Interpolation-Based Smart Video Stabilization
Enterpolasyon Tabanlı Akıllı Video Stabilizasyonu

Semiha Dervişoğlu[1,*], Mehmet Sarıgül[1], Levent Karacan[1]
[1]Department of Computer Engineering, Iskenderun Technical University, Hatay Turkey
ORCIDs: 0000-0002-3677-8792, 0000-0001-7323-6864, 0000-0003-2764-5258
E-mails: semihadervisoglu.mfbe19@iste.edu.tr, mehmet.sarigul@iste.edu.tr, levent.karacan@iste.edu.tr
*Corresponding author.

***Abstract*—Video stabilization is the process of eliminating unwanted camera movements and shaking in a recorded video. Recently, learning-based video stabilization methods have become very popular. Supervised learning-based approaches need labeled data. For the video stabilization problem, recording both stable and unstable versions of the same video is quite troublesome and requires special hardware. In order to overcome this situation, learning-based interpolation methods that do not need such data have been proposed. In this paper, we review recent learning-based interpolation methods for video stabilization and discuss the shortcomings and potential improvements of them.**

Keywords*—*video stabilization; deep learning; unsupervised learning; interpolation methods

***Özetçe*—Video stabilizasyonu, kaydedilen bir videoda istenmeyen kamera hareketlerini ve titremeyi ortadan kaldırma işlemidir. Son zamanlarda, öğrenme tabanlı video sabitleme yöntemleri oldukça popüler hale geldi. Denetimli öğrenme temelli yaklaşımların etiketlenmiş verilere ihtiyacı vardır. Video stabilizasyon problemi için aynı videonun hem stabil hem de stabil olmayan versiyonlarını kaydetmek oldukça zahmetlidir ve özel donanım gerektirir. Bu durumun üstesinden gelebilmek için bu tür verilere ihtiyaç duymayan öğrenme tabanlı enterpolasyon yöntemleri önerilmiştir. Bu yazıda, video sabitleme için en son öğrenmeye dayalı enterpolasyon yöntemlerini gözden geçiriyoruz ve bunların eksikliklerini ve potansiyel iyileştirmelerini tartışıyoruz.**

Anahtar Kelimeler*—*video stabilizasyon; derin öğrenme; öğreticisiz öğrenme; enterpolasyon yöntemleri

I. Introduction

Videos, which have become a habit of our daily life, are used in many fields such as military (unmanned aerial vehicles), education (scientific research, etc.), healthcare (in order to determine the size and location of the problem in endoscopy and colonoscopy videos) and film industry (in movie shootings, etc.). A video is defined as a set of consecutive images taken with a handheld camera or a camera positioned on a vehicle. The shaking of the hand or the shaking of the vehicle in these videos can cause visual problems. Although some hardware has been produced to solve this problem, this equipment can be expensive. For this reason, the interest in the field of digital video stabilization has increased [1].

Digital video stabilization can be applied in three steps. Motion path estimation, motion path smoothing and producing the stable video (Fig. 1). In the majority of the digital video stabilization studies, motion is detected by using the similarity and feature extraction between the frames, camera path is corrected and video is stabilized. Pixel and block mapping-based motion detection methods use inter-frame distance metric and similarity measurement to estimate the motion of each pixel and block. In the pixel-based approach a pixel is represented with 3 color values and an invariant brightness value. It is hard to estimate the pixel movement between the consecutive frames due to the similarity of neighboring pixels. A pixel can match with several pixels in the next frame. To overcome this problem a holistic approach is applied. A 2-dimensional transform is applied to one of the frames as a whole and the number of matching pixels is tried to be maximized. Block-based mappings reduce the negative effects of pixel-based methods and block-pixels are used for displacements between two scenes. It also only considers blocks at a certain distance from the block to be matched to avoid over-matching. On the other hand, the feature-matching methods find the easily recognizable points in the scene. Only the displacements of the relevant points are calculated. All videos are processed frame-by-frame, and the positions of these relevant points are determined by tracking the selected features in the video frames, and the motion trajectories are determined. There are several feature extraction algorithms such as SIFT, SURF and FAST. SIFT does direction-independent feature extraction [2]–[4], SURF approach is an optimized version of the SIFT algorithm [5], and FAST is a corner detection algorithm reported to give more successful results than alternatives [2]. Although feature matching methods are successful, they have limitations. Feature extraction algorithms may not be able to extract enough features in certain parts of the scene and this may lead to poor results in stabilization. This is one of the reasons learning-based methods become popular in video stabilization. Supervised learning-based methods require stable and unstable pairs for each video in the learning set. Although some datasets are suggested on this subject, the quality of these studies is limited as it is dependent on the training dataset (DeepStab). On the other hand, self-supervised or semi-supervised methods may reach a better generalization quality [6]–[8]. Recently, semi-supervised or unsupervised methods based on interpolation are frequently encountered in the literature.

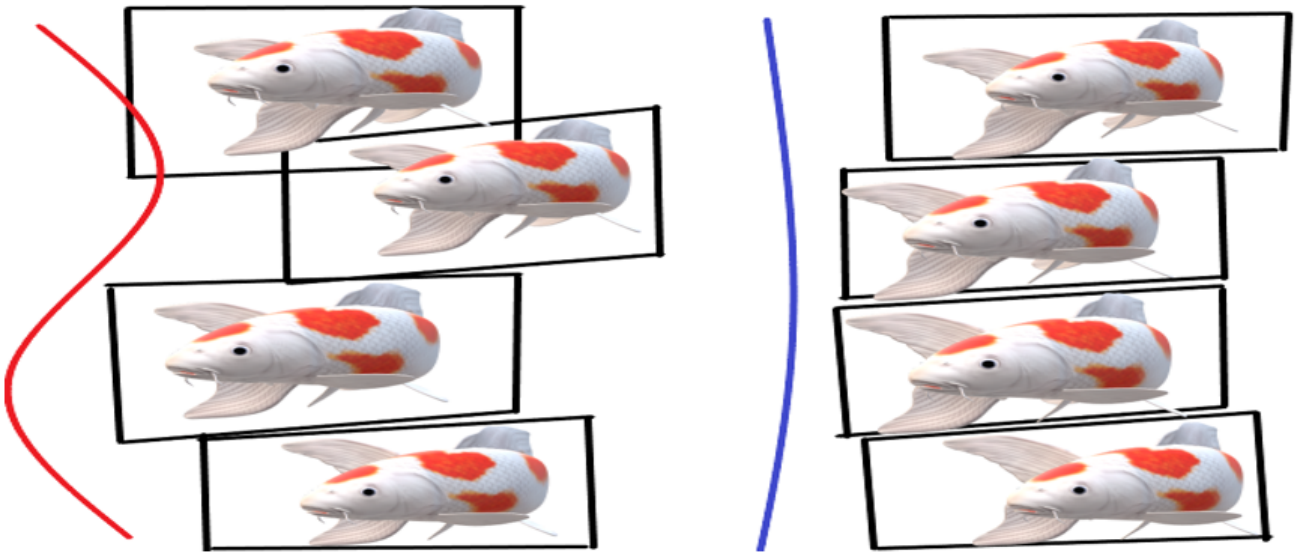

Figure 1: Principle of video stabilization

II. Development of Interpolation-Based Video Frame Interpolations

Interpolation is a technique that creates pixels instead of unknown pixels in image amplifications. There are widely used image interpolation techniques. The first of these methods is the Nearest-Neighbor interpolation. In this method, the distances between the nearest K neighbors and the unknown p pixel are found and the closest value is set to p. The second is bilinear interpolation. Four pixels are indexed as A, B, C, D according to the distance to unknown pixel, p. Using these 4 close points, the value of the unknown pixel is adjusted. Third, Bicubic interpolation is calculated in the similar way to bilinear interpolation. However, it is calculated by evaluating 16 close pixels instead of 4. In this way, it spreads the effect with more pixels. Another method, Cubic B-SPline makes the interpolation curve smoother and creates better image edges [9].

Mahajan et al. proposed a path-based interpolation method. This method preserves the operating frequency content, which takes two input images and generates an intermediate frame, and produces a natural-looking animation sequence. If this sequence is considered as a single scan line among the input frames A and B (shifted form of A), the pixels are copied from A initially and the image starts to be copied from B at some point. This is the interpolation path for unknown pixel p. In this way, a smooth image is obtained [10].

Classical video frame interpolation algorithms include optical flow and frame interpolation [11]. In these studies, the quality of frame interpolation is largely dependent on optical flow. Werlberger et al obtained natural looking intermediate frames (t+1/2) between two input (t, t+1) frames by propagating the optical flow using linear interpolation. This study provides advantages such as finding lost frames and partially correcting the distorted frame. However, optical flow cannot be found in the restored parts. Interpolation techniques are used for the restored parts of the image [12]. Similar work based on optical flow was performed in 3D [13]. There are also different approaches without optical flow. The phase-based method for estimating the motion is the alternative method which uses Euler as a phase-based approach for interpolation. This method reduces the limitations of motion analysis that could be interpolated. It was successful at frame interpolation and retiming of high-resolution high frame rate video [14].

There are also studies combining the video frame interpolation steps of motion estimation and pixel synthesis into a single local convolutional network. Pixel interpolation is performed by convolution on pixel patches instead of optical flow formulation. Its advantages are to handle occlusion, blur and sudden brightness change and it provides high frame interpolation. This method is more flexible than optical flow methods. However, it is not effective for large motions [15]. Another study used large kernels to handle large motions. Due to this memory demand, it limits the number of pixels the kernels can predict in a frame. To solve this problem, the model has formulated the frame interpolation as local separable convolution on input frames using synchronous 1D kernel pairs [16]. Classical optical flow-based methods fail when flow estimation is difficult. Recent neural network-based methods handling pixel values as hallucinations produce blurry solutions. Deep Voxel flow combines two methods. In this work, a network that learns to synthesize video by streaming pixel values from existing ones is trained. It showed that the Deep Voxel flow approach improves the latest CNN techniques to interpolate and extrapolate both optical flow and video.

This method is still limited in adapting to inaccuracies in the motion/voxel flow estimation [17]. Niklaus et al proposed another method solving the occlusion problem of video frame interpolation by estimating and using bidirectional flow. This flow is used to warp and blend the input frames. This approach warps not only the input frames but also the pixel-based contextual information. It used a pre-trained neural network to extract contextual information specifically. Finally, the difference from other methods is that pre-warped frames and context maps are fed into a video frame synthesis network to generate the interpolated frame in a context sensitive manner [18].

III. CNN-Based Development of Video Stabilizatoin

In order to achieve more efficient and consistent video stabilization, learning-based studies have been suggested, recently. Wang created the stable/unstable dataset called DeepStab and proposed a solution with a Siamese convolutional neural network called Stabnet [19]. With a new formulation of the stabilization problem, instead of estimating and correcting a virtual camera path, it produced a stable output in an online fashion that gradually learned the transformation parameters along the timeline for each unstable scene. This approach has been a breaking point in the deep learning approach for video stabilization. However, the 2D homomorphic transformations used in this study may fail on sequential images with heavy movements. Another study proposed a new online deep learning framework to learn the stabilization transformation for each given unstable frame and the generated stable frames, instead of classical stabilization [20]. The network consisted of a generative network with spatial transformer networks embedded in different layers. A stable frame was created for the incoming unstable frame by calculating an appropriate affine transformation. It has limits for severe tremors. Yu et al. (2019) eliminated the training step by learning enough 2D models with a convolutional network (CNN) used only as a regression tool [21]. However, since CNN is a computationally difficult approach. Choi et al. (2020) proposed an unsupervised approach focusing on image interpolation, which has a significant effect on video stabilization to avoid the cropping effect [22]. This method takes two consecutive frames of the video, performs a one-sixth transformation over one of these frames, and a network is trained to predict this transformation. With the help of this network, they proposed an algorithm that takes two frames and creates an intermediate-frame by iterative interpolation method and performs the stabilization process in this way. However, it causes blurring while stabilizing the process [23], [24].In the CAIN approach are used channel attention end-to-end splice video frame interpolation network. They proposed a simple and low computational method without using the optical flow method which reduces the effects of large movements involving occlusion. Instead of optical flow, PixelShuffle was used with channel attention in another study [25]. PixelShuffle progressively distributes motion-related information across channels. Creates the transformed feature map to capture variations between linked frames combined with a channel attention to capture motion. In this way it generalizes movements that are not visible. It provided video frame interpolation, which synthesizes high-quality images without motion estimation. Ali et al. (2020) used DIFRINT motivation instead of optical flow to detect local and global movements in video [26]. Perspective inconsistency in video pairs in the DeepStap [19] dataset has been detected. For this, training video pairs with similar perspectives but different movements were created. With the help of this dataset, motion-blind video has been tried to be stabilized. It offers unsupervised and expandable video frame interpolation-based strategy to produce iso-perspective videos. The first motion blind deep video stabilization network is presented with the help of iso perspective dataset.

IV. Conclusion & Discussion

Interpolation-based stabilization methods stabilize between frames, reducing inter-frame shake without cropping and distortion. However, the quality of the images created due to the iterative approach may decrease as the number of iterations increases. The most important advantage of these methods is that the scenes can be trained end-to-end without supervision.

In situations where flow estimation is difficult, traditional optical flow-based methods fail, while recent neural network methods that view pixel values as hallucinations can produce blurry solutions. Some recent interpolation approaches have used deep network structure, which learns to synthesize video scenes by streaming pixel values from existing fields of neighboring frames. Although these methods are similar to the optical flow method, they are used as an intermediate layer, therefore, their accuracy does not directly affect the result and these methods do not require surveillance like optical flow. However, 2D transformations may also create blur in the images.

Existing methods, handling 2D information, create gaps in distant and near objects. For this reason, perspective can be insufficient. Newer methods processing 3D information promise more efficient results in video stabilization.

As a result, there are shortcomings in video stabilization, such as occlusion, inaccurate estimation of the motion, and poor frame synthesis. For this reason, video stabilization has been up to date for almost 20 years. The fact that the problems in the video stabilization mentioned above have not been fully resolved yet is proof that the method will be popular in the near future.

Appendices

Author Contributions

All authors equally contributed on writing the paper. This article was presented in the 2021 4th International Conference on Artificial Intelligence towards Industry 4.0 (ICAII4'2021), November 11-12, 2021, Iskenderun, Hatay, Turkey.

Acknowledgments

This study was supported by project number 120E447 from the TUBITAK.

Conflicts

None declared.

Ethical Declaration

This article does not contain any studies involving human participants and/or animals performed by any of the authors.

REFERENCES

[1] Kir Savas B, Becerikli Y. Development of driver fatigue detection system by using video images. Journal of Intelligent Systems with Applications 2019; 2(1): 26-29.

[2] Chiu LC, Chang TS, Chen JY, Chang NYC. Fast SIFT design for real-time visual feature extraction. IEEE Transactions on Image Processing 2013; 22(8): 3158-3167.

[3] Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004; 60(2): 91-110.

[4] Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features. In European Conference on Computer Vision (ECCV), May 7-13, 2006, Graz, Austria, pp. 404-417.

[5] Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. In 2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain, pp. 2564-2571.

[6] Lee YC, Tseng KW, Chen YT, Chen CC, Chen CS, Hung YP. 3D video stabilization with depth estimation by CNN-based optimization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA, pp. 10621-10630.

[7] Xu Y, Zhang J, Maybank SJ, Tao D. DUT: Learning video stabilization by simply watching unstable videos. arXiv preprint, 2011.14574, 2020.

[8] Choi M, Kim H, Han B, Xu N, Lee KM. Channel attention is all you need for video frame interpolation. Proceedings of the AAAI Conference on Artificial Intelligence 2020; 34(7): 10663-10671.

[9] Han D. Comparison of commonly used image interpolation methods. In Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE), 2013, pp. 1556-1559.

[10] Mahajan D, Huang FC, Matusik W, Ramamoorthi R, Belhumeur P. Moving gradients: A path-based method for plausible image interpolation. ACM Transactions on Graphics (TOG) 2009; 28(3): 1-11.

[11] Tran LT, Ly NQ. Learning video stabilization using optical flow. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA, pp. 8159-8167.

[12] Werlberger M, Pock T, Unger M, Bischof H. Optical flow guided TV-L 1 video interpolation and restoration. In International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, July 25-27, 2011, St. Petersburg, Russia, pp. 273-286.

[13] Yu Z, Li H, Wang Z, Hu Z, Chen CW. Multi-level video frame interpolation: Exploiting the interaction among different levels. IEEE Transactions on Circuits and Systems for Video Technology 2013; 23(7): 1235-1248.

[14] Meyer S, Wang O, Zimmer H, Grosse M, Sorkine-Hornung A. Phase-based frame interpolation for video. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA, pp. 1410-1418.

[15] Niklaus S, Mai L, Liu F. Video frame interpolation via adaptive convolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA, pp. 670-679.

[16] Niklaus S, Mai L, Liu F. Video frame interpolation via adaptive separable convolution. In 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy, pp. 261-270.

[17] Liu Z, Yeh RA, Tang X, Liu Y, Agarwala A. Video frame synthesis using deep voxel flow. In 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy, pp. 4463-4471.

[18] Niklaus S, Liu F. Context-aware synthesis for video frame interpolation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA, pp. 1701-1710.

[19] Wang M, Yang GY, Lin JK, Zhang SH, Shamir A, Lu SP, Hu SM. Deep online video stabilization with multi-grid warping transformation learning. IEEE Transactions on Image Processing 2018; 28(5): 2283–2292.

[20] Xu SZ, Hu J, Wang M, Mu TJ, Hu SM. Deep video stabilization using adversarial networks. Computer Graphics Forum 2018; 37(7): 267-276.

[21] Yu J, Ramamoorthi R. Robust video stabilization by optimization in CNN weight space. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA, pp. 3800-3808.

[22] Choi J, Kweon IS. Deep iterative frame interpolation for full-frame video stabilization. ACM Transactions on Graphics (TOG) 2020; 39(1): 1-9.

[23] Sarigul M, Karacan L. Classifying stable and unstable videos with deep convolutional networks. Journal of Intelligent Systems with Applications 2020; 3(2): 90-92.

[24] Guilluy W, Oudre L, Beghdadi A. Video stabilization: Overview, challenges and perspectives. Signal Processing: Image Communication 2021; 90: 116015.

[25] Shi W, Caballero J, Huszar F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA, pp. 1874-1883.

[26] Ali MK, Yu S, Kim TH. Learning deep video stabilization without optical flow. arXiv preprint, 2011.09697, 2020.