Journal of Visual Communication and Image Representation 54 (2018) 123–132

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Image quality assessment in first-person videos[☆],[☆☆]

Chen Bai[*], Amy R. Reibman

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

ARTICLE INFO

Keywords:
First-person videos
Local Visual Information (LVI)
Mutual reference
Image quality assessment
Pseudo-reference

ABSTRACT

First-person videos (FPVs) or egocentric videos provide a huge amount of data for visual lifelogs. The quality assessment of frames in FPVs serves as an important tool, feature or evaluation baseline for not only structuring but also analyzing lifelogs. To develop a frame-quality measure for FPVs, we introduce a new strategy for image quality estimation, called mutual reference (MR), which uses one or more pseudo-reference images to evaluate a test image. We then propose a MR quality estimator, called Local Visual Information (LVI), that primarily measures the relative blur between two images. To apply the MR strategy to FPVs, we propose a mutual reference frame quality assessment for FPVs (MRFQAFPV) framework which incorporates LVI. Our results, using both real and synthetic distortions and objective and subjective tests, demonstrate both methods perform better than existing NR QEs at measuring the quality of frames in FPVs.

1. Introduction

Wearable cameras (Pivothead, Looxcie Camera, Mobius, Gopro, Google Glass) mounted on human bodies can record videos at any time and place without length limitation. These so-called first-person videos (FPVs) or egocentric videos can record continuous data about personal daily life. People are increasingly using FPVs to document activities, share experiences, record trips, and more [4]. The huge amount of information from long-time and unstructured FPVs is a rich source for visual lifelogs [5]. Recent research on assessing lifelogs in FPVs involves two aspects: structuring and analysis. Methods to structure visual lifelogs consists of informative-image detection [6], temporal segmentation [7,8], egocentric summarization [9,10] and content-based search and retrival [11,12]. Analysis of lifelogs involves object discovery [13], activity recognition [14] and spatial localization [15].

The visual quality of individual frames influences the ability to both structure and analyze FPVs. First, image quality is one important indicator when searching for informative images, which are defined in [6] as “intentional” images and can be used to summarize FPVs. Second, image quality provides an evaluation tool for applications related to viewing experience, including fast-forward and stabilization [16,17]. Third, it can be used to filter out useless frames before applying methods for content search [12] and activity recognition [14]. In addition, it can provide information about the wearer’s motion as well as environmental cues regarding fog, over-exposure or under-exposure.

FPVs have significantly different attributes than typical broadcast and mobile videos. Broadcast videos are often captured by stably-mounted cameras with high-quality frames, and mobile videos are captured from hand-held mobile devices. In both cases, a filmmaker captures scenes guided by real-time feedback from a screen, so the camera can be intentionally controlled to be reasonably stable and have the desired field of view. However, wearable cameras rarely are stably mounted nor have real-time feedback. Video is often gathered passively, without attending to composition. Even if there is an intention to record a high-quality video, the camera may not capture a well-composed high-quality video. This occurs not only because the wearer may be unaware of the field of view, but also because external factors may temporarily influence body actions as well. As a result, FPVs as recorded from camera rarely tell an effective story that is attractive from an aesthetic perspective, which are two attributes of professional videos [18]. An experienced filmmaker can learn to capture professional-quality video using a mobile camera. However, the passive nature of FPVs, as well as their lack of organization and shot boundaries, limits their ability to tell an effective story. Even with a high spatial resolution and high quality, FPVs would rarely be considered professional.

Camera motions due to head or body movement of the camera wearer can significantly degrade the quality of individual frames in an FPV [1,2]. The motion-induced distortions of images in FPVs can be mainly classified as blur and the geometric distortions of rolling shutter artifacts and rotation. Blur could be caused by any camera movement, and arises when motion is sufficiently large during the exposure period [19]. Rolling shutter artifacts mainly arise from camera panning and

[☆] This paper has been recommended for acceptance by Cathal Gurrin.

[☆☆] Portions of this paper have appeared in [1–3].

[*] Corresponding author.
E-mail address: baichen@purdue.edu (C. Bai).

https://doi.org/10.1016/j.jvcir.2018.05.005
Received 15 March 2017; Received in revised form 27 October 2017; Accepted 6 May 2018

tilt, and produce skew or wobble in an image. Skew appears when the camera moves at a constant speed; wobble occurs when the frequency of motion is greater than the frame rate of the recording video [20]. Finally, image rotation is a combination of translational camera motion and roll. For example, when camera is mounted on the hat of the wearer and the head tilts to left or right, the camera rotates around an axis with some distance to the camera center.

To evaluate the quality of individual frames, it is typical to apply Image Quality Estimators (IQEs). Existing IQEs are normally classified into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. FR and RR methods [21–24] need a high-quality corresponding reference image that is the source of the distorted image to be evaluated. These types of IQEs are not applicable for assessing frames in a FPV, because no reference image exists. Moreover, since the image might already be degraded, the results of FR and RR methods will not meaningfully reflect any additionally introduced degradations.

In contrast, NR methods estimate the quality of a single image without relying on any reference [25]. However, most existing NR methods are content dependent [26–29]. As a result, it is often difficult to interpret the output of a NR method [30]. For example, setting a quality threshold in a system is challenging; all five NR QEs considered in [30] are unable to consistently partition high-quality images from heavily degraded images. In addition, these IQEs are rarely evaluated on the types of degradations present in individual frames of an FPV [2].

In this paper, we propose a new strategy of quality estimation, called mutual reference (MR), which does not fit into the previous categorization of FR, RR or NR methods. A MR QE estimates the quality of a test image based on one or more pseudo-reference image. Unlike FR and RR QEs, perfect pixel alignment is not necessary; instead the pseudo-reference image and the test image are constrained only to have sufficient overlapping content. For example, the pseudo-reference could be a high-quality image captured by a stably-mounted camera from one viewpoint, and test images can capture the same scene from different points of view using a moving camera. Another example is a group of temporally-adjacent video frames, where one or more frames can be a pseudo-reference for the remaining frames.

The MR strategy is a natural choice to assess the quality of frames in a FPV. First, MR provides a *relative* quality estimation that allows degradations to be present in any images. A relative score can be used to select the image with the best quality from a set of images. Second, MR uses information from the overlapping regions between two or more images. This minimizes content dependency in quality scores, so that scores are more easily interpretable in a system.

We apply the mutual reference approach to design a MR QE, called Local Visual Information (LVI) [1], to measure the relative blur. The principle of LVI is to locally measure the effective visual information in the human visual system (HVS), and to evaluate the quality difference based on the information ratio. Based on LVI, we design a framework of mutual reference frame quality assessment for FPVs (MRFQAFPV), which measures the LVI score of each frame in a FPV [3].

Section 2 describes prior works in FR QEs and NR QEs. Section 3 presents a detailed description of the strategy for MR. Our proposed MR QE, LVI, is described with its basic principle and reliability check in 4. Our MRFQAFPV is described in Section 5. The framework has three steps: temporal partitioning, reference search and quality estimation. In Section 6, we demonstrate our framework is effective at assessing quality of individual frames in FPVs, and outperforms existing NR QEs in this context. Our results include demonstrating temporal partitioning methods, as well as two subjective tests that include synthetic distortions and real frames captured from FPVs. Section 7 summarizes this paper and discusses future work.

2. Prior work on QEs

2.1. Full-reference QEs

FR QEs use a pixel-aligned reference image to estimate the quality of distorted versions of the same image. They can be categorized by whether they apply models of the human visual system, image structure, or image statistics [31]. Two common QEs are the Structural Similarity Index (SSIM) [21], which is based on structure, and Visual Information Fidelity (VIF) [32], which is based on statistics.

SSIM computes means and variances of each image, applies a similarity measure to each,

$$S(x,y) = \frac{2f_x f_y}{f_x^2 + f_y^2}, \tag{1}$$

and combines these with a correlation term to quantify distortions in the luminance and contrast. In Eq. (1), x is the reference image and y is the test image, and f_x and f_y are extracted features from x and y, respectively. The same quality score will be unchanged if we swap the order and instead consider the distorted image to be the reference x. This type of symmetry does not allow SSIM to be used to determine which image has better quality. In addition to SSIM, Feature Similarity (FSIM) [22], Gradient Magnitude Similarity (GSM) [33] and Spectral Residual based Similarity (SR-SIM) [34] employ the same similarity measure in Eq. (1) using other features. Therefore, these QEs also are incapable of determining whether a test image is better than its reference image. While, some other QEs, for example, VSNR [35] and MAD [36], use a non-symmetric structure to compute quality scores, reversing the order of the reference image and the test image still does not lead to a meaningful comparison.

VIF [32] is an information-based QE. It assumes that the two images are from the exact same source field, which it models using the statistics of the reference image. Since VIF does not depend on the similarity of features or error images, it is able to distinguish which image is better among the two images despite having no prior information. Another QE that can compare the quality of two images is Visual Distortion Gauge (VDG) [37]. However, neither VIF nor VDG have been designed to measure two images with geometric changes.

2.2. No-reference QEs

No-reference (NR) QEs use only the information of the input image to be evaluated. One specific subset of NR QEs are NR blur metrics, which were summarized in [38,25]. One uses the histogram of DCT coefficients [39]. Edge-based blur QEs have also been proposed and comprise the majority of blur QEs: [40,41], JNBM [38], CPBD [42]. Non-edge blur metrics using the discrimination between re-blurred versions of an image [43,44] and local phase coherence [45] were also proposed. However, blur estimation developed from these strategies depends heavily on the image content. If we have two images that share only a portion of their content, then because blur metrics may show very different behaviors in their non-common areas, the overall blur scores of the two images cannot accurately reflect their visual difference. NR QEs may also be based on statistics. Specifically, BRISQUE [27], NIQE [28], and IL-NIQE [29] all use natural scene statistics (NSS) to compute quality. These QEs are still content-dependent, and do not often have bounded range of their quality scores. Moreover, they are less effective when applied to images that differ in spatial resolution from the images that were used to train them [30].

In [30], the question is considered of whether a QE can distinguish between badly degraded images and relatively undistorted images. Their results indicate that it is challenging for NR QEs. In particular, there exists a large overlap between the historgrams of the quality scores for undistorted and badly degraded images using BRISQUE, NIQE and IL-NIQE. In addition, our results in Section 6 demonstrate

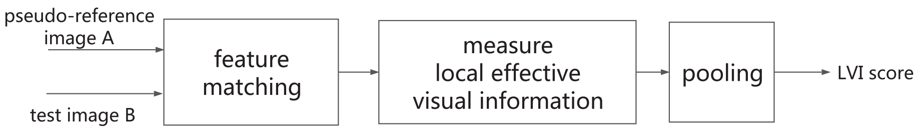


Fig. 1. Block diagram of Local Visual Information (LVI) quality estimator.

that the state-of-the-art NR QEs are source-dependent, and our proposed method in Section 3 significantly reduces the source dependency when estimating the quality of First-Person images.

3. Mutual reference

Mutual reference (MR) is a strategy of image quality estimation whose basic idea is to use a collection of “similar enough” images that can provide each other with effective information for quality assessment. To define ”similar enough”, we introduce the concept of a near-set, which is a group of images that share common content. One example is a group of images captured from nearby locations. In addition, images in the near-set do not need to have the same spatial resolution. For example, [46] considers quality estimation for downsampled images, while [47] considers quality of image super-resolution techniques.

Within the MR strategy, there are two approaches: pairwise and group-based measures. The pairwise approach uses a single pseudo-reference image to estimate the quality of a test image. The pseudo-reference does not need to be pixel aligned with the test image, but can be classified into the same near-set as the test image. Typically, the pseudo-reference image needs to be the best image in an identified near-set. One way of creating a MR QE using the pairwise measure is that the QE is able to distinguish which of two images is better. Such a MR QE can identify the pseudo-reference by pairwise comparison in a near-set. One example is the MR QE, Local Visual Information (LVI), described in Section 4 and first presented in [1].

The group measure approach for MR QE estimates the quality of an image using more than one pseudo-reference. One example is the quality assessment of image fusion, for which the goal is to integrate complementary information from a group of images into a new image, in order to obtain more complete and useful information for image-processing tasks [48]. To evaluate the quality of a fused image, all source images are used as references [49,50]. The near-set consists of all source images and the fused image.

MR provides a relative quality estimation, which allows quality degradations to be present in all images in the near-set. The best image in a near-set does not necessarily need to be a high-quality image. Also, a new image can easily be added into an existing near-set. If the added image has better quality than all other images in the near-set, the new image can be set to be the pseudo-reference.

MR methods do not fit into the typical categorization of FR, RR or NR methods. Specifically, MR uses the effective information from the overlapping regions between different images. The overlapping area could differ in a geometric transformation or distortions. As a comparison, FR and RR uses a high-quality reference image that is also the source of the distorted image to provide information for quality assessment. NR uses implicit knowledge of distorted image versus high-quality image.

MR quality assessment has two major application areas. The first application is quality assessment for image fusion, as discussed above [48–50], and including the quality metric for exposure fusion techniques [51]. The second application is to assess images captured either from, or of, nearby locations. For example, in this paper, we consider quality assessment of individual frames in a video using temporally nearby frames. Another example would be to assess the quality of frames in two videos taken in nearby locations on, say, two different days. The third example is to assess images considered in [52], which implemented a subjective test using images captured of the same scene by either different cameras or the same camera with different settings.

4. Local Visual Information

In this section, we describe our proposed MR QE, Local Visual Information (LVI) [1], which primarily measures relative blur between two images.

4.1. Basic principle

LVI is derived from the approach of VIF [32]. VIF quantifies the visual quality of an image using the mutual information between the test image and its reference. VIF uses natural scene statistics (NSS) [53] to model the reference image, and uses the model obtained from the reference plus a distortion channel to model the test image. First, it decomposes the two images into blocks and sub-bands. Second, it computes the mutual information between the reference and the test image in each block and subband using a NSS model. Third, the VIF score is pooled from all blocks and subbands.

LVI has two major changes. First, instead of computing a global measure of information in an image, LVI measures patch-based local information. Second, LVI models the source field of the two input images separately, which enables LVI to compare the quality of any two images in a near-set. One assumption behind LVI is that the image has consistent spatial quality.

The quality measure LVI has three procedures, shown in Fig. 1. The input of LVI is a pseudo-reference image A and a test image B, where A and B are in the same near-set. In the first step, LVI computes the pixel relationship between A and B using feature matching. All matching points are filtered by a ratio test, a symmetric test and a RANSAC test to remove outliers. A matching patch is defined to be the square block centered around a matching point in the image. The output of the first procedure is the locations of all corresponding patches.

The second step measures the effective local visual information between A and B for all corresponding patches. High-quality images can be described by Gaussian scale mixtures (GSMs) in the wavelet domain based on natural statistics. LVI approximately models either sharp or blurry images by GSMs, whose shapes are determined by the statistics of the image content. The effective visual information is quantified by the amount of mutual information between the input and output images in human visual system (HVS).

Let the index for each matching image patch be l. A_l and B_l are two matching image patches from A and B, respectively. GSMs describe an image according to its content, so A_l and B_l have different shapes of GSMs in the wavelet domain. We describe the GSMs of A_l and B_l in the pth subband as

$$A_{lp} = S_{lp}^{A}\cdot U_{lp}^{A} \tag{2}$$

$$B_{lp} = S_{lp}^{B}\cdot U_{lp}^{B} \tag{3}$$

where S_{lp} is a scalar random variable in the pth subband modeling the source field, and U_{lp} is a zero mean Gaussian random vector. A_{lp} and B_{lp} are the wavelet coefficients of the patch in the pth subband for image patch A_l and B_l, respectively.

The HVS model in [32] uses a Gaussian channel to model the uncertainty that image information flows through it. The model can be expressed as

$$C_{lp} = A_{lp} + \mathscr{X} \tag{4}$$

$$D_{lp} = B_{lp} + \mathscr{X}' \tag{5}$$

where C_{lp} and D_{lp} are the outputs of A_{lp} and B_{lp} after flowing through the HVS model, respectively. $\mathscr{X}$ and $\mathscr{X}'$ are Gaussian noise drawn from $\mathscr{N}(0,\sigma_x^2)$ to model the noise from HVS.

The amount of mutual information between input image signals and output image signals of the HVS can be calculated as

$$I(C_{lp};A_{lp}|S_{lp}^A) = \frac{1}{2}\sum_m \log_2\left(1+\frac{(s_{lp}^A)^2\lambda_m^A}{\sigma_x^2}\right) \tag{6}$$

$$I(D_{lp};B_{lp}|S_{lp}^B) = \frac{1}{2}\sum_n \log_2\left(1+\frac{(s_{lp}^B)^2\lambda_n^B}{\sigma_x^2}\right) \tag{7}$$

where λ are the eigenvalues of U_{lp}, and m and n is the indices of eigenvalues. s_{lp}^A and s_{lp}^B are the realizations of S_{lp}^A and S_{lp}^B, respectively.

The third step is to pool the LVI score using the local visual information in all corresponding patches. By computing the sum of the information from all corresponding local regions of A and B, LVI takes the ratio of the total amount of information from the two images as the output.

$$Q_{LVI} = \frac{\sum_l\sum_p I(D_{lp};\, B_{lp}|S = S_{lp}^B)}{\sum_l\sum_p I(C_{lp};\, A_{lp}|S = S_{lp}^A)} \tag{8}$$

The output score of Eq. (8) represents the quality of B relative to the pseudo-reference A. If B has worse quality than A, LVI varies from 0 to 1, which indicates that B has less visual information pooled than A. Otherwise, the LVI score is larger than 1, which indicates our selected pseudo-reference A is worse than B. The value of LVI score between two images represents their relative quality, and provides a quality comparison.

Fig. 2 shows an example of the LVI measure between a pseudo-reference image and a test image, extracted from a captured FPV. The connected lines are the center of matching patches. Two corresponding patches are enlarged to display the difference.

4.2. Reliability check

LVI fails to provide an effective quality measure at all cases. To ensure we only apply LVI in those situations when its score is meaningful, we design a reliability check to verify that neither of the two known issues are present to reduce the accuracy of the computed LVI score.

The first known limitation is that LVI cannot measure quality when there are insufficient feature matching points between the pseudo-reference and the test image. For example, when the test image is heavily blurred, there are very few feature matching points between the two images.

The second known limitation is that LVI is sensitive to scaling, although it is insensitive to other affine transformations [1]. This allows LVI to measure quality degradations almost independently of geometric distortions when the image is sheared or rotated relative to the pseudo-reference. However, when the two images have similar quality but have objects in very different sizes or scales, their LVI scores often have a large difference. Our reliability check is designed to identify these unreliable scores.

Within a near-set, we expect the geometric relationship between two images to be approximately modeled by a homography. This homography can be estimated [54,55] using matching feature points. Specifically, we apply point-based homography [54] using the result of the feature matching step in Fig. 1. Then by decomposing the homography matrix M_H, as described below, we can independently extract scale changes both horizontally and vertically.

First, M_H is decomposed into the product of an affine transform M_A and a projective transform M_P, given by

$$M_H = M_A M_P = \begin{bmatrix} u_a & u_b & u_c \\ v_a & v_b & v_c \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ w_a & w_b & 1 \end{bmatrix}, \tag{9}$$

where and w_b are projective parameters in M_P. The affine matrix M_A has six degrees of freedom corresponding to parameters, u_a, u_b, u_c, v_a, v_b, v_c. When and w_b are very small, M_H is approximated well by M_A.

Further, M_A is a combination of five independent transformations, translation, shear, rotation, scaling and aspect ratio. In FPVs, shear and rotation artifacts often occur in frames from a near-set. Focusing only on horizontal shear and rotation, M_A can be decomposed as

$$M_A = M_s M_r M_k M_t = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & k_s & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{10}$$

where M_s,M_k,M_r and M_t are scale, shear, rotation and translation matrices, respectively. s_x and s_y are scaling factors in horizontal and vertical directions, respectively, and s_x/s_y is the aspect ratio. k_s is the shear value, θ is the rotation angle, and t_x and t_y are translation distances in horizontal and vertical directions, respectively. Using the parameters estimated from M_A, we can calculate M_s as

$$M_s = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{u_a v_b - u_b v_a}{\sqrt{v_a^2 + v_b^2}} & 0 & 0 \\ 0 & \sqrt{v_a^2 + v_b^2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

When either one of s_x and s_y exceeds the range bounded between $[a,\frac{1}{a}]$, where a is the threshold experimentally set to be 0.95, the LVI score is

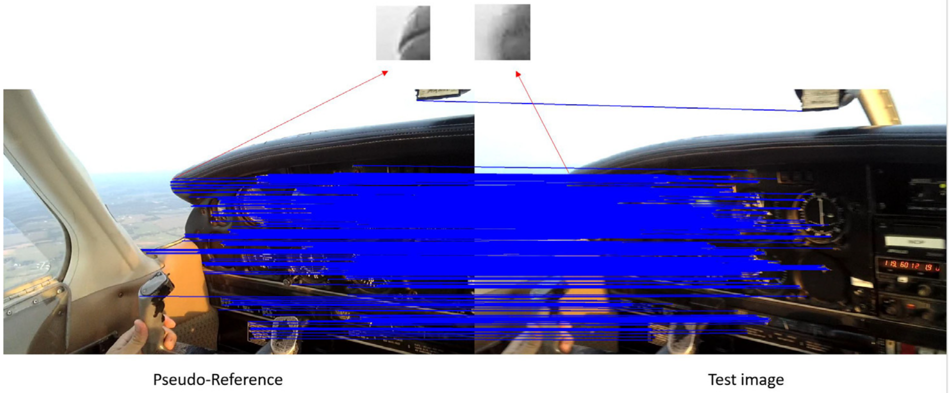


Fig. 2. Left: Pseudo-reference. Right: Test image. LVI score = 0.771.

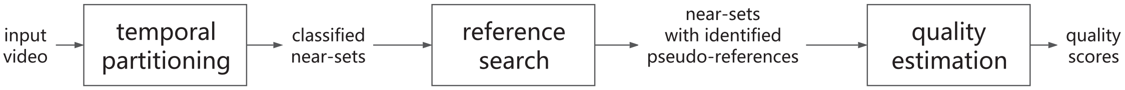


Fig. 3. Framework of quality assessment for First Person Video.

Fig. 4. Sample test images: (0) basketball (1) run (2) walk (3) billiards (4) cat (5) eat (6) ping pong (7) talk (8) car (9) flight.

considered to be unreliable.

This reliability check ensures that an effective LVI score is calculated between two images that are neither too blurry nor have significant scale differences. In the next Section 5, we will describe how LVI can be incorporated into a quality assessment framework for FPVs using the strategy of mutual reference.

5. Framework of MR quality assessment of FPVs

Our framework of mutual reference frame quality assessment of FPVs (MRFQAFPV) can be separated into three steps: temporal partitioning, reference search and quality estimation. Fig. 3 shows the block diagram of MRFQAFPV. In the first step, frames from the input FPV are temporally partitioned into different near-sets. In the second step, the system searches for one pseudo-reference image in each near-set using the pairwise approach of MR. In the third step, the LVI quality score of each frame is calculated based on the identified pseudo-reference.

The *temporal partitioning* shown in the first block of Fig. 3 is designed to temporally partition frames within different time intervals into near-sets, in which all images have similar scale. Let *k* be a near-set index. An initial partitioned near-set *k* is represented as (B_1^k, B_2^k), where B_1^k is the start frame and B_2^k is the end frame. The basic procedure is: (1) Set $k = 1, B_1^k = 1$. (2) *Boundary Search* for B_2^k starting from B_1^k. (3) Set $k = k + 1, B_1^k = B_2^k + 1$, and then go to (2).

Method 1 NFP

```
1: get the start frame number $B_1^k$
2: Let $n = 1$, $\delta = 20$, $T = 50$
3: do feature matching between $B_1^k$ and $B_1^k + 10$
4: if the number of matching points $<T$ then
5:     set $B_2^k = B_1^k$, break
6: else
7:     do feature matching between $B_1^k$ and $B_1^k + n\cdot\delta$, store the
       number of matching points after RANSAC as $N$
8:     if $N < T$ and $n = 1$ then
9:         do binary search from $B_1^k + 10$ to $B_1^k + 20$ using the same
           decision rule $N < T$, break when the search interval $\leqslant 1$, and set
           $B_2^k$ to be start frame of the search interval
10:    else if $N < T$ and and $n > 1$ then
11:        do binary search from $B_1^k + 10$ to $B_1^k + 20$ sing the same
           decision rule $N < T$
12: else
13:        set $T = max(\frac{N}{2}, T)$ and $n = n + 1$, goto 3
14:    end if
15: end if
```

Method 2 FMA

```
1: get the start frame number $B_1^k$
2: do feature matching between $B_1^k$ and $B_1^k + 10$, and store the
   locations of all matching points by a bounding box $S_{10}$
3: Let $n = 1$, $\delta = 20$
4: do feature matching between $B_1^k$ and $B_1^k + n\cdot\delta$, get the bounding
   box $S_{n\cdot\delta}$
5: if $|S_{10} \cap S_{n\cdot\delta}| < \frac{1}{4}|S_{10}|$ then
6:     do binary search between $B_1^k + (n-1)\cdot\delta$ and $B_1^k + n\cdot\delta$ using the
       same decision rule, break when the search interval $\leqslant 1$ and set $B_2^k$
       to be start frame of the search interval
7: else
8:     set $n = n + 1$, goto 4
9:     if $B_2^k - B_1^k < 10$ then
10:        set $B_2^k = B_1^k$
11:    end if
12: end if
```

For *Boundary Search* in the basic procedure, we introduce two different methods, as shown in Method 1 and Method 2. Method 1 is based on the number of feature matching *points* between frames, denoted by NFP. Method 2 is based on the feature matching *area* between frames, denoted by FMA. Both methods rely on feature matching, during which we incorporate the scale check detailed in Section 4.2 to guarantee that we have reliable LVI measures in the following steps. Note that the parameter δ is empirically set to be 20, since we often have near-sets from 20 to 40 frames. If we increase or decrease δ, the near-set length is similar. The threshold for the number of matching points T is set to be 50. If we increase T, it will introduce more uncategorized frames. If we

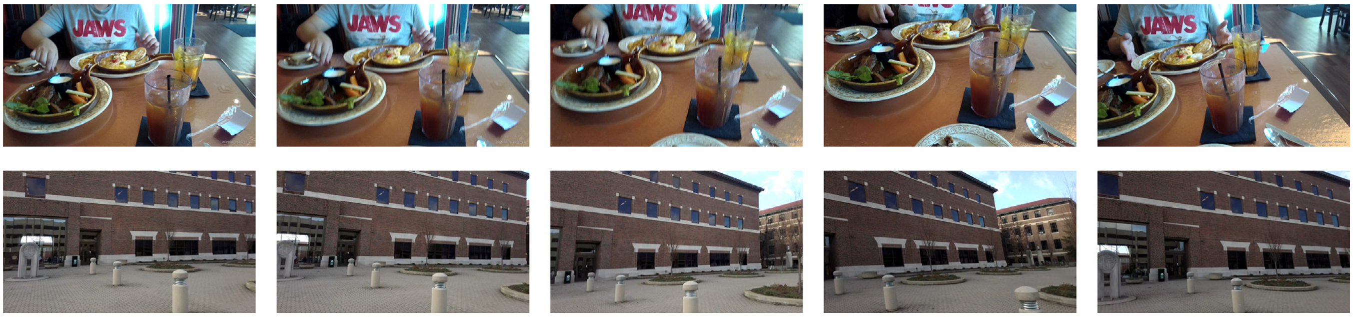


Fig. 5. Sample partitioned near-sets.

decrease T, the percentage of unreliable matching points increases significantly. We empirically set the minimum length of a partitioned near-set to be 10 frames. If the partitioning does not satisfy the length constraint, the current B_1^k is considered to be an uncategorized frame, and we repeat the basic procedure with $B_1^k = B_1^k + 1$.

The *reference search* in the second block of Fig. 3 finds the pseudo-reference image in each near-set iteratively. Let R_k be the pseudo-reference in the k_{th} near-set. Initially, let $R^k = B_1^k$, and use it as the initial pseudo-reference in the k_{th} near-set. Then, we calculate the LVI scores from $B_1^k + 1$ to B_2^k using the current R^k. Those frames with better quality than the current R_k have LVI scores larger than 1. We reset the frame with the largest LVI score in the k_{th} near-set to be our new R^k. A typical output of the k_{th} near-set is (B_1^k,B_2^k,R_k).

The *quality estimation* in the third block of Fig. 3 calculates the frame quality score. The input is the representation of the k_{th} near-set, (B_1^k,B_2^k,R^k). Let $k^{(n)}$ be the n_{th} frame in the k_{th} near-set. The quality estimation uses R^k as the pseudo-reference to measure the quality of all remaining frames in the k_{th} near-set, and stores the LVI score as $Q_{LVI}^{k^{(n)}}$, the quality measure for frame $k^{(n)}$.

6. Experiments and results

In this section, we present experimental results of applying our LVI and MRFQAFPV to First-Person Videos captured from a Pivothead camera at 1080p30. Our experiments explore two aspects: design considerations, and evaluating the performance for quality assessment. For the first, we explore six design choices for the temporal partitioning step in MRFQAFPV shown in Fig. 3, and two feature detectors for the first step of LVI shown in Fig. 1. For the second, we explore performance of our methods using both synthetically injected distortions as well as images taken from actual FPV containing real, so-called authentic, distortions. In addition, we explore performance of quality assessment not only using objective comparisons, but also using two subjective tests. The first demonstrates that MRFQAFPV provides an effective quality assessment for individual frames in FPVs, while the second shows that not only does LVI outperform existing NR QEs, but both LVI and other existing QEs that are insensitive to geometric distortions can be generalized to better estimate overall frame quality in FPVs. Finally, by applying LVI to images from the typical image quality databases [36,56,57], we demonstrate that LVI is also effective to assess the quality for some distortions that are not typically present in FPVs.

6.1. Implementation design comparisons

In this section, we explore the performance of several design options for both LIVE and MRFQAFPV. Specifically, we compare and select the FMA method with affine estimation as the scale check to be our temporal partitioning method in MRFQAFPV. Also, we show SIFT and ORB have similar performance in LVI and MRFQAFPV, so ORB is a better design choice because it is less time-consuming.

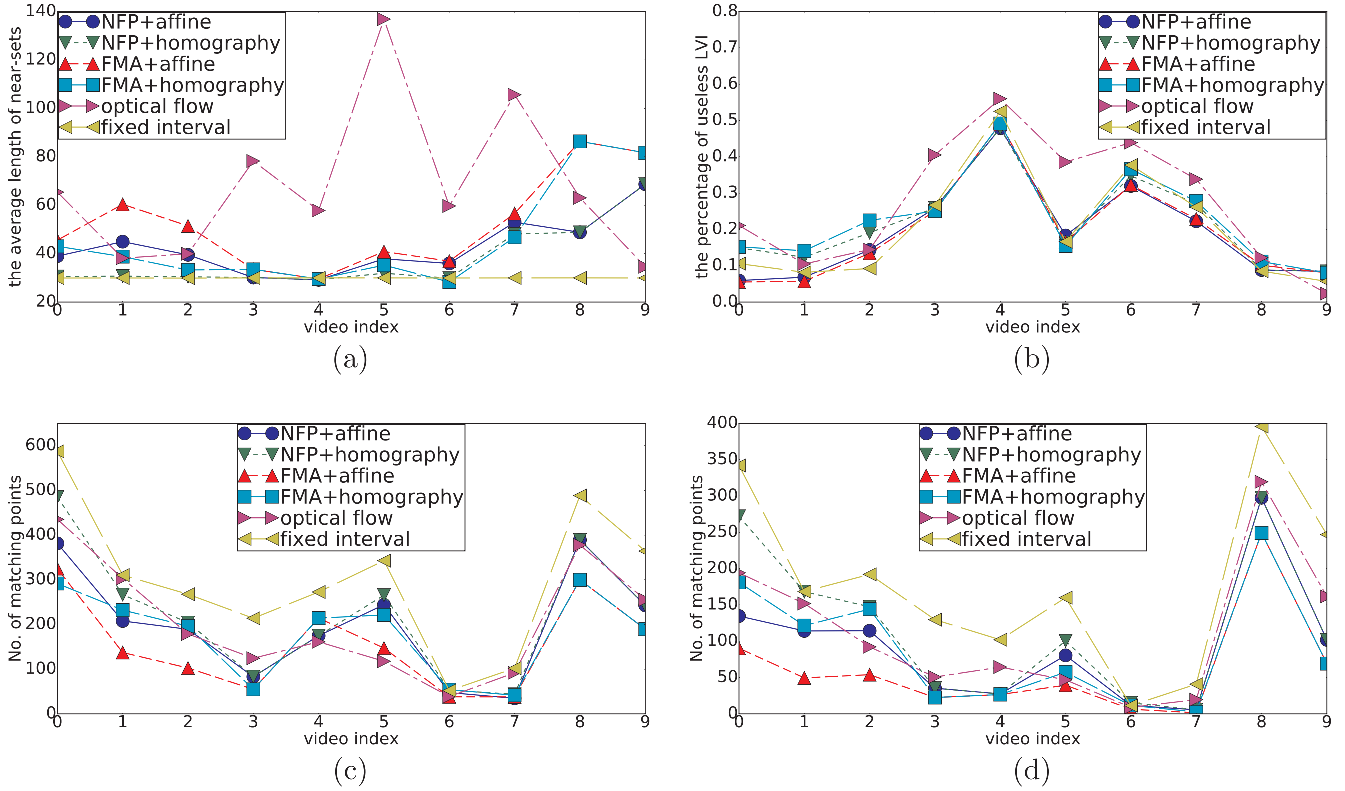
Temporal partitioning: We compare six approaches to form near-sets for the temporal partitioning step in Fig. 3. Section 5 presents two methods, NFP and FMA. In addition, the scale check detailed in 4.2 incorporated in NFP or FMA can be implemented using either affine or homography estimation. Thus, our experiments compare four proposed methods: NFP + affine, NFP + homography, FMA + affine and FMA + homography. In addition to these four methods, two baseline methods are introduced. One baseline method uses a fixed time interval (30 frames) to separate frames into each near-set. Another baseline method partitions using displacements computed by optical flow as in [8], such that each partitioned interval has a cumulative displacement of 10% of a frame width. Note that the shot boundary detection method [58] is not effective to segment FPVs, because it typically classifies the entire video into only one shot.

A good partitioning for a near-set has three criteria:

1. The length of the near-set is long enough so that most frames captured in the same scene are included.
2. Frames with a useless LVI are rare in the entire FPV. Three types of frames are considered to have useless LVI: uncategorized frames, frames that failed the reliability check, and frames with LVI score greater than 1.
3. The shared content between two frames in different temporally adjacent near-sets is small. We estimate the degree of overlap between any two frames by counting the number of matching points.

Fig. 6 presents the performance of the six methods using these three criteria. The first and second criteria are demonstrated by the average length of the near-set and the percentage of useless LVI, as shown in Fig. 6a) and and (b), respectively. The third criterion is demonstrated with two values, the average number of matching points between pseudo-references and between start frames in temporally adjacent near-sets, as shown in Fig. 6(c) and (d). The video indexes represent videos with different content. Outdoor videos are indexed from 0 to 2, indoor videos are indexed from 3 to 7, and 8, 9 are in-vehicle videos. Sample frames for each video are shown in Fig. 4, and frames in two partitioned near-sets are shown in Fig. 5. The test dataset is available at [59].

The first baseline method, fixed interval, has the shortest near-set length and third least percentage of useless LVI. The second baseline method, optical flow, has the longest average near-set length, but the highest percentage of useless LVI. Actually, compared to all methods, FMA + affine method shows the best performance among the six methods; it has the second longest near-set length, the least percentage of useless LVI, and the least or the second least number of matching points either for pseudo-references or for start frames in all videos. The effectiveness of the other three methods can be successively ordered as follows: NFP + affine, FMA + homography, NFP + homography. According to the results, FMA creates a better partitioning than NFP. Affine estimation outperforms homography estimation using the same partitioning method according to the percentage of useless LVI, so the former is more effective at estimating scale change than the latter. Given the performance comparison, we uses the FMA + affine, the best among the six methods, as our temporal partitioning method in



(a)

(b)

(c)

(d)

Fig. 6. The performance of six temporal partitioning methods in 10 FPVs: (a) criteria 1: the average length of near-sets (b) criteria 2: the percentage of useless LVI (c) criteria 3: the average number of matching points between pseudo-references in temporally adjacent near-sets (d) criteria 3: the average number of matching points between start frames in temporally adjacent near-sets.

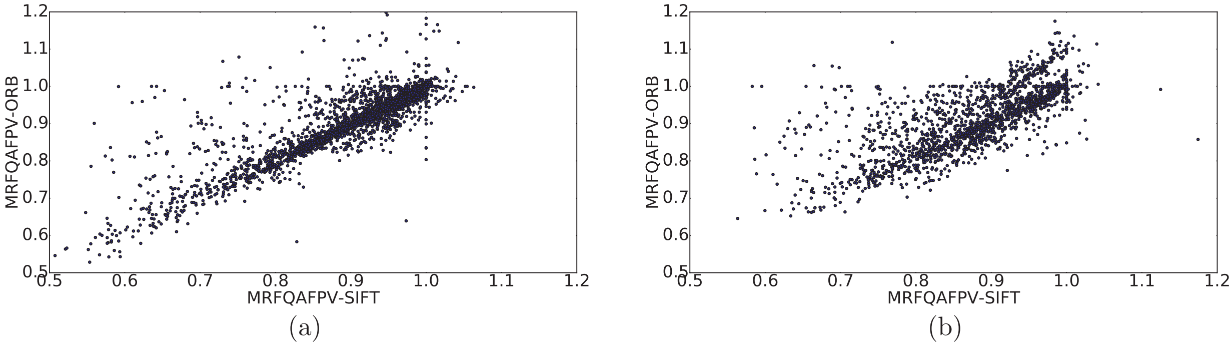
MRFQAFPV in the following sections.

Feature detector: Next, we explore the performance of LVI using two different feature detectors for step 1 of Fig. 1. Specifically, we compare the quality scores of LVI using SIFT [60] (SIFT-LVI) and using ORB [61] (ORB-LVI). Their results are similar in most images, but there are large difference in a few pairs of images. We apply MRFQAFPV as in Section 5 by incorporating either SIFT and ORB as the feature matching detector. Fig. Fig. 7(a) and (b) shows scatter plots of the LVI scores for MRFQAFPV-SIFT versus MRFQAFPV-ORB from outdoor and indoor videos, with average mean square error (MSE) 0.03 and 0.05, respectively. Note that we do not consider those frames that have too few matching points using either SIFT or ORB.

In addition, we also apply SIFT-LVI and ORB-LVI on three image-quality datasets: the LIVE image quality database [56], the Categorical Subjective Image Quality (CSIQ) [36] and the Tampere Image Quality Database (TID2013) [57]. The MSE between all calculated quality scores of SIFT-LVI and ORB-LVI are 0.156, 0.049 and 0.071, respectively. The advantage of using ORB instead of SIFT is that ORB is computationally much faster than SIFT [61]. Given the small performance differences between using SIFT and ORB, we choose ORB as a more computationally efficient feature detector in LVI and MRFQAFPV.

6.2. Performance evaluation

In this section, we explore performance of our methods using both synthetically injected distortions as well as images taken from actual FPV containing real distortions. We begin by with objective comparisons on images with synthetically-generated distortions to show that LVI is effective at measuring blur, but insensitive to geometric distortions, including shear and rotation. Next, we present results of a subjective test using images extracted from FPVs, which demonstrate that MRFQAFPV outperforms existing NR QEs for quality assessment of individual frames with “similar enough” content in FPVs. A second subjective test demonstrates that LVI and existing NR QEs can be generalized to measure images with both blur and geometric distortions simultaneously. Finally, we apply LVI to subjective data with



(a)

(b)

Fig. 7. The distribution of MRFQAFPV-SIFT versus MRFQAFPV-ORB: (a) outdoor content (b) indoor content.

distortions other than those in FPVs [36,56,57], to demonstrate that LVI is able to characterize quality of some of these distortions as well.

Synthetic distortions: LVI is sensitive to motion blur, but insensitive to affine transformation. To demonstrate this, we introduce synthetic distortions including motion blur, shear and rotation into 13 manually-selected high-quality FPV frames with different content [1]. We apply different 1-D box filters with lengths 1 to 30 to simulate different amounts of motion blur. The LVI scores of all test images decreases significantly, from 1 to an average of 0.461 as the blur increases. Synthetic shear and rotation are also created using an affine transformation. For these geometric distortions, LVI decreases to no less than 0.947 when the shear difference increases from 0 to 0.4, and decreases to no less than 0.965 when the rotation increases from 0° to 90°.

Subjective test for MRFQAFPV: Next, we implemented a subjective test to evaluate the performance of a quality measure within the MRFQAFPV framework. The goal of this test is to evaluate the effectiveness of MRFQAFPV to characterize frame quality within an identified near-set.

The test material are frames selected from the 10 videos tested in Section 6.1, and all images are rescaled to 1280 × 720 both for computing in MRFQAFPV and for presentation to viewers in the test. The selection procedure of frames from one FPV has three steps, with the goal to find five images that have similar content but distinct quality. First, we identify all near-sets that have frames with LVI scores located in [0,9,1),[0,8,0.9),[0.7,0.8),[0.6,0.7), respectively. Second, we choose the near-set $\mathscr{X}$ with the most frames among all near-sets found in the first step. Third, we choose the pseudo-reference frame and four frames with LVI score closest to each of 0.95, 0.85, 0.75 and 0.65 in $\mathscr{X}$. In total, we have 10 test groups, each with five test images.

The test methodology is paired comparison. In each of the 10 test groups, we implement full paired comparisons for all five frames. The platform of this test is Amazon Mechanical Turk. The number of participants is 30 with no record of gender. The instruction presented before each test is as follows: *In the test, there will be some pairs of images for you to compare, and please select the image with better technical quality in each pair. The technical quality mainly refers to blur, noise and compression artifacts, and does not include composition. For each pair of images, you can view both images back and forth to a maximum of five times and then make your decision anyway.* Any accepted answer is not allowed to have at more than one circular triad [62], defined as a situation that $I_1 > I_2, I_2 > I_3$ and $I_3 > I_1$, where I_1, I_2, I_3 are three different images, and ">" means "better".

The subjective score of each image is calculated based on the Bradley-Terry Model [63]. We apply LVI and five NR QEs, NIQE [28], IL-NIQE [29], a perceptual blur metric (Blurriness) [44], JNBM [38] and CPBD [42] to all test images. Table 1 shows the PLCC and SROCC between subjective scores with LVI and the five NR QEs. LVI shows the best performance in five near-sets, "basketball", "walk", "eat", "ping pong", and "flight", with PLCC greater than 0.9. The PLCC is relatively low in four near-sets, "run", "billiards", "talk" and "car" with PLCC less than 0.8. In terms of the overall performance of the five NR QEs, the best is outdoor videos, next is indoor videos, the worst is in-vehicle videos. Among the five NR QEs, blurriness and JNBM show better performance than the other three QEs. LVI outperforms the five NR QEs in six near-sets, and shows intermediate performance in the other four near-sets.

Discussion: Content influences all tested QEs; however, LVI is less influenced by content than the other five QEs. All QEs have somewhat inconsistent performance across different contents. This content dependency is apparent from the fact that the PLCC has large variations when evaluating the ten near-sets. Compared to the five NR QEs, LVI shows more consistent performance indicating a reduction in content-dependency.

In addition, there are three challenging contents for all the QEs: "talk", "car" and "run". First, the set of "talk" is captured in a small room with apparent geometric distortions. LVI shows the best performance among all QEs with PLCC 0.72. Second, the set of "car" is difficult for most participants to distinguish quality variations in the subjective test. Third, there exists spatially inconsistent motion blur in the set of "run" that significantly influences the LVI measure.

Subjective test for LVI and geometric distortions: Next, we implemented a subjective test using paired comparison in [2] to validate the performance of LVI and to evaluate the overall quality of images with both blur and geometric distortions. The test mainly has three components: motion blur, motion blur with shear, motion blur with rotation. Recall these are the dominant types of distortions in FPV frames. The subjective scores are calculated by Bradley-Terry Model [63]. The motion blur test uses temporally nearby captured frames of three contents. Each content contains test images of five levels, which is partitioned based on their LVI scores. Compared with seven NR QEs, JNBM [38], BIQI [64], CPBD [42], BRISQUE [27], CORNIA [65], IL-NIQE [29] and NIQE [28], only LVI correctly ranks all test images. In the motion blur with shear test, we evaluate images with multiple distortions using four levels of synthetic motion blur and four levels of synthetic shear. We use the same number of distortion levels in motion blur as in the rotation test; the difference here is the four different levels of rotation are captured using real images. The results indicate that both shear and rotation introduce quality degradations to images, and the overall quality of an image is a combined effect of blur and geometric distortions. We proposed a form of quality mapping function, Eq. (12), to map LVI or existing NR QEs that are insensitive to geometric distortions with estimated shear and rotation value to the overall quality. Eq. (12) is the mapping function to calculate the overall quality of an image with motion blur and geometric distortions simultaneously.

$$Q(\mathscr{D},q) = q\cdot\left(1-p\cdot\exp\left(-\frac{|q-q_{best}|}{|q_{best}-q_{worst}|}\right)\cdot\mathscr{D}^2\right). \tag{12}$$

where $\mathscr{D}$ is the measured value of shear or rotation (k_s or θ in Eq. 10). q is the QE score of the image, q_{best} and q_{worst} indicate the quality scores for the best- and the worst-quality images based on the corresponding quality measure q, respectively. p is a constant parameter. In terms of

Table 1
PLCC(SROCC) of LVI and five NR QEs with subjective scores.

Video type	Video name	LVI	NIQE	IL-NIQE	Blurriness	JNBM	CPBD
outdoor	basketball	**0.9936**(1.0)	0.9351(1.0)	0.8846(0.7)	0.9862(1.0)	0.9814(1.0)	0.9385(1.0)
	run	0.7096(0.5)	0.4899(0.2)	0.4392(0.1)	**0.9933**(1.0)	0.9739(1.0)	0.9430(0.9)
	walk	0.9052(0.9)	0.7547(0.7)	0.1326(0.3)	0.9398(1.0)	**0.9721**(0.9)	0.8881(0.7)
indoor	billiards	0.7468(0.7)	0.5513(0.7)	0.5523(0.1)	0.7834(0.7)	**0.8377**(0.7)	0.7063(0.7)
	cat	**0.8823**(0.9)	0.8142(0.8)	0.8150(0.6)	0.8396(0.9)	0.8202(0.7)	0.5610(0.4)
	eat	0.9265(0.9)	**0.9911**(0.9)	0.9253(0.9)	0.9732(0.9)	0.8162(0.9)	0.8242(0.8)
	ping pong	**0.9735**(1.0)	0.7010(0.7)	0.6255(0.6)	0.9014(0.8)	0.9095(1.0)	0.8331(0.8)
	talk	**0.7247**(0.7)	0.6045(0.6)	0.6408(0.6)	0.3901(0.6)	0.5937(0.7)	0.5023(0.7)
in-vehicle	car	**0.6765**(0.7)	0.2105(0.3)	0.2865(0.1)	0.5501(0.4)	0.4644(0.4)	0.1801(0.3)
	flight	**0.9527**(0.9)	0.7019(0.7)	0.2869(0.3)	0.7718(0.9)	0.9449(0.9)	0.7263(0.9)

Table 2
SROCC of LVI and five FR QEs for the LIVE, CSIQ and TID2013 image databases.

database name	distortion type	LVI	SSIM	VIF	FSIM	VSNR	SR-SIM
LIVE	Gaussian blur	0.9651	0.9516	0.9728	0.9707	0.9413	0.9660
	JPEG	0.8291	0.9764	0.9849	0.9834	0.9656	0.9822
	JPEG2000	0.9427	0.9614	0.9716	0.9716	0.9551	0.9701
	Fastfading	0.9176	0.9556	0.9650	0.9499	0.9027	0.9467
CSIQ	Gaussian blur	0.9630	0.9609	0.9745	0.9729	0.9446	0.9768
	JPEG	0.7466	0.9553	0.9705	0.9654	0.9174	0.9668
	JPEG2000	0.9371	0.9605	0.9672	0.9686	0.9486	0.9774
	Contrast	0.9404	0.7924	0.9347	0.9421	0.8720	0.9530
TID2013	Gaussian blur	0.9430	0.9633	0.9649	0.9569	0.9526	0.9619
	JPEG	0.8211	0.9111	0.9191	0.9303	0.9037	0.9377
	JPEG2000	0.9265	0.9010	0.9516	0.9584	0.9270	0.9675
	Image denoising	0.8727	0.9101	0.8912	0.9313	0.9116	0.9401
	Contrast change	0.8519	0.4551	0.8386	0.4718	0.3514	0.4704

the optimized p values based on SROCC between subjective and objective quality scores, both shear and rotation are highly dependent on content. Specifically, shear is less sensitive to content variations than rotation.

Overall, LVI outperforms existing NR QEs in evaluating actual captured frames in FPVs. Also, both LVI and NR QEs that are insensitive to geometric changes can be generalized to incorporate measurements of geometric quality degradations.

Scenarios other than FPVs: LVI is effective at measuring distortions other than blur in FPVs; however, LVI cannot be used to measure distortions caused by any type of noise. We apply LVI to three image databases designed for evaluating IQEs, LIVE [56], CSIQ [36] and TID2013 [57]. Note that the images in these databases only contain synthetically created distortions, and are in perfect pixel alignment. We use Spearman correlation coefficients (SROCC) to compare the performance of LVI with 5 FR methods: SSIM [21], VIF [32], FSIM [22], VSNR [35] and SR-SIM[34]. Table 2 lists some distortions that LVI can measure in the three image databases. The results indicate that LVI demonstrates acceptable performance in the scenarios shown in Table 2, despite the fact that it has not been designed for those cases. Note that LVI works much better for JPEG2000 than JPEG. The reason is that JPEG introduces block boundary effects in the matching patches used in the LVI measure. The block boundaries have the potential to increase the information measure in a single patch. In addition, in [32], the results also show that VIF performs better in JPEG2000 than JPEG.

7. Conclusions

In this paper, we introduce a new image quality assessment strategy, mutual reference, that uses effective information provided by the overlap between images, without relying on pixel alignment. This mutual reference strategy does not fit into the typical categorization of FR, RR or NR methods. We then propose a mutual reference QE, Local Visual Information (LVI), that primarily measures the relative blur between two images. LVI is effective for comparing two images that have similar scales and are not too blurry. To apply the MR strategy to assess the quality of frames within a First-Person Video, we propose a framework, MRFQAFPV, which uses a pairwise measure and incorporate LVI as the quality estimator.

MRFQAFPV provides several effective tools for assessing lifelogs. First, the temporal partitioning in MRFQAFPV partitions FPVs into different segments such that each segment contains different content. The pseudo-references in each segment provide information for video summarization using shots. Second, the quality estimation in MRFQAFPV is an effective assessment tool for video fast-forward. It can help to avoid using frames with heavy quality degradations. Third, from the perspective of analysis, the quality score of each frame provides an indication of useful and useless frames for applications such as object detection and activity recognition.

We experimentally explore and validate several properties of LVI. First, LVI primarily measures blur, and is insensitive to shear and rotation. Second, LVI outperforms existing NR QEs at measuring the quality of actual frames in FPVs. Third, LVI has acceptable performance in measuring some additional distortions, such as contrast change. Also, we implement a subjective test to demonstrate that MRFQAFPV is an effective framework to estimate the quality of individual frames with similar content in FPVs.

The future work for our framework is to (1) remove the scaling constraint so that the quality measure can be applied to images with different scales, (2) develop a quality estimator between images that have no overlapping content and incorporate it into our present framework, and (3) incorporate measures of more varieties of quality degradations, such as hazing, over-exposure and under-exposure.

Conflict of interest

No conflict.

References

[1] C. Bai and A.R. Reibman, Characterizing distortions in first-person videos, in: IEEE International Conference on Image Processing, 2016, pp. 2440–2444.

[2] C. Bai, A.R. Reibman, Subjective evaluation of distortions in first-person videos, Human Vision ElectronicImaging (2017).

[3] C. Bai and A.R. Reibman, Mutual reference frame-quality assessment for first-person videos, in: IEEE International Conference on Image Processing, 2017.

[4] A. Betancourt, P. Morerio, C.S. Regazzoni, M. Rauterberg, The evolution of first-person vision methods: a survey, IEEE Trans. Circuits Syst. Video Technol. 25 (5) (2015) 744–760.

[5] M. Bolanos, M. Dimiccoli, P. Radeva, Toward storytelling from visual lifelogging: an overview, IEEE Trans. Human-Mach. Syst. (2016).

[6] B. Xiong and K. Grauman, Detecting snap points in egocentric video with a web photo prior, in: European Conference on Computer Vision, 2014, pp. 282–298.

[7] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[8] Y. Poleg, C. Arora, and S. Peleg, Temporal segmentation of egocentric videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2537–2544.

[9] J. Ghosh, Y.J. Lee, and K. Grauman, Discovering important people and objects for egocentric video summarization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1346–1353.

[10] A.L. Baulida, Semantic and diverse summarization of egocentric photo events, Erasmus Mundus M.Sc. in Visions and Robotics Thesis, 2012.

[11] P. Wang, A.F. Smeaton, Semantics-based selection of everyday concepts in visual lifelogging, Int. J. Multimedia Inf. Retrieval 1 (2) (2012) 87–101.

[12] V. Chandrasekhar, W. Min, X. Li, C. Tan, B. Mandal, L. Li, and J. Hwee Lim, Efficient retrieval from large-scale egocentric visual data using a sparse graph representation, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 527–534.

[13] H. Kang, M. Hebert, and T. Kanade, Discovering object instances from scenes of daily living, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 762–769.

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

[15] H. Wannous, V. Dovgalecs, R. Mégret, and M. Daoudi, Place recognition via 3d modeling for personal activity lifelog using wearable camera, in: International Conference on Multimedia Modeling, 2012, pp. 244–254.

[16] M.M. Silva, W.L.S. Ramos, J.P.K. Ferreira, M.F.M. Campos, and E.R. Nascimento, Towards semantic fast-forward and stabilized egocentric videos, in: European Conference on Computer Vision, 2016, pp. 557–571.

[17] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, Egosampling: Fast-forward and stereo for egocentric videos, in: Computer Vision and Pattern Recognition, 2015, pp. 4768–4776.

[18] Y. Niu, F. Liu, What makes a professional video? a computational aesthetics approach, IEEE Trans. Circuits Syst. Video Technol. 22 (7) (2012) 1037–1049.

[19] H. Jin, P. Favaro, R. Cipolla, Visual tracking in the presence of motion blur, IEEE Conf. Comput. Vision Pattern Recogn. 2 (2005) 18–25.

[20] S. Baker, E. Bennett, S.B. Kang, R. Szeliski, Removing rolling shutter wobble, IEEE

Conf. Comput. Vision Pattern Recogn. (2010) 2392–2399.
[21] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[22] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.
[23] L. Zhang, L. Zhang, X. Mou, D. Zhang, A comprehensive evaluation of full reference image quality assessment algorithms, IEEE Int. Conf. Image Process. (2012) 1477–1480.
[24] A. Rehman, Z. Wang, Reduced-reference image quality assessment by structural similarity estimation, IEEE Trans. Image Process. 21 (8) (2012) 3378–3389.
[25] S.S. Hemami, A.R. Reibman, No-reference image and video quality estimation: applications and human-motivated design, Signal Process.: Image Commun. (2010).
[26] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.
[27] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[28] A. Mittal, R. Soundararajan, A.C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.
[29] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.
[30] H. Liu and A.R. Reibman, Software to stress test image quality estimators, in: Quality of Multimedia Experience (QoMEX), 2016.
[31] D.M. Chandler, Seven challenges in image quality assessment: past, present, and future research, ISRN Signal Process. (2013).
[32] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[33] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.
[34] L. Zhang, H. Li, Sr-sim: A fast and high performance iqa index based on spectral residual, IEEE Int. Conf. Image Process. (2012) 1473–1476.
[35] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[36] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) http://vision.okstate.edu/index.php?loc = csiq.
[37] W. Lin, L. Dong, P. Xue, Visual distortion gauge based on discrimination of noticeable contrast changes, IEEE Trans. Circuits Syst. Video Technol. 15 (7) (2005) 900–909.
[38] R. Ferzli, L.J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), IEEE Trans. Image Process. 18 (4) (2009) 717–728.
[39] X. Marichal, W.-Y. Ma, H. Zhang, Blur determination in the compressed domain using DCT information, IEEE International Conference on Image Processing, vol. 2, IEEE, 1999, pp. 386–390.
[40] P. Marziliano, F. Dufaux, S. Winkler, Perceptual blur and ringing metrics: application to jpeg2000, Signal Process. Image Commun. 19 (2) (2004) 163–172.
[41] X. Wang, B. Tian, C. Liang, D. Shi, Blind image quality assessment for measuring image blur, Image Signal Process. 1 (2008) 467–470.
[42] N.D. Narvekar, L.J. Karam, A no-reference image blur metric based on the cumulative probability of blur detection (CPBD), IEEE Trans. Image Process. 20 (9) (2011) 2678–2683.
[43] H. Hu, G. De Haan, Low cost robust blur estimator, IEEE Int. Conf. Image Process. (2006) 617–620.
[44] F. Crete, T. Dolmiere, P. Ladret, M. Nicolas, The blur effect: perception and estimation with a new no-reference perceptual blur metric, Electron. Imaging (2007).
[45] R. Hassen, Z. Wang, M.M. Salama, Image sharpness assessment based on local phase coherence, IEEE Trans. Image Process. 22 (7) (2013) 2798–2810.
[46] A.M. Demirtas, A.R. Reibman, H. Jafarkhani, Full-reference quality estimation for images with different spatial resolutions, IEEE Trans. Image Process. 23 (5) (2014) 2069–2080.
[47] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[48] T. Stathaki, Image Fusion: Algorithms and Applications, Academic Press, 2011.
[49] G. Piella, H. Heijmans, A new quality metric for image fusion, IEEE Int. Conf. Image Process. 3 (2003) III–173.
[50] C. Yang, J.-Q. Zhang, X.-R. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2) (2008) 156–160.
[51] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[52] M.A. Saad, M.H. Pinson, D.G. Nicholas, N.V. Kets, G.V. Wallendael, R.V. Jaladi, P.J. Corriveau, Impact of camera pixel count and monitor resolution perceptual image quality, Colour Visual Comput. Symp. (2015) 1–6.
[53] M.J. Wainwright, E.P. Simoncelli, A.S. Willsky, Random cascades on wavelet trees and their use in analyzing and modeling natural images, Appl. Comput. Harmonic Anal. 11 (1) (2001) 89–123.
[54] A. Agarwal, C. Jawahar, P. Narayanan, A survey of planar homography estimation techniques, Centre Visual Inf. Technol. (2005).
[55] P. Kellnhofer, T. Ritschel, K. Myszkowski, H.-P. Seidel, A transformation-aware perceptual image metric, Electron. Imaging (2015).
[56] H.R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “LIVE image quality assessment database release 2,” 2005, http://live.ece.utexas.edu/research/quality/subjective.htm.
[57] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C.J. Kuo, Image database TID2013: peculiarities, results and perspectives, Signal Process. Image Commun. 30 (2015) 57–77.
[58] J. Mas and G. Fernandez, Video shot boundary detection based on color histogram, Notebook Papers TRECVID2003, 2003.
[59] Test video dataset for Mutual Reference Frame Quality Assessment of First-Person videos, https://engineering.purdue.edu/VADL/resources/MRFQAFPV/test_videos.zip.
[60] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[61] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to sift or surf, IEEE Int. Conf. Comput. Vision (2011) 2564–2571.
[62] J.-S. Lee, F. De Simone, T. Ebrahimi, Subjective quality evaluation via paired comparison: application to scalable video coding, IEEE Trans. Multimedia 13 (5) (2011) 882–893.
[63] J.C. Handley, Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment, in: PICS, 2001.
[64] A.K. Moorthy, A.C. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Process. Lett. (2010).
[65] L.K.D.D. Peng Ye, Jayant Kumar, Unsupervised feature learning framework for no-reference image quality assessment, IEEE Conf. Comput. Vision Pattern Recogn. (2012) 1098–1105.