VIDEO STABILIZATION PERFORMANCE ASSESSMENT

Matti Niskanen, Olli Silvén

Machine Vision Group, Infotech Oulu
P.O.Box 4500, FIN-90014 Univ. of Oulu,
Finland

Marius Tico

Nokia Research Center
P.O.Box 100, FIN-33721 Tampere,
Finland

ABSTRACT

Shooting videos with a hand-held camera introduces shaking, which incontrovertibly reduces video quality. Digital video stabilization is a process to compensate for camera motion by means of image processing. In the best case, it not only removes the image motion, but also reduces image distortion caused by unintentional camera motion.

In practice, removing solely unwanted jitter cannot be achieved precisely. Furthermore, the stabilization process itself often introduces some additional distortion in images instead of removing it. In this paper, various means to automatically evaluate the performance of the video stabilization process are proposed, based on measuring the divergence and jitter of the remaining unintentional motion and blurring using point spread function (PSF). This helps, for example, in tuning the system parameters for better quality.

1. INTRODUCTION

When a scene is imaged with a hand-held or a vehicle-mounted video camera, the output is most likely not what was intended; it is rather a distorted representation of the view. Shaking of the camera leads to a shaking video sequence, where a lot of unwanted motion exists. A video stabilization process tries to remove this unintentional, typically high-frequency motion, known as jitter, and thus ought to provide more satisfactory video sequences.

Digital video stabilization is typically considered to contain three successive steps: motion estimation, motion filtering, and motion compensation. Success in each of these phases affects the quality of the resulting video. For example, image motion that is caused by camera motion has to be separated from other motion seen in a view, and only the unintentional part of this motion should be removed. Compensating for the motion should not decrease the image quality. However, typically a decrease in resolution is inevitable, and the interpolation required by rotation, scaling and translation in sub-pixel accuracy causes additional blurring to the image.

Video stabilization is a common feature in much video processing software and many camcorders today. Assessment of the video stabilization performance is important in order to tune and compare different methods. Nevertheless, there is little reported work about this in a literature. This paper proposes a number of means to automatically assess the performance of the video stabilization process, and provides the following original contributions:

- **A review** of existing video stabilization performance assessment methods is given.
- **Motion estimation and filtering** is assessed by decomposing the remaining image motion into *jitter* and *divergence*, providing more informative objective measurements.
- **Blurring** is recognized as a problem of digital stabilization, and a means to assess it via PSF is proposed.

The remainder of this paper is organized as follows. Section 2 discusses the overall performance of video stabilization and is mainly a review of the existing approaches. Section 3 divides performance assessment into sub-problems and proposes new informative attributes to the measure. Section 4 describes the experiments and tabulates some example results. Finally, Section 5 concludes the work.

2. OVERALL VIDEO STABILIZATION PERFORMANCE

One way of comparing stabilization algorithms is to compare the performance of an application that uses stabilized videos [1, 2]. Applications range from object tracking to bit rate reduction in video compression [1, 2, 3]. However, mainly stabilization is performed to satisfy the human eye, and it is ultimately a subjective opinion as to how good the resulting quality is. A mean opinion score (MOS) is a value obtained from a number of subjective opinions (for example, see [4, 5]). It is clear, that such a slow and expensive arrangement is often not possible, and thus, MOS is more often seen as a way to validate automatic criteria [5].

Typically automatic criteria for image quality are based on a simple pixel-by-pixel comparison between the ground truth image and the reference image. The most common of such criteria are the peak signal-to-noise ration (PSNR) and the related mean square error (MSE) (e.g. in [6, 7]). They are widely used, primarily because of their mathematical tractability (simple to calculate and differentiable). However, it is well known that they do not correlate well with perceived quality measurements [5, 8, 7]. Despite the problems, these pixel-wise approaches are often extended to video quality assessment, simply by comparing still images on a frame-by-frame basis [5].

Most of the video quality assessment work reported in the literature is intended for video compression applications, where typical distortion differs from that in video stabilization, and most of the criteria proposed for compression applications cannot be utilized for stabilization. For video stabilization, the most important quality affecting factor is misalignment between the frames.

Studies of the human visual system (HVS) support the intuition that the amount of displacement does not contain enough information alone. Human sensitivity to motion depends on a combination of motion frequency and amplitude [9], but also on spatial image frequency, color and intensity, and even the context (see [8, 5, 4] for details). Some objective image and video quality (of compressed video) assessment methods do indeed try to incorporate the foundations of HVS [5, 4]. We believe that, for video stabilization, the HVS is best incorporated by measuring the more informative attributes.

In the next section, we propose criteria to measure both the amount and nature of displacement. Another criterion is suggested that is suitable for measuring the blurring, which is a typical decrease in image quality with digital video stabilization. An overall performance could be then pooled from all the individual measurements, if necessary.

3. TOWARDS INFORMATIVE OBJECTIVE CRITERIA

Figure 1 summarizes the whole chain of video stabilization assessment using artificial videos, as will be proposed in this chapter. The left part of the figure sketches the generation of artificial videos and camera shaking caused distortions, that should be removed. The middle part shows the main steps of digital video stabilization and distortions that are reduced, but also some additional ones that might be introduced. The right part of the figure summarizes how these attributes can be separately assessed using the proposed criteria.

3.1. Motion estimation and filtering

The amount and type (frequency content) of estimated and compensated motion form clearly the most interesting characteristics to measure when video stabilization performance is being assessed. To evaluate the quality of the motion estimation and compensation, some authors show graphically the motion of the original and stabilized videos, either in the time [10, 11, 12] or in frequency domain [3, 9]. Similarly, only a error component might be shown [13]. A professional may obtain a great deal of information from these, but they are suitable only for a subjective evaluation of the results.

3.1.1. Amount of misalignment

Motion estimation and filtering phases can be assessed separately only if the amount of estimated motion is known. Typically, one has no access to the estimated parameters, but can only evaluate the remaining motion from the videos, and these two phases are assessed simultaneously.

PSNR gives some indication about the misalignment between two otherwise equal frames. Two PSNR based criteria that indicate the long and very short term stabilization performance for a fixed view were proposed for a video surveillance application by [14], and further utilized by [2].

Misalignment between the frames can, however, also be computed in a more intuitive manner. It should be noted, that typically a view does not have to be fixed, but there can be intentional camera motion which has to be separated from unwanted motion. While with synthetic videos the intended motion is unambiguous, for real data, this is a subjective, task and video specific matter, and some assumption has to be made. We have used a $1Hz$ cut-off frequency as such an assumption, as this seems to be close to what most users find to be a natural threshold. The deviation between the stabilized and intended frame position over time (i.e. *remaining unintentional motion*) is then measured in terms of translation, rotation and scaling, using the Fourier-Mellin transform. In the following subsection, we propose a method for decomposing these deviations into more meaningful measurements.

3.1.2. Interpreting the motion

Especially if there is intended motion in the camera, stabilized video may follow behind the original video. It can be very stable in nature, but the average *divergence* might be large. On the other hand, it is possible that the stabilized video is averagely very well aligned with the intended one, but contains a lot of unwanted *jitter*. Such a case can produce smaller average displacement, but is more irritating from a human point of view. There is always a trade-off between jitter and divergence.

We decompose unintentional motion to a divergence (bias term), and to jitter. These are obtained with a low/high pass filter with a certain cut-off frequency c that decomposes the signal. The low frequency part $e_{f\leq c}(i)$ is the expected error during the frame i, and its square forms the divergence:

$$D_c = \frac{1}{frames} \sum_{i=1}^{frames} \{e_{f\leq c}(i)\}^2. \quad (1)$$

Similarly, the square of the high frequency part forms the jitter:

$$J_c = \frac{1}{frames} \sum_{i=1}^{frames} \{e_{f>c}(i)\}^2. \quad (2)$$

Misalignment e may indicate the difference between the obtained and optimal parameter of position along the x- and y-axis, roll angle or scaling, for example. The measurements of equations 1 and 2 can be alternatively computed from the power spectral density (PSD) functions.

$$\textit{Jitter attenuation} = J_{c\_stabilized} / J_{c\_original} \quad (3)$$

indicates the amount of remaining jitter relative to the original jitter, providing a value that is more independent of the original motion. It is common in signal processing to give attenuation in decibels. However, in this paper we have used the direct formula above for the sake of intuition.

As divergence (eq. 1) is the square of the low frequency components, its square root indicates the *expected amount of displacement*,

$$\mathrm{E}\{|e|\} \approx \sqrt{D}. \quad (4)$$

3.2. Compensation

With digital video stabilization, it is common that the observed image is a blurred version of the optimal one. A human observer often finds this blurring irritating, even if images were perfectly aligned. Thus, a measure to indicate the blurring is needed. While PSNR for aligned frames is somewhat dependent also on this, we propose a more accurate method to assess directly the blurring process itself.

In the case of linear blurring, the observed image I_{obs} is obtained by convoluting the original image I_{orig} with a convolution mask h, $I_{obs} = I_{orig} * h$, where h is known as a point spread function (PSF). It defines how a single bright spot in the original image is observed in the other.

This kind of image deformation is common with motion blur, where camera pan and tilt cause the image to move on a sensor during the exposure. Linear convolution occurs also in the interpolation required by image scaling, rotation and translation in sub-pixel accuracy. The amount of additional blurring introduced by a stabilization process itself is affected by an interpolation filter and other implementation details. For example, if translation, rotation and scaling are not combined but performed sequentially, the expected amount of blurring is clearly higher.

To estimate a PSF (solve h from the equation above), we take sample images containing spatially high frequency components, and a PSF between two aligned windows is then assessed using the frequency domain utilizing standard signal processing techniques. The

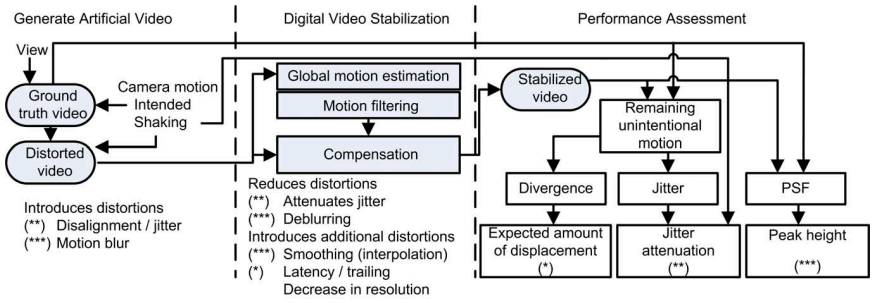


Fig. 1. Assessing video stabilization performance

Table 1. Motion in test sequences

Run	Artificial motion of planar view
	Heavy artificial (run) jitter
Window	Almost planar + moving objects
	Small artificial (still) jitter
Buildings	Some parallax motion
	Heavy high frequency jitter
Walking	Lot of parallax, camera dolly
	Heavy medium frequency jitter

normalized *peak height* of the PSF is taken as a measurement to indicate the preservation of high frequency image components. Before the PSF is computed, the frames need to be aligned and the effect of additional interpolation possibly required by this can be removed by convoluting the resulting PSF with this same interpolation filter. Without the reference frame, one can get hints of blurring by directly comparing the amount of high frequency components.

3.3. Other performance criteria

Various interesting issues can be solved by special purpose test material. For example, we have measured jitter attenuation to solve the ranges of compensated motion, amplitude responses of motion filters, and noise sensitivities of stabilization methods. Also other special videos, such as videos with high periodic texture content or videos with an exaggerated amount of blurring and other image deformations can be used to distinguish between the different methods.

Some performance criteria are very much application specific. If stabilization is performed off-line by computer, memory usage, computational cost or latency are of secondary importance, while with small hand held devices that stabilize videos on-line, these issues are clearly very important.

4. EXPERIMENTS

We have compared two different video stabilization methods, A and B, with the algorithms proposed. Fig. 2 shows an example, where translation about the y-axis of a test sequence 'Run' is observed. The upper part of the figure represents the position of the original, ideal, and two stabilized videos. The middle part shows an error between the ideal position and the position obtained by stabilization. A bias is the low frequency part of the error ($e_{f \leq c}$) indicating the expected error. System B is clearly not aligned with the ideal signal, containing large divergence. However, the remaining motion is on average even slightly smoother than for system A. The expected amount of displacements (on average 2.2 and 16.5 pixels) and jitter (average of 1.2 and 0.8 pixels^2), for systems A and B respectively, are decomposed from error signals using $1Hz$ cut-off frequency and are shown at the bottom of the figure.

Fig. 3 (a) shows an average PSF between the optimal video and the stabilized video. Video stabilization is made for an artificially generated shaking video, which contains also some motion blur. From the PSF function (above) or from its amplitude response below, one can see that on average a lot of high frequency components are lost during the shaking/stabilization process. As this blurring is not evenly distributed among all the frames, in some frames, it can be very irritating.

As we know that neither of the systems under evaluation tries to correct the motion blur, it is sufficient to measure the additional blurring that is caused by the stabilization process itself. PSFs between the shaking and the stabilized videos are not affected by motion blur, and are shown in Fig. 3 (b) and (c) for systems A and B, respectively. The peak values, 0.53 and 0.59 indicate that system B is slightly better here. However, unlike system B, system A scales the cropped image to the original size. If the scaling were performed afterward also on system B, it would require another interpolation and reduce the value to below the former one. Nevertheless, both of these values are rather good, being close to the theoretical average 0.56 of bilinear interpolation. The nearest neighbor interpolation (i.e. full pixel translation) would produce a peak of height 1 and a bicubic interpolation of about 0.71.

Four test videos are characterized in Table 1. The first two are artificial videos. Run is generated by shaking the still image, and the second (Window) by shaking the stable ground truth video. Also other distortions, such as motion blur and the effect of a rolling shutter, are simulated. Last two videos (Buildings and Walking) are real videos, where the ground truth motion is assumed by a low-pass filtering the original motion. The complete results for these are shown in Table 2.

The results reveal that translational jitter attenuation is about the same with both methods, and stabilization indeed smoothes the jitter. Rotational motion is compensated only by system A. It is also capable of following the intended position much more closely than system B (5 pixels compared to over 20 pixels).

Preservation of high-frequency components, the peak of PSF, is slightly better for system B. This is mainly due to the fact that it neither scales nor rotates the image, and thus produces more frames with full pixel translations only. As a whole, the results seem to favor system A. Subjective evaluations by the authors conform the results. All the criteria seem reasonable, while their relative importance depends much on the video content.

5. CONCLUSIONS

The stabilization obtained consists of two kinds of errors in motion: divergence and jitter. Divergence is related to the expected error and increases, for example, if there is latency between the videos, or if the motion compensation filter is not capable of reacting quickly enough to the desired motion. Jitter is the remaining high frequency motion component. In this work, we proposed a measurement method for these two properties from video sequences.

In addition, we proposed a method to estimate the blurring process of video stabilization. With a ground truth video, it provides means of accurately estimating the motion blurring process. The

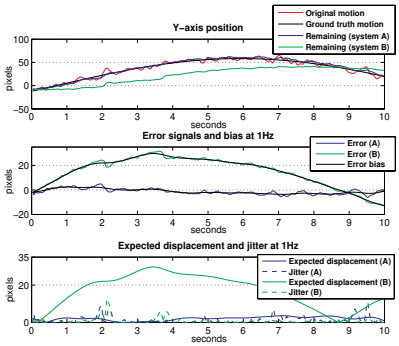
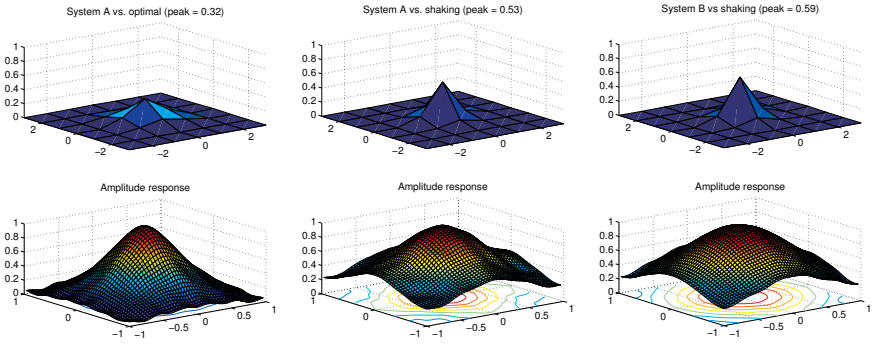


Fig. 2. Decomposing error signal for two stabilization methods.



(a) (b) (c)

Fig. 3. Average PSFs in the top row and their corresponding amplitude responses below, (a) from optimal to stabilized, (b) From shaking to system A, and (c) From shaking to system B.

Table 2. Comparison of two stabilization methods

Run		**Window**		**Buildings**		**Walking**		Total	
A	B	A	B	A	B	A	B	A	B
Translational jitter attenuation, $c = 1Hz$									
0.09	**0.05**	**0.5**	1.5	**0.21**	0.51	**0.17**	0.18	**0.12**	0.14
Rotational jitter attenuation, $c = 1Hz$									
0.12	1.00	1.35	**1.00**	**0.60**	1.00	**0.20**	1.00	**0.15**	1.00
Translational expected displacement (pixels), $c = 1Hz$									
2.6	27.9	**4.6**	13.9	**9.0**	32.5	**4.1**	12.2	**5.1**	21.6
Rotational expected displacement (degrees), $c = 1Hz$									
0.25	**0.18**	0.04	**0.01**	0.35	**0.02**	0.25	**0.06**	0.22	**0.07**
Normalized PSF peak height									
0.53	**0.59**	0.47	**0.68**	0.52	**0.77**	0.54	**0.81**	0.52	**0.71**

same procedure, with a shaking video, can be used for estimating the additional blurring caused by the stabilization process itself.

Comparison between the different systems or tuning of the parameters in order to improve the quality of the stabilization method, require informative and objective criteria to be measured. The proposed intuitive measurements about divergence, jitter, and blurring, fulfill this requirement.

6. REFERENCES

[1] S.B. Balakirsky and R. Chellappa, "Performance characterization of image stabilization algorithms," *Real-Time Imaging*, vol. 2, pp. 297–313, 1996.

[2] L. Marcenaro, G. Vernazza, and C.S. Regazzoni, "Image stabilization algorithms for video-surveillance applications," in *International Conference on Image Processing*, 2001, pp. I: 349–352.

[3] A. Engelsberg and G. Schmidt, "A comparative review of digital image stabilising algorithms for mobile video communications," in *IEEE Transactions on Consumer Electronics*, 1999, pp. 591–597.

[4] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, no. 2, pp. 231–252, 1999.

[5] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furth and O. Marqure, Eds., chapter 41, pp. 1041–1078. CRC Press, 2003.

[6] R.M. Kinape and M.F. Amorim, "A study of the most important image quality measures," in *Proceedings of the 25th Annual International Conference of the IEEE EMBS, Cancun, Mexico*, 2003, vol. 1, pp. 934–936.

[7] Ahmet Eskicioglu and Paul S. Fisher, "Image quality measures and their performance," in *IEEE Transactions on Communications*, 1995, pp. 2959–2965.

[8] Yao Wang, Ya quin Zhang, and Joern Ostermann, *Video Processing and Communications*, Prentice Hall PTR, 2001.

[9] M. Oshima, T. Hayashi, S. Fujioka, T. Inaji, H. Mitani, J. Kajino, K. Ikeda, and K Komoda, "Vhs camcorder with electronic image stabilizer," *IEEE Transactions on Consumer Electronics*, vol. 35, no. 4, pp. 749–758, 1989.

[10] S. Erturk, "Image sequence stabilisation: motion vector integration (mvi) versus frame position smoothing (fps)," in *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis, Pula, (Croatia)*, 2001, pp. 266–271.

[11] K. Ratakonda, "Real-time digital video stabilization for multimedia applications," in *Proceedings of the IEEE International Symposium on Circuits and Systems, Monterey, (USA)*, 1998, vol. 4, pp. 69–72.

[12] Jesse S. Jin, Zhigang Zhu, and Guangyou Xu, "Digital video sequence stabilization based on 2.5d motion estimation and inertial motion filtering," *Real-Time Imaging*, vol. 7, no. 4, pp. 357–365, 2001.

[13] Pyung Soo Kim, "Fir filtering based image stabilization mechanism for mobile video appliances.," in *Computational and Information Science, First International Symposium, Shanghai, (China)*, 2004, pp. 1106–1113.

[14] C. Morimoto and R. Chellappa, "Evaluation of image stabilization algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, (USA)*, 1998, vol. 5, pp. 2789–2792.