CHARACTERIZING DISTORTIONS IN FIRST-PERSON VIDEOS

Chen Bai and Amy R. Reibman

School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA

ABSTRACT

First-person videos (FPVs) captured by wearable cameras often contain heavy distortions, including motion blur, rolling shutter artifacts and rotation. Existing image and video quality estimators are inefficient for this type of video. We develop a method specifically to measure the distortions present in FPVs, without using a high quality reference video. Our local visual information (LVI) algorithm measures motion blur, and we combine homography estimation with line angle histogram to measure rolling shutter artifacts and rotation. Our experiments demonstrate that captured FPVs have dramatically different distortions compared to traditional source videos. We also show that LVI is responsive to motion blur, but insensitive to rotation and shear.

***Index Terms*—** first-person videos, image quality, video quality, motion blur, rolling shutter

1. INTRODUCTION

Many so-called first person videos (FPVs) or egocentric videos recorded by wearable video cameras (Pivothead, Looxcie Camera, Gopro, Google Glass) have been widely shared on Twitter, Youtube and other personalized streaming. Research related to FPVs have explored activity recognition [1], video summarization [2], interaction detection [3] and "snap points" prediction [4].

FPVs differ from traditional videos which are captured by stably-mounted cameras. Most frames captured in traditional videos have high quality, free from distortions. These videos usually provide a comfortable viewing experience. However, typical camera wearers rarely record FPVs with an intention to control the camera. Therefore, most frames are subject to distortions due to random camera motion. Continuously shaking scenes due to body or head movement dramatically reduce viewability and cause discomforts for viewers.

To measure quality of videos, full-reference (FR) and no-reference (NR) metrics are commonly used. In FR metrics, source videos are used as references for distorted videos [5]. However, FPVs often have such low quality that they can not provide adequate reference information. NR metrics overcome the absence of reference information, but they have significant content variation (see [6]) and ignore the available information from neighboring frames with similar content. On the other hand, distortions in FPVs often contain motion blur and geometric distortions, including rolling shutter artifacts and rotation. These types of distortions are rarely considered. Moreover, distortions in FPVs are also such that different types of distortions could exist simultaneously and the amounts may vary spatially. An image may have motion blur with heavy rolling shutter artifacts in its top half, but be clear with few rolling shutter artifacts in its bottom half.

Existing no-reference blur metrics (see [7, 8]) cannot accurately compare two images that share only half of their content, because the blur metrics are inherently content-dependent and half of the two images have different content. Quality metrics considering geometric distortions have been studied in [9, 10]. However, these two metrics are based on a full-reference image metric, SSIM [11], and neither are robust when images suffer from large motion blur. Therefore, a new metric for distortion measurement should be designed specifically for FPVs.

Our proposed method classifies different distortions in FPVs. It separates distortion classification into blur measurement and geometric measurement. Blur measurement considers motion blur, and applies an information-based algorithm, called local visual information (LVI). The algorithm mathematically measures the information received by the human visual system for two images, and uses the information ratio to estimate their relative blur. Geometric measurement considers rolling shutter artifacts and rotation. Homography estimation [12] and line angle histogram [13] are two basic methods. The line angle histogram detects the rotation and shear, and the homography estimation measures geometric transformation parameters between two images. In section 2, we describe different distortions in FPVs. In section 3, we describe the overall classification framework and illustrate the LVI algorithm, homography estimation and line angle histogram. In section 4, we present test results of synthetic distortions and video statistics of FPVs and traditional videos.

2. DISTORTIONS

Distortions in FPVs mainly result from camera panning, both horizontal and vertical, and camera shaking due to head and body movement of the camera wearer.

Motion blur is mainly caused by rapid movement of the wearable camera. During one finite exposure time, the objects

change positions continuously relative to the camera. This type of distortion is present in most frames, and often appears with geometric distortions simultaneously.

Rolling shutter artifacts mainly arises from camera panning, both vertically and horizontally, and from camera shaking. In wearable cameras, one frame is exposed from the top row to the bottom row sequentially during one exposure time. During fast camera motion, skew and vertical scaling distortions are introduced [14]. Figure 1 demonstrates the impact of rolling shutter. The arrows indicate the direction of camera motion. Solid lines surround the captured image in a camera. Dashed lines indicate the corresponding area in the real scene for that captured image. Motion in (a) and (c) contribute to skew distortions, corresponding to shear in geometric transformation. Motion in (b) and (d) result in vertical scaling, corresponding to the scaling difference between horizontal and vertical direction. Because of camera motion, rolling shutter artifacts are usually accompanied by motion blur.

Rotation comes from camera rotation due to head movement. Camera wearers rarely keep their head horizontal; they shake their heads randomly whether sitting or walking.

3. PROPOSED METHOD

3.1. Overall Framework

Our distortion classification method has three components: the LVI algorithm, the homography estimation and the line angle histogram. The overall framework is shown in figure 2. Both the LVI and homography estimation are based on feature matching between two images. They measure the geometric relationship between two nearly adjacent frames, which are separated by a small time interval. Affine estimation is used to approximate the homography estimation.

In the first step, the input video is classified into static frames, non-static frames and useless frames. This preliminary classification is based on an affine estimation using consecutive frames. We classify those frames captured when the camera had very little motion to be static frames. The remaining frames with large motion are classified as non-static frames. All static frames are potentially free from distortions. A few frames in the video may fail during affine estimation due to heavy motion blur or meaningless content. These frames that have few edges or corners are classified to be useless frames.

After the preliminary classification, static frames and non-static frames are evaluated by our proposed blur measurement and geometric measurement. Blur measurement is based on the LVI algorithm, which uses potential distortion-free images as reference to evaluate blur degradations in non-static frames. As such, the LVI values indicate the relative blur. Geometric measurement uses the line angle histogram and affine estimation. The line angle histogram detects whether the image is rotated or sheared. Frames without rotation and shear

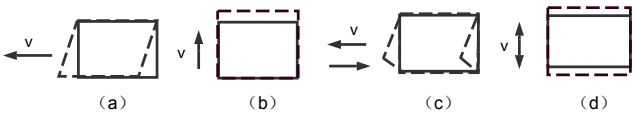


Fig. 1. (a) horizontal camera panning (b) vertical camera panning (c) horizontal camera shaking (d) vertical camera shaking

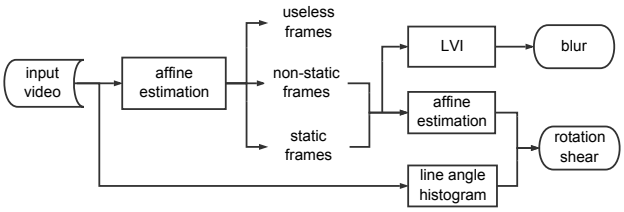


Fig. 2. Framework flowchart

are used as references for the affine estimation, which quantifies geometric transformation of rotation and shear.

3.2. Local Visual Information

Visual information fidelity (VIF) [15] is a full-reference quality metric designed to evaluate image quality. Our local visual information (LVI) is based on the essential idea of measuring the information ratio as in VIF, but is designed for local image patches. This method can measure blur between two images that differ by a geometric transformation and overcomes the limitation of content dependence. Because LVI is based on local measurement, it compares the shared area between two images but discards the non-common area. LVI indicates a relative evaluation of blur instead of an absolute value.

To measure motion blur independently, first, LVI should be invariant to subpixel shift, because image patches selected by matching feature points have subpixel resolution. Second, LVI should be invariant to rotation and shear. This allows it to accommodate geometric distortions.

To satisfy these two properties, LVI uses an information-based measurement to distinguish between sharp images and blurry images. According to natural scene statistics, images captured in high quality can be approximately expressed by Gaussian scale mixtures (GSMs) in the wavelet domain [16]. For images captured using wearable cameras, the GSMs can describe all clear images that are free from distortions. For images with distortions, the distribution of their wavelet coefficients can be approximately described by the GSMs. Similar to VIF, LVI measures the extracted visual information based on a HVS model.

Corresponding image patches between two images are selected according to matching feature points. The ORB feature [17] is used to find matching points, and RANSAC is applied to remove outliers. LVI is the extracted information ratio between two images. The amount of mutual information between input and output image signals of the HVS is

quantified to be the extracted visual information.

Let p indicate the matching patches. A pair of corresponding image patches, denoted by A_p and B_p, are selected from image A and image B, respectively. A_p and B_p are modeled by GSMs separately in the wavelet domain. A_{pi} and B_{pi} are the wavelet coefficients in the ith subband for A_p and B_p, respectively. S_{Api} and S_{Api} are corresponding different random fields for A_{pi} and B_{pi} in GSMs.

We use the same HVS model in [15]. C and D are denoted the outputs of A and B of the HVS, respectively. So LVI can be expressed by the ratio of mutual information between A and C and mutual information beween B and D:

$$LVI = \frac{\sum_p \sum_i I(C_{pi}; A_{pi}|S = S_{Api})}{\sum_p \sum_i I(D_{pi}; B_{pi}|S = S_{Bpi})} \quad (1)$$

If LVI is smaller than 1, it indicates B is more blurred than A; otherwise A is more blurred than B. By modeling the source field of two image patches separately, LVI is very close to 1 if two images only differ in a geometric transformation.

3.3. Homography Estimation

Optical flow has been used to measure geometric transformation between two images in FPVs. However, it fails when images have been subjected to heavy motion blur, which destroys gradient information. Instead, we use a homography to estimate the geometric relationship between two frames based on matching feature points.

In our algorithm, we use the ORB feature, since it is relatively robust to blur and geometric transformation. Let the homography matrix between two frames be H. H can be decomposed into

$$H = H_a H_p = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ w_a & w_b & 1 \end{bmatrix} \quad (2)$$

where H_p is a projective transform and H_a is an affine transform. When projective parameters w_a and w_b are very small, the homography matrix can be approximately by H_a. To measure geometric distortions, the parameters of shear, rotation, scale and translation are separated, by decomposing H_a as

$$H_a = H_s H_r H_k H_t \quad (3)$$

$$= \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

where H_s, H_r, H_k and H_t are the scale, shear, rotation and translation matrices, respectively. Using this decomposition, all parameters in the four matrices can be estimated based on matching feature points between the two frames. $\sqrt{t_x^2 + t_y^2}$ is a translation parameter, k is the shear parameter, $|1 - \frac{s_y}{s_x}|$ is the vertical scaling parameter, and θ is the rotation parameter. The translation and rotation parameters indicate the degree of motion between the two frames. The shear and vertical scaling parameters are used to measure rolling shutter artifacts.

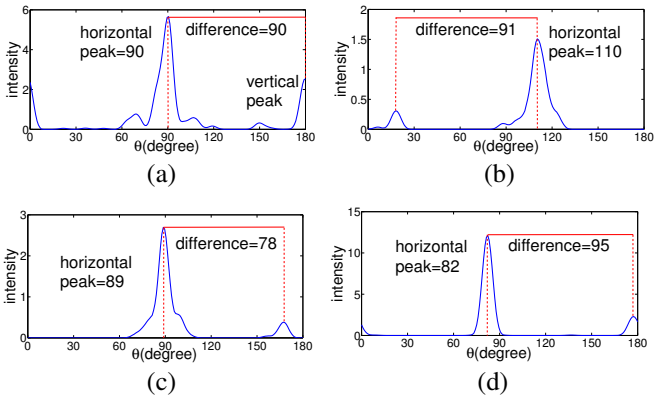


Fig. 3. Line angle distributions: (a) image free from shear and rotation (b) image with rotation only (c) image with shear only (d) image with both rotation and shear

3.4. Line Angle Histogram

Line angle histogram [13] is used to detect shear and rotation. The line angle distributions of different images are shown in figure 3. Horizontal is at 90°, and vertical is at 0° and 180°. The peaks closest to horizontal and vertical are denoted the horizontal peak and the vertical peak, respectively. The deviation of the horizontal peak from 90° indicates the rotation during capture. The difference between the horizontal and vertical peaks should be close to 90°. When the two peaks deviate from orthogonality, the image are sheared.

4. EXPERIMENTS AND RESULTS

4.1. Distortion Measurement

We choose 13 images almost free from distortions from 7 FPVs recorded ourselves by Pivothead (resolution 1080p, frame rate 30fps). All test images are created from these distortion-free images.

Synthetic distortions including motion blur, shear and rotation are tested to show the performance of the LVI algorithm, the affine estimation and the line angle histogram. We first demonstrate that LVI is responsive to synthetically-created motion blur, but is insensitive to synthetic rotation and shear. Motion blur is created by a 1-D box filter where the length of the filter controls the degree of blur. As the length of filter increases from 1 to 30, LVI performs similarly for all content. The LVI value drops from 1 to average 0.461 with little variation across different content. Note that feature matching fails when one image is much blurrier; we exclude these cases here. Our results demonstrates LVI is negatively correlated with motion blur.

In addition, we create synthetic rotation and shear to test LVI. Test pairs are created by symmetrically rotating or shearing a distortion-free image. This process introduces geometric distortions without introducing an asymmetric filtering effect. All images are cropped to the original size. As the shear

distortions	affine estimation	line angle histogram
rotation	0.0001	96.03%
motion blur+rotation	0.1689	96.03%
shear	0.0003	75.82%
motion blur+shear	0.0034	38.10%

Table I. Measurement of geometric distortions

difference k increases to 0.4, the fluctuation of LVI for all images is below 0.053. As each image is rotated by 45° in opposite direction, LVI produces results no smaller than 0.965. These results show that LVI is insensitive to shear and rotation; although more so for shear than rotation.

Next, we show that the line angle histogram and the affine estimation are effective methods to detect and quantify rotation and shear, respectively. Our experiment measures rotation from $-45°$ to $45°$ with $\Delta\theta = 3°$, and shear from -0.2 to 0.2 with $\Delta k = 0.02$. In table I, the results below affine estimation show the mean square error (MSE) between the actual shear or rotation and the measured geometric parameters using the affine estimation. Motion blur is added to each rotated or sheared image with filter length 30. The results show that our method performs well at detecting and quantifying rotation, even accompanied by motion blur. The line angle histogram is tested to determine whether the image is rotated or sheared. Consider images with rotation larger than 2° and with shear k greater than 0.04 to be rotated images and sheared images, respectively. Table I shows the accuracy of this measurement. The percentage below line angle histogram shows the proportion of correct detection for each distortion. The measurement of shear has relatively lower accuracy, especially for blurry images.

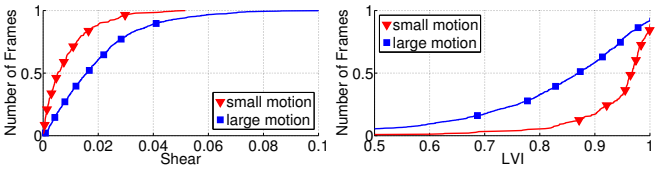
4.2. Video Statistics

We now apply our classification method to compare the differences between traditional videos and FPVs, we present statistics of distortions for the two types of videos in Table II. We selected six traditional videos from LIVE Video Quality Database [18, 19], and recorded six types of FPVs using the Pivothead. In Table II, the "talking","ping pong" and "eating" videos are recorded indoors, while other three FPVs are recorded outside. The comparison indicates a few frames are subject to distortions in the LIVE database, while most frames in FPVs are distorted images. Our results demonstrate FPVs have dramatically different distortions immediately after capture compared to traditional videos.

The six FPVs share common properties. First, all of them have more than 69% of frames with rotation, indicating that camera wearers keep their heads rotated most of the time. Second, the percentage of blurry images is in the range from 55% to 83%. Third, shear is less likely to exist in FPVs compared to rotation and blur. However, each FPV also shows some differences. The three indoor videos have more than

content	some blur	heavy blur	rotation	shear
LIVE	3.94%	0.33%	17.25%	4.36%
running	53.61%	7.08%	69.76%	13.67%
walking	50.01%	4.98%	76.19%	16.26%
basketball	50.00%	14.27%	69.55%	22.37%
talking	58.52%	22.37%	87.66%	6.07%
ping pong	52.86%	30.76%	75.61%	38.32%
eating	62.76%	13.19%	91.73%	51.84%

Table II. Comparison between FPVs and traditional videos



(a) (b)

Fig. 4. Cumulative distributions: (a) shear (b) LVI

75% of their frames with blur, while the percentages of the other three outdoor videos are no more than 65%. So indoor videos have worse quality compared to outdoor videos.

Figure 4 shows two cumulative distributions of frames in the "running" video. We extract two groups of frames: small motion and large motion. To partition them, we use the translation parameter from the affine estimation, and declare those with translation greater than 50 to be large motion, and those with translation smaller than 10 to be small motion. In frames with small motion, 89% have shear change smaller than 0.02 and 68% have LVI greater than 0.95; but for frames with large motion, only 60% have shear smaller than 0.02 and 23% have LVI larger than 0.95.

We also applied our LVI algorithm on the image quality database TID2013 [20], for two distortions; Gaussian blur and contrast change. The Pearson correlation coefficients for Gaussian blur and contrast change are 0.9320 and 0.9018, respectively. The correlations of Gaussian blur for other image quality metrics, SSIM, FSIM [21] and VIF, are 0.9191, 0.8905 and 0.9530, and the correlations of contrast change are 0.6385, 0.6924 and 0.8730, respectively. This demonstrates that LVI is useful to measure more distortions than motion blur; and the performance of LVI can compete with other image quality metrics.

5. CONCLUSION

We present different distortions in images of FPVs including motion blur, rolling shutter artifacts and rotation. Then we propose a measurement method for classification and quantification of these types of distortions. Our proposed algorithm provides information about how to design an image or video quality metric for FPVs. For our future work, first, how to describe and quantify an image with spatially varying quality is an open question. Second, other than Pivothead cameras, some wearable cameras (i.e. GoPro) have a fisheye effect, which has not been considered in our method.

6. REFERENCES

[1] Michael S Ryoo and Larry Matthies, "First-person activity recognition: What are they doing to me?," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2730–2737.

[2] Joydeep Ghosh, Yong Jae Lee, and Kristen Grauman, "Discovering important people and objects for egocentric video summarization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1346–1353.

[3] Alireza Fathi, Jessica K Hodgins, and James M Rehg, "Social interactions: A first-person perspective," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1226–1233.

[4] Bo Xiong and Kristen Grauman, "Detecting snap points in egocentric video with a web photo prior," in *European Conference on Computer Vision*, 2014, pp. 282–298.

[5] Zhou Wang, Ligang Lu, and Alan C Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[6] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi, "A no-reference perceptual blur metric," in *IEEE International Conference on Image Processing*, 2002, vol. 3, pp. III–57.

[8] Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *Electronic Imaging*, 2007.

[9] Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel, "A transformation-aware perceptual image metric," in *Electronic Imaging*, 2015.

[10] Omer Barkol, Hadas Kogan, Doron Shaked, and Mani Fischer, "A robust similarity measure for automatic inspection," in *IEEE International Conference on Image Processing*, 2010, pp. 2489–2492.

[11] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Elan Dubrofsky, *Homography estimation*, Ph.D. thesis, University of British Columbia, 2009.

[13] Jana Košecká and Wei Zhang, "Video compass," in *European Conference on Computer Vision*, pp. 476–490. 2002.

[14] Wei Hong, Dennis Wei, and Aziz Umit Batur, "Video stabilization and rolling shutter distortion reduction," in *IEEE International Conference on Image Processing*, 2010, pp. 3501–3504.

[15] Hamid Rahim Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[16] Martin J Wainwright, Eero P Simoncelli, and Alan S Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, no. 1, pp. 89–123, 2001.

[17] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: an efficient alternative to sift or surf," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[18] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[19] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan C Bovik, and Lawrence K Cormack, "A subjective study to evaluate video quality assessment algorithms," in *Electronic Imaging*, 2010.

[20] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[21] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.