A New Approach to Video Stabilization with Iterative Smoothing

Yuchi Xu, Shiyin Qin
School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
Email: xuyuchi@asee.buaa.edu.cn, qsy@buaa.edu.cn

***Abstract*—A new approach to video stabilization is proposed in this paper. Firstly, SURF (Speeded-Up Robust Features) is employed to extract feature points from input frames. Then, the local feature matching strategy is presented to speed up the global motion estimation procedure. Secondly, a weak Gaussian smoother is constructed to smooth the camera motion iteratively. In order to make the weak smoother adapt to different videos, the acceleration of feature trajectory is calculated to control the iteration of smoothing. Finally, a series of comparisons with other methods, such as spline interpolation, Kalman filter, are carried out to demonstrate the well performance of the proposed approach, which also depend less on the selection of parameters.**

***Keywords*—video stabilization; iterative smoothing; feature trajectory; global motion estimation**

I. Introduction

Hand-held digital camera becomes more and more popular with the development of camera technology. However, camera shake is one of the biggest factors which reduce the quality of captured videos. Therefore, to enhance the quality of these videos, video stabilization, removing unwanted camera jitter, plays an important role.

Usually, motion estimation and motion compensation are two crucial components of video stabilization approach. Motion estimation aims estimate the global motion between different frames. The concept of motion compensation refers to methods that reduce the unwanted camera jitter by removing high-frequency fluctuations.

Nowadays, many cameras are equipped with hardware stabilizers, but such stabilizers are not sufficient to remove unwanted camera jerks from the videos taken by non-professional users [1]. Some online methods such as Kalman filter [2][3] were introduced to solve this problem. The approaches based on Kalman filter performed well just when precise motion model of the camera carrier can be easily constructed, which is however not the case for hand-held cameras. Thus, offline video stabilization algorithms are still required for making these videos steady [1]. In typical offline approaches such as spline interpolation [4][5], the selection of knots will affect the result of video stabilization. It might fail to stabilize the video when the shake is severe. Moreover, though a group of knots fit one video sequences well, they may not fit others. In addition, low pass filter with fixed parameters [6] might lead to the problem of over-smooth or under-smooth.

978-1-4244-5900-1/10/$26.00 ©2010 IEEE

In this paper, a new video stabilization algorithm is proposed. In order to estimate the camera motion accurately, affine model is adopted and SURF [7] is used for its accuracy and robustness in different conditions. Extracted robust features are matched by considering the space distance and the difference of descriptors. After utilizing RANSAC [8] algorithm, the globe motion is estimated and the robust feature trajectories can be constructed. For the smoothing purpose, a weak Gaussian smoother is constructed to smooth the motion parameters frame by frame iteratively. The acceleration of feature trajectory, which can reflect the smoothness of the video, is calculated to control the iteration of smoothing procedure before each smoothing iteration. Finally, the video is stabilized by frame warping with the compensation matrix.

II. Global Motion Estimation

A. Inter-frame Parametric Motion Model

The relationship that map pixel coordinates between two adjacent frames can be described as a kind of mathematical transformation, and it can be expressed as follow:

$$\mathbf{p}_t^i = \mathbf{T}_{t-1}^t \mathbf{p}_{t-1}^j, \tag{1}$$

where $\mathbf{p}_{t-1}^j = [x_{t-1}^j, y_{t-1}^j, 1]^T$ is a pixel point coordinates in frame $t-1$ and $\mathbf{p}_t^i = [x_t^i, y_t^i, 1]^T$ is corresponding point in frame t. In this paper, we consider $\mathbf{T}_{t-1}^t$ as the affine transfer matrix between frame $t-1$ and frame t. It can be written as follow:

$$\mathbf{T}_{t-1}^t = \begin{bmatrix} \mathbf{A} & \mathbf{\Delta} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \Delta_{t-1,x}^t \\ a_{21} & a_{22} & \Delta_{t-1,y}^t \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{A}$ is a 2×2 non-singular matrix. $\Delta_{t-1,x}^t$ and $\Delta_{t-1,y}^t$ represents the horizontal and vertical displacement respectively.

B. Feature Extraction

In this paper, SURF (Speeded-Up Robust Features) [7] is adopted to extract the features in the frames. SURF is a fast and robust scale- and rotation-invariant interest point detector and descriptor that can be used in different computer vision tasks such as object recognition and 3D reconstruction [7]. Compared with SIFT [9], SURF uses integral images and box filter to speed up the process of feature detection, and uses only 64 dimensions descriptor to reduce the computation complexity of descriptor constructing and feature matching.

C. Feature Matching

According to the method proposed by H.Bay *et al.*[7], the interest points are found at blob-type structures. The sign of the Laplacian (i.e. the trace of Hessian matrix) distinguishes bright blobs on dark backgrounds and vice versa. For speeding up the feature matching stage, comparing only happened between the features having the same type of contrast. Fig.1 shows the condition that features don't match.

Fig. 1. If the contrast between two interest points is different, these two points are not considered a valuable match[7].

Feature matching in video stabilization field is different from others, such as 3D reconstruction and image registration. Because adjacent frames in a video are very similar, searching region can be constrained in a circular neighborhood with radius γ around the interest point to speed up the feature matching procedure.

To search the corresponding feature point of $\mathbf{p}_{t-1}^{j}$ in frame t quickly, the local feature matching procedure is presented. This procedure can be divided into four steps as follow:

Step 1: Find an unselected feature point $\mathbf{p}_t^i$ in frame t. If $\mathbf{p}_t^i$ and $\mathbf{p}_{t-1}^j$ has the same sign of Laplacian, go to Step 2. Otherwise, repeat Step 1. If there is no unselected feature point, go to Step 4.

Step 2: Calculating the Euclidean distance between $\mathbf{p}_t^i$ and $\mathbf{p}_{t-1}^j$.

$$dis(\mathbf{p}_t^i, \mathbf{p}_{t-1}^j) = ||\mathbf{p}_t^i - \mathbf{p}_{t-1}^j||. \tag{3}$$

If $dis(\mathbf{p}_t^i, \mathbf{p}_{t-1}^j) < \gamma$, go to Step 3. Otherwise, go to Step 1. Here $\gamma = \sqrt{(\frac{FrameWidth}{10})^2 + (\frac{FrameHeight}{10})^2}$.

Step 3: Calculating the Euclidean distance between the descriptor of $\mathbf{p}_t^i$ and the descriptor of $\mathbf{p}_{t-1}^j$. If the value is smaller than the previously obtained minimal distance, set minima equals to current value. Otherwise keep the minima. Then, go to Step 1.

Step 4: If the distance between $\mathbf{p}_t^k$'s descriptor and $\mathbf{p}_{t-1}^j$'s is the minima and it is below the threshold τ, the point $\mathbf{p}_t^k$ is $\mathbf{p}_{t-1}^j$'s corresponding feature in current frame t. Otherwise, there is no such feature point match $\mathbf{p}_{t-1}^j$ in frame t.

After applying this feature matching strategy to all feature points in frame $t-1$, a lot of point correspondences can be extracted between frame $t-1$ and t.

D. Computing Motion Parameters and Constructing Feature Trajectories

For the affine transformation, which has six degrees of freedom, the parameters can be computed from three point correspondences. The transfer relationship described by (1) can be written as:

$$\begin{bmatrix} 0 & x_{t-1}^j \\ 0 & y_{t-1}^j \\ 0 & 1 \\ -x_{t-1}^j & 0 \\ -y_{t-1}^j & 0 \\ -1 & 0 \end{bmatrix}^T \begin{bmatrix} a_{11} \\ a_{12} \\ \Delta_{t-1,x}^t \\ a_{21} \\ a_{22} \\ \Delta_{t-1,y}^t \end{bmatrix} = \begin{bmatrix} -y_t^i & x_t^i \end{bmatrix}^T. \tag{4}$$

This equation shows one point correspondences, but any number of point correspondences can be added, with each point correspondences contributing two columns to the first and last matrix. If three different point correspondences are provided, the formula can be solved and $\mathbf{T}_{t-1}^t$ will be known.

After feature extraction and feature matching, plenty of point correspondences are obtained. Any 3 different point correspondences can provide a solution for (4). But some false matches still exist in the extracted features. To obtain the accurate globe motion, RANSAC (RANdom SAmple Consensus) [8] algorithm is used to select the inliers of point correspondences in this paper. Then, Levenberg-Marquardt algorithm is adopted to find out the optimal affine transformation $\mathbf{T}_{t-1}^t$ between the two consecutive frames in the inliers of point correspondences.

When the RANSAC algorithm is adopted to improve the accuracy of globe motion parameters, the false matches will be discarded. It is certainly possible to concatenate point correspondences frame-by-frame to obtain long range feature trajectories after applying algorithm mentioned above.

III. Motion Compensation

Intentional camera motion is usually smooth with slow variations. On the other hand, unwanted camera jitter involves rapid variations over time. The feature trajectories in casual videos are usually winding, which are illustrated on left one of Fig.2. The feature trajectories in stabilized video, which are illustrated on right one of Fig.2, should go through the video smoothly and still keep their own trend path. In this paper, the weak Gaussian kernel smoother is constructed to smooth the camera motion iteratively and acceleration of the feature trajectory is adopted to control the iteration of the smoothing.

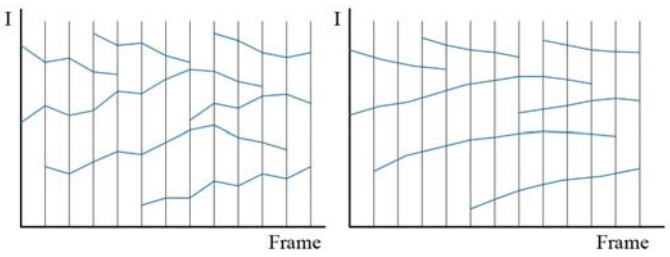

Fig. 2. Left: The feature trajectories in shaky hypothetical video sequence. Right: The feature trajectories in stabilized hypothetical video sequence.

Firstly, the motion parameters need to be smoothed should be extracted from affine transformation matrix. The jitter in a casual video is usually caused by rigid body rotation θ and 2D movement $\mathbf{\Delta}$. According to the method proposed in [10], the matrix $\mathbf{A}$ in (2) can be decomposed as follow:

$$\mathbf{A} = \mathbf{R}(\theta)(\mathbf{R}(-\phi)\mathbf{D}\mathbf{R}(\phi)), \tag{5}$$

where the $\mathbf{R}(\theta)$ and $\mathbf{R}(\phi)$ are rotations by θ and ϕ respectively. $\mathbf{D}$ is a diagonal matrix with λ_1 and λ_2 on main diagonal. $\mathbf{R}(-\phi)\mathbf{D}\mathbf{R}(\phi)$ is deformation.

The original transformation chain of the video that need to be stabilized is $\mathbf{T}_0^1, \mathbf{T}_1^2, \cdots, \mathbf{T}_{t-1}^t, \cdots, \mathbf{T}_{n-1}^n$. We denote the indices of neighboring frames as $\mathcal{F} = \{l | t-k \leq l \leq t+k, l \neq 0\}$. Then parameters need to be smoothed are $\theta_{t-1}^t, \Delta_{t-1,x}^t, \Delta_{t-1,y}^t$, and can be written as follow:

$$\mathbf{V}_{t-1}^t = \left[\begin{array}{ccc} \theta_{t-1}^t & \Delta_{t-1,x}^t & \Delta_{t-1,y}^t \end{array} \right]. \tag{6}$$

The smoothed parameter vector is denoted as $\tilde{\mathbf{V}}_{t-1}^t$, and calculated as follow:

$$\tilde{\mathbf{V}}_{t-1}^t = \sum_{j=t-k}^{t+k} \mathbf{V}_{j-1}^j \mathbf{G}[j-t], \tag{7}$$

where $\mathbf{G}[n]$ is 3×1 matrix, and each row is a discrete Gaussian kernel, which is $g[n] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}}$. In this paper, $k = 3$ and $\sigma = 1$ is used to avoid over-smooth in one smoothing iteration.

Then we reconstruct the smoothed transformation chain $\tilde{\mathbf{T}}_0^1, \tilde{\mathbf{T}}_1^2, \cdots, \tilde{\mathbf{T}}_{t-1}^t, \cdots, \tilde{\mathbf{T}}_{n-1}^n$ from smoothed parameter vector. The compensation transformation $\mathbf{C}_t$ from the original frame to the motion compensated frame is computed as follow:

$$\mathbf{C}_t = \prod_{j=t}^{1} \tilde{\mathbf{T}}_{j-1}^j \prod_{j=1}^{t} (\mathbf{T}_{j-1}^j)^{-1}. \tag{8}$$

The smoothness of the video can be reflected by accelerations of feature trajectory that illustrated in Fig.2. If the feature trajectory η_j goes through frame $t-1, t$ and $t+1$, the acceleration of the feature trajectory η_j at frame t is defined as follow:

$$\mathbf{a}_t^{\eta_j} = \mathbf{C}_{t+1}\mathbf{p}_{t+1}^i - 2\mathbf{C}_t\mathbf{p}_t^i + \mathbf{C}_{t-1}\mathbf{p}_{t-1}^i, \tag{9}$$

where $\mathbf{p}_{t-1}^i, \mathbf{p}_t^i$ and $\mathbf{p}_{t+1}^i$are corresponding feature points in frame $t-1, t$ and $t+1$, and they are all on feature trajectory η_j. $\mathbf{a}_t^{\eta_j}$ also denotes the acceleration of $\mathbf{p}_t^i$.

We just calculate the acceleration of feature trajectory whose length is longer than threshold ζ at each iteration of the smoothing process for accelerating the program speed. Here $\zeta = 0.4 \times Z$. Z denotes the total number of frames in a video. After each iteration, current acceleration will be compared with the acceleration calculated at preceding iteration. If the acceleration of 90% feature points which are on selected feature trajectories become invariant, the iterative smoothing procedure will be ended. Otherwise, the smoothed transformation chain $\tilde{\mathbf{T}}_0^1, \tilde{\mathbf{T}}_1^2, \cdots, \tilde{\mathbf{T}}_{t-1}^t, \cdots, \tilde{\mathbf{T}}_{n-1}^n$ is set to be the original transformation chain $\mathbf{T}_0^1, \mathbf{T}_1^2, \cdots, \mathbf{T}_{t-1}^t, \cdots, \mathbf{T}_{n-1}^n$ in next iteration.

Finally, the motion compensated frame I_t' can be warped from the original frame I_t as follow:

$$I_t' = \mathbf{C}_t I_t. \tag{10}$$

IV. Experiment Results

The proposed algorithm has been applied to a number of video sequences. All source videos are from paper [11], and the frame resolution is 480×272. The computational cost of our approach is about 3.2 frames per second with a Pentium4 3.0 GHz CPU.

For comparison purpose, Kalman filter and spline interpolation are also utilized. When spline interpolation is applied, knots selection has significant impact on video stabilization result. If the knots located on 'noisy' points, it will deteriorate the stabilization effect. In this paper, we use Kalman filter with similar model with paper [2] and [3]. If the hand-held camera motion model and noise model are not accurate, this approach can't handle it well.

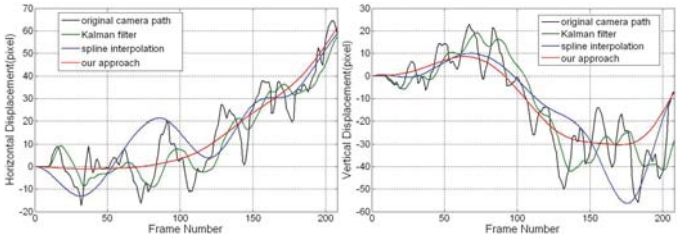

Fig. 3. Left: Horizontal displacement of camera and smoothed result by different approaches. Right: Vertical displacement of camera and smoothed result by different approaches.

Fig.3 shows the results of smoothing by different approaches. When compare with spline interpolation and Kalman filter, our approach is more robust when dealing with shaky video.

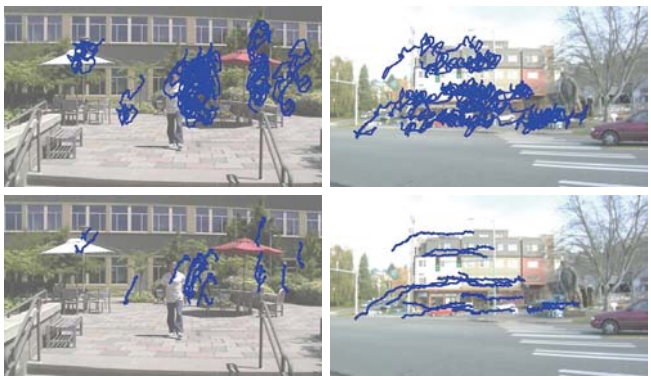

Fig. 5. Feature Trajectories (Left column: Length≥90 frames. Right column: Length≥140 frames) Top row: Original feature trajectories. Bottom row: Our result. After stabilization by our method, the feature trajectories become smoother.

The results stabilized by different approaches are listed in Fig.4. The four frames stabilized by our approach are

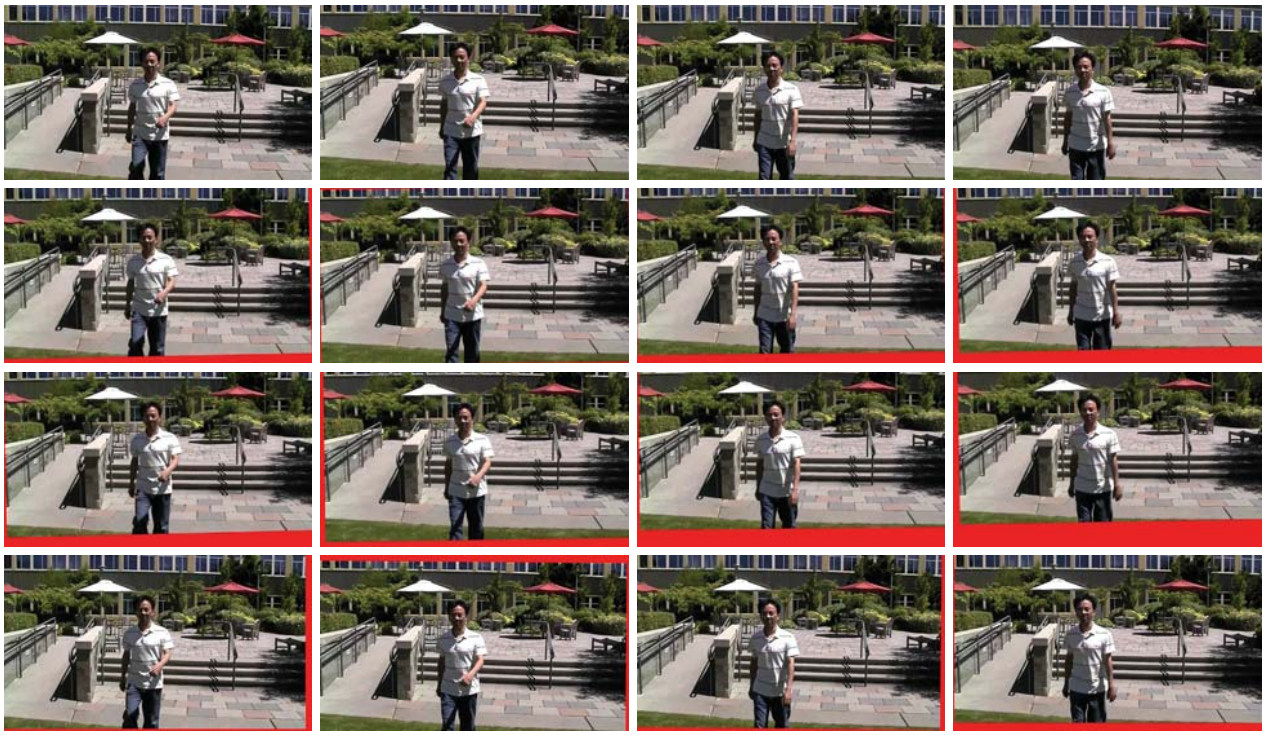

Fig. 4. Top row: Four frames of the original video. Second row: Stabilized frames by Kalman filter. The red regions show the uncovered area. Third row: Stabilized frames by spline interpolation. Bottom row: Stabilized frames by our approach. It shows that our approach stabilizes the video and leaves less uncovered area.

more satiable, and have less uncovered area (red regions). For illustrating overall effect, feature trajectories before and after smoothing procedure are shown in Fig.5. It is obvious that the feature trajectories become smoother after stabilization by our approach.

V. Conclusion

A new video stabilization approach is proposed in this paper. By considering the smoothness of the robust feature trajectories, we stabilized the video by iterative smoothing method. Compared with existing methods, the stabilized camera motion path which is smoothed by our approach is smoother, and it is closer to the ideal trend path. Hence, our approach leaves less uncovered area in the stabilized video. Our approach requires neither analysis of camera motion nor the cause of the jitter. We have also used many different video clips to test our approach. It shows that our approach is more adaptive, and does not suffer from the insufficient motion analysis. It is more robust and adaptive than other methods.

The limitation of our approach is like all 2D video stabilization methods. First, a full-frame warp cannot model the parallax; second, the 2D warp cannot describe camera paths in 3D[11]. Our next phase of research is use 3D approach to improve the precision of stabilized video.

Acknowledgment

This work was supported by International Cooperation Program of Science and Technology of China (No.2007DFA11530) and the National Natural Science Foundation of China (No.60875072).

References

[1] K.-Y Lee, Y.-Y.Chuang, B.-Y. Chen and M. Ouhyoung, "Video Stabilization using Robust Feature Trajectories," IEEE International Conference on Computer Vision, pp. 1397-1404, 2009. (ICCV2009)

[2] A. Litvin, J. Konrad, and W. Karl, "Probabilistic Video Stabilization Using Kalman Filtering and Mosaicking," In Proceedings of IS&T/SPIE Symposium on Electronic Imaging, Image and Video Communications and Proc, pp. 663-674, 2003.

[3] J. Zhu and B. Guo, "Electronic Image Stabilization System Based on Global Feature Tracking," Journal of Systems Engineering and Electronics, vol. 19(2), pp. 228-233, 2008.

[4] L. Meng, X. Lin, L. Xu and F. Fu, "Video Stabilizing System for Digital Camera," In Proceedings of 7th International Conference on Singal Processing, vol.2, pp. 1119-1122, 2004.

[5] S. Wu, Z. Ren, "Video Stabilization by Multi-Trajectory Mapping and Smoothing," Fifth International Conference on Information, Communications and Signal Processing, pp. 542-545, 2005.

[6] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame Video Stabilization with Motion Inpainting," IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(7), pp. 1150-1163, 2006.

[7] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, "Speeded-Up Robust Features (SURF)," Journal of Computer Vision and Image Understanding, 110(3), pp. 346-359, 2008.

[8] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Commun. Assoc. Comp. Mach., vol. 24:381-395, 1981.

[9] D. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," International Journal of Computer Vision, 60:91-110, 2004.

[10] R. I. Hartley and A. Zisserman, "Multiple View Geometry," Cambridge, UK: Cambridge University Press, March 2004.

[11] F. Liu, M. Gleicher, H. Jin and A. Agarwala, "Content-Preserving Warps for 3D Video Stabilization," ACM Transactions on Graphics, 28(3),43:1-43:9, 2009.