SCIENCE CHINA
Information Sciences

• MOOP •

January 2020, Vol. 63 114101:1–114101:3
https://doi.org/10.1007/s11432-018-9764-0

Spatiotemporal consistency-based adaptive hand-held video stabilization

Xiao LI[1], Shuai LI[1,2]*, Hong QIN[3] & Aimin HAO[1]

[1] *State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191*, China;*
[2] *Beihang University Qingdao Research Institute, Qingdao* 266000*, China;*
[3] *Department of Computer Science, Stony Brook University, New York* 11794*, USA*

Received 27 June 2018/Revised 16 October 2018/Accepted 29 December 2018/Published online 11 September 2019

Citation Li X, Li S, Qin H, et al. Spatiotemporal consistency-based adaptive hand-held video stabilization. Sci China Inf Sci, 2020, 63(1): 114101, https://doi.org/10.1007/s11432-018-9764-0

This study develops a novel, joint optimization method to successively respect spatial structure consistency and temporal feature constraints. Our method is more flexible and adaptive as we regard the camera multitrack motion to be noisy signals and adaptively suppress the noise without manual intervention. In addition, our method can achieve stabilization effects similar to that obtained by 3D methods while retaining the efficiency and robustness of 2D methods over challenging videos. As shown in Figure 1, our method could be divided into three components: salient region preserving-based homography estimation of spatial structure consistency, self-adaptive intrinsic mode function (IMFs) and feature-centric empirical mode decomposition (EMD). Salient region preserving-based homography estimation of spatial structure consistency is content-aware, which can preserve salient and visually prominent regions. The frames in the wobbly video are warped based on a saliency map. Each frame is uniformly divided into multiple grids. The local homography in each grid is estimated to spatially build mesh-wise inconsistent camera paths. The method can eliminate the parallax between different grids in the same video frames while preserving salient regions. Self-adaptive IMFs can read all frames in the wobbly video and find their scale-invariant feature transform (SIFT) features. Moreover, geometric transformation can be estimated from the matching point pairs. The original signal is decomposed into a finite component and few components. Finally, we can calculate the optimizing ratio of the IMF using CVX (Matlab software for disciplined convex programming). The feature-centric EMD aims to retain the centric feature of EMD. Our method brings in weighted Gaussian distribution to ensure that the new path retains the trend of the original path while suppressing jitters.

Salient region preserving-based spatial structure consistency optimization. We frame-wise extract the SIFT features from the wobbly video and further match the features. The geometric transformation (the second sub-figure in Figure 1(a)) maps the inliers in the matched points of the left frame to the inliers in those of the right frame. Using geometric transformation algorithm [1], we could compute and denote the transformation with the 3×3 matrix T_t. The relative camera motion at time t can be represented by a 2D Euclidean transformation T_t, satisfying $S_t = S_{t-1}T_{t-1}$. A uniform grid is overlaid on the image with $\hat{N}$ columns and $\hat{M}$ rows [2]. The target is to compute a deformed grid for the resized image. Consistent with common image re-targeting methods, the saliency map $\Psi(\hat{x}, \hat{y})$ is used to assign an importance value between 0 and 1 to each pixel of the image. We average the saliency values inside each cell of the grid for the original image so that the saliency vector Ψ_i could be obtained. The optimization further reduces the influences of spatially mesh-wise inconsistency, which greatly decreases the paral-

* Corresponding author (email: lishuai@buaa.edu.cn)

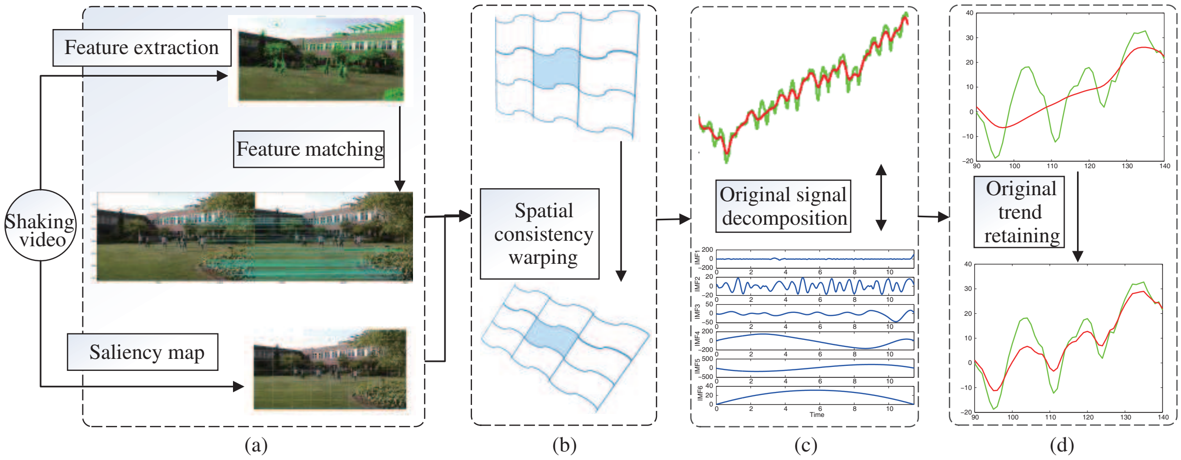

Figure 1 (Color online) Pipeline of our framework. (a) Feature extraction, matching, and saliency map construction; (b) salient region preserving-based spatial structure consistency optimization; (c) self-adaptive IMFs; (d) feature-centric EMD.

lax. Based on the previous camera path $S_i(t-1)$ and the local homographies $T_i(t-1)$, we can define spatially mesh-wise inconsistent camera paths for the entire video. Taking $S_i(1)$ as the identity matrix, let $S_i(t)$ be the camera pose of the grid cell i at frame t. It can be formulated as $S_i(t)=\prod_{\hat{t}=1}^{t-1}T_i(\hat{t})$. We uniformly divide the frame into multiple grids. $D(t)$ denotes the smoothed path, and $B(t)$ denotes the transformation from the original path $S(t)$ to the smoothed path $D(t)$. Each grid has one trajectory, which is denoted by $S_i(t)$. $T_i(t-1)$ denotes the estimated local homographies at the same grid cell i from $S_i(t-1)$ to $S_i(t)$. The camera trajectories of spatial structure consistency could be smoothed by

$$\mathcal{O}(D(t)) = \text{argmin}(\Sigma_i(\|D_i(t)-S_i(t)\|^2 \\ +\lambda_t\Psi_i\Sigma_{j\in\Omega(i)}\|D_i(t)-D_j(t)\|^2)). \quad (1)$$

Here, $S=\{S(t)\}$ denotes the original path and $D=\{D(t)\}$ denotes the optimized path. $\Omega(i)$ represents the eight neighbors of the grid cell i. To reduce cropping and distortion, the term $\|D_i(t)-S_i(t)\|$ guarantees the new camera path to be close to the original one, whereas $\|D_i(t)-D_j(t)\|$ can keep the current grid cell consistent with the nearby neighbors. The parameter λ_t is used to balance the above two terms. For the marginal grid cell, we set its value to be the same as those of its in-existent neighbors; namely, it can be formulated as $D_j(t)=D_i(t)$ when j is nonexistent. This optimization is quadratic, and its optimum result can be obtained by a Jacobi-based iteration [3]. Then, we can obtain the optimized paths $D_i(t)$. Using $B(t)=S^{-1}(t)D(t)$, the original video frames could be transformed into ones with spatial structure consistency while preserving salient regions. Subsequently, the parallax among the spatially variant grid cells of each frame is eliminated.

SIFT-based motion signal construction. We can convert a 3×3 transform T_t to an SRT (scale, rotation, translation) transform, which returns the scale, rotation, and translation parameters as well as the reconstituted transform T_t. This study focuses on more comprehensive parameters, such as the scale, angle, x-coordinate, and y-coordinate. The rotation parameter contains the angle. The translation parameter contains the x-coordinate and the y-coordinate. We then concatenate the scale, rotation, and translation parameters into a 4D vector $\hat{S}_t$ to represent the camera pose at time t. We regard the component in the vector $\hat{S}_t$ as a motion signal.

Self-adaptive IMFs. EMD can decompose any complicated signal to generate IMFs via a sifting process [4]. Specifically, it can decompose the original signal $\hat{S}$ via $\hat{S}=\sum_{k=1}^{N}f_k+r_N$. Here, $f_k(k=1,\ldots,N)$ are IMFs, and r_N is the corresponding residual. Figure 1(c) demonstrates the $\text{IMF}_k(k=1,\ldots,5)$, and IMF_6 denotes the residuals. We set $f_{N+1}=r_N$ for easy expression and calculation. The original signal is decomposed into the IMFs and residual. To stabilize the video, the high-frequency signals should be smoothed. The optimal camera trajectory is obtained by minimizing the following objective function:

$$\mathcal{O}(\alpha)=\left\|\nabla\left(\sum_{k=1}^{N+1}\alpha_k f_k\right)\right\|_1+\left\|\nabla^2\left(\sum_{k=1}^{N+1}\alpha_k f_k\right)\right\|_1 \\ +\left\|\nabla^3\left(\sum_{k=1}^{N+1}\alpha_k f_k\right)\right\|_1+W\left\|\sum_{k=1}^{N+1}\alpha_k f_k-\hat{S}\right\|_1. \quad (2)$$

Here, α denotes the ratio of the IMF. We use X to denote the variable. When $X=\sum_{k=1}\alpha_k f_k$,

$\|\nabla(X)\|_1$, $\|\nabla^2(X)\|_1$, and $\|\nabla^3(X)\|_1$ are the $L1$ norms of the first-order, second-order and third-order derivatives of X, respectively. The minimum of the sum of $\|\nabla(X)\|_1$, $\|\nabla^2(X)\|_1$, and $\|\nabla^3(X)\|_1$ smooths the IMFs to remove jitters in the unstable video. $\hat{S}$ denotes the original signal. The minimum of the difference between $\sum_{k=1}\alpha_k f_k$ and $\hat{S}$ makes the original signal close to the optimized signal to avoid excessive cropping. W is the adaptive equilibrium factor, which is used to balance the above four terms. This study empirically sets W to 0.1. In summary, our optimization method comprehensively considers multiple competing factors such as eliminating vibration, excluding excessive cropping, and minimizing distortional deformation. Then, the optimized motion signal could be calculated via $\hat{T} = \sum_{k=1}^{N+1}\hat{\alpha}_k f_k$. $\hat{T}$ is the optimized motion signal. $\hat{\alpha}_k$ is the new ratio of the IMF. Figure 1(c) shows the camera trajectories before and after smoothing, marked with green and red lines, respectively. Our method could compute the ratio of each IMF by autonomic learning and can thus improve the smoothness via repeated iterations.

Feature-centric EMD. The green line denotes the motion signal of the original path, and the red line denotes the motion signal of smoothing path without feature-centric EMD (Figure 1(d)). The original is over-smoothed and loses the original trend of the motion, thereby leading to excessive cropping. To retain the tendency of the original EMD motion signal, we define the extreme point of original motion signal as the feature. To retain the centric feature of EMD signals while smoothing signals, our feature-centric EMD is formulated as follows:

$$\widetilde{T}_t = (1-\widetilde{W})\frac{\Sigma_{\widetilde{t}\in\omega_t}(G_t(\|S_{\widetilde{t}}-S_t\|)\hat{T}_{\widetilde{t}})}{\Sigma_{\widetilde{t}\in\omega_t}G_t(\|S_{\widetilde{t}}-S_t\|)} + \widetilde{W}\frac{\Sigma_{\widetilde{t}\in\omega_t}(G_t(\|\hat{T}_{\widetilde{t}}-\hat{T}_t\|)S_{\widetilde{t}})}{\Sigma_{\widetilde{t}\in\omega_t}G_t(\|\hat{T}_{\widetilde{t}}-\hat{T}_t\|)}. \tag{3}$$

Here, we empirically set ω_t to denote 60 neighboring frames. We bring in Gaussian functions $G_t(\cdot)$ and empirically set the standard deviation of $G_t(\cdot)$ to 10. S_t denotes the original value at frame t without the feature-centric EMD. $S_{\widetilde{t}}$ denotes the original value at frame $\widetilde{t}$ without the feature-centric EMD. $\hat{T}_t$ denotes the optimized value at frame t. $\hat{T}_{\widetilde{t}}$ denotes the value at frame $\widetilde{t}$. Eq. (3) ensures that the new path retains the trend of the original path while successfully suppressing both high-frequency jitters and low-frequency bounces of the original path. $\frac{\Sigma_{\widetilde{t}\in\omega_t}(G_t(\|S_{\widetilde{t}}-S_t\|)\hat{T}_{\widetilde{t}})}{\Sigma_{\widetilde{t}\in\omega_t}G_t(\|S_{\widetilde{t}}-S_t\|)}$ mainly suppresses the shaky components of the original path, simultaneously retaining its initial trend. Meanwhile, $\frac{\Sigma_{\widetilde{t}\in\omega_t}(G_t(\|\hat{T}_{\widetilde{t}}-\hat{T}_t\|)S_{\widetilde{t}})}{\Sigma_{\widetilde{t}\in\omega_t}G_t(\|\hat{T}_{\widetilde{t}}-\hat{T}_t\|)}$ mainly ensures that the new path retains the trend of the original path while suppressing its trembling signal. $\widetilde{W}$ is the adaptive equilibrium factor ranging from 0 to 1, which is used to balance the above two terms. We set $\widetilde{W}$ as 0.4 for all our examples. When smoothing the signal, fast panning or scene transition may cause rapid signal motion. In this case, some excessive cropping may be yielded owing to inappropriate smoothing. The motion signal may significantly deviate from its original path, as indicated by the green lines in Figure 1(d). The result from our adaptive smoothing produces much less cropping. The green line denotes the motion signal of the original path. The red line denotes the motion signal of the smoothing path with feature-centric EMD, which gives rise to better results.

Experimental results. To validate our joint optimization approach for adaptive video stabilization, we conduct comprehensive experiments on public benchmarks and perform extensive and quantitative evaluations with available state-of-the-art methods as well as popular commercial software. All our experiments demonstrate the advantages of the spatiotemporal optimization method in terms of its versatility, accuracy, and efficiency. All results and evaluations are provided in our supplementary videos.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61672077, 61532002) and Applied Basic Research Program of Qingdao (Grant No. 161013xx).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge: Cambridge University Press, 2004

2 Liu S C, Yuan L, Tan P, et al. Bundled camera paths for video stabilization. ACM Trans Graph, 2013, 32: 78

3 Bronshtein I N, Semendyayev K A. Handbook of Mathematics. Berlin: Springer, 2013

4 Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc London Ser A-Math Phys Eng Sci, 1998, 454: 903–995