MUTUAL REFERENCE FRAME-QUALITY ASSESSMENT FOR 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 are explored for applications of sharing experiences, recording daily lives, measuring social interactions and behaviors. These applications can be improved by an accurate quality assessment. To maximally use the information present in a FPV, we introduce a new strategy for image quality assessment, called mutual reference (MR). MR does not fit into the previous categorization of full-reference, reduced-reference and no-reference. It uses the overlapping content between images to provide effective information for quality estimation. We propose a framework of mutual reference frame-quality assessment for FPVs (MRFQAFPV) to implement the MR strategy based on a MR quality estimator (QE), LVI. The effectiveness of MRFQAFPV is demonstrated in a subjective test by comparing with 3 no-reference QEs and frame-to-frame motion.

Index Terms— first-person videos, image quality assessment, mutual reference, LVI, near-set

1. INTRODUCTION

Wearable cameras provide a new way to record videos in first-person perspective without holding any device in hand. First-person videos (FPVs) captured by these cameras are therefore becoming a widely spread type of videos that can document activities, share experiences and record trips without length limitation or specific structure. Numerous applications of FPVs have emerged using object tracking, activity recognition, video summarization and retrieval [1]. Recently, some topics related to viewing experience of FPVs have been proposed, involving fast-forward and stabilization [2, 3], engagement detection [4] and quality evaluation [5, 6].

Quality assessment of FPVs is very important. First, it can identify whether frames have high enough quality for applications using object tracking and activity recognition. Second, it serves as an evaluation tool for improving the viewing experience of FPVs [3]. Third, the visual quality of frames is a considerable factor for keyframe or snap points detection [7], and can be incorporated into frameworks for video summarization [8, 9].

To evaluate the quality of every frame in a FPV, image quality estimators (IQEs) can be applied. Existing IQEs are normally classified into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. FR and RR methods estimate the quality of a distorted image based on its high-quality corresponding reference image that is also the source of the distorted image. One limitation for most FR and RR QEs [10–12] is that they cannot evaluate a test image that is better than its reference image. Two exceptions are VIF [13] and Visual Distortion Gauge [14]. Another related limitation is that FR and RR methods assumes that the reference image is not degraded, otherwise their results are not meaningful.

NR methods estimate the quality of a single image without relying on any reference. One subset of NR QEs is blur metrics [15, 16]; another subset is natural scene statistics based QEs [17–19]. However, most existing NR methods are content dependent so that it makes sense to compare their quality scores only when the two images have almost the same content.

We propose a new image quality estimation strategy, called mutual reference (MR), which does not fit into any of the previous classification of FR, RR or NR methods. The basic strategy for MR is to estimate the quality of a test image based on one or more pseudo-reference images. As implied by the name pseudo-reference, this image needs to share overlapping content with the test image but does not necessarily need to be pixel-aligned. Compared to FR, RR and NR methods, the advantages to apply a MR method for FPVs are: (1) MR provides a ***relative*** quality estimation that allows degradations to be present in the pseudo-reference image. (2) MR uses information provided from overlapping content between images to minimize content dependency in quality scores.

To apply the MR strategy to FPVs, we design a framework of mutual reference frame-quality assessment for FPVs (MRFQAFPV) based on a MR QE, the local visual information (LVI), proposed in [5]. This paper is organized as follows: Section 2 presents a detailed description of the mutual reference strategy. Section 3 describes the MRFQAFPV, which has 3 steps: temporal partitioning, reference search, and quality estimation. In Section 4, we first evaluate the performance of our proposed temporal partitioning method. Then, a subjective test is implemented to demonstrate the effectiveness of MRFQAFPV. Section 5 summarizes this paper and discusses the future work.

2. 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 collection of images that share common content. One example is a group of images captured either from or of nearby locations.

There are two approaches for MR image quality assessment. The first is a pairwise measure: use a single pseudo-reference image to estimate the quality of a test image when both images belong to the same near-set. Ideally, the pseudo-reference image should have the best quality in the identified near-set. One example of this method is the Local Visual Information (LVI) [5], which uses a pseudo-reference to estimate the quality of a test image with pixel misalignment [6].

The second approach is a group measure: evaluate the quality of an image using more than one image in the near-set as pseudo-references. An example is the quality assessment of image fusion, in which complementary information from a group of images is integrated to form a new image [20]. A common strategy to estimate the quality of the fused image [21, 22] is to use all source images for fusion, possibly with misalignment. Source images and the fused images created by different image fusion algorithms [22] can all be classified into the same near-set.

MR methods cannot be classified into any of the FR, RR or NR methods. In particular, MR uses the effective information provided by the overlapping portion of the images. The overlaps between those images could differ because of any geometric transformation. In contrast, FR and RR uses an exact high-quality source image as reference to provide information, while NR uses implicit knowledge of distorted image versus high-quality image.

MR provides a relative quality estimation that allows degradations to be present in all images in the near-set. As a comparison, FR and RR methods estimate any distorted image relative to an undistorted reference image of entirely same content. NR methods provides absolute quality scores, not relative to any other images. It is designed to be used for comparing two images with completely different content.

One application for MR is to assess images captured of the same scene from slightly different locations, as has been considered in [23]. Another application is to estimate the quality of temporally nearby frames in a video.

3. MUTUAL REFERENCE FRAMEWORK FOR FIRST-PERSON VIDEOS

3.1. Overview of LVI

As described in [5], Local Visual Information (LVI) has 3 steps as shown in Figure 1. In the first block, LVI builds pixel

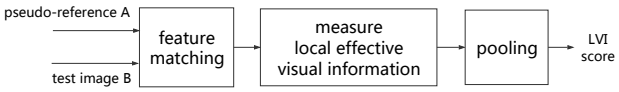

Fig. 1. Block diagram of local visual information (LVI)

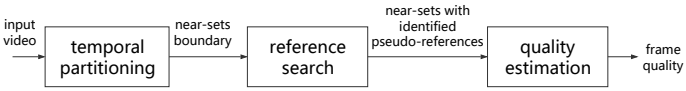

Fig. 2. Framework of quality assessment

correspondences between A and B using feature matching with the ORB feature [24]. In the second block, it measures the effective local visual information between A and B for all corresponding patches using a method similar to VIF [13]. See more details in [5]. In the final block, the LVI score is pooled from all local information measures.

However, there exists two limitations when applying LVI. First, LVI is not effective at measuring quality when there are insufficient feature matching points between the pseudo-reference and the test image. In this case, we consider the test image to have a zero LVI score. Second, LVI is sensitive to scaling change, although it is insensitive to shear and rotation [5]. When the test image has objects in very different scales relative to the pseudo-reference, the LVI score is unreliable. In the implementation of LVI, we apply affine or homography estimation [5] to measure the scale change between the pseudo-reference and the test image in horizontal direction, s_x, and vertical direction, s_y. If s_x or s_y exceeds the bounded range $[a, 1/a]$, we reject the LVI score as unreliable. a is experimentally set to be 0.95.

3.2. Framework

Our framework of mutual reference frame-quality assessment for FPVs (MRFQAFPV) consists of 3 steps: temporal partitioning, reference search and quality estimation, as shown in Figure 2. Note that the framework is designed based on LVI, so we use the approach of pairwise measurement as described in Section 2.

The step of *temporal partitioning* shown in Figure 2 is to partition frames within a time interval into different near-sets. The k_{th} near-set 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. Algorithm 1 describes our general temporal partitioning method, which includes a boundary search strategy detailed in Algorithm 2. Algorithm 2 relies on feature matching, after which a scale check process is incorporated to guarantee that frames within any classified near-set have small scale change. This allows us to effectively apply LVI in our near-sets in the following steps. Note that if the partitioned near-set has fewer than 10 frames, then we discard the near-set and the current B_1^k is considered to be an uncategorized frame.

The step of *reference search* in Figure 2 is an iterative

approach to find the best pseudo-reference frame in each near-set. Let R_k be the pseudo-reference frame in near-set k. First, initialize R^k to be the start frame B_1^k with initial LVI score 1. Second, we calculate the initial LVI scores from frame $B_1^k + 1$ to the end frame B_2^k using R^k as pseudo-reference. LVI indicates that any frame with score greater than 1 has better quality than R_k. Therefore, we choose the frame with the largest initial LVI score to replace the current R^k. Finally, R^k is added to the representation of the near-set k, which is (B_1^k, B_2^k, R_k).

The step of *quality estimation* in Figure 2 is to estimate the frame LVI score of each frame. The input is a near-set k, represented by (B_1^k, B_2^k, R^k). Let index n be the frame number. In near-set k, the system uses R^k as the pseudo-reference to measure the quality of all remaining frames in near-set k, and stores the LVI score of frame n as Q_{LVI}^n, which is the final quality measure.

Algorithm 1 temporal partitioning

1: set $k = 1$, $B_1^k = 1$
2: **boundary search** for B_2^k based on B_1^k
3: set $k = k+1$, set $B_1^k = B_2^k + 1$, **break** when B_1^k exceeds the last frame in the video
4: go to 2

Algorithm 2 boundary search

1: get the start frame 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 bisection search between $B_1^k + (n-1) \cdot \delta$ and $B_1^k + n \cdot \delta$ using the same decision rule, **break** when the bisection interval ≤ 1 and set B_2^k to be start frame of the bisection 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**

4. EXPERIMENTS AND RESULTS

We evaluate both the performances of our temporal partitioning algorithm and frame quality estimation of MRFQAFPV. Our test resources are 10 FPVs with different content captured by a Pivothead camera (1080p, 30fps).

4.1. Evaluation of Temporal Partitioning

The performance of our proposed temporal partitioning method described in Section 3.2 is compared with two baseline methods. The first baseline method uses a fixed time interval (30 frames) to partition frames into different near-sets. The second baseline method uses block-based optical flow to compute cumulative displacement [25], where each partitioned temporal interval has a cumulative displacement of value 0.1. We also considered the boundary detection method [26], but because it often creates only one segment for the entire FPV, we do not present its results here.

The evaluation of temporal partitioning is based on 3 criteria: (1) The length of the near-set is long enough that it covers most frames captured in the specific scene. (2) The percentage of frames with useless LVI is low. We consider 3 types of frames to have useless LVI: frames that failed in feature matching, frames with LVI score greater than 1, and uncategorized frames. (3) Temporally adjacent near-sets contain frames that have little shared content. We measure the shared content between near-sets by counting the number of matching points between any two frames.

We compare our proposed method and the 2 baseline methods in Figure 3. Figure 3(a) and Figure 3(b) compare the 3 methods using the average length of near-sets and the percentage of frames with useless LVI, based on the first and the second criteria, respectively. Figure 3(c) and Figure 3(d) compare the methods using the third criterion demonstrated by both the average number of matching points between pseudo-references and between start frames in temporally adjacent near-sets. The video indexes in Figure 3 represent different FPVs. Specifically, 0 to 2 are outdoor, 3 to 7 are indoor, and 8, 9 are in-vehicle videos.

As can be seen in Figure 3, our proposed method has the intermediate near-set length, the lowest percentage of useless LVI, and the fewest matching points between either pseudo-references or start frames in temporally adjacent near-sets. As a comparison, the method using fixed interval has the shortest near-set length and the intermediate percentage of useless LVI. The method using optical flow has the longest near-set length and the largest percentage of useless LVI. Based on the 3 criteria mentioned above, our proposed method outperforms the two baseline methods.

4.2. Subjective Test of Frame Quality

We implement a subjective test to evaluate the quality of frames in identified near-sets and test if the quality measure of MRFQAFPV correctly ranks frame quality.

Individual test images are selected from FPVs to have specific LVI scores within the context of our MRFQAFPV framework. The content of test images are listed in Table I. In the implementation of LVI and the presentation of our subjective test, we rescale all frames to 1280×720. The procedure to select test frames in one FPV is as follows: (1) All near-sets that

video type	video content	LVI	NIQE	IL-NIQE	Blurriness	motion
outdoor	basketball	**0.9936(1.0)**	0.9351(1.0)	0.8846(0.7)	0.9862(1.0)	0.9931 (1.0)
	run	0.7096(0.5)	0.4899(0.2)	0.4392(0.1)	**0.9933(1.0)**	0.6407(0.8)
	walk	0.9052 (0.9)	0.7547(0.7)	0.1326(0.3)	**0.9398(1.0)**	0.6024(0.6)
indoor	billiards	0.7468(0.7)	0.5513(0.7)	0.5523(0.1)	**0.7834(0.7)**	0.6399(0.6)
	cat	**0.8823(0.9)**	0.8142(0.8)	0.8150(0.6)	0.8396(0.9)	0.6958(0.5)
	eat	0.9265(0.9)	**0.9911(0.9)**	0.9253(0.9)	0.9732(0.9)	0.8988(0.9)
	ping pong	0.9735**(1.0)**	0.7010(0.7)	0.6255(0.6)	0.9014(0.8)	**0.9743**(0.9)
	talk	0.7247(0.7)	0.6045(0.6)	0.6408(0.6)	0.3901(0.6)	**0.8172(0.9)**
in-vehicle	car	**0.6765(0.7)**	0.2105(0.3)	0.2865(0.1)	0.5501(0.4)	0.2511(0.2)
	flight	**0.9527(0.9)**	0.7019(0.7)	0.2869(0.3)	0.7718(0.9)	0.8126(0.7)

Table I. PLCC(SROCC) between LVI, 3 NR QEs and motion and subjective scores

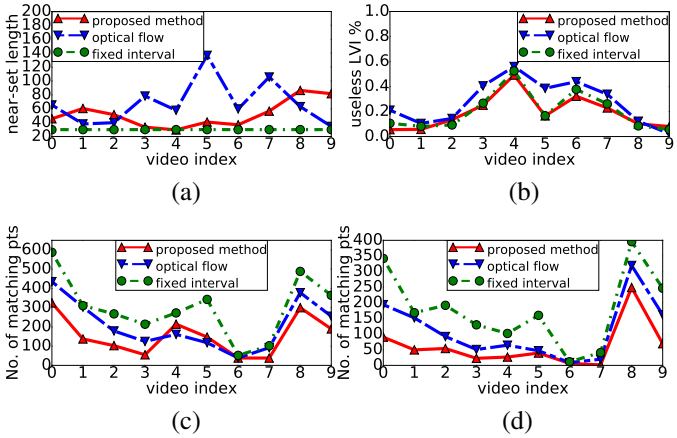

Fig. 3. Comparison of our proposed temporal partitioning method and two baseline methods: (a) the average length of near-sets (b) the percentage of frames with useless LVI (c) the average number of matching points between pseudo-references in temporally adjacent near-sets (d) the average number of matching points between start frames in temporally adjacent near-sets

have frames with LVI scores located respectively in $[0, 9, 1)$, $[0, 8, 0.9)$, $[0.7, 0.8)$, $[0.6, 0.7)$ are identified. (2) The near-set with the longest length among all identified near-sets is selected as our test near-set S. (3) We choose the pseudo-reference frame and 4 frames with LVI scores closest to each of 0.95, 0.85, 0.75, 0.65 in S as our test frames. Hence, we get 5 test frames chosen from one near-set within one FPV.

The subjective test is implemented on Amazon Mechanical Turk by paired comparison using 30 participants. The instruction before each test is presented as: *In the test, there will be some pairs of images for you to compare. 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 5 times and then make your decision.* Any accepted answer is allowed to have at most one circular triad [27], defined as a situation that $A > B$, $B > C$ and $C > A$, where A, B, C are 3 different images in a test group, and "$>$" means the choice of "better".

The subjective scores are calculated using the Bradley-Terry Model [28]. We evaluate the quality measure of MRFQAFPV and 3 NR QEs (NIQE [18], IL-NIQE [19], a perceptual blur metric [29] using as feature blurriness in [7]).

In addition, we compare these methods with simply using the estimated frame-to-frame motion. Specifically in videos, motion often introduces quality degradations. Characterizing frame quality using frame-to-frame motion is a MR method, since it uses the effective information provided by neighboring frames. In our experiment, the motion is characterized by the optical flow magnitude from the previous frame to the current frame. First, we use method in [25] to calculate optical flow between neighboring frames. Second, we employ self-tuning spectral clustering [30] to separately cluster horizontal and vertical optical flow vectors, and the centroid of the dominant cluster is considered as the aggregate optical flow vector of the frame. Finally, the motion is quantified as the magnitude of the aggregate optical flow vector.

Table I shows the PLCC and SROCC between subjective scores with LVI, 3 NR QEs and motion. LVI shows the best or the second best performances in all near-sets. Among the 3 NR QEs, the performance is successfully ranked as follows: bluriness, NIQE, IL-NIQE. The relation between frame-to-frame motion and frame quality varies across different contents, and shows intermediate performances in most near-sets among these metrics.

5. CONCLUSION

In this paper, we propose a new strategy of image quality assessment, called mutual reference, which does not fit the typical categorization of FR, RR and NR methods. Then, we propose a framework of mutual reference frame-quality assessment for FPVs (MRFQAFPV), in which we estimate the frame quality by incorporating the MR QE, LVI [5]. To evaluate the performance of MRFQAFPV, we implement a subjective test to validate its effectiveness by comparing with existing NR QEs and frame-to-frame motion. Remaining issues for future work are how to compare the quality between different near-sets and how to incorporate motion features into the framework of MR quality assessment for FPVs.

6. REFERENCES

[1] Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni, and Matthias Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 744–760, 2015.

[2] Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg, "Egosampling: Fast-forward and stereo for egocentric videos," in *Computer Vision and Pattern Recognition*, 2015, pp. 4768–4776.

[3] Michel Melo Silva, Washington Luis Souza Ramos, Ferreira, et al., "Towards semantic fast-forward and stabilized egocentric videos," in *European Conference on Computer Vision*, 2016, pp. 557–571.

[4] Yu-Chuan Su and Kristen Grauman, "Detecting engagement in egocentric video," *European Conference on Computer Vision*, 2016.

[5] Chen Bai and Amy R. Reibman, "Characterizing distortions in first-person videos," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2440–2444.

[6] Chen Bai and Amy R. Reibman, "Subjective evaluation of distortions in first-person videos," in *Human Vision and Electronic Imaging*, 2017.

[7] 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.

[8] 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.

[9] Zheng Lu and Kristen Grauman, "Story-driven summarization for egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[10] 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.

[11] 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.

[12] Damon M. Chandler and Sheila S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE transactions on image processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[13] 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.

[14] Weisi Lin, Li Dong, and Ping Xue, "Visual distortion gauge based on discrimination of noticeable contrast changes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 900–909, 2005.

[15] Hao Hu and Gerard De Haan, "Low cost robust blur estimator," in *IEEE International Conference on Image Processing*, 2006, pp. 617–620.

[16] Niranjan D. Narvekar and Lina J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678–2683, 2011.

[17] 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.

[18] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[19] Lin Zhang, Lei Zhang, and Alan C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[20] Tania Stathaki, *Image fusion: algorithms and applications*, Academic Press, 2011.

[21] Gemma Piella and Henk Heijmans, "A new quality metric for image fusion," in *IEEE International Conference on Image Processing*, 2003, vol. 3, pp. III–173.

[22] Cui Yang, Jian-Qi Zhang, Xiao-Rui Wang, and Xin Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, no. 2, pp. 156–160, 2008.

[23] Michele A Saad, Margaret H Pinson, Nicholas, et al., "Impact of camera pixel count and monitor resolution perceptual image quality," in *Colour and Visual Computing Symposium (CVCS), 2015*, 2015, pp. 1–6.

[24] 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.

[25] Yair Poleg, Chetan Arora, and Shmuel Peleg, "Temporal segmentation of egocentric videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2537–2544.

[26] Jordi Mas and Gabriel Fernandez, "Video shot boundary detection based on color histogram," *Notebook Papers TRECVID*, 2003.

[27] Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi, "Subjective quality evaluation via paired comparison: application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[28] John C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *PICS*, 2001.

[29] 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.

[30] Lihi Zelnik-Manor and Pietro Perona, "Self-tuning spectral clustering," *Advances in Neural Information Processing Systems*, 2004.