5D VIDEO STABILIZATION THROUGH SENSOR VISION FUSION

Binnan Zhuang, Dongwoon Bai, Jungwon Lee

Samsung SoC R&D Lab, San Diego, USA

ABSTRACT

We propose a novel 5D video stabilization solution based on sensor vision fusion. Traditional gyroscope based video stabilization approaches only stabilize the 3D rotation of a camera. They often suffer in scenes with highly dynamic translational movements. In this paper, we model the residual camera motion after 3D rotation based stabilization as residual 2D translation in the image plane, which can be estimated using visual information. The proposed 5D stabilization is achieved by carefully combining sensor based 3D rotation stabilization and vision based residual 2D translation stabilization. The 5D stabilization is prototyped on a smartphone for real time recording. We demonstrate its advantages over two state-of-the-art methods. The 5D stabilization can also be used to solve the challenging foreground object stabilization problem.

Index Terms— Video Stabilization, Sensor Vision Fusion, Motion Estimation, Rolling Shutter Compensation

1. INTRODUCTION

The goal of video stabilization is to mitigate the shaky visual artifacts due to undesired motion jitter in order to satisfy the cinematographic perception of viewers. Video stabilization techniques can be divided into two main categories, namely, optical image stabilization (OIS) and digital image stabilization (DIS). OIS is usually achieved by mechanically moving the camera lens or sensor based on the instantaneous camera motion measured using a gyroscope. Due to mechanical limitation, OIS is only capable of filtering out high frequency motion jitter with small magnitude.

DIS is widely used for video stabilization on smartphones. In a DIS approach, the camera motion trajectory over multiple frames can be first estimated. A desired smooth path is then computed based on the raw estimate. Through image warping, the output video can be stabilized as if the camera was moving along the smooth path. Our proposed solution belongs to the DIS class.

The DIS solutions can be further divided into vision based and sensor based methods, depending on the source of camera motion estimation. Most sensor based methods use a gyroscope to estimate the 3D rotation of a camera [1, 2, 3, 4, 5]. Hence, only 3D rotation is stabilized, and the video may still look shaky due to uncompensated translational movements. On the other hand, the vision based video stabilization methods usually estimate the camera motion from an image sequence. Both feature based approaches [6, 7, 8] and global intensity alignment approaches [9] can be used for vision based motion estimation. Most existing works [10, 6, 11, 12, 13, 14] model the transformation between two consecutive frames as a homography, which cannot model parallax and will suffer from the ambiguity in visual information[1].

A more comprehensive model has been considered in [15, 16], where both 3D rotation and 3D translation are obtained by solving the classical structure from motion problem [17]. However, estimating 3D camera poses using pure visual information has been pointed out difficult in [15]. Note that proper utilization of the estimated 3D translation requires dense 3D mapping. Alternative vision based solutions [18, 19] consider stabilizing individual feature trajectories over many frames.

In this work, we design a smartly integrated architecture, which utilizes both vision and sensor information. We first estimate the raw 3D rotation path of a camera using a gyroscope. The corresponding smoothed 3D path is obtained via path optimization. We then apply the 3D compensation on motion vectors (MVs) extracted from adjacent image frames to estimate the remaining translational jitter, which are used to build a raw residual 2D translation path. After 2D path smoothing, the 5D raw and smoothed paths are used to stabilize the inter-frame motion. The intra-frame motion is also estimated during this process to correct the rolling shutter (RS) effect. The integrated 5D stabilization enjoys the merits of both gyroscope based 3D rotation compensation and the additional vision based translation compensation without complicated 3D mapping nor 3D translation estimation. Compared with the homography model (with maximum eight degrees of freedom), the 5D model precisely captured the 3D rotation and use simple 2D approximation to deal with the parallax related 3D translation, which makes the motion estimation more efficient and robust to outliers. Unlike the region/feature dependent motion estimation in [14, 18, 19], the 5D model estimates a global 3D rotation, since the 3D rotation compensation is common to the entire image. The solution is prototyped on a Galaxy S8 smartphone with limited buffers for real-time 5D stabilization under 30 fps.

Some recent works [20, 21] focus on stabilizing the fore-

[1]The homography model cannot well distinguish small rotation from small translation, although they should be compensated differently.

ground of the scene separately. The proposed 5D stabilization can also be used to stabilize the object of interest (OOI) in the foreground. The 3D rotation compensation stabilizes the background of the scene, while the residual 2D translation compensation is only applied on the OOI. As a result, the trade-off between background stabilization and foreground OOI stabilization can be easily achieved.

2. GYRO BASED 3D STABILIZATION

2.1. Camera Model

The simple pin-hole camera model is used in this paper. An 3D object point, $\boldsymbol{X}$, is projected to the image point, $\tilde{\boldsymbol{x}}$, in 2D homogeneous coordinate through:

$$\tilde{\boldsymbol{x}} = \boldsymbol{K}\left[\boldsymbol{R}\boldsymbol{X} + \boldsymbol{P}\right], \tag{1}$$

where $\boldsymbol{K}$ is the camera intrinsic matrix. The camera motion is modeled as the combination of 3D rotation $\boldsymbol{R} \in \boldsymbol{SO}(3)$ and 3D translation $\boldsymbol{P} \in \mathbb{R}^3$. Throughout this paper, we use $\tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ to differentiate the homogeneous and 2D Euclidean representations of a 2D point.

2.2. Gyro based 3D Stabilization

The gyro based 3D stabilization can be considered as part of the 5D stabilization system shown in Fig. 1. The camera 3D rotation is estimated in the 3D rotation estimation block using gyro data. The inter-frame rotation is obtained through integrating the corresponding angular velocity measurements as in [3]. Let $\Delta\boldsymbol{\theta}^{\text{Inter}}_{n-1\to n}$ be the rotation vector, representing the inter-frame rotation from frame $n-1$ to frame n. The corresponding 3D rotation matrix is denoted by $\Delta\boldsymbol{R}^{\text{Inter}}_{n-1\to n}$. The intra-frame 3D rotation is estimated analogously for RS correction. Denote the intra-frame rotation from the mid row of frame n to the yth as $\Delta\boldsymbol{\theta}^{\text{Intra}}_{n}(y)$. The raw 3D path can be obtained through accumulating inter-frame 3D rotations:

$$\boldsymbol{R}_n = \prod_{i=1}^{n} \Delta\boldsymbol{R}^{\text{Inter}}_{i-1\to i}, \tag{2}$$

The rotation vector corresponding to $\boldsymbol{R}_n$ is $\boldsymbol{\theta}_n$.

The 3D rotation smoothing block takes the raw 3D path as input, and outputs the desired smooth 3D path via path optimization, which will be explained in Section 3.2. We denote the rotation matrix and the rotation vector of the stabilized path as $\boldsymbol{R}'_n$ and $\boldsymbol{\theta}'_n$. The image warping for 3D rotation compensation and RS correction are calculated based on $\boldsymbol{R}_n$, $\boldsymbol{R}'_n$ and $\Delta\boldsymbol{\theta}^{\text{Intra}}_{n}(y)$ as described in [3]. The warped images are cropped to get rid of empty boarders with no image data.

3. 5D STABILIZATION

Two extra blocks are added for 5D stabilization in Fig. 1. The residual 2D translation estimation block estimates the residual motion due to 3D translation. The raw 2D path is smoothed in the residual 2D translation smoothing block. The 5D raw and smoothed paths are gathered in the distortion calculation block to generate the distorted grid for image warping.

3.1. Residual 2D Translation Estimation

The end effect of camera translation is captured in MVs. One problem is that the MVs are also affected by 3D rotation. Denote the kth MV as $(\boldsymbol{x}_{k,n-1}, \boldsymbol{x}_{k,n})$, where the image points $\boldsymbol{x}_{k,n-1}$ and $\boldsymbol{x}_{k,n}$ are projected from the 3D object point $\boldsymbol{X}_k$ in frame $n-1$ and frame n through:

$$\tilde{\boldsymbol{x}}_{k,n} = \boldsymbol{K}[\boldsymbol{R}_n\boldsymbol{X}_k + \boldsymbol{P}_n]. \tag{3}$$

The 3D rotation compensation is given by the difference between the raw and the smoothed 3D paths. The image point of $\boldsymbol{X}_k$ after 3D compensation becomes:

$$\tilde{\boldsymbol{x}}'_{k,n} = \boldsymbol{K}\boldsymbol{R}'_n\boldsymbol{R}_n^{-1}\boldsymbol{K}^{-1}\begin{bmatrix}\boldsymbol{x}_{k,n}\\ 1\end{bmatrix}. \tag{4}$$

Thus, $(\boldsymbol{x}'_{k,n-1}, \boldsymbol{x}'_{k,n})$ forms the MV after 3D compensation. The residual 2D translation cannot be directly computed as $\boldsymbol{x}'_{k,n} - \boldsymbol{x}'_{k,n-1}$. This is because the difference contains both the residual 2D translation and the stabilized 3D rotation.

To remove the stabilized rotation, we apply the extra rotation $\boldsymbol{R}'_n\boldsymbol{R}'^{-1}_{n-1}$ on the source point, $\tilde{\boldsymbol{x}}'_{k,n-1}$ to get $\tilde{\boldsymbol{y}}_{k,n-1}$:

$$\tilde{\boldsymbol{y}}'_{k,n-1} = \boldsymbol{K}\boldsymbol{R}'_n\boldsymbol{R}_{n-1}^{-1}\boldsymbol{K}^{-1}\begin{bmatrix}\boldsymbol{x}_{k,n-1}\\ 1\end{bmatrix}. \tag{5}$$

The residual 2D translation on the kth MV is defined as:

$$\Delta\boldsymbol{T}^{k}_{n-1\to n} = \boldsymbol{x}'_{k,n} - \boldsymbol{y}'_{k,n-1}. \tag{6}$$

Assuming the stabilized 3D path is ideal, i.e., the rotation $\boldsymbol{R}'_n\boldsymbol{R}'^{-1}_{n-1}$ is the desired motion by the viewers, $\Delta\boldsymbol{T}^{k}_{n-1\to n}$ is exactly the residual jitter due to unstabilized translation.

The residual 2D translation also has parallax. Precise compensation requires per-pixel residual 2D translation estimate, which is too complicated for implementation on smartphones. Here, we take an approximation to estimate the average inter-frame residual 2D translation over all MVs:

$$\Delta\boldsymbol{T}^{\text{Inter}}_{n-1\to n} = \frac{1}{K}\sum_{k=1}^{K}\Delta\boldsymbol{T}^{k}_{n-1\to n}. \tag{7}$$

The approximation works well in most scenarios, where the depths of the 3D points are much larger than their variation.

The accumulated raw 2D path is calculated from the inter-frame residual 2D translation:

$$\boldsymbol{T}_n = \sum_{i=1}^{n}\Delta\boldsymbol{T}^{\text{Inter}}_{i-1\to i}, \tag{8}$$

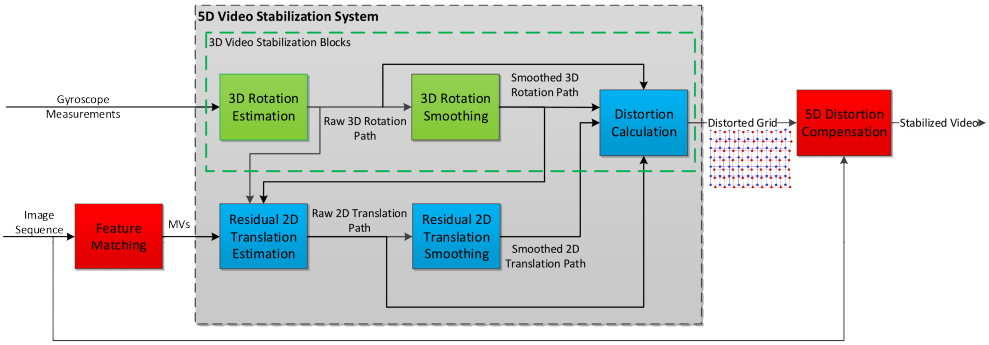

Fig. 1. Block diagram of the proposed 5D video stabilization system.

which will be used for 2D translation smoothing to generate the smoothed path $\boldsymbol{T}'_n$. The intra-frame residual 2D translation from the mid-row to the yth row in frame n is denoted by $\Delta\boldsymbol{T}_n^{\text{Intra}}(y)$, which can be easily calculated from the inter-frame estimates, assuming constant velocity within a frame.

3.2. Optimization based Path Smoothing

The 5D raw camera path consists of 3D raw path, $\boldsymbol{\theta}_n$, and residual 2D raw path, $\boldsymbol{T}_n$. Each dimension of the 5D raw path is smoothed by solving an optimization problem similar to [13]. Without loss of generality, denote one dimension of the 5D path in frame n as y_n. The input to the path optimization problem in frame n is $\boldsymbol{y} = [y_{n-a_1}, \cdots, y_{n+a_2}]^T$, which consists of a_1 previous frames, the current frame, and a_2 future frames. The path optimization problem is formulated as:

$$\underset{\boldsymbol{x}}{\text{minimize}} \;\; w_0\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \sum_{i=1}^{3} w_i\|\boldsymbol{D}_i\boldsymbol{x}\|_1^1 \tag{9a}$$

$$\text{subject to} \;\; \boldsymbol{l} \le \boldsymbol{x}-\boldsymbol{y} \le \boldsymbol{u}, \tag{9b}$$

where $\boldsymbol{D}_i$ is the ith order discrete derivative operator; w_i's are the combining weights of different cost terms; and the bounds $\boldsymbol{l}$ and $\boldsymbol{u}$ are to guarantee that the stabilized image covers the entire crop window.

In addition to the derivative terms that penalize different orders of motion jitters [13], we add the ℓ_2 distance between the raw and smoothed paths to improve robustness of the path optimization solution. The QP formulation also allows us to design an efficient problem-specific solver that solves the 5D path optimization problem within 15 milliseconds in each frame. Hence, we are able to implement the 5D stabilization on a Galaxy S8 smartphone for real time recording of 30 fps. Due to space limitation, the details are omitted here.

4. DISTORTION CALCULATION

The image warping for 5D compensation and RS correction requires calculating a distorted grid every frame. The distorted grid in the raw image corresponds to a regular grid in the stabilized image. Hence, the problem is to find the point $\boldsymbol{x} = (x[1], x[2])$ in the raw image that corresponds to a given a point $\boldsymbol{x}' = (x'[1], x'[2])$ in the stabilized image.

The distortion calculation is given by:

$$\tilde{\boldsymbol{x}} = \boldsymbol{K}\Delta\boldsymbol{R}_n^{\text{Intra}}(x'[2])\boldsymbol{R}_n\boldsymbol{R}_n'^{-1}\boldsymbol{K}^{-1}\begin{bmatrix}\boldsymbol{x}-\boldsymbol{T}_n^C(x'[2])\\ 1\end{bmatrix}, \tag{10}$$

where $\boldsymbol{T}_n^C(x'[2]) = (\boldsymbol{T}_n'^{\text{Inter}} - \boldsymbol{T}_n^{\text{Inter}}) - \Delta\boldsymbol{T}_n^{\text{Intra}}(x'[2])$. The global 3D rotation and residual 2D translation are stabilized using $\boldsymbol{R}_n\boldsymbol{R}'$ and $\boldsymbol{T}_n'^{\text{Inter}} - \boldsymbol{T}_n^{\text{Inter}}$, respectively; $\Delta\boldsymbol{R}_n^{\text{Intra}}(x'[2])$ and $\Delta\boldsymbol{T}_n^{\text{Intra}}(x'[2])$ are to correct the corresponding rolling shutter effects. Finally, $\boldsymbol{x}$ is obtained by normalizing $\tilde{\boldsymbol{x}}$.

5. 5D OOI STABILIZATION

We are sometimes interested in stabilizing specific object(s) in the scene. The proposed 5D stabilization can be easily generalized for this application. The 3D rotation compensation stabilizes the main oscillation due to camera rotation. The residual 2D translation compensation stabilizes the OOI, by estimating the residual 2D translation of the OOI.

For the result provided in Section 6.3, we use the learning adaptive discriminative correlation filter based tracker [22] for object tracking. The tracker outputs a bounding box over the OOI in the down-sampled QVGA image. FlowNet2 [23] is used to generate dense MVs within the bounding box for residual 2D translation estimation. The trade-off between background stabilization and foreground object stabilization is achieved by adjusting the combining weights in (9a), differently, for 3D path smoothing and residual 2D path smoothing.

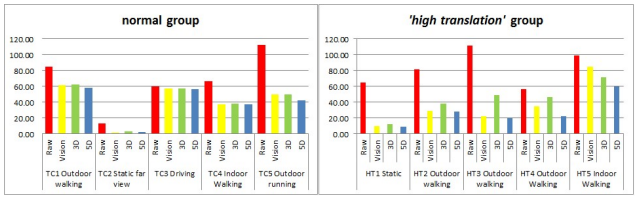

Fig. 2. Motion jitter score comparison.

6. RESULTS

The 5D stabilization is compared with two state-of-the-art solutions in [3] and [13] with source codes, such that we can control parameters such as smoothing window size and crop ratio for fair comparison. The approach in [3] is a typical gyro based 3D stabilization solution, which uses a simple Gaussian filter for path smoothing. The solution in [13] is a vision based stabilization method based on homography model.

6.1. Simulation Setup

The proposed 5D stabilization is implemented on a Galaxy S8 smartphone for real-time stabilization. The methods in [3] and [13] are applied offline using the captured uncompensated image sequence and gyro samples. Due to lack of shaky video data set with gyro data, we collect eighteen videos consisting of a variety of scenes and motions. Among the eighteen videos, ten of them are ordinary scenes; and the other eight are *'high translation'* scenes with additional high translational motion on top of ordinary movements. All stabilized videos are provided in the supplementary materials.

6.2. Performance Comparison

The quantitative comparison is based on *motion jitter score*. The motion titter is calculated as the weighted sum of the three derivative terms in (9a) using the optical flow generated from an output video. The motion jitter score is obtained through averaging the motion jitters across all frames and summing over the horizontal and vertical directions.

The motion jitter scores for five normal and five *'high translation'* test cases are shown in Fig. 2. The jitter scores for the raw video, the stabilized video by [13], the stabilized video by [3], and, the stabilized video by the 5D stabilization are labeled as 'Raw' (red), 'Vision' (yellow), '3D' (green), and, '5D' (blue), respectively. The proposed 5D stabilization achieves the lowest motion jitter score in almost all the cases, except for the very static scene in TC2. In TC2, the hand shaking are well compensated by all three solutions. The minor degradations by '3D' and '5D' are due to the gyro drifts.

Compared to [3], the proposed 5D stabilization are superior in both path smoothing and motion estimation. The optimization based path smoothing adapts to dynamic changes in camera motion trajectories and provides more natural smoothed motion. Therefore, the motion jitter score is reduced in all test cases. The additional residual 2D translation compensation achieves significant performance enhancement in the *'high translation'* test cases as shown in Fig. 2.

The motion jitter scores achieved by [13] usually lie in between '3D' and '5D'. However, the stabilized videos using the homography based motion estimation often have *wobbling* effects as shown in the supplementary materials. This global homography model cannot handle parallax. Although the residual 2D translation estimation also uses the approximation based on closely distributed depths, the 5D stabilization significantly improve the *wobbling* effect due to sequentially compensating 3D rotation and residual 2D translation.

6.3. 5D OOI Stabilization Results

The 5D OOI stabilization is demonstrated using a video consisting of a person playing skateboard as the OOI. We compare the 3D stabilization [3], the 5D stabilization and the 5D OOI stabilization. The motion jitter scores calculated using the entire image as well as using the region of the moving person are shown in Table 1. The 3D stabilization gives high jitter scores due to uncompensated translation. The 5D OOI stabilization achieves the lowest jitter score on the moving person, at the cost of slightly increased jitter score of the entire scene compared to the 5D stabilization. Based on the stabilized video shared in the supplementary materials, the 5D OOI stabilization achieves a good balance between stabilizing the background and the foreground moving object.

Table 1. Jitter score comparison between 3D stabilization, 5D stabilization and 5D OOI stabilization.

Jitter Score	3D	5D	5D OOI
Entire Image	44.09	22.26	22.64
Moving Person	55.36	29.17	24.98

7. CONCLUSION

We have proposed a novel 5D video stabilization solution through sensor vision fusion. The solution estimates the residual 2D translation using visual information after gyro based 3D rotation compensation. The combined 5D stabilization significantly improves the stabilization quality in scenes with high translation. Numerical experiments using shaky videos captured on a smartphone suggest the proposed 5D video stabilization outperforms the state-of-the-art solutions using either gyroscope or visual information alone. The 5D stabilization can also be used for stabilizing the OOI to reach a harmonious stabilization between the background and the foreground object.

8. REFERENCES

[1] Per-Erik Forssén and Erik Ringaby, "Rectifying rolling shutter video from hand-held devices," in *CVPR*, 2010.

[2] Gustav Hanning, Nicklas Forslöw, Per-Erik Forssén, Erik Ringaby, David Törnqvist, and Jonas Callmer, "Stabilizing cell phone video using inertial measurement sensors," in *ICCV Workshops*, 2011.

[3] Alexandre Karpenko, David Jacobs, Jongmin Baek, and Marc Levoy, "Digital video stabilization and rolling shutter correction using gyroscopes," *CSTR*, vol. 1, pp. 2, 2011.

[4] Steven Bell, Alejandro Troccoli, and Kari Pulli, "A non-linear filter for gyroscope-based video stabilization," in *EECV*, 2014.

[5] Sung Hee Park and Marc Levoy, "Gyro-based multi-image deconvolution for removing handshake blur," in *CVPR*, 2014.

[6] Alberto Censi, Andrea Fusiello, and Vito Roberto, "Image stabilization by features tracking," in *IEEE 10th Int. Conf. on Image Analysis and Processing*, 1999.

[7] Philip HS Torr and Andrew Zisserman, "Feature based methods for structure and motion estimation," in *International workshop on vision algorithms*, 1999.

[8] Sebastiano Battiato, Giovanni Gallo, Giovanni Puglisi, and Salvatore Scellato, "Sift features tracking for video stabilization," in *ICIAP*, 2007.

[9] James R Bergen, Patrick Anandan, Keith J. Hanna, and Rajesh Hingorani, "Hierarchical model-based motion estimation," in *ECCV*, 1992.

[10] Zhigang Zhu, Guangyou Xu, Yudong Yang, and Jesse S Jin, "Camera stabilization based on 2.5 D motion estimation and inertial motion filtering," in *IEEE Int. Conf. on Intelligent Vehicles*, 1998.

[11] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum, "Full-frame video stabilization with motion inpainting," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2006.

[12] Won-ho Cho and Ki-Sang Hong, "Affine motion based CMOS distortion analysis and CMOS digital image stabilization," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 3, pp. 833–841, 2007.

[13] Matthias Grundmann, Vivek Kwatra, and Irfan Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," in *CVPR*, 2011.

[14] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun, "Bundled camera paths for video stabilization," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 78, 2013.

[15] Guofeng Zhang, Wei Hua, Xueying Qin, Yuanlong Shao, and Hujun Bao, "Video stabilization based on a 3D perspective camera model," *The Visual Computer*, vol. 25, no. 11, pp. 997, 2009.

[16] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala, "Content-preserving warps for 3D video stabilization," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 44, 2009.

[17] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[18] Yu-Shuen Wang, Feng Liu, Pu-Sheng Hsu, and Tong-Yee Lee, "Spatially and temporally optimized video stabilization," *IEEE transactions on visualization and omputer graphics*, vol. 19, no. 8, pp. 1354–1361, 2013.

[19] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala, "Subspace video stabilization," *ACM Transactions on Graphics*, vol. 30, no. 1, pp. 4, 2011.

[20] Shuaicheng Liu, Binhan Xu, Chuang Deng, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj, "A hybrid approach for near-range video stabilization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1922–1933, 2017.

[21] Fang-Lue Zhang, Xian Wu, Hao-Tian Zhang, Jue Wang, and Shi-Min Hu, "Robust background identification for dynamic video editing," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 197, 2016.

[22] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking," *arXiv preprint arXiv:1807.11348*, 2018.

[23] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," *CVPR*, 2017.

[24] Giovanni Puglisi and Sebastiano Battiato, "A robust image alignment algorithm for video stabilization purposes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1390–1400, 2011.