Digital Video Stabilization: Algorithms and Evaluation

Marcos Roberto e Souza[1], **Helio Pedrini**[1]

[1]Institute of Computing – University of Campinas - UNICAMP
Caixa Postal 6176 – 13083-852 – Campinas – SP – Brazil

marcoosrs@gmail.com, helio@ic.unicamp.br

Abstract. *Several devices have allowed the acquisition and editing of videos in various circumstances, such as digital cameras, smartphones and other mobile devices. However, the use of cameras under adverse conditions usually results in non-precise motion and occurrence of shaking, which may compromise the stability of the obtained videos. To overcome such problem, digital stabilization aims to correct camera motion oscillations that occur in the acquisition process, particularly when the cameras are mobile and handled in adverse conditions, through software techniques, without the use of specific hardware, to enhance visual quality either with the intention of enhancing human perception or improving final applications, such as detection and tracking of objects. This is important in order to avoid hardware cost and indispensable for videos already recorded. This work proposed three methods to perform digital video stabilization and two other techniques to evaluate video stabilization quality.*

1. Introduction

In this work, we are particularly interested in investigating two-dimensional (2D) video stabilization methods, in which geometric transformations are employed to represent frame-to-frame motion and stabilize the videos. The reason for this interest is that even though 3D methods allow higher quality stabilization, 2D methods have a lower computational cost and are more robust to a variety of situations, which causes them to be constantly preferred in practice. The 2D digital video stabilization process is usually divided into three main steps: (i) camera motion estimation, where the motions performed by the camera are estimated, constructing a path that corresponds to the one traveled by the camera, (ii) removal of unwanted motion, which smooths the unstable video motion, and (iii) generation of the corrected video, which transforms the video frames according to the remaining motion.

This work aimed to investigate and evaluate digital video stabilization methods for correcting disturbances and instabilities that occur during the process of video capture. It also proposed novel methods for digital video stabilization and for qualitative evaluation of the video stabilization process. Experiments were conducted on several video sequences. A comparative analysis of the results obtained with the proposed method and with other approaches of the literature were presented and discussed.

The main contributions of this work, which provided the following publications, are: (i) a consensual approach to combining different methods of local features in motion estimation. We experimentally demonstrated that the results of individual methods could be improved by combining different methods. This method was published in the Signal, Image and Video Processing journal (Qualis B1) [Souza and Pedrini 2018a]; (ii) an

approach that detected failures in the global motion estimation obtained through local features and proposed an optimization technique to calculate a new estimate of the corrected motion. Experiments showed that estimation of the optimization method is considerably superior when compared to the individual use of local features. The state-of-the-art stabilization method used in YouTube [Grundmann et al. 2011] was also used for comparison, which presented typical flaws when using local features to motion estimation, obtaining in these cases a worse result than the method proposed. Although there are newer approaches, the literature predominantly uses YouTube's method as a reference. In addition, the other methods have few changes in motion estimation, typically made through local features. This method was published in the IET Image Processing journal (Qualis B1) [Souza et al. 2018]; (iii) a new technique for removing unwanted motion based on the Gaussian filter to smooth the camera path. Experiments demonstrated the effectiveness of the method, which generated videos with proper stabilization rate while maintaining a reasonable amount of frame pixels. This method was published in the EURASIP Journal on Image and Video Processing (Qualis B1) [Souza and Pedrini 2018b]; (iv) new techniques for the qualitative evaluation of video stabilization through visual representations based on visual rhythms and motion energy image. We proposed a visualization scheme based on visual rhythms to represent the behavior of the motion present in a video. In addition, a visualization based on motion energy image was used to represent the amount of motion present in a video. Both proposed evaluation approaches were intended for human beings to assess the quality of the stabilization. Experimental results demonstrated that the both visual representations were effective to evaluate the stability of camera motion by differentiating stable and unstable videos. From these two methods, the paper about the motion energy image was published in The Visual Computer journal (Qualis A2) [Souza and Pedrini 2017]. A paper related to visual rhythms has been submitted to the EURASIP Journal on Image and Video Processing (Qualis B1) [Souza and Pedrini 2018c].

2. Video Stabilization Methods

The first proposed video stabilization method is a consensual approach to combining different methods of local features in motion estimation. Initially, M local features methods are applied for each pair of frames. Then, the local features are matched considering the local features of each method separately. A pre-evaluation is performed to remove the methods with a potentially large number of outliers. A consensual combination is then applied in the remaining methods, such that only local features that are consistent with their transformation are considered as final local features. This combination can be seen as a method based on RANSAC, which makes use of different sources of information rather than considering random samples from the same source. The motivation of this combination is to use different methods and types of local features to confirm the motion between two frames.

The second method is an approach that detects failures in the global motion estimation obtained through local features and proposes an optimization technique to calculate a new estimate of the corrected motion. Initially, the motion estimation between two consecutive frames is made through local features methods. Next, we applied a consistency check on the estimated matrix, comparing it with the estimated (and final) in the previous frame pair. If there is inconsistency, the motion estimation is done again using

the proposed optimization method. The structural similarity index (SSIM) is the basic evaluation metric for the detection and optimization steps. Although we use the local feature method as basis, our optimization method can be applied to any other approach that estimates the motion between pairs of frames. In this method, the Powell method is used to minimize our objective function.

We also proposed a new adaptive technique for removing unwanted motion based on the Gaussian filter to smooth the camera path. We consider a vertical translation factor, a horizontal translation factor, a rotation factor and a scaling factor. Each factor f of the matrix is decomposed and the trajectory of each of them is calculated in order to accumulate its previous values. In the adaptive Gaussian filter, the trajectory will be smoothed by considering a distinct value for σ_i at each point i. The value of σ_i is greater when there is greater variation in values around a neighborhood. After applying the motion filtering, it is necessary to recalculate the value of each factor for each transformation matrix. In order to do that, the transformation matrix value of a given factor is calculated by the difference between each point of its smoothed trajectory and its predecessor.

In the evaluation based on visual rhythms, two different path directions are considered: horizontal and vertical. The vertical rhythm extracts the information from the columns of each frame, while the horizontal rhythm takes the information from the lines of each frame. For both path directions, the rhythm is obtained from the sequential concatenation of the information, so that the j-th column of the visual rhythm image corresponds to the information in the j-th frame. The width of a visual rhythm corresponds to the number of frames of the video, whereas its height corresponds to the height or width of the frames for the vertical or horizontal rhythm, respectively. The use of only one column or row in the extraction of information from each frame, as it is done in the literature, may be inadequate since it considers little information of the frame. In addition, it makes horizontal and vertical separation less accurate. Thus, the average of the columns or rows is adopted in our work to compensate for this difference. As post-processing, we apply an adaptive histogram equalization technique through the Contrast Limited Adaptive Histogram Equalization (CLAHE).

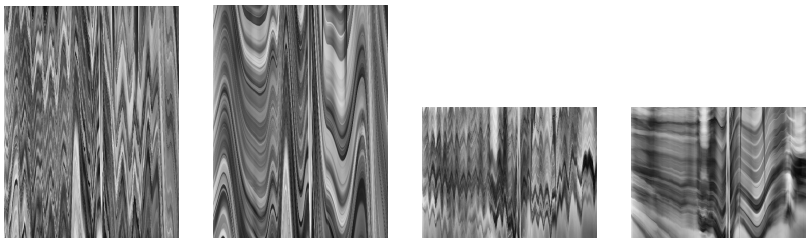
We conjecture that the stabilization evaluation can be done through the amount of motion present in the video, which complements the analysis of the motion behavior. For each video frame i, the difference of the gray level intensities of each pixel is calculated. In this step, a binary image is obtained, in which 1 is assigned to the pixel with difference greater than a certain threshold, and 0 otherwise. We consider a Motion Energy Image (MEI) for each frame i, which is obtained through the differences of the frames within a sliding window of size W_{MEI}, centered in i. By taking the MEI of each frame, the average image of the MEIs is calculated, where each pixel (x, y) is taken as the arithmetic mean of the pixels (x, y) of all the MEIs of the video. A pseudocolor transformation is applied, so that high gray-level intensity values are mapped to red, whereas lower intensities to blue. A detailed description of all the methods briefly presented in this section can be found in Chapter 3 of the Master's Dissertation [Souza 2018].

3. Experiments

Three databases are used to evaluate the effectiveness of the proposed video stabilization methods. The first consists of eleven videos available in the GaTech

VideoStab [Grundmann et al. 2011] dataset and three others collected separately. The second, available by Liu et al. [Liu et al. 2013], consists of 139 videos divided into categories. Finally, we create a dataset that is complementary to the others, in which four videos are collected separately. From the original videos, excerpts with moving objects in the foreground and with little representative background are extracted, generating a total of eight videos.

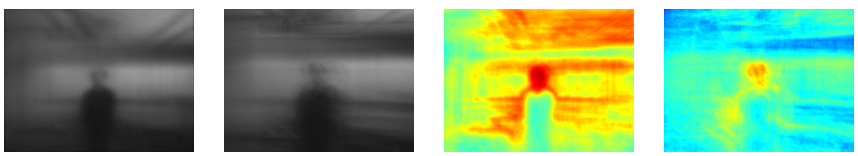
Figure 1 presents the visual rhythms generated for the video #12 of the first dataset. In order to obtain the stabilized version of the video, we submit it to YouTube, which applies one of the state-of-the-art digital video stabilization approaches [Grundmann et al. 2011]. We can notice the twitches and irregularities present in the lines of original rhythms. On the other hand, there are more continuous, well defined and softer lines in the rhythms of the stabilized video.



(a) (b) (c) (d)

Figure 1. Visual rhythms for video #12. (a) horizontal rhythm, original video. (b) horizontal rhythm, stabilized video. (c) vertical rhythm, original video. (d) vertical rhythm, stabilized video.

Figure 2 shows the results of a simple frame average and the average of the MEIs for video #7 of the first dataset. From the frame average, it is not so easy to differentiate the unstable video from the stabilized one. In fact, the stabilized video seems to have more amount of motion. On the other hand, the stabilized video presents an average MEI image with bluer tones, correctly indicating a smaller amount of motion. This visual representation is efficient to show the amount of motion present in a video, making possible the evaluation and comparison of different stabilization methods. Our technique is more effective than the simple average of the gray levels of the video frames, which can generate inaccurate results when considering the intentional motion of the camera and small changes in the scene.



(a) (b) (c) (d)

Figure 2. Average grayscale and colored MEIs for video #7. (a) average, original video. (b) average, stabilized video. (c) MEI, original video. (d) MEI, stabilized video.

From the results obtained with the consensual local feature combination method, we can see that, when we combine methods with inferior results, the combination leads to better performance for most videos for both PSNR and SSIM, so that weaker methods produce superior results. These results show that the application of the combination

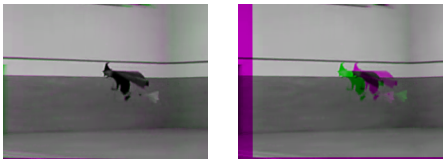
strategy obtains a greater robustness in the motion estimation. We notice that the results obtained with the combination are lower than with the SURF method. This occurs because the results achieved with SURF for these two datasets are already very good, correctly estimating the global motion between two frames in practically all cases.

Figure 3 presents a failure situation of local features for different videos, as well as the correction performed with our optimization method. Matches considered as inliers by the RANSAC method are drawn in blue and green, whereas the outlier matches are drawn in pink and yellow. We can see that the matches of the object were considered as inliers, which made the movements of the object, not the camera, compensated. On the other hand, our optimization-based method obtained excellent results, finding the transformation matrix that matches the motion performed by the camera.

(a) (b) (c)

Figure 3. Motion estimation for the 481th frame of video $\texttt{ours}_2$. (a) local features; (b) warped frame; (c) our result.

Higher values of similarity measures, such as PSNR or SSIM, may indicate a better quality in the motion estimation for most cases. However, there are cases where such measures do not indicate the correct estimate and, therefore, a simple optimization that takes the measures into account would not be efficient. Figure 4 presents different matches where different values of PSNR and SSIM are obtained. It can be observed that higher values are obtained in incorrect cases. Since background is unrepresentative, higher similarity is obtained if object matching is done. However, this is semantically incorrect since the object is in motion.



(a) (b)

Figure 4. Different matches for the 40th frame of video $\texttt{ours}_7$. (a) PSNR = 30.576 and SSIM = 0.932. (b) PSNR = 23.491 and SSIM = 0.896.

From the results obtained with the removal of unwanted motion, the proposed adaptive Gaussian achieved values comparable to the original version, maintaining considerably more pixels (up to 50 percentage points more). From the results obtained in the second dataset, the gain in the percentage of pixels held was more significant in the `QuickRotation`, `Zooming` and `Running` categories. Compared to the Youtube method, we can observe a certain parity for both methods in terms of ITF and ITF_{SSIM} metrics, with a slight advantage of the YouTube, while the maintained pixels are in general comparable and, when lower, they do not differ much. We also present the final results of the stabilization of the videos considering the process of spatio-temporal optimization, after the process of removal of unwanted motion. We can notice that the visual rhythms

obtained in the stabilization through estimation based on local features and in the version obtained with the `YouTube` method have several discontinuities, which represent abrupt movements in the videos, both due to the problem in the motion estimation. However, the rhythm generated with the spatio-temporal optimization method is significantly more regular, representing a better quality in the video stabilization process. All the experiments conducted on the datasets, as well as the performed comparisons, are presented and discussed in detail in Chapter 4 of the Master's Dissertation [Souza 2018].

4. Conclusions and Future Work

The main objective of this work was to investigate the problem of video stabilization. We then developed and evaluated 2D methods for digital stabilization of videos. The 2D video stabilization process is usually divided into three main steps: estimation of camera motion, removal of unwanted motion, and generation of the corrected video. This work presented five novel methods related to digital video stabilization. Experiments were conducted on three distinct sets of videos.

From the investigation conducted on this work, we have identified some directions that can be explored in future work: (i) construction of a local motion estimation technique for the frames in which the optimization is applied, extending our method to deal with local motion, (ii) development of a new method for removing unwanted motion through a constrained optimization, which obtained the smoothest camera path possible considering a certain minimum amount of frame pixels to be held, and (iii) proposition of objective metrics calculated from the visual representations proposed in our work, using them for the characterization and evaluation of video stabilization.

References

Grundmann, M., Kwatra, V., and Essa, I. (2011). Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 225–232. IEEE.

Liu, S., Yuan, L., Tan, P., and Sun, J. (2013). Bundled Camera Paths for Video Stabilization. *ACM Transactions on Graphics*, 32(4):78.

Souza, M. R. (2018). Digital Video Stabilization: Algorithms and Evaluation. Master's thesis, Institute of Computing, University of Campinas.

Souza, M. R., da Fonseca, L. F. R., and Pedrini, H. (2018). Improvement of Global Motion Estimation in Two-Dimensional Digital Video Stabilisation Methods. *IET Image Processing*, 12(12):2204–2211.

Souza, M. R. and Pedrini, H. (2017). Motion Energy Image for Evaluation of Video Stabilization. *The Visual Computer*, pages 1–13.

Souza, M. R. and Pedrini, H. (2018a). Combination of Local Feature Detection Methods for Digital Video Stabilization. *Signal, Image and Video Processing*, 12(8):1513–1521.

Souza, M. R. and Pedrini, H. (2018b). Digital Video Stabilization Based on Adaptive Camera Trajectory Smoothing. *EURASIP Journal on Image and Video Processing*, 2018(1):37.

Souza, M. R. and Pedrini, H. (2018c). Visual Rhythms for Qualitative Evaluation of Video Stabilization. *EURASIP Journal on Image and Video Processing*. (submitted).