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