cn; wandingrui00@mails.tsinghua.edu.cn).

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/TCSVT.2011.2154810

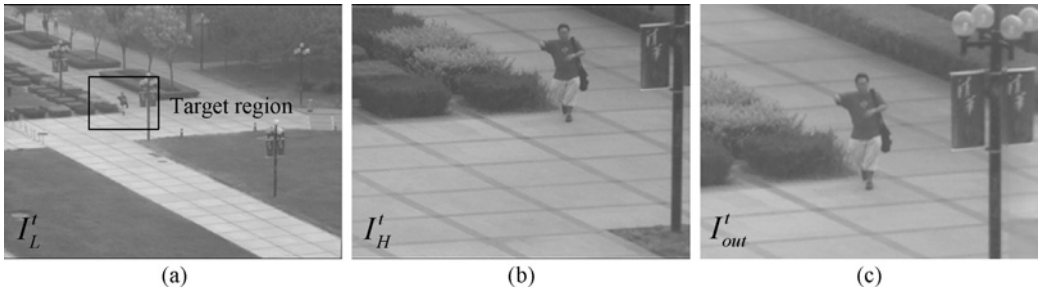


Fig. 1. Example: (a) I_L^t with the target region, (b) I_H^t, and (c) optimal output, I_{out}^t. $k_o = 5$.

II. Framework Overview

We denote I_L^t and I_H^t as the low-resolution and high-resolution image at the tth frame, I_{out}^t as the output image. The video stabilization and completion has three goals as follows.

1) The center of interesting target should be kept near the image center of resulted video, and the target's motion should be smooth.
2) The image of each frame should be intact, i.e., there is no unfilled part in the image.
3) The resulted video should contain as more high-resolution information from the high-resolution camera as possible rather than low-resolution contents from the low-resolution camera.

The FOV of I_{out}^t corresponds to a rectangle region in I_L^t, which is called the target region. We first determine the target region (the initial target region is assumed as known, which can be marked manually by the human inspector or determined by using automatic human detection and behavior analysis technologies), and then fill it with high-resolution information as more as possible. The scale between I_{out}^t and the target region is called the "output magnification factor" which is denoted by k_o. An example is provided in Fig. 1.

We denote M_{LH}^t as the mapping model between I_L^t and I_H^t. If M_{LH}^t is known, the completion can be achieved by warping the high-resolution images into I_{out}^t. In many cases, I_H^t cannot fill in all pixels in I_{out}^t, when the FOV of I_H^t does not cover the target region, or I_H^t is considered to be invalid due to blurriness. So image completion should be carried out to keep the output video intact.

The flowchart of the proposed algorithm is shown in Fig. 2. The main procedures of stabilization (M_{LH}^t's estimation) and completion (region inpainting) are described as follows.

Stabilization: M_{LH}^t's estimation

1. *feature-based method is used to calculate a rough affine model between I_L^t and I_H^t;*
2. *pixel-based alignment is adopted to refine the model; and*
3. *neighborhood information is used to smooth the model.*

Completion: region inpainting

1. *foreground and background segmentation in I_L^t;*
2. *estimating high-resolution background I_{HB}^t using I_H^i, $i = 1, 2, \cdots, t+N$ and M_{LH}^i;*
3. *step-by-step inpainting according to their priority levels:*
   1) *inpainting with current I_H^t and M_{LH}^t;*
   2) *for background region, use high-resolution background I_{HB}^t;*
   3) *for foreground region, use sample patch based motion inpainting algorithm;*
   4) *for other non-filled region, inpainting with the interpolated I_L^t;*
4. *post-processing.*

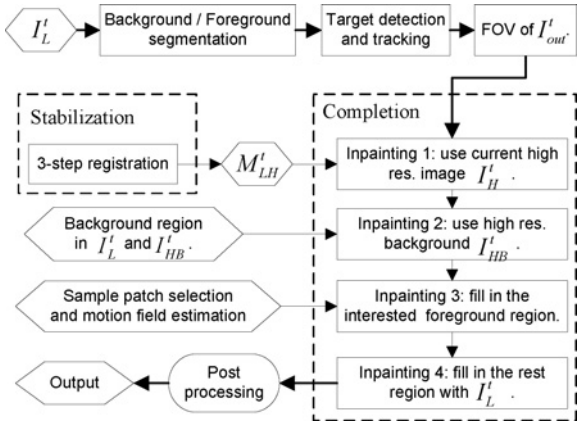


Fig. 2. Flowchart of the proposed algorithm.

III. Stabilization

In our study, we align images from two cameras at each time-stamp. The main difficulties of the algorithm are as follows: 1) the pose parameters of cameras are usually unknown, as a result, the searching range for the registration algorithm might be huge without prior-knowledge; 2) for different FOVs, a same camera may do different illuminant adjustments, which may cause intensity gaps; and 3) it is hard to obtain a quite precise registration model because of the large discrepancy in image resolution.

Traditional image registration methods can be mainly classified into two categories: feature-based approaches and pixel-based approaches (also called the "direct method"). References [6]–[8] made some extensive reviews and comparisons. In these researches, feature-based approaches are regarded to be less accurate than the pixel based ones, because the distribution

of feature points in the overlapping region is unpredictable. On the other hand, pixel-based approaches usually need a good initial model and demand that the intensities in two images to be comparable, which might not hold for real applications [6]–[8]. Here, we will propose a method to combine feature-based approach and pixel-based approach. We first use feature-based approach to get a coarse model, which can be used to do intensity adjustment and as the initial model for pixel-based ones. In this way, we can overcome the shortcomings of these two methods.

Roughly speaking, there are two objectives for video stabilization: 1) the interesting target should be located near image center, and 2) the target's motion should be as continuous and smooth as possible. We use the mean-shift tracking algorithm [9] to obtain the trace of the interesting object. In order to obtain more accurate locations of the interesting target and smooth the variation of the target's location, we average the centers of the interesting object within 50 neighboring frames to decrease the computational errors. The mean center is set to be the center of target. Since this system is designed for visual surveillance, the size of the same interesting object does not change a lot in the low-resolution camera. So the size of target region can be set as constant. The output magnification factor, k_o, is about 5 in our experiments.

Since I^t_{out} and the target region only have one scaling relation with scaling factor k_o, we calculate the mapping model (M^t_{LH}) between I^t_H and I^t_L instead. For long-distance surveillance, the disparity between two views can be neglected, because the baseline width is much smaller than the distance of the scene to the cameras (it should be noted that the affine model will be not accurate enough when the target object is near the camera). Actually, if we assume that the distance between the target and camera is about 100 m; zoom levels of two cameras are less than 10, which is the highest zoom factor of the high-resolution camera in our experiments); the baseline is about 0.4 m and the depth varies more than $20m$, the corresponding disparity varies only 1-pixel [10]. Therefore, we can choose an affine model for the mapping from I^t_H and I^t_L.

We utilize a three-step algorithm to estimate the registration model. First, we use the sparse feature points matching method to get a rough registration model, which can be used as an initial guess for the following refinement. The rough model also provides a rough overlapping FOV in which the intensity mapping between two images will be estimated to solve the intensity inconsistence problem. Then, a refined model can be obtained by using the pixel-based approaches. Finally, we adopt a post process to smooth the refined model using neighboring high-resolution images to improve the stability of the estimated model among frames.

A. Step 1: Rough Model Estimation

Since the zoom ratio between I^t_L and I^t_H is unknown, we choose scale-invariant feature transform (SIFT) [11] feature descriptor for the registration. We only compute these key points in the target region of I^t_L to reduce computation. In key points matching, the approximate nearest neighbors kd-tree package [12] is utilized. The random sampling consensus (RANSAC) [13] strategy is employed to estimate an affine model M^t_{LH1} from I^t_L to I^t_H by matching these points (the subscript "1" indicates that it is a rough model). If the number of matches is less than 10, we set M^t_{LH1} to be invalid, and skip the next two steps. A matching example is shown in Fig. 3(a).

We use a 1-D mapping ($[0, 255] \rightarrow [0, 255]$) to make the intensities in I^t_L and I^t_H comparable, so that most traditional pixel-based image registration methods can be applied. The convex hull of matched key points is defined as the testing region for each image. A histogram equalization method is used to compare the two cumulative intensity histograms in testing regions. We choose a three-piece linear mapping model. The middle part contains 90% pixels [see Fig. 3(b)]. Fig. 3(c) shows an example of both original and adjusted intensity histograms.

B. Step 2: Refined Model Estimation

The rough model, M^t_{LH1}, estimated in Step 1, is used as an initial value in the iterative algorithm of pixel-based estimation (direct method). In order to reduce the computation, first, we convert I^t_H into $I^t_{H\_adj}$ via the reverse transform of M^t_{LH1}. The registration model between $I^t_{H\_adj}$ and I^t_L should be close to a 3×3 identity matrix. Second, the intensity of I^t_L is adjusted according to the intensity mapping model, and we denote it by $I^t_{L\_adj}$. After that, the gradient based Hessian matrix is utilized to iteratively solve the following optimization problem [14]:

$$M_I = \arg\min_M \sum_i \left\| I^t_{H\_adj}(Mx_i) - I^t_{L\_adj}(x_i) \right\| \tag{1}$$

where M is a 3×3 affine matrix with an initial value $M_0 = I_{3\times3}$. The range with respect to the summation is the target region in $I^t_{L\_adj}$. In our system, M_I will be considered to be invalid, if it does not satisfy the two constraints: 1) rotation and scale constraint: $\|R^M_{2\times2} - I_{2\times2}\|_\infty < 0.3$, and 2) translation constraint $\|t^M_{2\times1}\|_\infty < 4$, where $\left[R^M_{2\times2}\ t^M_{2\times1}\right]$ is the first two rows of M_I. If M_I is valid, we have the refined registration model as $M^t_{LH2} = M^t_{LH1} M_I$; otherwise, M^t_{LH2} is also invalid, and the next step will be skipped. Fig. 3(d) shows an example of warping I^t_H onto I^t_L via the estimated M^t_{LH2}.

C. Step 3: Model Smoothing

Considering the uncertainty of M^t_{LH2} mentioned above, we smooth the refined registration model to improve the stability. The final smoothed model is denoted by M^t_{LH}.

Take the ith frame for example. We consider $2N + 1$ neighboring frames ($N = 5$ in our experiment). We denote $j = i - N, i - N + 1, \cdots, i + N$ as the indexes of neighboring frames, M^j_{LH2} as the refined model at the jth frame, and M^i_j as the homographic model from I^j_H to I^i_H. The smoothed model M^i_{LH} can be computed by

$$M^i_{LH} = \sum_{j=i-N}^{i+N} \omega_j \delta_j M^i_j M^j_{LH2} \tag{2}$$

where ω_j is Gaussian weight and δ_j is the characteristic function satisfying

$$\delta_j = \begin{cases} 1, \text{if } M^i_j \text{ and } M^j_{LH2} \text{ are both valid} \\ 0, \text{otherwise} \end{cases} \tag{3}$$

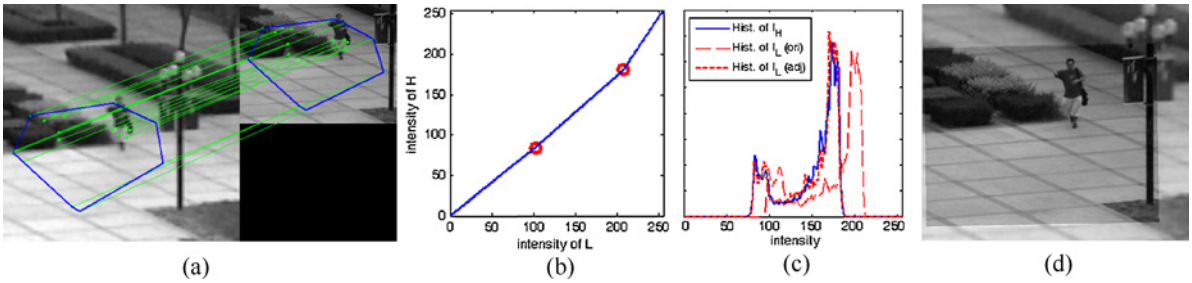


Fig. 3. Image registration result for the images in Fig. 1. (a) Feature points matching with the left image is the magnified target view of I_L^t with a magnification factor 5, and the right one is I_H^t. (b) Calculated piece-wise linear mapping model. (c) Intensity histogram in (polygonal) testing region of I_H^t and I_L^t (original and adjusted). (d) Warping I_H^t onto I_L^t by the refined mapping model, M_{LH2}^t.

and

$$\sum_{j=i-N}^{i+N} \omega_j \delta_j = 1. \tag{4}$$

Physically, the smoothing method means that, for a stationary point P, assume p_H^j is the image coordinates in I_H^j; $p_H^{j,i}$ is the transformed point from p_H^j by M_j^i. After model smoothing, the final location of P in I_H^i will be the Gaussian average of all $p_H^{j,i}$.

The only unknown parameter in (2) is M_j^i. Since I_H^j and $I_H^i (i \neq j)$ are captured at different time, foreground motion, especially independent movement may affect the precision of M_j^i calculation. So we remove the foreground in I_H^t in advance. This procedure can be done easily. While tracking the object in I_L^t, we use the running average method [15] to obtain the background model. So the foreground region in I_L^t can be detected. The corresponding foreground in I_H^t can be also located by the refined model M_{LH2}^t. If M_{LH2}^j is invalid, $\delta_j = 0$, and therefore, there is no need to calculate M_j^i.

The alignment between two images only concerns background image regions. We use the SIFT features again (which have already been extracted in the previous procedures) to estimate an affine homographic model, M_j^i. Note that, as we take more frames' information into consideration, even when the estimation of some M_j^i fails (e.g., too few matched points), the smoothing algorithm can still work. In the worst case, $M_{LH}^i = M_{LH2}^i$. If M_{LH2}^i is invalid, the smoothing step will be skipped.

Blurring often happens in high-resolution image sequence due to fast camera movement. Severe blurriness can intensely affect the estimation of M_{LH}^t and M_t^i. Sometimes these frames might also yield valid M_{LH}^t. However, these high-resolution information is not what we need. Since the absolute blurriness is difficult to calculate, we use the relative blurriness [1], that is

$$b_t = \frac{1}{\sum_{p_t} \left[dx^2(p_t) + dy^2(p_t) \right]} \tag{5}$$

where $dx(\cdot)$ and $dy(\cdot)$ are gradients along x-direction and y-direction. The greater the gradient, the smaller the relative blurriness will be. We only consider p_t in the background region of I_H^t. The tth frame will be considered to as blurred, if $b_t > 1.3 * \min\{b_{t-1}, b_{t+1}\}$. In this case, we set M_{LH}^t invalid.

IV. Completion

The goal of completion is to obtain a complete video output. To this end, we have designed a four-step strategy: 1) direct inpainting with current high-resolution image; 2) background inpainting with the updating high-resolution mosaic background; 3) foreground inpainting based on a reference sample patch and the corresponding motion field; and 4) inpainting with the scaled low-resolution wide-view-angle image for the remainder regions. After that, a post-processing step is taken to remove the artifacts between blocks.

The current high-zoom image is the best source to fill in I_{out}^t. However, I_H^t might not cover all pixels in I_{out}^t. This could happen when: 1) the FOV of I_H^t does not cover all the target region; 2) I_H^t is considered to be invalid because of large blurriness; or 3) M_{LH}^t is invalid. So it is necessary to consider using different source image information to fill in I_{out}^t.

Intuitively, high-resolution information and credible information has precedence over others. In our study, we propose four inpainting priority levels to complete I_{out}^t. In order to make an intuitive explanation, two examples are provided in Fig. 4 to illustrate the four kinds of inpainting. The four kinds of textures in the fourth column indicate the inpainting priority level from 1 to 4, respectively.

A. Priority-1: Direct High-Resolution Inpainting

The priority-1 inpainting is based on current high-zoom image I_H^t and its corresponding M_{LH}^t. As we discussed in the previous section, if M_{LH}^t is valid, the homography between I_H^t and I_{out}^t will be available, then we can directly warp I_H^t onto I_{out}^t, and the overlapping region in I_{out}^t will be filled by current high-resolution information.

This inpainting step will be skipped in the following cases: 1) M_{LH}^t is invalid; 2) the warped I_H^t has no overlapping region with I_{out}^t; and 3) I_H^t is severely blurred (blurriness is assessed as we discussed). In Fig. 4, region $R1$ is inpainted with this type. Fig. 4(a) has $R1$ because I_H^t contains parts of the target region. Fig. 4(b) has no $R1$ because I_H^t is considered as a blurred one.

B. Priority-2: Background Inpainting

After inpainting with I_H^t, a two-layer inpainting strategy is used: the foreground (moving objects) layer and background (static) layer [2]. The priority-2 inpainting is for background layer.

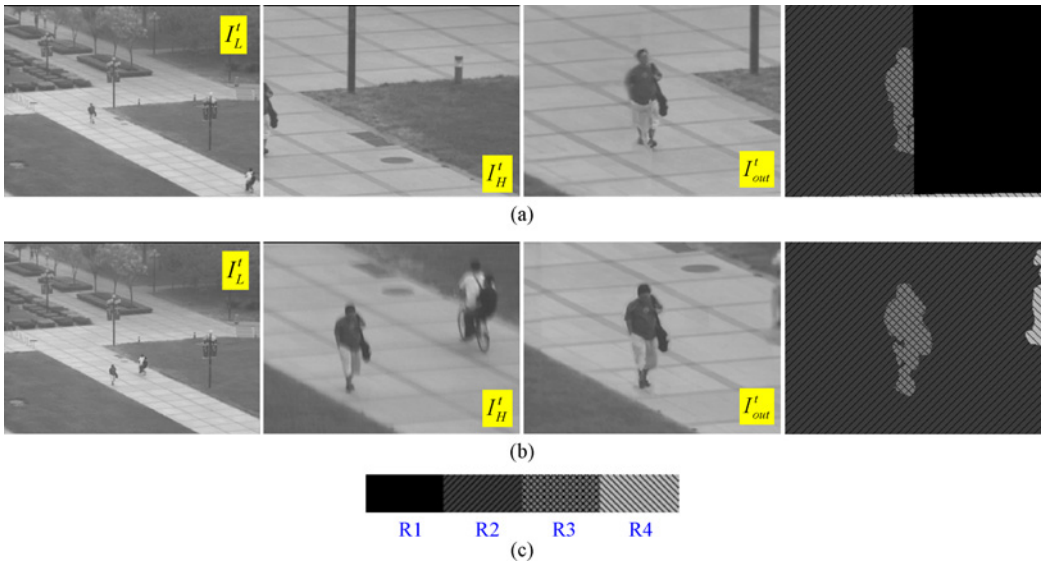


Fig. 4. Two examples to illustrate the four kinds of inpainting. (a) I_H^t does not fully contain the interesting object. (b) I_H^t is severely blurred. In (a) and (b), images from left to right are wide-view-angle image (I_L^t), high-resolution image (I_H^t), output image (I_{out}^t), and the inpainting type mask. Each type has a different mask shown in (c), where Ri, $i = 1, 2, 3, 4$ indicates the four inpainting types, respectively.

If M_{LH}^t is available, the background region in I_H^t can be obtained from I_L^t. This background information can be used to update the high-resolution background model, I_{HB}^t. I_{HB}^t contains the high-resolution background information of all the past frames and the next N neighboring frames, i.e., frame $1, 2, \cdots, t+N$. In our experiments, we set $N = 50$. The scaling factor from I_L^t to I_{HB}^t is the same as the output magnification factor, k_o.

For each high-zoom image, I_H^{t+N+1}, if M_{LH}^{t+N+1} is valid, we warp the background pixels of I_H^{t+N+1} into I_{HB}^t. An attenuation-weighted updating strategy with attenuation factor 0.5 is used to update I_{HB}^{t+1}. Fig. 5 shows a high resolution background image updated by the whole image sequence. After priority-1 inpainting, if the unfilled region in I_{out}^t contains some background pixel, we directly use the corresponding image information in I_{HB}^t to fill in. In Fig. 4, region $R2$ indicates priority-2 inpainting.

C. Priority-3: Foreground Inpainting

For the unfilled regions belonging to foreground layer, we use the reference sample patch with motion field based method to implement the priority-3 inpainting. Different from conventional image inpainting methods [1]–[3], our algorithm utilizes two image sequences with different resolution. We will describe its details in Section V.

D. Priority-4: Low-Zoom Image Inpainting

After the above three inpainting steps, some regions might still be unfilled, such as the non-interesting foreground region in Fig. 4(b) (R4), and the background which is not covered by high-resolution background model in Fig. 4(a) (R4), and so on. We use the magnified low-resolution image with bilinear interpolation to fill in these regions. The reason why we use bilinear interpolation rather than other super-resolution methods is mainly due to its low computational cost. This inpainting step can be viewed as a safeguard to maintain the integrity of the output image.

V. Foreground Inpainting

Particularly, foreground inpainting is the most difficult among the above four steps. Some relative techniques have been reported in the previous research on video stabilization and completion (using single image sequence). Image mosaicing, which is a simple way for inpainting, does not consider the non-planar scene and foreground motion [16], [17]. So it can only be used in small hole filling with small motion. Jia *et al.* [2] used a two-layer approach to inpaint foreground with the most similar patch in the previous frames. However, it needs the cyclic motion assumption. Wexler *et al.* [3] used a nonparametric sampling-based approach to deal with this problem, which divided the target patch into smaller pieces and inpainting each piece from all previous stored patches. Compared to the previous approaches, it does not need cyclic motion assumption or depend on a single frame. However, it is computationally expensive. Matsushita *et al.* [1] proposed a motion inpainting method using a neighboring patch and a local motion field. Current local motion is estimated in a neighborhood. An equivalent constraint condition is to preserve objects' boundaries. One advantage of this approach is that, it does not need a long sequence to find the best similar patch or sample pieces. Instead, it needs available neighboring frames to propagate the motion within the first order approximation of Taylor series expansion. This approximation may not hold for large motion between non-neighboring frames.

In our study, a novel method based on reference sample patch (SP) and relative motion filed is proposed to inpaint the foreground. The main difference from the previous approaches is that, an SP contains both high-resolution and low-resolution image information in previous frames, so that the foreground

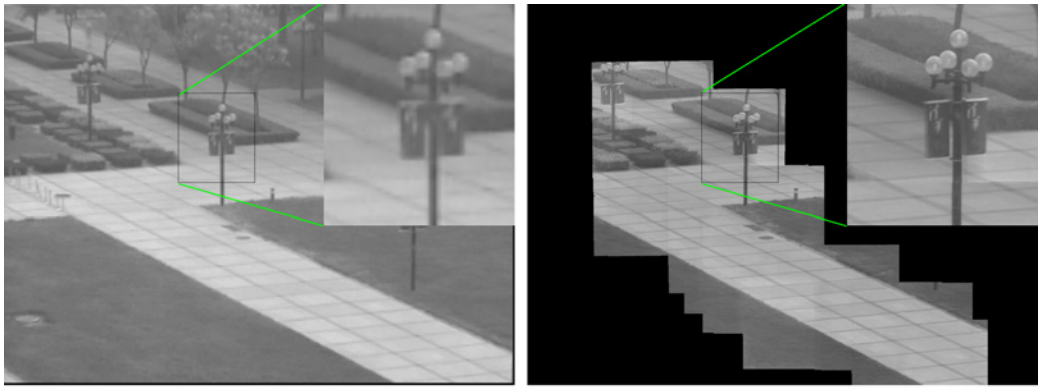


Fig. 5. (a) Magnified low-resolution background image. (b) High-resolution background image.

inpainting will be more robust and efficient, even for the case that several successive images need inpainting.

An SP is a pair of image blocks $\{\mathrm{SP}_L, \mathrm{SP}_H\}$ with the same FOV from I_L^t and I_H^t. Both SP_L and SP_H contain the whole interesting target with background removed. For long-distance surveillance, we assume that the size of interested target does not change significantly. So we fix the block size of SP_L to be 40×40, and the size of SP_H is k_o (output magnification factor) times that of SP_L. An example of SP is shown in Fig. 6(a).

The SP pool, denoted by $\{\mathrm{SP}^i\} = \{\mathrm{SP}_L^i, \mathrm{SP}_H^i\}$ (i is the index of samples), is formed by SPs from those frames satisfying the following three conditions (taking the t-frame for example): 1) M_{LH}^t should be valid; 2) I_H^t should contain the whole interesting target; and 3) I_H^t is not blurred. We model the SP pool as a FIFO queue with the size of N_{SP}. Note that if the motion of the interesting target is cyclic, it will be better to set N_{SP} to be larger than the period of the movement, so that the latest periodic motion is likely to be preserved. In our study, the period is about 25 frames (for human walking), and we set $N_{\mathrm{SP}} = 60$.

The motion field is represented by optical-flow field between the reference frame and destination frame with the same scale as SP_H. In our study, we consider not only the information of current frame with its corresponding reference SP, but also that of neighboring frames, so that both spatial accuracy and temporal continuity can be guaranteed to some extent.

Assume that the jth frame is the inpainting target. The foreground inpainting procedure includes the following three steps: 1) find a proper reference SP, i.e., SP^{ref_j}; 2) estimate the motion field, F_H^j, from $\mathrm{SP}_H^{ref_j}$ to the goal image, I_{out}^j; and 3) construct I_{out}^j by $\mathrm{SP}_H^{ref_j}$ and F_H^j with proper interpolation and post processing.

A. Producing a Reference

Take the jth frame for example. Since we know the location of target in I_L^j, we only consider the image region containing the whole target, which is denoted by $\mathrm{Sub}(I_L^j)$. We compute the similarities between $\mathrm{Sub}(I_L^j)$ and all SP_L^i $(i = 1, 2, \cdots, N_{\mathrm{SP}})$ in the SP pool. As the SP pool is timely updated, the difference in rotation and scaling can be ignored. We align two images with a translation model for simplicity, and then use the mean absolute difference (MAD) criterion to calculate the similarity.

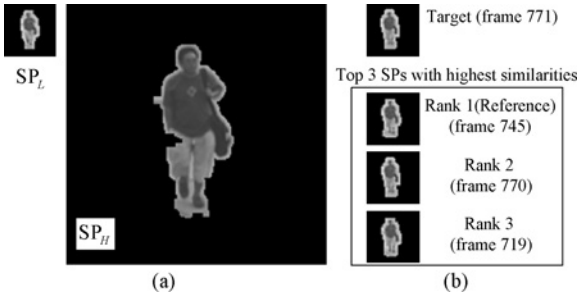


Fig. 6. (a) Example of SP. (b) Top three SPs with the highest similarities.

For a rigid object, the center of the object can be used to calculate the translation parameters. However, for a non-rigid object, the center has less consistency, such as the pedestrian. The object center may not be precise enough. Fortunately, it can be used as an initial value for iterative estimation of the translation parameters [14], [18]. This computation cost is low, since the size of image patch is very small. In order to improve the efficiency, the gradient information of each SP is pre-calculated and saved in company with SP_L.

For the ith SP, we apply the calculated translation model on $\mathrm{Sub}(I_L^j)$, and calculate the MAD score between transformed image and SP_L^i for all overlapping pixels. If the total amount of overlapping pixels is less than 60% of the foreground area of $\mathrm{Sub}(I_L^j)$ or SP_L^i, we set the MAD score to be infinity. If the smallest MAD score among all SPs is smaller than Th_{MAD} (in our experiment, $Th_{\mathrm{MAD}} = 20$), the corresponding SP will be selected as the reference patch, which is denoted by SP^{ref_j}. The corresponding translation model will be recorded as M_j^{ref}; otherwise, we deem that frame-j has no reference SP, i.e., SP^{ref_j} is invalid.

An example is shown in Fig. 6(b). We list three SPs with the smallest MAD scores. From the frame index, we can see that these similar SP_L are from the very neighboring frames or another motion period.

B. Estimating F_H^j

Assume that the jth frame needs foreground inpainting. When SP^{ref_j} is valid, we estimate F_H^j so that I_{out}^j can be recovered from $\mathrm{SP}_H^{ref_j}$ by F_H^j.

If we only use the information of frame-j and SP^{ref_j}, the problem will be simple, but have two drawbacks: 1) inter-frame information is not considered, so temporal continuity might not be preserved well, and 2) this motion field is calculated in low resolution, small error might cause large displacement in the output high-resolution image. In our study, neighboring information is used in estimating F_H^j so that both temporal continuity and spatial accuracy are considered to some extent. We propose a global optimization framework to estimate F_H^j

$$\begin{aligned}\min E = \alpha &\sum_{(x,y)\in V} \omega_1(x,y)\left[(u-u_H)^2+(v-v_H)^2\right] \\ &+ \beta \sum_{(x,y)\in V} \omega_2(x,y)\left[(u-u_L)^2+(v-v_L)^2\right] \\ &+ \sum_{(x,y)\in V}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2+\left(\frac{\partial v}{\partial x}\right)^2+\left(\frac{\partial v}{\partial y}\right)^2\right]\end{aligned} \tag{6}$$

where V is the valid image region, (x, y) is a pixel in V. u and v represent $u(x, y)$ and $v(x, y)$ for short, which are the x and y-components of F_H^j at pixel (x, y), respectively.

The first part considers the inter-frame high-resolution information. We use V^H to represent the estimated high-resolution optical-flow field which contains local relative motion with respect to $\mathrm{SP}_H^{ref_j}$, and (u_H, v_H) indicates the optical flow at (x, y). $\omega_1(x, y)$ is the weight, which is defined as $\omega_1(x, y) = \exp(-\|(u_H, v_H)\|/10)$. V^H is estimated from $\mathrm{SP}_H^{ref_i}$ $(i = j-1, j, j+1)$. We first remove the global motion (e.g., an affine model) from $\mathrm{SP}_L^{ref_{j-1}}$ and $\mathrm{SP}_L^{ref_{j+1}}$ to $\mathrm{SP}_L^{ref_j}$, respectively; then we calculate the local motion field, $V_{j,j-1}$ and $V_{j,j+1}$; finally, we take the 1-order temporal continuity assumption, i.e., $V_{j,j-1}(x, y) = -V_{j,j+1}(x, y)$, to calculate $V^H = \frac{1}{2}\left(V_{j,j-1} + V_{j,j+1}\right)$ [see Fig. 7(a)], so the temporal continuity is considered. Note that the global motion can be efficiently computed by multiplying several 3×3 matrix via $\mathrm{SP}_L^{ref_i}$ and $\mathrm{Sub}(I_L^i)$ $(i = j-1, j, j+1)$.

The second part considers the inner-frame low-resolution information. We use V^L to represent the magnified image from F_L^j using bilinear interpolation [see Fig. 7(b)]. Then it has a same resolution with F_L^j. (u_L, v_L) indicates the optical flow at (x, y) in V^L. $\omega_2(x, y)$ is the corresponding weight. In our experiment we set $\omega_2(x, y) = 1$. V^L could supply the local information with a larger scale than V^H because of the limitation of image resolution. Although this seems to be redundant, when neighboring SP is not available, V^L will play a dominant role in the estimation of F_H^j. α and β are utilized to adjust the weights of first two parts in (6). When neighboring SPs are valid, α should have a greater value, such as $\alpha = 2\beta$; otherwise, we set $\alpha = 0$, i.e., the degenerated case. In our study, we use the pyramidal Lucas–Kanade optical flow algorithm [19] to calculate V_H and V_L.

The third part considers the smoothness of the estimated F_H^j, so that spatial continuity can be guaranteed. A general way to solve this problem is to calculate the partial derivatives of (6) with respect to u and v, and then use the 3×3 Laplacian operator for discretization [20]. In our implementation, we have already considered the smooth factor in calculation of both V_H and V_L, so we ignore this part for simplicity.

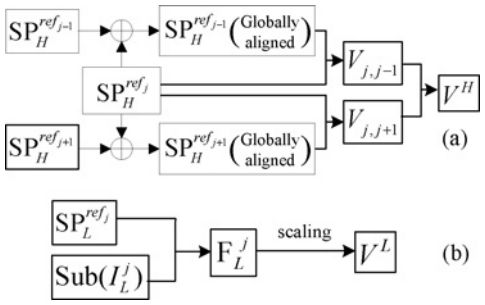


Fig. 7. Block diagram for computing (a) V_H and (b) V_L.

C. Recovering Output Image

After F_H^j is estimated, we apply this motion field on $\mathrm{SP}_H^{ref_j}$, and then bilinear interpolation is employed to fill I_{out}^j with the high-resolution foreground information. Fig. 8 shows an example of foreground inpainting.

VI. Post Processing

Post processing is needed to adjust the intensities after I_{out}^t inpainting, because even when all pixels in I_{out}^t are perfectly inpainted, the intensity might still be inconsistent in two aspects: 1) the spatial inconsistence near the junction among neighboring regions with different inpainting types, and 2) the temporal inconsistence between successive frames. This phenomenon might affect the visual effect sometimes.

There are four kinds of source inpainting information corresponding to the four inpainting types, which belong to I_H^t, $\mathrm{SP}_H^{ref_t}$, I_{HB}^t, and I_L^t, respectively. In order to smooth the intensities from one inpainting region to another, it is necessary to set a benchmark, so that we can adjust the intensity according to its inpainting type. While it is difficult to build an exact benchmark, we choose one from I_H^t, $\mathrm{SP}_H^{ref_t}$, I_{HB}^t and I_L^t as an approximation.

In our study, we take I_{HB}^t as the benchmark. I_L^t is in low resolution, and it is unsuitable to be the benchmark for high-resolution output. I_H^t and $\mathrm{SP}_H^{ref_t}$ are with a high resolution, but the image intensity is not stable when FOV changes. The high-resolution background image, I_{HB}^t, is constructed by several I_H^i, and it can be approximately regarded as an average of high-resolution images. So it is much better to choose I_{HB}^t as the benchmark.

We use different ways to adjust the intensity for different inpainting types by using the benchmark. For regions inpainted from I_H^t, we calculate the intensity mapping using the piecewise linear model. For regions inpainted from I_L^t, we use a similar method, except for the regions inpainted from both $\mathrm{SP}_H^{ref_t}$ and I_{HB}^t, where $\mathrm{SP}_H^{ref_t}$ belongs to the foreground and it has no comparable pixels with respect to I_{HB}^t.

In addition, the boundary between two regions with different inpainting types where the texture might be unsmoothed. In order to solve this problem, we define a transition region (which is dilated from this boundary with a 5×5 structuring element), and smooth it with a 3×3 mean filter. Note that, this smoothing will damage the high-resolution information, so those pixels belong to foreground and are inpainted with high-resolution information will remain unchanged.

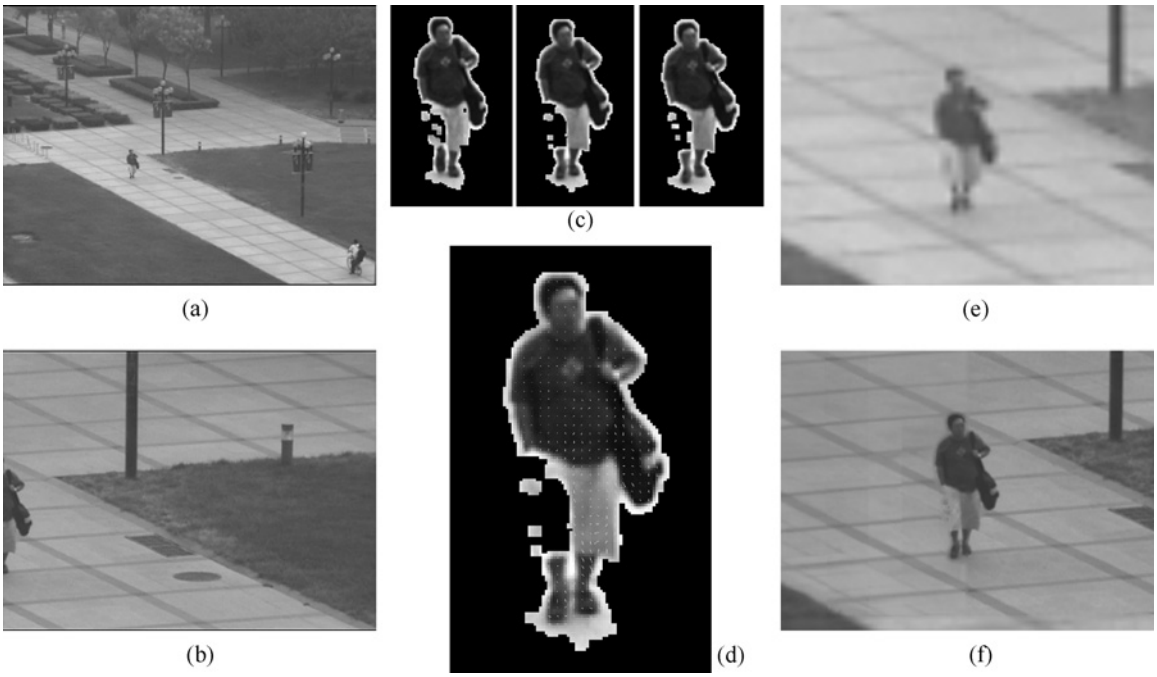


Fig. 8. Example of foreground inpainting. (a), (b) Original low and high-resolution images (at frame $j = 705$). (c) From left to right: $\mathrm{SP}_H^{ref_{j-1,j}}$, $\mathrm{SP}_H^{ref_j}$, and $\mathrm{SP}_H^{ref_{j+1,j}}$, where the reference SPs are obtained from frames 677, 679, and 680. (d) F_H^{705} with $\mathrm{SP}_H^{ref_{705,705}}$. (e) Scaled image of (a) with bilinear interpolation. (f) Final output image with foreground inpainting.

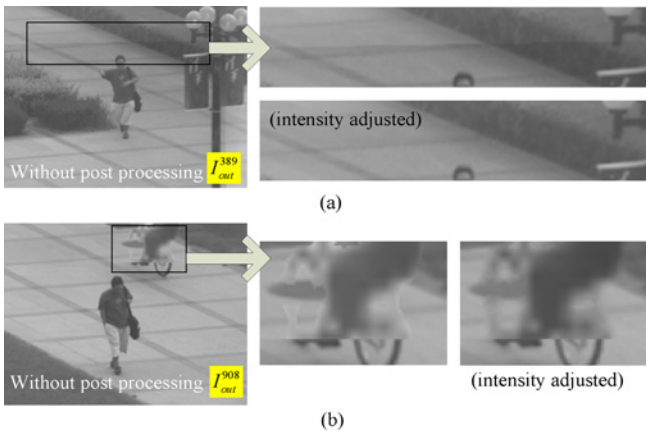


Fig. 9. Examples of post processing. (a) I_H^t adjustment. (b) I_L^t adjustment.

Fig. 9 shows two examples of post-processing. After intensity adjustment, the output images seem more clear and more sharp. In Fig. 9(a), there is a significant gap in the middle of the image before intensity adjustment, and after intensity adjustment, it is better. In Fig. 9(b), two wheels of the bicycle are clearer after intensity adjustment.

VII. Experimental Results

In our experiment, the system runs on one computer with Intel 3.0G CPU and 1.5 G memory. Two SONY EVI D70 cameras are utilized as the video capture device. The size of captured images is 320×240. We choose the outdoor scene for long-diatance surveillance. The usage of output video is for activity analysis (however, the output video can reach a higher resolution for human face recognition if using cameras with higher resolution (e.g., 1280*960). The width of baseline (distance between two cameras) is 0.4 m, the distance from target to camera center is about 100 m.

We have carried out experiments on two real data sets. Some frames from these two data sets are shown in Fig. 10. The experimental parameters are kept same for all these experiments.

The first row in Fig. 10 is low-zoom wide-view-angle images which contain the interesting targets for all frames. The second row shows the corresponding high-zoom images. In the experiments, the high-zoom active camera is manually controlled. Actually, as the interesting target is well tracked by the other camera, it is feasible to automatically control the active camera. Since the proposed approach should be tested under different situations, we manually simulate the following cases: the interesting target is invisible or half-visible in high-zoom image for some frames, and the high-zoom image is severely blurred due to fast camera movement, and so on. The output image I_{out}^n is shown in the third row, and the output magnification factor is $k_o = 5$. The size of interesting target (e.g., pedestrian) in the output image is about 20×40 pixels, which is enough for human activity analysis. The corresponding visual field is denoted by a rectangle in the first row. From the output videos, we can observe that the interesting target can be kept near the image center and the target's motion is very smooth. For quantitative demonstrations, we compute the average Euclidean distance between the interesting target and the image center, $\overline{d}$, and we also use the standard variations of target's locations relative to the image center in the horizontal and vertical directions, σ_x and σ_y, to measure the video's smoothness. In the original

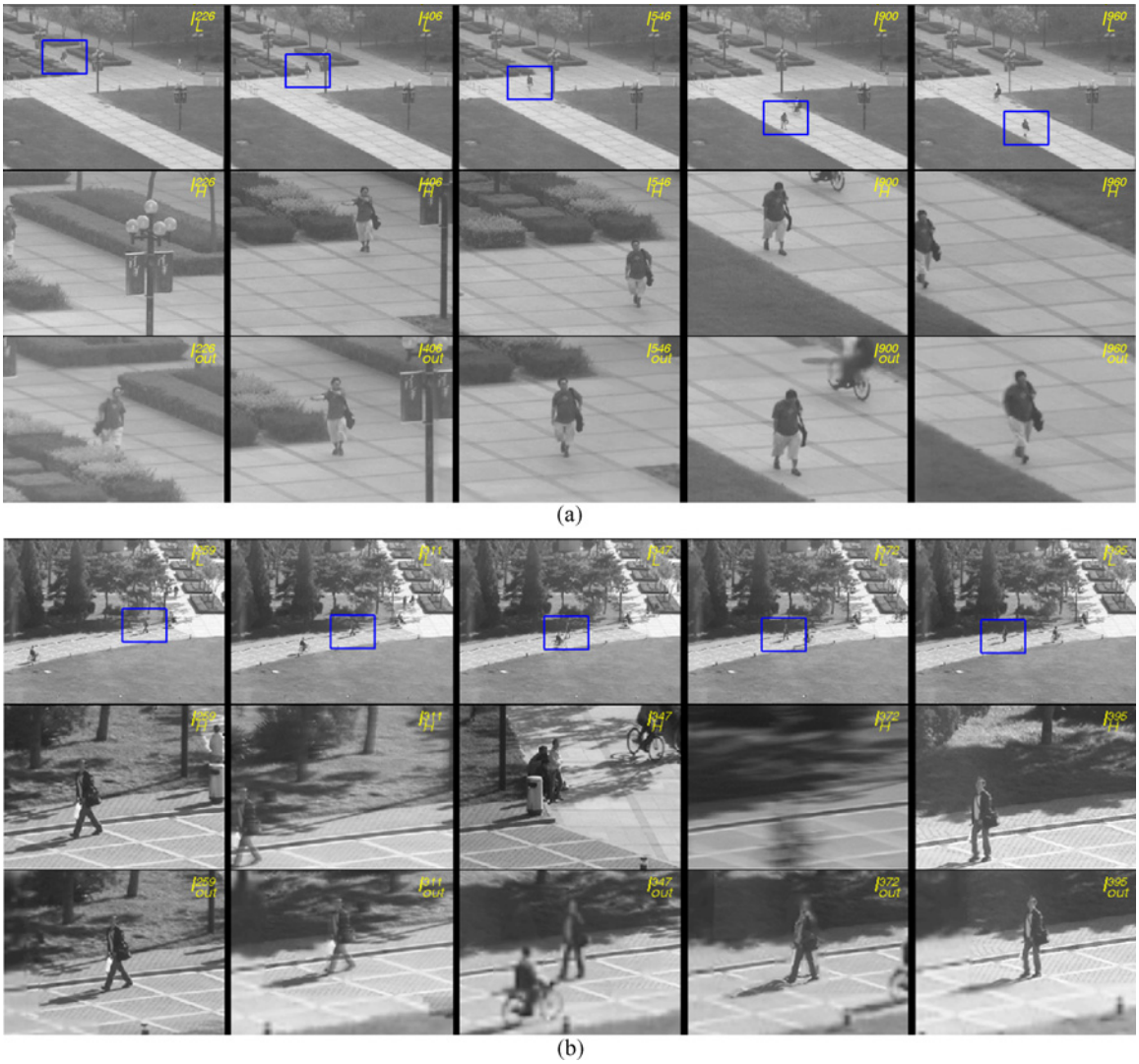


Fig. 10. Experimental results: five frames from two sets of experimental sequences, respectively, in (a) and (b). The first row is the panorama low-zoom view, I_L^n, the second is the high-zoom view, I_H^n, and the third is the output stabilized and completed image, I_{out}^n.

high-resolution videos, $\overline{d}$ = 68.8 pixels, σ_x = 64.3 pixels, and σ_y = 34.3 pixels. By using the proposed algorithm, $\overline{d}$ = 4.1 pixels, σ_x = 2.4 pixels, and σ_y = 2.3 pixels for the resulted videos.

A. Impact of Zoom Variation on Video Stabilization

The precision of estimating of M_{LH}^t is related to the scale (or zoom) ratio between I_L^t and I_H^t. For the two data sets (Data1 and Data2) in Fig. 10, the scale ratios are about 1:4.2 and 1:5.2, respectively. Generally speaking, the greater the ratio, the registration will be easier. So we only test the performance on the same data sets with smaller ratio. We manually reduce the size of I_L^t before alignment, and count the frames with valid M_{LH}^t. For the two data sets, the total numbers of frame are 1023 and 610, respectively. Table I shows the experimental result.

This experiment shows that when the discrepancy of scale ratio between I_L^t and I_H^t becomes larger, the probability of obtaining valid M_{LH}^t is likely to decrease. When the reduction factor is 0.6, the scale ratios of two data sets are about 1:7 and 1:8.7 (the size of interesting target in I_L^t is about 5×12 pixels), so the computation M_{LH}^t fails in many frames. The extreme case is that no frame has valid M_{LH}^t. It means that the relationship between the two image sequences will be unavailable. As a result, the problem degrades to the single camera video based stabilization and completion problem. When the number of frames with invalid M_{LH}^t increases, Priority-1 inpainting will be less frequently used, and the other three inpainting types will be used more often.

TABLE I
PROPORTION OF FRAMES WITH VALID M_{LH}^t WHEN DIFFERENT REDUCING FACTOR OF I_L^t IS CHOSEN

Reducing factor	Data1	Data2
1.00	0.9844	0.9016
0.90	0.9179	0.8262
0.75	0.7937	0.6098
0.60	0.1642	0.1459

Fig. 11. Testing of high-resolution inpainting. (a) Ground truth. (b) Inpainting result by assuming that M_{LH}^t is invalid. (c) Absolute difference between (a) and (b).

B. Accuracy of High-Resolution Inpainting

We select several successive frames with good I_H^t (i.e., it is not blurred and the interesting target is fully visible). We set the corresponding M_{LH}^t to be invalid so that I_H^t will not contribute to video completion, including supplying SP, updating high-resolution background model and directly inpainting. In order to quantitatively evaluate the performance, we use I_H^t to generate a ground truth. In this experiment, we chose 35 frames from Data1. One result is shown in Fig. 11: part (a) shows the ground truth which is warped from I_H^t with M_{LH}^t, and (b) shows the inpainting result. Since we only considered the accuracy of high-resolution inpainting, we compared the gray-level difference between these two images in those regions with Priority-2 and 3 inpainting method.

We define the inpainting error as the average absolute gray-level difference between the inpainted image and the ground truth per pixel. Among the 35 frames, the inpainting error of Priority-2 is 2.88, and for Priority-3, it is 9.10. The total inpainting error is 3.33, and Fig. 11(c) shows an error image. In [1], the authors also used the mean absolute difference of intensity to evaluate their method and the reported best difference is about 7.5. So, this result shows that the proposed high-resolution inpainting method is effective.

C. Comparisons with Single-Camera Based Approaches

A major difference between traditional single-camera based stabilization approaches and our framework is the definition of stabilization. Since motion segmentation is difficult for monocular active camera video, many traditional single-camera based stabilization algorithms are designed to remove high frequency camera motion. As a result, the FOV of each stabilized frame is determined by its neighboring frames. These stabilization algorithms are also called "camera motion driven." But in our framework, we constrain the interesting target should be near the image center under smoothed camera motion. The FOV of stabilized frame is determined by the trace of interesting target. So, we call this "both object and camera motion driven." This difference will cause the stabilized videos of the two categories of approaches to be very different and incomparable.

We compare the (foreground) inpainting method with one state-of-the-art video stabilization method, the motion inpainting approach [1]. Some results are shown in Fig. 12.

Motion inpainting method uses information from neighboring frames. It has two assumptions: 1) neighboring N_n frames should contain enough information for inpainting, and 2) the motion in inpainted area should coincide with that in overlapping area. If the inpainted region satisfies the above two conditions, motion inpainting based method is competent for image completion, e.g., Fig. 12(a) with $N_n = 6$; otherwise, the inpainted image could be incomplete [e.g., Fig. 12(b) with $N_n = 6$] or distorted [e.g., Fig. 12(c) with $N_n = 12$]. In the proposed framework, even if there is no reference SP found or the motion field is unable to compute (e.g., non-cyclic object motion), the case of incomplete image will never happen because of the Priority-4 inpainting.

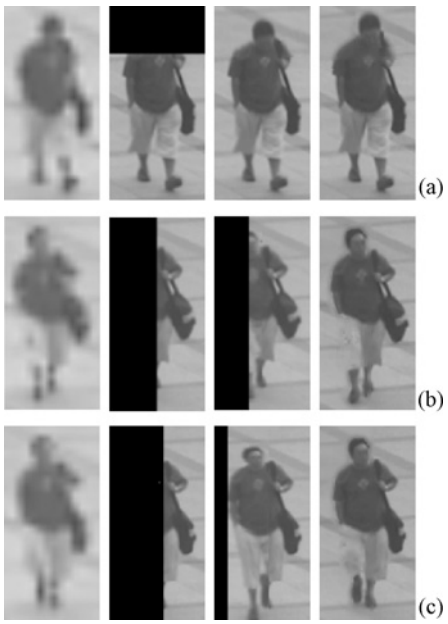


Fig. 12. Comparison between motion inpainting approach [1] and proposed foreground inpainting method. From left to right, the four images are magnified low-zoom image, original high-zoom image, the result of motion inpainting approach, and the result of our approach. (a) Both approaches work well. (b), (c) Two failed cases for motion inpainting approach.

On the other hand, in our approach, the calculation of motion field on reference SP is more like a kind of motion smoothing, but not prediction. As a result, local large-scale motion will hardly happen, and consequently, the inpainted foreground is unlikely to distort much.

Most single-camera based stabilization approaches do not consider some video segments are missing or contains totally irrelevant content. So most of these algorithms are based on the conjunctions of neighboring frames. If some conjunctions are interrupted, the stabilization will be intermitted. When moving

object is captured with high-zoom camera in long-distance surveillance, this case might happen sometimes because of the unpredictable object motion or unprecise camera control, e.g., the three frames shown in Fig. 10(b). In this case, it is unreliable to compute the global motion between neighboring frames using only high-zoom image sequence. For example, we calculate the global motion with neighboring size $N_n = 6$, the middle 65 frames are totally irrelative. This will cause both temporal and spatial discontinuity. In our framework, since the low-zoom image sequence is in use, this temporary blindness of high-zoom video will not interrupt the whole stabilization.

The experimental data (including input and output ones) can be downloaded from the website http://ivg.au.tsinghua.edu.cn/Datasets/Datasets.aspx.

VIII. Conclusion

In this paper, we proposed a new framework to solve the high-zoom video stabilization and completion problem by using a static low-zoom wide-view-angle camera and a synchro high-zoom active camera. It is very suitable for long-distance surveillance situation where the high-zoomed view is necessary.

In the proposed framework, the static view can easily provide the trace of interesting target, which will greatly facilitate video stabilization, and it will efficiently improve the accuracy of alignment among high-zoom views, which can help extracting more available high-resolution information for the completing. We designed four types of completing methods to collect as much high-resolution information as possible to fill the output video and ensure overall video integrity as well.

However, there are also some limitations of the proposed framework.

1) When the scale difference between the wide-view-angle image and high-zoom image becomes too large, the precision of the mapping model is likely to drop.
2) Although the output video is complete and smoothed by post-processing, temporal discontinuity may still exist between frames inpainted with different resolution information. It might need global spatial-temporal constraint to achieve a better visual performance. These problems will be considered in our future study.

References

[1] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1150–1163, Jul. 2006.

[2] J. Jia, T.-P. Wu, Y.-W. Tai, and C.-K. Tang, "Video repairing: Inference of foreground and background under severe occlusion," in *Proc. CVPR*, vol. 1. 2004, pp. 364–371.

[3] Y. Wexler, E. Shechtman, and M. Irani, "Space-time video completion," in *Proc. IEEE CVPR*, vol. 1. Jun.–Jul. 2004, pp. 120–127.

[4] M. Pilu, "Video stabilization as a variational problem and numerical solution with the Viterbi method," in *Proc. IEEE CVPR*, vol. 1. Jun.–Jul. 2004, pp. 625–630.

[5] C. Buehler, M. Bosse, and L. McMillan, "Non-metric image-based rendering for video stabilization," in *Proc. CVPR*, vol. 2. 2001, pp. 609–614.

[6] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, 1992.

[7] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[8] R. Szeliski, "Image alignment and stitching: A tutorial," Microsoft Corporation, Redmond, WA, Tech. Rep. MSR-TR-2004-92, 2004.

[9] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. CVPR*, 2000, pp. 2142–2149.

[10] D. Wan and J. Zhou, "Stereo vision using two PTZ cameras," *Comput. Vis. Image Understanding*, vol. 112, no. 2, pp. 184–194, 2008.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] H.-Y. Shum and R. Szeliski, "Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment," *Int. J. Comput. Vis.*, vol. 36, no. 2, pp. 101–130, 2000.

[15] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[16] M. Bertalmío, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. SIGGRAPH*, 2000, pp. 417–424.

[17] A. Levin, A. Zomet, and Y. Weiss, "Learning how to inpaint from global image statistics," in *Proc. ICCV*, 2003, pp. 305–312.

[18] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. ECCV*, 1992, pp. 237–252.

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

[20] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, 1981.

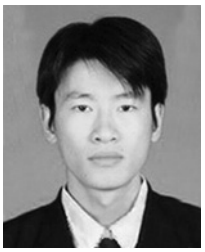
Jie Zhou (M'01–SM'04) was born in November 1968. He received the B.S. and M.S. degrees, both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995.

From 1995 to 1997, he was a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, T-IP, and CVPR. His current research interests include computer vision, pattern recognition, and image processing.

Dr. Zhou is an Associate Editor for the *International Journal of Robotics and Automation*, *Acta Automatica*, and two other journals.

Han Hu was born in June 1988. He received the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2008. He is currently pursuing the M.S. degree from the Department of Automation, Tsinghua University.

He has published three papers in peer-reviewed conferences, including CVPR and ICIP. His current research interests include pattern recognition, image processing, and computer vision.



Dingrui Wan was born in September 1981. He received the B.S. and Ph.D. degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2004 and 2009, respectively.

He is currently with the Department of Automation, Tsinghua University. He has published five papers in peer-reviewed journals. His current research interests include computer vision and pattern recognition.