IMAGE AND VIDEO QUALITY ASSESSMENT WITH APPLICATIONS IN FIRST-PERSON VIDEOS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Chen Bai

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Amy R. Reibman, Chair

School of Electrical and Computer Engineering

Dr. Jan P. Allebach

School of Electrical and Computer Engineering

Dr. Charles A. Bouman

School of Electrical and Computer Engineering

Dr. Mary L. Comer

School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis

Head of the School of Electrical and Computer Engineering

To my parents with deepest gratitude.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my doctoral advisor Professor Amy R. Reibman for her support of my research in image and video processing area. Her innovative tutoring and rigorous research attitude helped me to come up with new ideas and overcome difficulties in my works. She led the path for me that how to find a new problem, define the problem, create new methods and validate my work. I truly enjoyed our communication not only about projects but also research attitude and life. I am very honored to be her student and a member of the video analytics for daily living lab (VADL).

Second, I would like to thank my parents for their love and support, especially for these years that I was away from home. Their valuable advice and deep care helped a lot in pursuing my academic career. I also thank my parents for giving me life and the opportunity to grow, learn and pursue dreams.

Third, I would like to say thank you to the rest of my committee members, Professor Jan P. Allebach, Professor Charles A. Bouman, and Professor Mary L. Comer for their insightful advice, consistent encouragement and meaningful comments despite their busy schedules.

I am also thankful to all my lab mates in VADL, Biao Ma, He Liu, Chengzhang Zhong and Haoyu Chen, for their help of my work, communication in image and video processing area, and memorable time we spent in the lab.

I would like to thank all my friends from Purdue University. Our friendship along my journey to finish my PhD helped me to overcome difficulties, build confidence and discard unhappiness.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

ABBREVIATIONS xii

ABSTRACT xiv

1 INTRODUCTION 1

1.1 First-Person Video Quality Assessment 1

1.2 Validation of Proposed Quality Assessment Methods 8

1.3 Contributions and Outline 12

2 IMAGE QUALITY ASSESSMENT FOR FIRST-PERSON VIDEOS 15

2.1 Introduction 15

2.2 Related Works 17

2.2.1 Image Quality Estimator 17

2.2.2 Subjective Test 20

2.3 Mutual Reference 20

2.4 Local Visual Information 22

2.4.1 Basic Principle 22

2.4.2 Reliability Check 25

2.5 Framework of MR Quality Assessment of FPVs 27

2.6 Experiments and Results 30

2.6.1 Implementation Design Comparisons 30

2.6.2 Performance Evaluation 34

2.7 Video Statistics 40

2.7.1 Distortion Classification 41

2.7.2 Classification Results 42

Page

2.8 Subjective Test for Blur and Geometric Distortions 45
2.8.1 Test Overview 45
2.8.2 Test Method and Setup 46
2.8.3 Test Sources 48
2.8.4 Results and Discussion 51
2.9 Conclusions 62
3 CONTROLLABLE ILLUMINATION ENHANCEMENT 68
3.1 Introduction 68
3.2 Related Works 71
3.2.1 Over-exposure and Under-exposure 71
3.2.2 Existing Enhancement Methods 72
3.2.3 Existing Enhancement Evaluation Metrics 72
3.3 Controllable Illumination Enhancement 73
3.4 Over-enhancement Measure 75
3.5 Experiments and Results 78
3.6 Conclusions 81
4 VISIBILITY-INSPIRED TEMPORAL POOLING WITH APPLICATION TO VIDEO STABILIZATION 89
4.1 Introduction 89
4.2 Related works 91
4.2.1 The Window of Visibility 91
4.2.2 Temporal Pooling Methods 92
4.2.3 Objective Quality Assessment 94
4.2.4 Subjective Quality Assessment 94
4.3 Visibility Measurement 95
4.4 Visibility-Inspired Temporal Pooling Modelling 97
4.4.1 Pooling Method 97
4.4.2 Estimating the Function $\lambda(\cdot)$ 97

Page

4.4.3 Data Gathering Strategy 98
4.5 Subjective test for visibility-inspired temporal pooling 100
4.5.1 Test Video Sets 100
4.5.2 Test Setup 102
4.5.3 Subjective Test Results 102
4.5.4 Estimating $\lambda(\cdot)$ 103
4.5.5 Evaluating the Overall Method 104
4.6 Measuring perceptual blurriness after video stabilization 106
4.6.1 Test Motivation and Strategy 106
4.6.2 Synthetic creation of shaky videos 108
4.6.3 Test Description 110
4.6.4 Subjective Results 111
4.6.5 Method Validation 112
4.7 Conclusions 115
5 CONCLUSIONS AND FUTURE WORK 119
5.1 Summary 119
5.2 Future Work 120
REFERENCES 123
VITA 134
PUBLICATIONS 135

LIST OF TABLES

Table Page

2.1 PLCC(SROCC) of LVI and five NR QEs with subjective scores 39

2.2 SROCC of LVI and five FR QEs for the LIVE, CSIQ and TID2013 image databases 40

2.3 Comparison between FPVs and traditional videos 44

2.4 Test images in subjective test 47

2.5 Symmetric transformation method to create shear images 51

2.6 QE performance: motion blur - billiards 53

2.7 QE performance: motion blur - eating 53

2.8 QE performance: motion blur - flight 54

2.9 QE performances: rotation - winter Hovde Hall 56

2.10 QE performances: rotation - bell tower 57

2.11 QE performances: shear - autumn Hovde Hall 59

2.12 QE performances: shear - parking lot 60

3.1 Negative subjective quality ("0" indicates the best) and average processing time of the 6 enhancement methods 81

4.1 PLCC (SROCC) between objective pooling scores and subjective scores. 117

4.2 SROCC and PLCC between objective video quality scores and subjective scores. 118

LIST OF FIGURES

Figure Page

1.1 (a) horizontal camera panning (b) vertical camera panning (c) horizontal camera shaking (d) vertical camera shaking 4

1.2 (a) motion blur (b) rolling shutter artifacts (c) tilt (d) fisheye (e)(f) exposure distortions 5

1.3 Existing QE structure 7

1.4 Generalized QE structure 7

2.1 Block diagram of local visual information (LVI) quality estimator 23

2.2 Left: Pseudo-reference. Right: Test image. LVI score = 0.771 25

2.3 Framework of quality assessment for First Person Video. 27

2.4 Sample test images: (0) basketball (1) run (2) walk (3) billiards (4) cat (5) eat (6) ping pong (7) talk (8) car (9) flight 33

2.5 Sample partitioned near-set 1 34

2.6 Sample partitioned near-set 2 35

2.7 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 36

2.8 The distribution of MRFQAFPV-SIFT versus MRFQAFPV-ORB: (a) outdoor content (b) indoor content 37

2.9 Blockdiagram of estimating video statistics 41

2.10 Line angle distributions: (a) image free from shear and rotation (b) image with rotation only (c) image with shear only (d) image with both rotation and shear 43

2.11 Cumulative distributions: (a) shear (b) LVI 44

Figure Page

2.12 Reference images for each content: (a) billiards (b) eating (c) flight (d) bell tower (e) winter Hovde Hall (f) parking lot (g) autumn Hovde Hall (h) apartment building (i) parking garage 46

2.13 Test images captured in FPVs to have different amounts of motion blur . . 49

2.14 Test images intentionally captured to have different amounts of rotation . . 64

2.15 Test images with different amounts of synthetic shear created from one reference image 64

2.16 Test images with different amounts of synthetic fisheye created from one reference image 65

2.17 Subjective test - rotation and blur: (a) winter Hovde Hall (b) bell tower . 66

2.18 Curve fitted with logistic function between subjective scores and rotation-LVI: (a) winter Hovde Hall (p=4.53) (b) bell tower (p=1.13) 66

2.19 Subjective test - shear and blur: (a) autumn Hovde Hall (b) parking lot . . 66

2.20 Curve fitted with logistic function between subjective scores and generalized shear-LVI (g=4.07): (a) autumn Hovde Hall (b) parking lot 66

2.21 Subjective test - fisheye and blur: (a) parking garage (b) apartment building67

2.22 Parking garage: group 1 prefers non-fisheye, group 2 prefers fisheye 67

2.23 Apartment building: group 1 prefers non-fisheye, group 2 prefers fisheye . . 67

3.1 (a)(d) Motion induced lighting variation (b)(e) Bad environmental lighting (c)(f) Combination of both 69

3.2 Illumination enhancement block diagram 74

3.3 (a) $p = 1000, \beta = 20$ (b) $p = 10, \beta = 50$ (c) $p = 0.1, \beta = 200$ 76

3.4 (a) original image (b) under-exposed map M_u (c) over-exposed map M_o . . 77

3.5 (a) original image (b) $\beta = 2$ (c) $\beta = 4$ (d) $\beta = 8$ (e) $\beta = 12$ (f) $\beta = 16$ (g) $\beta = 20$ (h) $\beta = 24$ 83

3.6 (a) $LOM = 0.07$ (b) $LOM = 0.10$ (c) $LOM = 0.13$ 84

3.7 Test images: (1) $P_u = 0.35$ (2) $P_u = 0.57$ (3) $P_u = 0.58$ (4) $P_u = 0.76$ (5) $P_u = 0.76$ (6) $P_u = 0.82$ 85

3.8 Example enhanced images: (a) LDR (b) CVC (c) WAHE (d) SRIE (e) LLCRM (f) ours 86

3.9 9-level enhanced images: subjective quality with (a) LOM (b) SMO (c) LOE87

Figure Page

3.10 Video enhancement example: left frames are original, right frames are enhanced. 88

4.1 Green: the window of visibility (u_0, w_0) boundary. Red: spatio-temporal content of $\boldsymbol{u}$ in which the solid line is visible, and the dashed line is invisible 93

4.2 Comparison between Blur profile B_0, B_0' and visibility profile $1 - P_0$ . . . 100

4.3 Subjective scores (0: best quality in each test set) 103

4.4 Comparison between $\boldsymbol{V}$ and estimated $\lambda(\hat{\boldsymbol{V}})$ 104

4.5 Function $\lambda(\cdot)$ in Equation 4.1: x-axis is measured visibility V_i, y-axis is $\lambda(V_i)$. 104

4.6 Diagram of creating test set: Λ_u and Λ_s 108

4.7 (a) Campus (b) Grocery (c) Apartments (blur level j refers to V_s^j). The vertical bar is the corresponding confidence interval. 113

4.8 Objective scores versus subjective scores (circle points: stable videos Λ_s, triangle points: shaky videos Λ_u, black line: fitting curve) 115

ABBREVIATIONS

BRISQUE	Blind Image Spatial Quality Evaluator
CSIQ	Categorical Subjective Image Quality
CI	Confidence Interval
DCT	Discrete Cosine Transform
FMA	Feature Matching Area
FPV	First-Person Video
FR	Full Reference
GSM	Gradient Magnitude Similarity
HVS	Human Visual System
IQE	Image Quality Estimator
KROCC	Kendall Rank-order Correlation Coefficient
LMV	Large Motion Video
LOM	Lightness Order Measure
LVI	Local Visual Information
MOS	Mean Opinion Score
MR	Mutual Reference
MRFQAFPV	Mutual Reference Frame Quality Assessment of FPVs
NFP	Number of Feature Matching Points
NIQE	Naturalness Image Quality Evaluator
NR	No Reference
NSS	Natural Scene Statistics
QE	Quality Estimator
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal to Noise Ratio

RR	Reduced Reference
SROCC	Spearman Correlation Coefficients
SR-SIM	Spectral Residual based Similarity
SSIM	Structural Similarity Index
TID	Tampere Image Quality Database
VIF	Visual Information Fidelity
VQE	Video Quality Estimator
VSNR	Visual Signal-to-Noise Ratio
VTP	Visibility-inspired Temporal Pooling

ABSTRACT

Bai, Chen Ph.D, Purdue University, August 2019. Image and Video Quality Assessment with Applications in First-Person Videos. Major Professor: Amy R. Reibman.

First-person videos (FPVs) captured by wearable cameras provide a huge amount of visual data. FPVs have different characteristics compared to broadcast videos and mobile videos. The video quality of FPVs are influenced by motion blur, tilt, rolling shutter and exposure distortions. In this work, we design image and video assessment methods applicable for FPVs.

Our video quality assessment mainly focuses on three quality problems. The first problem is the video frame artifacts including motion blur, tilt, rolling shutter, that are caused by the heavy and unstructured motion in FPVs. The second problem is the exposure distortions. Videos suffer from exposure distortions when the camera sensor is not exposed to the proper amount of light, which often caused by bad environmental lighting or capture angles. The third problem is the increased blurriness after video stabilization. The stabilized video is perceptually more blurry than its original because the masking effect of motion is no longer present.

To evaluate video frame artifacts, we introduce a new strategy for image quality estimation, called mutual reference (MR), which uses the information provided by overlapping content to estimate the image quality. The MR strategy is applied to FPVs by partitioning temporally nearby frames with similar content into sets, and estimating their visual quality using their mutual information. We propose one MR quality estimator, Local Visual Information (LVI), that estimates the relative quality between two images which overlap.

To alleviate exposure distortions, we propose a controllable illumination enhancement method that adjusts the amount of enhancement with a single knob. The knob

can be controlled by our proposed over-enhancement measure, Lightness Order Measure (LOM). Since the visual quality is an inverted U-shape function of the amount of enhancement, our design is to control the amount of enhancement so that the image is enhanced to the peak visual quality.

To estimate the increased blurriness after stabilization, we propose a visibility-inspired temporal pooling (VTP) mechanism. VTP mechanism models the motion masking effect on perceived video blurriness as the influence of the visibility of a frame on the temporal pooling weight of the frame quality score. The measure for visibility is estimated as the proportion of spatial details that is visible for human observers.

1. INTRODUCTION

The measurement of image and video quality plays an important role in the process of image and video capture, pre-processing, compression, transmission, post-processing and displaying. Currently, new multimedia has emerged that one typical example is First-Person videos (FPVs) that are captured by wearable cameras. First-person videos (FPVs) are becoming a widely spread type of videos that can document activities, share experiences and record trips. Numerous applications of FPVs have emerged using object tracking, activity recognition, video summarization and retrieval [1].

Quality assessment of FPVs is important because of three reasons. 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 [2]. Third, the visual quality of frames is a considerable factor for key frames or snap points detection [3], and can be incorporated into frameworks for video summarization [4].

1.1 First-Person Video Quality Assessment

A First-person video itself has unique First-Person characteristics that is different from any other videos or simulating environments, such virtual reality. First, it provides an unconstrained egocentric perspective. The camera angle of view is not restricted to specific location or direction, and sometimes faces against meaningless or temporally unrelated scenes. One example is that the camera wearer makes an unpredictable motion due to distracted events. Second, it often contains violent First-Person motion. When we move our head or body, our brain has a compensation mechanism that cancel out most self-motion influences. However, FPVs preserve the self-motion that the wearer are not fully aware of.

First-Person videos 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 [5]. 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.

From the comparisons with other types of videos above, we see that FPVs are faced with more severe quality problems. There exist three questions that we want to ask about FPVs: (1) what are the quality problems in FPVs? (2) what existing methods or new proposed methods can we use to measure the quality of FPVs? (3) How to quantify the newly generated artifacts if we improve FPVs?

First-Person Video Quality Problems

The quality problems in FPVs can be classified into two types, motion-induced distortions and non-motion distortions. Motion-induced distortions come from camera motions that record head or body movement of the camera wearer [6, 7]. The motion-induced distortions of frames in FPVs can be mainly classified as blur and

the geometric distortions of rolling shutter artifacts and tilt. *Blur* could be caused by any camera movement, and arises when motion is sufficiently large during the exposure period [8]. See Figure 1.2(a) for an example. *Rolling shutter artifacts* mainly arise from camera panning and 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 [9]. Figure 1.1 demonstrates the impact of rolling shutter. The arrows indicate the direction of camera motion. Solid lines surround the captured image in a camera. Dashed lines indicate the corresponding area in the real scene for that captured image. Motion in (a) and (c) contribute to skew distortions, corresponding to shear in geometric transformation. Motion in (b) and (d) result in vertical scaling, corresponding to the scaling difference between horizontal and vertical direction. Finally, *tilt* 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. See Figure 1.2(b) for example images. In addition, camera motion introduces visually induced motion sickness (VIMS) which occurs when there exists a sensory conflict. Since viewers is shown fast visual motions while the actual body is static so that their visual and vestibular information differ from the normal situations when they walk or run. The VIMS causes dizziness so that the motion stabilization for FPVs is often necessary.

Another type of distortions in FPVs are non-motion distortions. *Exposure distortion* [10] is introduced by the motion and captured environments. Since the wearer is unaware of adjusting lighting direction for the camera, the captured video is often badly exposed. FPVs always suffer from exposure distortions. Since the wearers are often not fully aware of the lighting conditions during capture without real-time feedback, they have no intention to adjust the camera direction or location to the best illumination condition. In addition, FPVs are recorded with random motion so that the lighting condition changes violently. Therefore, there exists a large number of frames that are badly exposed with spatially inconsistent exposure distortions. The

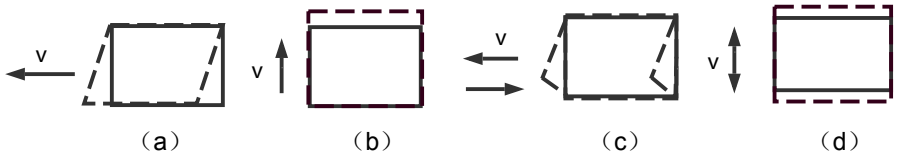


Fig. 1.1.: (a) horizontal camera panning (b) vertical camera panning (c) horizontal camera shaking (d) vertical camera shaking

example images are shown in Figure 1.2(e)(f). *Fisheye* is an internal property of many wearable cameras with ultra wide-angle lens (i.e. Gopro, Looxie Camera, Mobius). Instead of capturing a rectilinear image, the content appears to be convex. Fisheye is one of the lens distortions, called barrel transformation. The transformation warps the image to be bent; the magnification decreases from center to margins [11]. See Figure 1.2(d) for example.

Available Objective Quality Assessment Methods

To evaluate the quality of FPVs, it is typical to apply quality estimators (QEs). Existing QEs are normally classified into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods.

FR methods assumes the existence of a pristine image or video as the reference. FR methods interprets the image or video quality as the difference or the similarity compared to the original. The simplest computational method is to compute pixel-wise differences between the original image or video and the test image or video. The typical method in FR image quality is a two-stage strategy that first compute a local distortion or quality map by comparing test image with the reference image using measures of similarity or difference. The local quality map is then spatially pooled to an image score. The typical method of FR video quality metric is built on the image quality metrics that by combining image-level scores into a video-level scores.

(a) (b)

(c) (d)

(e) (f)

Fig. 1.2.: (a) motion blur (b) rolling shutter artifacts (c) tilt (d) fisheye (e)(f) exposure distortions

RR methods relax the constraints of FR methods by only comparing partial information between the reference and the test. A minimal set of extracted features from the original image or video is used as side information for evaluating the quality of the test image or video. RR still needs the existence of a pristine reference.

One limitation for most FR and RR QEs [12–14] is that they cannot evaluate a test image that is better than its reference image. Two exceptions are Visual Information Fidelity (VIF) [15] and Visual Distortion Gauge [16]. 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 assumes no other information except the image or video to be evaluated exists. It interprets the image or video quality as the perception of human observers, since human observers do not use any reference to evaluate the quality of an image or video. NR often uses implicit knowledge of the criteria for human to quantify how "good" is an image or video. One subset of NR QEs is blur metrics [17, 18]; another subset is natural scene statistics based QEs [19–21]. However, most existing NR methods are content dependent so that it provides consistent measure to compare their quality scores mostly in cases that the two images or videos have almost the same content.

Existing QE structures as discussed above can be summarized in Figure 1.3. Specifically, FR QEs are used when the "pristine" reference image is available, RR are used when only information from the "pristine" reference image is available, NR are used when no reference information is used.

However, QE structure can be generalized as Figure 1.4 to be available for more quality assessment scenarios. The distorted image and a collection of "similar enough" images are the inputs, the outputs are relative QE scores and QE confidence. A collection of "similar enough" images can provide each other with effective information for quality assessment. "Similar enough" can be interpreted as a group of images that share common content. One example is a group of images captured from nearby locations. In the generalized QE structure, the reference image is replaced with "similar enough" images that do not need to be unimpaired and pixel-aligned. The output QE score can provide relative scores that do not constrain the upper bound of quality scores. Another output, QE confidence, can help to avoid acting on an inaccurate measure, once you know the weaknesses of the QE. For example, a QE

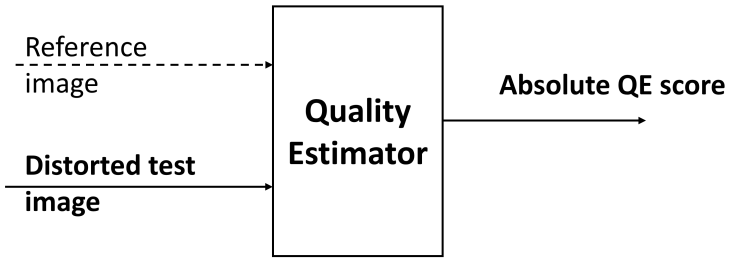


Fig. 1.3.: Existing QE structure

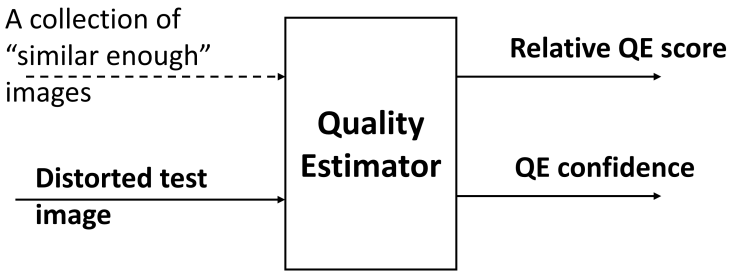


Fig. 1.4.: Generalized QE structure

may be known to be more accurate for one type of distortion, compression, than another, blur.

New Quality Problems after Enhancement

Because of the low-quality nature of FPVs, the enhancement of their video quality is necessary to provide a good viewing experience for human observer. However, improving the quality of FPVs needs a quality monitoring mechanism to guarantee the enhancement process will not destroy the quality by introducing new visible artifacts. Here, two enhancement quality issues, illumination enhancement and motion stabilization, will be discussed.

Illumination enhancement often introduces contrast distortions, new exposure distortions and newly-generated artifacts into the image [22], the quality monitoring during enhancement is necessary to avoid introducing new quality problems. Typical

artifacts after enhancement include loss of edges, textures [23] or unnaturalness [24]. Illumination enhancement and image visual quality has a concave relationship. Assume we have a knob to control the amount of enhancement. If we continue to increase the amount of enhancement starting from zero, the visual quality increases. Until a certain amount of enhancement, the best achievable quality for the image is reached. If we still continue to increase the amount of enhancement, the visual quality will decrease due to newly-generated artifacts. To avoid that the artifacts generated during the enhancement process destroy the visual quality, a possible solution is to design a metric to monitor the visual quality during enhancement. Then we can adjust the knob to control the enhancement until the image reaches the best achievable quality.

Another issue is the increased blurriness after video stabilization. Since FPVs are quite unstable, stabilization is often necessary for FPVs to them watchable. However, motion stability is only one aspect of the stabilized video quality, the perceptual blurriness should also be considered. The stabilized video is perceptually more blurry than its original for two reasons. First, the stabilization process applies geometric transformations into frames that introduce spatially-varying blur. The second more important reason is that the motion masking effect on blurriness is reduced when motion decreases. The amount of perceive blurriness becomes smaller as motion increases has been studied and explained in [25–27]. Therefore, the quality assessment of the increased visual blurriness is an important evaluation for a stabilized FPV.

1.2 Validation of Proposed Quality Assessment Methods

Subjective image or video assessment is the common gauge for measuring video quality. The gathered subjective quality data can be used to validate newly proposed QEs [28–30] or help design new QEs [7, 31].

Subjective quality assessment is defined to measure the image or video quality through the observation from human observer since human visual system (HVS) is the receiver for any visual stimuli. To gather the subjective assessment data, the

design and implementation of subjective test is critical. The test generally requires the human observer to compare or rate the perceived quality of images or videos according to specific criteria.

The validation of proposed image and video QEs uses systematically designed subjective test to gather subjective quality assessment data. The design of a subjective test has four steps:

1. Establish the test goal:

   The subjective data can be successfully gathered and used for validation of effectiveness of QEs in its design scenario.

2. Create test images or videos:

   The creation process follows two design principles: (1) human observers should be able to perceive the differences between images or videos, (2) extraneous quality factors that are not assessed should be equal. For example, assuming video blurriness is to be assessed, the test videos should avoid content with exposure, color or compression artifacts. Then, to obtain the test videos, we can either intentionally record or synthetically create the video based on the type of artifact to be evaluated, and then we can maintain those extraneous factors to have imperceivable differences. Therefore, to validate a video blurriness metric, test videos should have few motion differences to avoid the case that the observer judges the blurriness from motion information.

   Most existing subjective tests [32–34] explored quality degradations starting from a reference that is considered to be distortion-free. By synthetically adding distortions to the reference with different levels of severity, a series of distorted images or videos of the same content but different amounts of degradations are created. Since subjective data if distortions existed in FPVs have not been specifically gathered in existing image and video subjective quality databases, we should design methods to synthetically create these distortions and gather desired subjective data for the validation of any proposed FPV quality assessment method.

The synthetic method of quality degradations in FPVs are not well developed. Motion blur, tilt, rolling shutter, exposure distortions are not often considered in existing image or video datasets. Besides, the motion differences between FPVs and existing video datasets of broadcast videos are very large. The first step to gather subjective data for FPV cases is to define the distortion type and design the synthesis method to independently add the type of distortion.

FPVs contain geometric distortions and large motion that sometimes we are not able independently add one type of distortions while keep all other influencing factors to be zero. The construction of different amounts of geometric distortions is to keep the same content in the image center to maintain a consistent focus of attention. For motion influence, we need to systematically design the test that avoid the comparison between motion while our goal is compare other distortions.

3. Gather subjective data:

   The standard methods and procedures to gather subjective data are described in ITU recommendations [35, 36]. Several representative methods for subjective test are described:

   (a) Single stimulus method: One test stimulus is shown to the observer each time. The observer needs to rate the perceived quality of the test stimulus based on the provided criteria which includes the attributes of the test stimulus to be considered and the scale type (continuous or categorical).

   (b) Double stimulus method: A reference stimulus and test stimulus are shown to the observer at the same time or in sequence. The observer needs to rate the test stimulus by comparing with the reference stimulus based on provided criteria. The reference is typically considered to be the maximum quality in the provided scale.

   (c) Paired comparison method: A pair of test stimulus are shown to the observer simultaneously or one after another. The observer needs to judge which stimulus have better quality, sometimes a tie option is also provided.

Note that Paired comparison method is our priority, since the data it gathered avoids the internal variability between human observers when using absolute rating values and the comparison is an easier question so that the results are more reliable.

4. Subjective Data analysis: The gathered subjective data needs to be transformed into subjective quality, and then can be used to evaluate the performance of QEs.

   The typical methods to for data analysis are as follows:

   (a) Single Stimulus Method and Double Stimulus Method: the mean of the gathered subjective image quality ratings from human observers are used as subjective quality scores, called Mean Opinion Scores (MOS) or Differential Mean Opinion Scores (DMOS). The DMOS refers to the differential subjective ratings, computed as the difference between the reference stimulus and the test stimulus. To deal with the potential variability between the absolute values of human observers' ratings, the typical way is to normalize the scales across ratings from different observers by applying a transform under the assumption that observers share the same mean and standard deviation during their evaluation.

   (b) Paired comparison method: The results of a paired comparison test is a winning frequency matrix representing the frequencies that which is preferred against another. The typical models to transform winning frequency matrix into continuous quality scale are Thrustone and Bradley-Terry models [37].

In this thesis, we design target-specific subjective tests to evaluate individual distortions including motion blur, tilt, rolling shutter artifacts, over-enhancement distortions in images or videos. By gathering the target subjective data, our proposed image or video quality estimators are demonstrated their effectiveness in their application scenarios.

1.3 Contributions and Outline

In order to solve the image and video quality problems discussed in Section 1.1, we separately design three quality assessment framework to deal with the following three quality problems.

1. To evaluate video frame artifacts, we introduce a new strategy for image quality estimation, called mutual reference (MR), which uses the information provided by overlapping content to estimate the image quality. The MR strategy is applied to FPVs by partitioning temporally nearby frames with similar content into sets, and estimating their visual quality using their mutual information. 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. LVI is designed with several properties. 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 propose a frame-quality assessment framework and demonstrate the framework is very effective to estimate the quality of individual frames with similar content in FPVs. In addition, we also compare the statistics of distortions between FPVs and traditional videos, and implement a systematic subjective test to study geometric distortions existed in frames of FPVs.

2. To alleviate exposure distortions, we propose a controllable illumination enhancement method that adjusts the amount of enhancement with a single parameter. Our single parameter has a concave relationship with image quality. In our method, we model under-exposure and over-exposure differently to assign under-exposed and over-exposed probabilities for each pixel. We then design a system that applies logarithmic mapping in the identified under-exposed pixels with boundary-artifact compensation. Our mapping uses the assigned under-

exposed probabilities, the artifact compensation weights and the single adjustment parameter together to calculate mapping coefficients. We also propose an over-enhancement measure, Lightness Order Measure (LOM) to quantify the unnaturalness in the enhanced image. We consider the unnaturalness to be related to the inversion of relative lightness order between neighboring pixels, and which is influenced by both the proportion of inversions and the inversion magnitude. Since the visual quality is an inverted U-shape function of the amount of enhancement, our design is to control the amount of enhancement so that the image is enhanced to the peak visual quality.

3. To estimate the increased blurriness after stabilization, we propose a visibility-inspired temporal pooling (VTP) mechanism. The mechanism uses weighted average pooling strategy to combine frame quality scores to a video quality score, in which the pooling weight is computed as a function of visibility. We propose a visibility measure that estimates the perceived content under any magnitude of motion based on the window of visibility [25, 38]. The function that transforms the estimated visibility into the pooling weight for each frame is measured by a systematically designed subjective test that uses videos with temporally shifted blur but temporally similar visibility. The VTP mechanism can be effectively applied to measure the relative perceived blurriness between the stabilized video and its original version. In the validation experiments, we design a synthesis method for shaky videos that allows a controllable motion being injected into the test videos.

The rest of this thesis is organized as follows. Chapter 2 discusses the frame quality assessment solutions for FPVs including QE design, application framework in FPVs and subjective test of blur and geometric distortions. Chapter 3 presents the controllable illumination enhancement framework with an over-enhancement measure. Chapter 4 describes the visibility-inspired temporal pooling mechanism and two validation subjective test, one for the mechanism, another for the application in

video stabilization. Chapter 5 summarizes the work in this thesis and discusses future works.

2. IMAGE QUALITY ASSESSMENT FOR FIRST-PERSON VIDEOS

2.1 Introduction

First-person videos (FPVs) captured by wearable cameras are becoming a widely spread type of videos that can share experiences and document activities without length limitation and specific structure. Numerous applications of FPVs have emerged using object tracking, activity recognition, video summarization and retrieval [1]. Recently, research topics related to the viewing experience and the visual quality of FPVs have been also proposed, involving First-Person motion measuring and improving [39] and visual quality assessment [6, 7].

Because of the capture process of FPVs, quality degradations are not limited to transmission or post-processing, and the resulting distortions in frames have not been subjectively evaluated. Motion blur and geometric distortions are two major distortions in FPVs [6]. Motion blur mainly arises from fast motion of the camera. While the camera keeps changing its positions in the scan time of one frame, the scene captured is blurred. Geometric distortions can be classified into 3 categories: rotation, shear and fisheye. Rotation results from head or body rotation. Wearers regularly move their bodies and shake their heads, and rarely are aware whether or not the camera is kept horizontal while recording videos. Shear is caused by camera panning. When the camera changes its positions in the scan time of one frame, the top rows of the frame are not vertically align with the bottom rows. For example, architecture in a sheared image is visually skewed. Fisheye images are captured by wearable cameras with ultra wide-angle lens (i.e. Gopro, Looxie Camera, Mobius). Instead of capturing a rectilinear image, the content appears to be convex. Fisheye

is one of the lens distortions, called barrel transformation. The transformation warps the image to be bent; the magnification decreases from center to margins [11].

To evaluate the quality of individual frames in FPVs, 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 [12, 13, 40, 41] 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 [42]. However, most existing NR methods are content dependent [19–21, 43]. As a result, it is often difficult to interpret the output of a NR method [44]. For example, setting a quality threshold in a system is challenging; all five NR QEs considered in [44] 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 [7].

In this work, we propose a new strategy of quality estimation, called mutual reference (MR) [45, 46], 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.

Section 2.2 describes prior works in FR QEs and NR QEs, and discusses related works about subjective test. Section 2.3 presents a detailed description of the strat-

egy for MR. Our proposed MR QE, LVI, is described with its basic principle and reliability check in 2.4. Our MRFQAFPV is described in Section 2.5. The framework has three steps: temporal partitioning, reference search and quality estimation. In Section 2.6, 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 2.7 propose a distortion classification method to classify frames in FPVs, and compare the distortion statistics between FPVs and traditional videos. Section 2.8 implements a systematic subjective test for motion blur, tilt (rotation), rolling shutter artifacts and fisheye. LVI and existing NR QEs can be generalized to measure images with both blur and geometric distortions. Section 2.9 summarizes this paper and discusses future work.

2.2 Related Works

2.2.1 Image Quality Estimator

Existing image quality estimators (QE) can be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. FR QEs and RR QEs estimate a distorted image using its corresponding "pristine" reference image. "Pristine" reference image is considered to be an unimpaired source image, and the distorted image is pixel aligned with the reference. In comparison, FR QEs use the whole reference image itself, while RR QEs use some supporting information from the reference image. NR QEs evaluate the distorted image without relying on any reference.

FR QEs and RR QEs have both the distortion image and the reference image to be the inputs. These QEs often measure how the distorted image is similar to the original image. The distorted image is considered to have better quality if it is more similar to its reference. The Mean Square Error (MSE) is the most widely used mathematical tool. By computing the value differences between every corresponding

pixel from both images, the MSE can clearly show the pixel differences. Peak Signal to Noise Ratio (PSNR) is derived from MSE value and PSNR is a commonly used FR QE. However, MSE and PSNR do not accurately predict perceived image quality, since they have not considered human visual system (HVS) [47].

Existing FR QEs and RR QEs can be categorized by whether they apply models of the human visual system, image structure, or image statistics [48]. Two common QEs are the Structural Similarity Index (SSIM) [12], which is based on structure, and Visual Information Fidelity (VIF) [15], which is based on statistics. SSIM computes means and variances of each image, applies a similarity measure to each,

$$S(x, y) = \frac{2f_x f_y}{f_x^2 + f_y^2}, \tag{2.1}$$

and combines these with a correlation term to quantify distortions in the luminance and contrast. In Equation (2.1), 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) [13], Gradient Magnitude Similarity (GSM) [49] and Spectral Residual based Similarity (SR-SIM) [50] employ the same similarity measure in Equation (2.1) 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 [14] and MAD [33], 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.

NR QEs use the distorted image itself as the only input. One specific subset of NR QEs are NR blur metrics, which were summarized in [42, 51]. One uses the histogram of DCT coefficients [52]. Edge-based blur QEs have also been proposed and comprise the majority of blur QEs: [53, 54], JNBM [51], CPBD [18]. Non-edge blur metrics using the discrimination between re-blurred versions of an image [17, 55] and local phase coherence [56] 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 [19], NIQE [20], and IL-NIQE [21] 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 [44].

In [44], 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 histograms of the quality scores for undistorted and badly degraded images using BRISQUE, NIQE and IL-NIQE. In addition, our results in Section 2.6 demonstrate that the state-of-the-art NR QEs are source-dependent, and our proposed method in Section 2.3 significantly reduces the source dependency when estimating the quality of First-Person images.

Geometric distortions have not been received much attention, since traditional videos rarely have issues regarding a large amount of tilt or rolling shutter artifacts. Existing metrics that consider the influence of geometric distortions mainly employ two approaches. The first approach is based on the measure of displacement field [57,58]. An improvement of this approach was proposed in [59] by considering human visual properties. Another approach can be described as the modified versions of SSIM that is robust to minor geometric changes [60,61]. One example is the transformation-aware metric [62]. It uses homography estimation to add the influence of geometric transformation to SSIM.

2.2.2 Subjective Test

Most existing subjective tests [32–34] explored quality degradations starting from an undistorted image, which is considered to be the reference image. By synthetically adding distortions (i.e. Gaussian blur, JPEG, JPEG 2000, noise) to the reference image with different degradation levels, a series of distorted images of the same content but different severity is created. To design a subjective test for FPVs, however, applying each type of distortions separately fails to consider two issues:

1. Actual images captured by a camera may be subtly different than those created synthetically using a model [63]. In our test, images with real, synthetic or real plus synthetic distortions are evaluated. The real images are extracted from frames in FPVs.

2. Many blurry images are subject to geometric distortions (i.e. rotation, shear or fisheye) simultaneously in FPVs. The question is what the visual impact is on the overall quality when one image has multiple distortions. Multiply-distorted images have been evaluated in the subjective test, but only for blur, JPEG and noise [64]. However, these distortions all have pixel-to-pixel correspondence, whereas geometric distortions do not. Therefore, when constructing images for the subjective test for different amounts of geometric distortions, we keep the same content in the image center to maintain a consistent focus of attention.

2.3 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, [65] considers the quality estimation

for downsampled images, while [66] considers the 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 3.

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 [67]. To evaluate the quality of a fused image, all source images are used as references [68, 69]. 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 [67–69], and including the quality metric for exposure fusion techniques [70]. The second application is to assess images captured either from, or of, nearby locations. For example, in this chapter, 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 [71], which implemented a subjective test using images captured of the same scene by either different cameras or the same camera with different settings.

2.4 Local Visual Information

In this section, we describe our proposed MR QE, Local Visual Information (LVI) [6], which primarily measures relative blur between two images.

2.4.1 Basic Principle

LVI is derived from the approach of VIF [15]. 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) [72] 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.

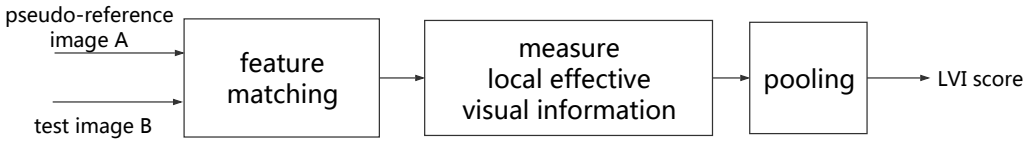


Fig. 2.1.: Block diagram of local visual information (LVI) quality estimator

The quality measure LVI has three procedures, shown in Fig. 2.1. 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

$$A_{lp} = S_{lp}^{A} \cdot U_{lp}^{A} \tag{2.2}$$

$$B_{lp} = S_{lp}^{B} \cdot U_{lp}^{B} \tag{2.3}$$

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

$$C_{lp} = A_{lp} + \mathcal{X} \tag{2.4}$$

$$D_{lp} = B_{lp} + \mathcal{X}' \tag{2.5}$$

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

$$I(C_{lp}; A_{lp}|S_{lp}^A) = \frac{1}{2}\sum_m \log_2(1 + \frac{(s_{lp}^A)^2\lambda_m^A}{\sigma_x^2}) \tag{2.6}$$

$$I(D_{lp}; B_{lp}|S_{lp}^B) = \frac{1}{2}\sum_n \log_2(1 + \frac{(s_{lp}^B)^2\lambda_n^B}{\sigma_x^2}) \tag{2.7}$$

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.

$$Q_{LVI} = \frac{\sum_l \sum_p I(D_{lp}; B_{lp}|S = S_{lp}^B)}{\sum_l \sum_p I(C_{lp}; A_{lp}|S = S_{lp}^A)} \tag{2.8}$$

The output score of Equation (2.8) 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.

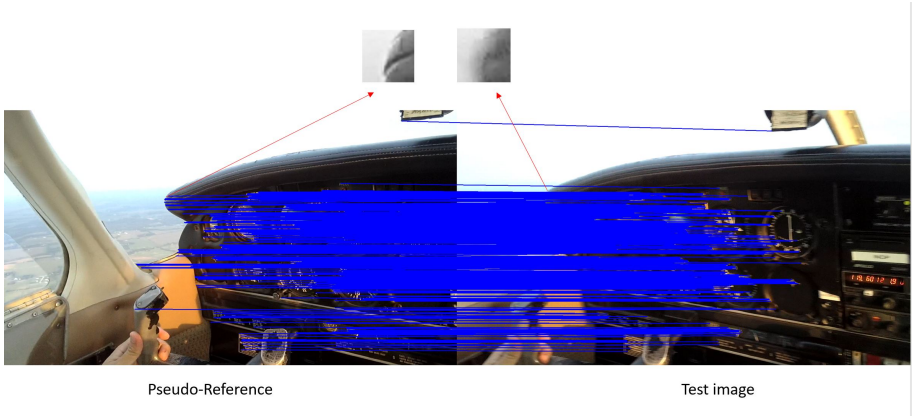


Fig. 2.2.: Left: Pseudo-reference. Right: Test image. LVI score = 0.771

Figure 2.2 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.

2.4.2 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 [6]. 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 [62,73] using matching feature points. Specifically, we apply point-based homography [73] using the result of the feature matching step in Fig. 2.1. 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

$$M_H = M_A M_P = \begin{bmatrix} u_a & u_b & u_c \\ v_a & v_b & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ w_a & w_b & 1 \end{bmatrix}, \tag{2.9}$$

where w_a 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 w_a 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

$$M_A = M_s M_r M_k M_t = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & k_s & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.10}$$

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

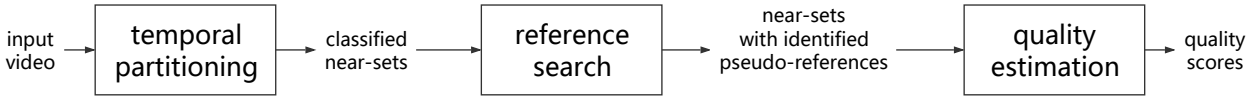


Fig. 2.3.: Framework of quality assessment for First Person Video.

t_y are translation distances in horizontal and vertical directions, respectively. Using the parameters estimated from M_A, we can calculate M_s as

$$M_s = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{u_a v_b - u_b v_a}{\sqrt{v_a^2 + v_b^2}} & 0 & 0 \\ 0 & \sqrt{v_a^2 + v_b^2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.11}$$

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 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 2.5, we will describe how LVI can be incorporated into a quality assessment framework for FPVs using the strategy of mutual reference.

2.5 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. 2.3 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. 2.3 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 ≤ 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**

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 2.4.2 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

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 ≤ 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**

the number of matching points T is set to be 50. If we increase T, it will introduce more uncategorized frames. If we 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. 2.3 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. 2.3 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 $k^{(n)}$.

2.6 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 Figure 2.3, and two feature detectors for the first step of LVI shown in Figure 2.1. 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 [33, 34, 74], we demonstrate that LVI is also effective to assess the quality for some distortions that are not typically present in FPVs.

2.6.1 Implementation Design Comparisons

In this section, we explore the performance of several design options for both LVI 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 Figure2.3. Section 2.5 presents two methods, NFP and FMA. In addition, the scale check detailed in 2.4.2 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 [75], such that each partitioned interval has a cumulative displacement of 10% of a frame width. Note that the shot boundary detection method [76] 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.

Figure 2.7 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 Figure 2.7(a) and Figure 2.7(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 Figure 2.7(c) and Figure 2.7(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 Figure 2.4, and frames in two partitioned near-sets are shown in Figure 2.6. The test dataset is available at [77].

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 use the FMA+affine, the best among the six methods, as our temporal partitioning method in MRFQAFPV in the following sections.

Feature detector: Next, we explore the performance of LVI using two different feature detectors for step 1 of Figure 2.1. Specifically, we compare the quality scores of LVI using SIFT [78] (SIFT-LVI) and using ORB [79] (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 2.5 by incorporating either SIFT and ORB as the feature matching detector. Figure Figure 2.8(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.

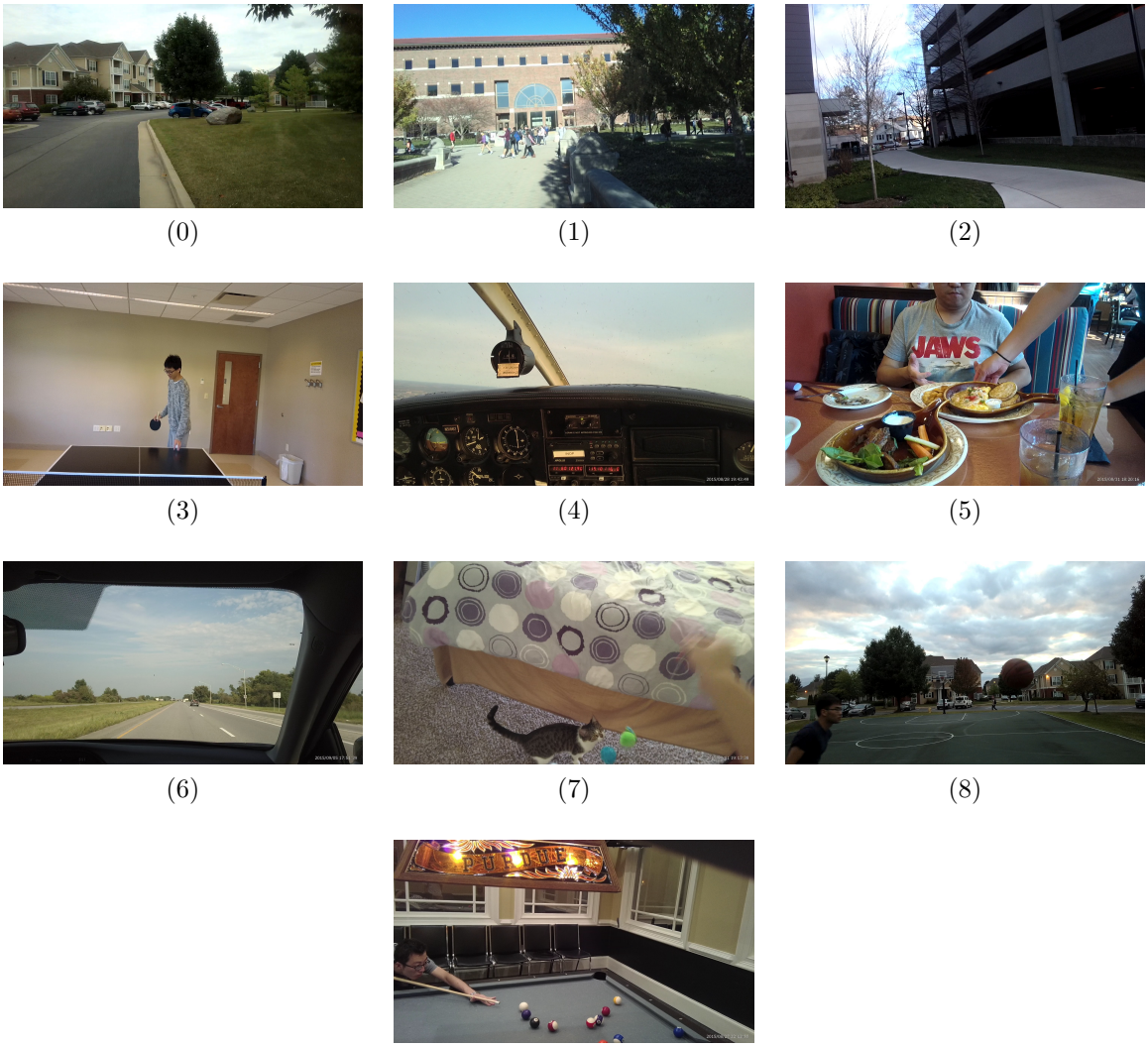


Fig. 2.4.: Sample test images: (0) basketball (1) run (2) walk (3) billiards (4) cat (5) eat (6) ping pong (7) talk (8) car (9) flight

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: LIVE image quality database [74], CSIQ [33], and TID2013 [34]. 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 computa-

Fig. 2.5.: Sample partitioned near-set 1

tionally much faster than SIFT [79]. Given the small performance differences between using SIFT and ORB, we choose ORB as a more computationally efficient feature detector in LVI and MRFQAFPV.

2.6.2 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.

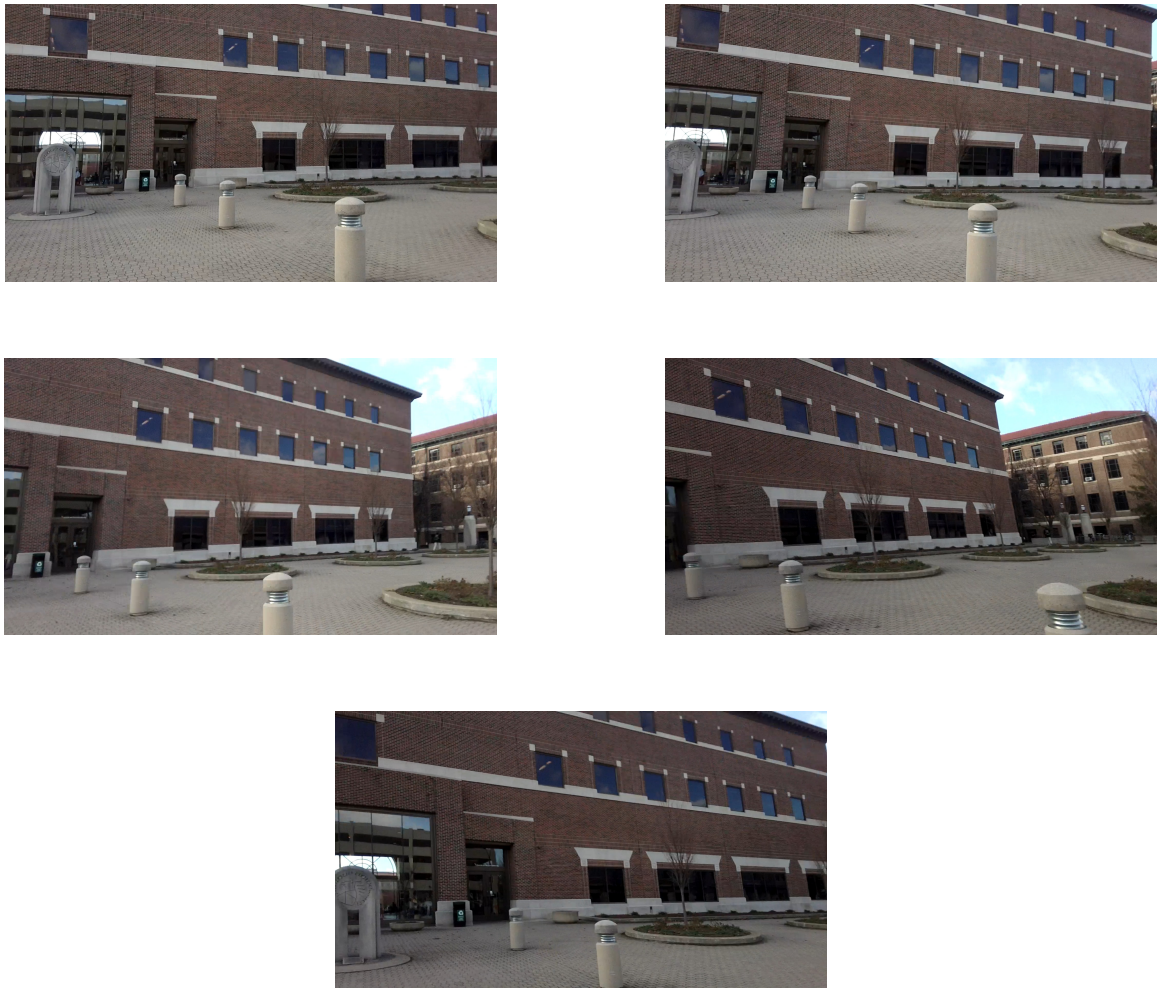


Fig. 2.6.: Sample partitioned near-set 2

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. Finally, we apply LVI to subjective data with distortions

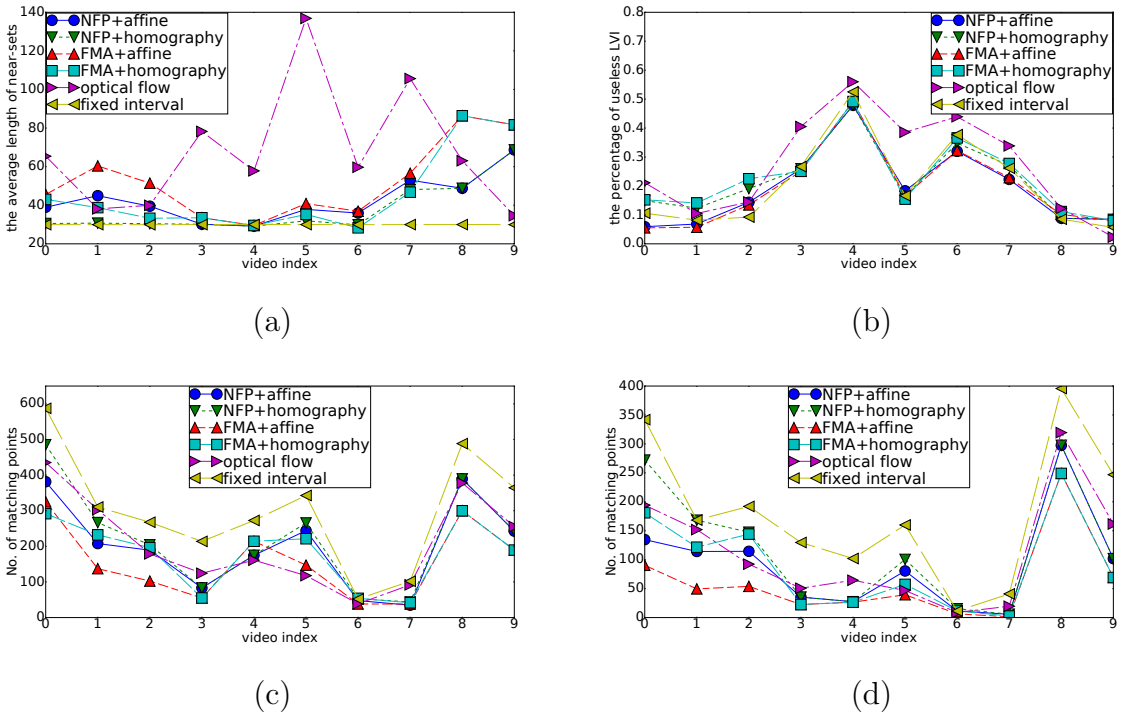
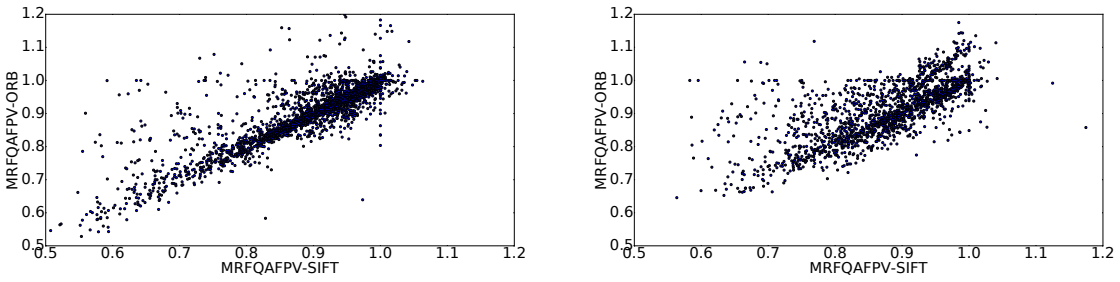


Fig. 2.7.: The performance of six temporal partitioning methods in 10 FPVs: (a) criteria 1: the average length of near-sets (b) criteria 2: the percentage of useless LVI (c) criteria 3: the average number of matching points between pseudo-references in temporally adjacent near-sets (d) criteria 3: the average number of matching points between start frames in temporally adjacent near-sets

other than those in FPVs [33, 34, 74], 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 [6]. 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



(a) (b)

Fig. 2.8.: The distribution of MRFQAFPV-SIFT versus MRFQAFPV-ORB: (a) outdoor content (b) indoor content

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 2.6.1, 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 $\mathcal{X}$ 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 $\mathcal{X}$. 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 [80], 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 [37]. We apply LVI and five NR QEs, NIQE [20], IL-NIQE [21], a perceptual blur metric (Blurriness) [55], JNBM [51] and CPBD [18] to all test images. Table 2.1 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,

Table 2.1.: 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)

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.

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 [74], CSIQ [33] and TID2013 [34]. 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 [12], VIF [15], FSIM [13], VSNR [14] and SR-SIM [50]. Table 2.2 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 2.2, 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

Table 2.2.: 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

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 [15], the results also show that VIF performs better in JPEG2000 than JPEG.

2.7 Video Statistics

In this section, we classify different distortions in FPVs. It separates distortion classification into blur measurement and geometric measurement. Blur measurement applies LVI to measure motion blur. Geometric measurement considers rolling shut-

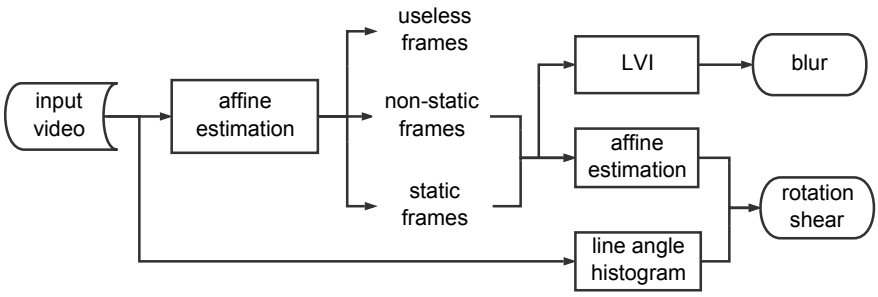


Fig. 2.9.: Blockdiagram of estimating video statistics

ter artifacts and rotation. Homography estimation method in Section 2.4.2 and line angle histogram [81] are two basic methods. The line angle histogram detects the rotation and shear, and the homography estimation measures geometric transformation parameters between two images. We then apply the distortion classification method to compare the differences between traditional videos and FPVs.

2.7.1 Distortion Classification

Our distortion classification method has three components: the LVI algorithm, the homography estimation and the line angle histogram. The overall framework is shown in figure 2.9. Both the LVI and homography estimation are based on feature matching between two images. They measure the geometric relationship between two nearly adjacent frames, which are separated by a small time interval. Affine estimation is used to approximate the homography estimation and its measuring method has been described in Section 2.4.2.

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

or meaningless content. These frames that have few edges or corners are classified to be useless frames.

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

The Line angle histogram [81] is used to detect shear and rotation. The line angle distributions of different images are shown in figure 2.10. Horizontal is at 90°, and vertical is at 0° and 180°. The peaks closest to horizontal and vertical are denoted the horizontal peak and the vertical peak, respectively. The deviation of the horizontal peak from 90° indicates the rotation during capture. The difference between the horizontal and vertical peaks should be close to 90°. When the two peaks deviate from orthogonality, the image are sheared.

2.7.2 Classification Results

We present statistics of distortions for the two types of videos in Table 2.3. We selected six traditional videos from LIVE Video Quality Database [82,83], and recorded six types of FPVs using the Pivothead. In Table 2.3, the "talking","ping pong" and "eating" videos are recorded indoors, while other three FPVs are recorded outside. The comparison indicates a few frames are subject to distortions in the LIVE database, while most frames in FPVs are distorted images. Our results demonstrate FPVs have dramatically different distortions immediately after capture compared to traditional videos.

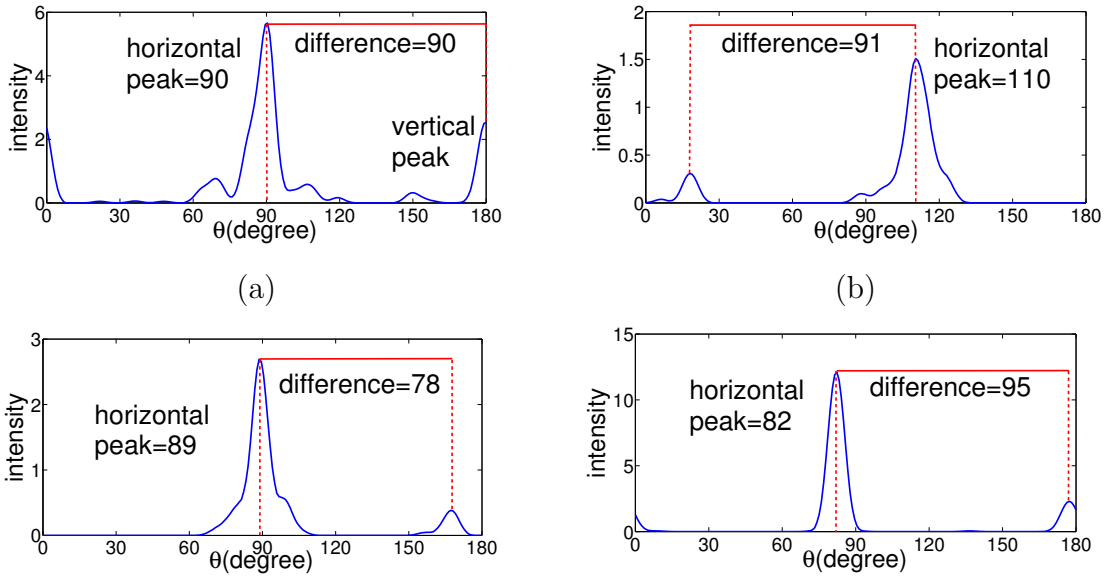


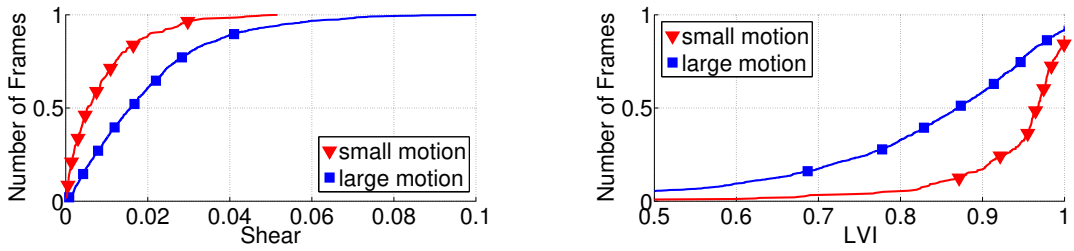
Fig. 2.10.: Line angle distributions: (a) image free from shear and rotation (b) image with rotation only (c) image with shear only (d) image with both rotation and shear

The six FPVs share common properties. First, all of them have more than 69% of frames with rotation, indicating that camera wearers keep their heads rotated most of the time. Second, the percentage of blurry images is in the range from 55% to 83%. Third, shear is less likely to exist in FPVs compared to rotation and blur. However, each FPV also shows some differences. The three indoor videos have more than 75% of their frames with blur, while the percentages of the other three outdoor videos are no more than 65%. So indoor videos have worse quality compared to outdoor videos.

Figure 2.11 shows two cumulative distributions of frames in the "running" video. We extract two groups of frames: small motion and large motion. To partition them, we use the translation parameter from the affine estimation, and declare those with translation greater than 50 to be large motion, and those with translation smaller than 10 to be small motion. In frames with small motion, 89% have shear change smaller than 0.02 and 68% have LVI greater than 0.95; but for frames with large motion, only 60% have shear smaller than 0.02 and 23% have LVI larger than 0.95.

Table 2.3.: Comparison between FPVs and traditional videos

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



(a) (b)

Fig. 2.11.: Cumulative distributions: (a) shear (b) LVI

We also applied our LVI algorithm on the image quality database TID2013 [34], for two distortions; Gaussian blur and contrast change. The Pearson correlation coefficients for Gaussian blur and contrast change are 0.9320 and 0.9018, respectively. The correlations of Gaussian blur for other image quality metrics, SSIM, FSIM [13] and VIF, are 0.9191, 0.8905 and 0.9530, and the correlations of contrast change are 0.6385, 0.6924 and 0.8730, respectively. This demonstrates that LVI is useful to measure more distortions than motion blur; and the performance of LVI can compete with other image quality metrics.

2.8 Subjective Test for Blur and Geometric Distortions

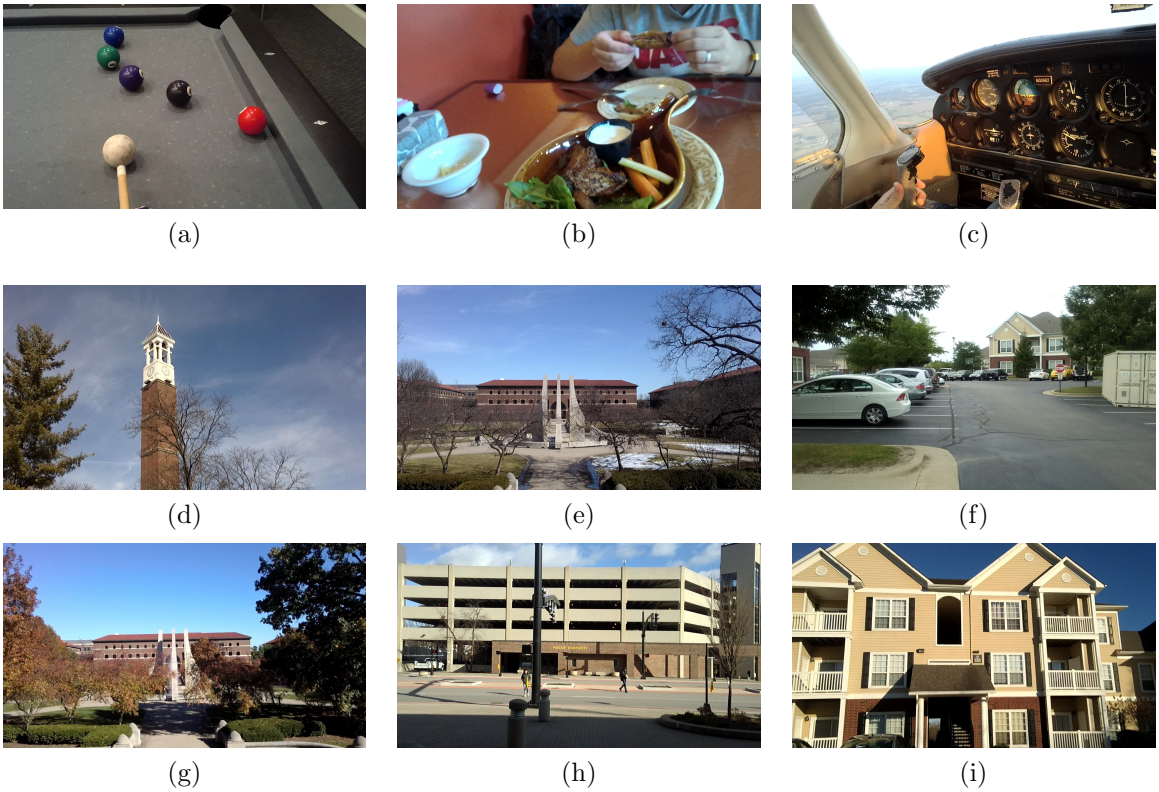
In this section, we implement a subjective test that demonstrates LVI and existing NR QEs can be generalized to measure images with both blur and geometric distortions including rotation and shear simultaneously.

We design a subjective test to evaluate motion blur and geometric distortions in FPVs. Our subjective test evaluates actual captured images with real distortions, synthetic distortions or a combination of both using the paired comparison method. The types of distortions include motion blur, rotation, shear and fisheye. We then propose two mapping functions for rotation and shear to compute the overall quality of images with blur and geometric distortions. Personal preferences and content dependence in fisheye are also discussed.

2.8.1 Test Overview

Our subjective test evaluates actual captured images with real distortions, synthetic distortions or a combination of both using the paired comparison method. The types of distortions include motion blur, rotation (tilt), shear and fisheye. In addition, we proposed two mapping functions for rotation and shear to compute the overall quality of images with blur and geometric distortions. Personal preferences and content dependence in fisheye are discussed.

Our test employs the paired comparison method for still images containing both actual and synthetic distortions that are typical of images extracted from FPVs. Each pair of test images are simultaneously displayed on two monitors side by side. Viewing distance is kept to be around 3 times the height of test images. 9 videos have been recorded by a Pivothead camera (frame rate: 30fps, resolution: 1920×1080) including "billiards", "eating", "flight", "bell tower","winter Hovde Hall", "parking lot", "autumn Hovde Hall", "apartment building" and "parking garage". 4 distortions including motion blur, rotation, shear and fisheye, are evaluated. Test images are either real frames from the 9 videos or created by adding synthetic distortions to



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 2.12.: Reference images for each content: (a) billiards (b) eating (c) flight (d) bell tower (e) winter Hovde Hall (f) parking lot (g) autumn Hovde Hall (h) apartment building (i) parking garage

selected frames. Because synthetic shear and fisheye change the image size compared to the original image, all test images are cropped to be 1600×900 to remove marginal regions with little content.

2.8.2 Test Method and Setup

In a paired comparison subjective test, the subject needs to indicate his or her preference among the two images according to their visual quality. 50 subjects including 43 males and 7 females participated in the test. All pairs of images are displayed in random order. To improve the efficiency of paired comparison, we use the "square design" in [84]. Given that we have n stimulus, a full comparison needs $0.5n(n-1)$

Table 2.4.: Test images in subjective test

video content	real distortions	synthetic distortions	number of images	SI
(a)	blur (5 levels)	-	5	6.58
(b)	blur (5 levels)	-	5	8.57
(c)	blur (5 levels)	-	5	9.40
(d)	rotation (4 angles)	blur (4 levels)	16	9.69
(e)	rotation (4 angles)	blur (4 levels)	16	15.75
(f)	-	blur (4 levels) + shear (4 levels)	16	13.30
(g)	-	blur (4 levels) + shear (4 levels)	16	16.78
(h)	-	blur (4 levels) + fisheye (3 levels)	12	13.91
(i)	-	blur (4 levels) + fisheye (3 levels)	12	18.50

pairs. By using the "square design", the comparison number can be decreased to $n(\sqrt{n}-1)$.

A pair of test images are presented on the two monitors (Dell U2415) side by side with time synchronization. The two monitors are calibrated (calibration tool: Spyder5ELITE) to have ignorable visual difference. The brightness after calibration is 120 cd/m^2. The monitor resolution is 1920×1200, and the test image size is 1600×900. The test image is displayed in the center of the monitor with a surrounding background that is uniformly gray 128.

The environment fixes the viewing conditions for each subject to minimize the influence from external stimuli other than the test images. For each pair of images, the

subject needs to indicate which image has better viewing quality according to his or her visual evaluation by keyboard ("1" for choosing left image, "0" for choosing right image). The maximum time for the comparison of one pair of images is 10 seconds. Whenever the subject fails to make a choice after 10 seconds, he or she must randomly choose one of the two test images as the better one. The time interval between each comparison is 1.5 seconds. The interface for this test is built on PsychoPy [85]. We also conducted informal post-test feedback discussions with some participants who were willing to share their opinions.

2.8.3 Test Sources

Figure 2.12 shows source images (actual captured frames) of each content and Table 2.4 lists all test images of different categories in our test. The index of each content is the same as in Figure 2.12. We take 3 distinct approaches to create test images for motion blur, rotation plus blur, shear plus blur and fisheye plus blur. First, for "billiards", "eating", "flight", 5 nearby frames of each content that have different amounts of blur are selected. Next, "Winter Hovde Hall" and "bell tower" are intentionally created by continuous head rotation in front of a scene. 4 sharp frames with different amounts of rotation are then selected from these two sequences and different amounts of synthetic motion blur are added. Finally, one frame is chosen to be the reference respectively from "Autumn Hovde Hall", "parking lot", "parking garage" and "apartment building": distortions are applied to the reference with controllable amount. We also measure and report in Table 2.4 the spatial information (SI) [86] of each source image. SI is calculated as the mean of the gray-scale image filtered with both vertical and horizontal Sobel kernels. We experimentally find that the 3 geometric distortions, rotation, shear and fisheye, have very small influence on the SI of images in the same blur level.

Fig. 2.13.: Test images captured in FPVs to have different amounts of motion blur

Motion Blur: Our test of motion blur uses both nearby frames from FPVs and synthetic distorted images. These nearby frames are chosen to share at least half of their content and have minor difference in rotation and shear, but they differ in the amount of blur. Figure 2.13 shows chosen frames with the most and the least motion blur for each content. Synthetic motion blur are created by the motion model in [87]. The model can be used to create nonlinear motion blur kernels by controlling motion trajectory and motion kernel size. In our test, the motion trajectory is clockwise 45°

diagonally up to the right in a straight line. The motion kernels are created with size 2×2, 4×4, 8×8 to apply 3 levels of motion blur.

Rotation: Synthetic generation of rotation would require a significant area of the rotated image to be cropped to maintain a rectangular image. Therefore, we uses real images selected out of videos which were purposely created to contain rotated frames. The center of each image from the same video is selected to be almost the same location of the scene. This is intended to avoid a change in the location of focus of attention. Figure 2.14 shows sample test images of different amounts of rotation.

Shear: In geometric transformation, shear between two images can be modeled as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{2.12}$$

where $[x \quad y]^T$ are points in the reference image, and $[x' \quad y']^T$ are corresponding points in the sheared image. k is the shear parameter and $atan(k)$ indicates the angle of shear.

Shear transformation is a spatially varying filter. To create synthetic images with different levels of shear but almost the same amount of blur, we introduces a "symmetric transformation method" to add shear to the reference image as is shown in Table 2.5. Let i be image index, k_i is shear parameter for each image. The initial shear is 0. In each step of the process, we add the same amount of shear to each test image. The sign in front of shear amount is angle direction (i.e. "+" indicates shear to the right, "-" indicates shear to the left). At the conclusion of all steps, each test image has nearly similar blur but a distinct amount of shear. Figure 2.15 shows sample test images with different amounts of synthetic shear.

Fisheye: Fisheye distortion is a barrel transformation, which can be modeled as [88]

$$r' = a * r + b * r^2 + c * r^3 + d * r^4 + \mathcal{O}(r^5), \tag{2.13}$$

where r and r' are the distances of pixels to the image center in non-fisheye and fisheye images, respectively. $\mathcal{O}(r^5)$ are the higher order terms of r, which can be ignored. a, b, c, d are coefficients depending on the camera lens.

Table 2.5.: Symmetric transformation method to create shear images

initial shear	$k_1 = 0.00$	$k_2 = 0.00$	$k_3 = 0.00$	$k_4 = 0.00$
incremental shear 1	$+0.105$	$+0.105$	$+0.105$	$+0.105$
incremental shear 2	-0.035	-0.035	$+0.035$	$+0.035$
incremental shear 3	-0.035	$+0.035$	-0.035	$+0.035$
incremental shear 4	-0.035	-0.035	$+0.035$	$+0.035$
final shear	$k_1 = 0.00$	$k_2 = 0.07$	$k_3 = 0.14$	$k_4 = 0.21$

Barrel transformation introduces spatially varying blur to images. Therefore, we again use the "symmetric transformation method" shown in Table 2.5 to create image pairs with different amounts of fisheye. We set $a = 1, b = d = 0$, ignore $\mathcal{O}(r^5)$ and vary c only to get different amount of fisheye distortions. Fisheye images have decreasing scale factors from center to the image edges; our synthetically distorted images have the same scale in the image center compared to the original image. Figure 2.16 shows sample test images with different amounts of synthetic fisheye.

2.8.4 Results and Discussion

Our subjective test evaluates 4 distortions: motion blur, rotation, shear and fisheye. The test of motion blur uses nearby frames extracted from FPVs. Rotation, shear and fisheye are evaluated simultaneously with synthetic motion blur. By applying the Bradley-Terry model with maximum likelihood estimation [37], paired comparison results can be converted to relative subjective scores. Note that we use a logarithmic scale for the final subjective scores and set the score of the best image in each test to be 0. We also calculate the 95 percent confidence interval (CI) of each subjective score using the method presented in [37]. Each subjective score is represented as $q_s \pm q_r$, where q_s is the estimated subjective score, q_r is half the range of

the 95 percent CI of q_s. If the CIs of two image scores overlap each other, the quality difference of the two images is not significant, or namely, their quality is similar.

The test results indicate (1) LVI is effective at measuring blur when two images are not perfectly aligned, in the absence of rotation and shear. (2) Rotation and shear degrade quality. With the amount of degradation, shear is less sensitive to different contents than rotation. (3) The preference of fisheye versus non-fisheye differs from person to person, and this preference is influenced by content.

Motion Blur: The test of motion blur uses all actual captured frames from 3 FPVs with different content. In each content, the 5 test frames are selected to be temporally nearby and have different amounts of blur as measured by LVI.

In particular, the results show that LVI is an effective metric to estimate blur among misaligned images with minor rotation or shear difference. Seven existing NR QEs (JNBM [51], BIQI [89] , CPBD [90], BRISQUE [19], CORNIA [91], IL-NIQE [21] and NIQE [20]) and LVI are evaluated by subjective scores, as is shown in Tables 2.6, 2.7 and 2.8. The Spearman rank-order correlation coefficient (SROCC), the Kendall rank-order correlation coefficient (KROCC) and the Pearson linear correlation coefficient (PLCC) are employed to measure the performance of these QEs. Note that the images with the most severe blur, in "billiards" and "eating" respectively, cannot be measured by LVI because they have too few matching feature points. The LVI scores of these two images are considered to be zero. Note that LVI is the only QE we evaluated that correctly rank-orders the subjective quality for these image sets.

Rotation: Test images with different amounts of rotation are captured in front of the same scene with minor differences of viewpoint. The images with different rotation are selected to have tiny differences as measured by LVI. Motion blur degradations are synthetically added to images with different rotation. Figure 2.17 shows the subjective scores of content "winter Hovde Hall" and "bell tower", in which the average range of CIs is 0.22 for each content.

First, we explore the intra-relationship of both motion blur and rotation. For a fixed rotation angle, subjective scores monotonically decrease when the blur filter

Table 2.6.: QE performance: motion blur - billiards

QE name	SROCC	KROCC	PLCC
JNBM	0.9000	0.8000	0.8248
BIQI	0.6000	0.4000	0.2342
CPBD	0.3000	0.2000	0.1905
BRISQUE	0.1000	0.0000	0.1130
CORNIA	0.6000	0.4000	0.7446
IL-NIQE	0.9000	0.8000	0.8599
NIQE	0.9000	0.8000	0.8514
LVI	**1.0000**	**1.0000**	**0.8752**

Table 2.7.: QE performance: motion blur - eating

QE name	SROCC	KROCC	PLCC
JNBM	0.4000	0.4000	0.5701
BIQI	0.8000	0.6000	0.7137
CPBD	0.2000	0.2000	0.2732
BRISQUE	0.5000	0.2000	0.5467
CORNIA	0.3000	0.2000	0.1610
IL-NIQE	0.9000	0.8000	**0.9597**
NIQE	0.8000	0.6000	0.8167
LVI	**1.0000**	**1.0000**	0.8719

size increases. Only when the blur filter size increases from 0 to 2, the CIs for the respective subjective scores have overlap. For a fixed blur level, rotation introduces quality degradations, and the influence becomes larger as the blur level increases. The rotated images of the lowest and the second lowest blur levels have closer subjective

Table 2.8.: QE performance: motion blur - flight

QE name	SROCC	KROCC	PLCC
JNBM	0.9000	0.8000	0.8914
BIQI	0.9000	0.8000	0.9637
CPBD	0.6000	0.4000	0.6068
BRISQUE	0.3000	0.2000	0.5452
CORNIA	0.7000	0.6000	0.6342
IL-NIQE	0.9000	0.8000	0.9028
NIQE	0.2000	0.2000	0.2052
LVI	**1.0000**	**1.0000**	**0.9958**

scores than rotated images of higher blur levels, which is reflected as the overlap between the confidence regions in the first two blur levels compared to the non-overlapping of the other two higher blur levels.

To elaborate, let the score of a test image be (r, b), where r is the rotation level, b is the blur level, and $r, b = 1, 2, 3, 4$. For content "winter Hovde Hall", (3,1) and (4,1) are -1.55 ± 0.29 and -1.98 ± 0.23, respectively. The overlap between their confidence regions indicates their quality is similar. As a comparison, (3,3) and (4,3) are -4.57 ± 0.20, and -5.12 ± 0.24, respectively. The non-overlapping indicates that they have significant quality difference.

Second, we explore the inter-relationship between motion blur and rotation. In both contents, the quality differences of images are not statistically significant when blur is small. One example is in content "bell tower", (2,2) and (3,1) are similar, with scores -0.69 ± 0.27 and -0.78 ± 0.30, respectively. As a comparison, (4,3) is -4.96 ± 0.22, while (1,4) is -5.99 ± 0.23 worse than the former. As an addition, the content "winter Hovde Hall" has higher spatial information (SI) than "bell tower", and its subjective quality is more sensitive to rotation.

To model the overall quality measure of an image with blur plus rotation, we propose a mapping function to combine LVI and rotation. The mapping function is given by

$$Q(\theta, q_{LVI}) = q_{LVI} \cdot (1 - p \cdot \exp{(q_{LVI} - 1)} \cdot \theta^2), \tag{2.14}$$

where the rotation angle θ (radian) is estimated relative to the reference image by affine estimation as described in [6]. q_{LVI} is the LVI score of the distorted image, and p is a constant parameter which needs to be optimized. equation (2.14) is called rotation-LVI.

From discussion with the participants in the subjective test, preference regarding rotation is content dependent. The same rotation angle for different content gives rise to different viewing quality. We optimize p for each content to maximize SROCC and KROCC. The optimized p is 4.53 for "winter Hovde Hall" and 1.13 for "bell tower". Figure 2.18 shows the nonlinear mapping curve between subjective scores and rotation-LVI with optimized p. The logistic function used for curve fitting is

$$f(x) = (t_0 - t_1)/(1 + \exp{(-(x - t_2)/|t_3|)}) + t_1 \tag{2.15}$$

where t_0, t_1, t_2 and t_3 are 4 unknown parameters for fitting.

For extension to other quality metrics, the term q_{LVI} can be replaced with any other quality measure q, given by

$$Q(\theta, q) = q \cdot (1 - p \cdot \exp{(-\frac{|q - q_{best}|}{|q_{best} - q_{worst}|})} \cdot \theta^2). \tag{2.16}$$

The term $\exp{(q_{LVI} - 1)}$ is replaced with $\exp{(-\frac{|q-q_{best}|}{|q_{best}-q_{worst}|})}$, where 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.

Table 2.9 and 2.10 show the performances of 7 NR QEs and LVI. "rotation-" indicates the QE is mapped by equation (2.14) with corresponding optimized p. Note that we use the self-reported best and worst QE values when available, otherwise the observed best and worst QE values in [44] are used. One exception is that JNBM has maximum value at infinity, so we set its best score as the QE score of the best

Table 2.9.: QE performances: rotation - winter Hovde Hall

QE name	p	SROCC	KROCC	PLCC
JNBM	-	0.8441	0.6833	0.8920
BIQI	-	0.6941	0.5167	0.7140
CPBD	-	0.8205	0.5833	0.8412
BRISQUE	-	0.6441	0.4500	0.7361
CORNIA	-	0.5059	0.3833	0.4201
IL-NIQE	-	0.9176	0.7500	0.9429
NIQE	-	0.8941	0.7667	0.8561
LVI	-	0.8529	0.7000	0.9042
rotation-JNBM	6.06	0.9706	0.8833	0.9102
rotation-CPBD	8.48	0.9559	0.8333	0.8955
rotation-IL-NIQE	-1.60	0.9529	0.8500	**0.9574**
rotation-NIQE	-5.14	0.9382	0.8333	0.9050
rotation-LVI	4.53	**0.9853**	**0.9333**	0.9480

image for each content. The comparison shows that LVI, JNBM, CPBD, IL-NIQE and NIQE all improve their performance after including rotation mapping. To get a generalized model, we fix p to be 1.16; the resulting SROCC of rotation-LVI are 0.9588 and 0.9529 for "winter Hovde Hall" and "bell tower", respectively. However, since two scenes are not enough to allow generalization of the model, more subjective data needs to be collected.

Shear: Test images distorted with shear and motion blur are synthetically created from one reference image. Figure 2.19 shows the subjective scores of content from two sequences "autumn Hovde Hall" and "parking lot", with average range of CIs to be 0.28 and 0.75, respectively. The difference of their CI range results from content difference. The content "parking lot" has lower SI than information "autumn

Table 2.10.: QE performances: rotation - bell tower

QE name	p	SROCC	KROCC	PLCC
JNBM	-	0.9117	0.7333	0.9109
BIQI	-	0.6471	0.4333	0.6413
CPBD	-	0.9441	0.8167	0.9009
BRISQUE	-	0.3088	0.1833	0.5503
CORNIA	-	0.3853	0.2833	0.2539
IL-NIQE	-	0.7411	0.5667	0.8557
NIQE	-	0.9558	0.8500	0.9269
LVI	-	0.8764	0.6500	0.9256
rotation-JNBM	0.47	0.9618	0.8500	0.9180
rotation-CPBD	1.1	0.9705	0.9000	0.9052
rotation-IL-NIQE	-1.64	0.8882	0.7667	0.9210
rotation-NIQE	-0.4	**0.9794**	**0.9333**	**0.9406**
rotation-LVI	1.13	0.9618	0.8833	0.9221

Hovde Hall", and many of its edges are highly curved or within texture. Since local orientation structure is visually more sensitive to straight edges than curved edges or textures [92], shear is visually less sensitive in "parking lot" than in "autumn Hovde Hall".

First, we explore the intra-relationship of both motion blur and shear. In both contents, subjective scores monotonically decreases as the blur level increases for any fixed shear level. Within each blur level, the image with the greater shear often has worse visual quality. Note that while in "parking lot", the CI of image scores from the same blur level has significant overlap, the score of the image with the least shear has no overlapping CI with that of the most shear.

Second, we explore the inter-relationship between motion blur and shear. Let the score of a test image be (s, b), where s is the shear level, b is the blur level, and $s, b = 1, 2, 3, 4$. We find that (s, b) often has similar value with $(s - 1, b + 1)$. For instance, in content "autumn Hovde Hall", (2,1) and (1,2) has respective scores -0.89 ± 0.31 and -0.59 ± 0.42 with no significant difference. (4,2) is a little better than (3,3), with scores -3.36 ± 0.34 and -4.49 ± 0.34, respectively.

To model the overall quality measure of an image with blur plus shear, we propose a mapping function to combine LVI and shear. The overall quality is modeled as:

$$Q(k, q_{LVI}) = q_{LVI} \cdot (1 - g \cdot exp(q_{LVI} - 1) \cdot k^2) \tag{2.17}$$

where k is the shear in equation (2.12), q_{LVI} is the LVI score of the distorted image; g is a constant parameter. The mapping by equation (2.17) is called shear-LVI.

To find the optimized value of g, we also maximize SROCC between shear-LVI scores and subjective scores. The optimized values of g are 5.34 and 2.59 for "autumn Hovde Hall" and "parking lot", respectively. To generalize equation (2.17) for all content without dramatic influence on SROCC and KROCC, g is experimentally chosen to be 4.07. Figure 2.20 shows the nonlinear mapping curve between the generalized shear-LVI and subjective scores. The fitted logistic function used is equation (2.15).

By using the same replacement as equation (2.16), we can extend equation (2.17) to to other quality metrics. The performances of 7 NR QEs and LVI are compared in Table 2.11 and 2.12. "shear-" indicates that the QE score is mapped by equation (2.17) with corresponding optimized g. JNBM, CPBD, IL-NIQE, NIQE and LVI improve their performances after mapping by equation (2.17). The generalized shear-LVI (g=4.07) shows competitive performance compared to other 7 QEs after mapping.

Fisheye:

Test images distorted with fisheye and motion blur are synthetically created from one reference image. Figure 2.21 shows the subjective scores from two content "parking garage" and "apartment building", with average range of CIs to be 0.30 and 0.31.

Table 2.11.: QE performances: shear - autumn Hovde Hall

QE name	g	SROCC	KROCC	PLCC
JNBM	-	0.9235	0.8000	0.8853
BIQI	-	0.5706	0.3166	0.7395
CPBD	-	0.7529	0.5166	0.7787
BRISQUE	-	0.6441	0.5500	0.7011
CORNIA	-	0.2235	0.1333	0.2270
IL-NIQE	-	0.9294	0.8167	0.9283
NIQE	-	0.9382	0.8333	0.8734
LVI	-	0.7108	0.4602	0.8456
shear-JNBM	5.88	0.9735	0.9000	0.9470
shear-CPBD	23.38	0.9471	0.8333	0.9016
shear-IL-NIQE	-6.78	0.9735	0.9000	0.9601
shear-NIQE	-10.09	0.9647	0.8667	0.9528
shear-LVI	5.34	**0.9912**	**0.9500**	**0.9672**
shear-LVI	4.07	0.9853	0.9333	0.9694

Variations in quality due to different blur levels are stronger than those due to differences in the degree of fisheye. Specifically in Figure 2.21(a), the variances of subjective scores for 4 levels of blur with same fisheye are 5.56, 6.29 and 5.48, while the variances for 3 levels of fisheye with the same blur are 0.09, 0.11, 0.08 and 0.062. In Figure 2.21(b), the variances of subjective scores for 4 levels of blur with the same fisheye distortion are 7.40, 5.00 and 5.85, while the variances for 3 levels of fisheye with the same blur are 0.00, 0.06, 0.04 and 0.27. We can also draw the same conclusion from the CI of scores. For example, in content "parking garage", the scores of the 3 levels of fisheye in blur level 3 are -2.98 ± 0.43, -3.61 ± 0.16 and -3.56 ± 0.05, which have overlapping regions. In content "apartment building", the scores of the

Table 2.12.: QE performances: shear - parking lot

QE name	g	SROCC	KROCC	PLCC
JNBM	-	0.8353	0.6333	0.8885
BIQI	-	0.6411	0.4000	0.6547
CPBD	-	0.8147	0.6167	0.8473
BRISQUE	-	0.8353	0.7000	0.8710
CORNIA	-	0.1529	0.1000	0.1963
IL-NIQE	-	0.9088	0.7667	0.9580
NIQE	-	0.9706	0.8667	0.9669
LVI	-	0.8992	0.7113	0.9481
shear-JNBM	2.46	0.9765	0.8833	0.9779
shear-CPBD	11.53	0.9735	0.9000	0.9189
shear-IL-NIQE	-1.7	0.9500	0.8333	0.9686
shear-NIQE	-0.37	0.9735	0.8833	0.9697
shear-LVI	2.59	**0.9853**	**0.9333**	**0.9858**
shear-LVI	4.07	0.9794	0.9000	0.9750

3 levels of fisheye in blur level 3 are -3.12 ± 0.38, -3.08 ± 0.12 and -3.48 ± 0.19 with no significant difference. As a comparison, in both contents, the lowest score in blur level s is statistically greater than the highest score in blur level $s+1$, when $s = 2, 3, 4$.

Not apparent from Figure 2.21, however, a personal preference exists for fisheye, and that preference is content dependent. The preference is extracted based on the percentage of times that the participant chose images with smaller fisheye levels in the same blur level. For content "parking garage", Figure 2.22 shows a comparison of subjective scores between two group of subjects: 35 people prefer non-fisheye while the remaining 15 prefer fisheye. Figure 2.23 shows that 33 people prefer non-fisheye

while another 17 prefer fisheye images for content "apartment building". Among all subjects in the test, 23 prefer non-fisheye and 5 prefer fisheye for both contents, and 22 people show different preferences for the two scenes.

"Apartment building" has a relatively close view and higher spatial information (SI) compared to "parking garage". From post-test feedback, the bend of the scene around the edges in the former is not as obvious as in the latter. This feedback indicates the field of view of the content influences the viewing quality of a fisheye image. Some participants also indicate that the broad view of fisheye images could convey more information about the scene compared to non-fisheye images, especially for "apartment building".

we implemented a subjective test using paired comparison in [7] 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 [37]. 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 [51], BIQI [89], CPBD [18], BRISQUE [19], CORNIA [91], IL-NIQE [21] and NIQE [20], 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, Equation(2.18), to map LVI or existing NR QEs that are insensitive to geometric distortions with estimated shear and rotation value to the overall quality. Equation(2.18) is the mapping function

to calculate the overall quality of an image with motion blur and geometric distortions simultaneously.

$$Q(\mathcal{D}, q) = q \cdot (1 - p \cdot \exp(-\frac{|q - q_{best}|}{|q_{best} - q_{worst}|}) \cdot \mathcal{D}^2). \tag{2.18}$$

where $\mathcal{D}$ is the measured value of shear or rotation (k_s or θ in Equation 2.10). 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 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.

2.9 Conclusions

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

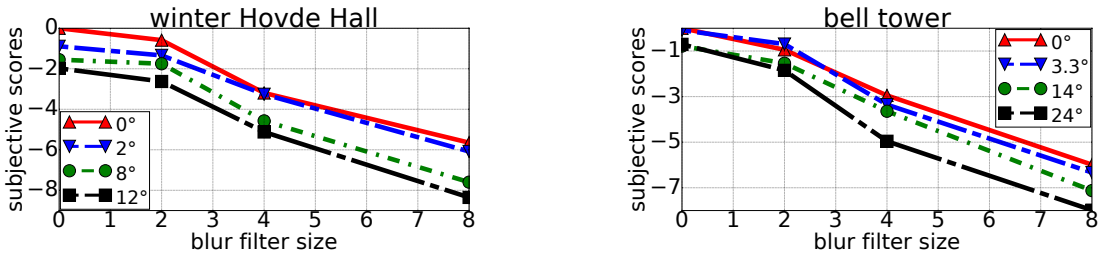
avoid using frames with heavy quality degradations. Third, from the perspective of analysis, the quality score of each frame provides an indication of useful and useless frames for applications such as object detection and activity recognition.

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.

Fig. 2.14.: Test images intentionally captured to have different amounts of rotation

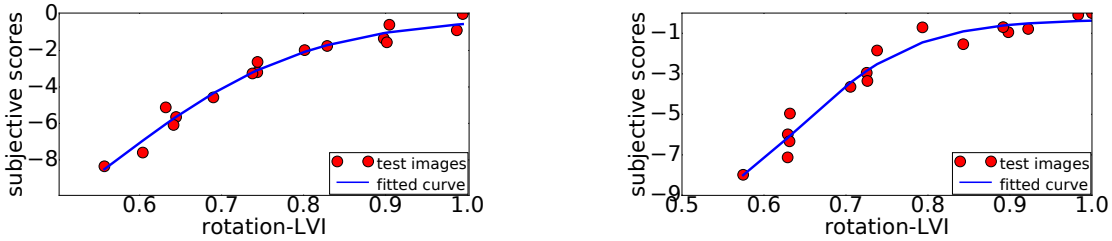
Fig. 2.15.: Test images with different amounts of synthetic shear created from one reference image

Fig. 2.16.: Test images with different amounts of synthetic fisheye created from one reference image



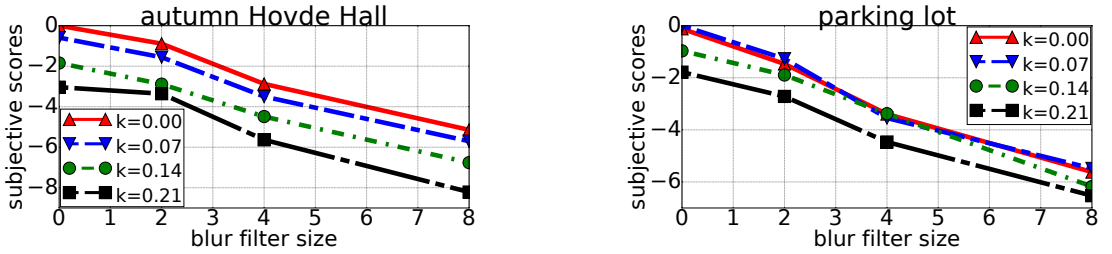
(a) (b)

Fig. 2.17.: Subjective test - rotation and blur: (a) winter Hovde Hall (b) bell tower



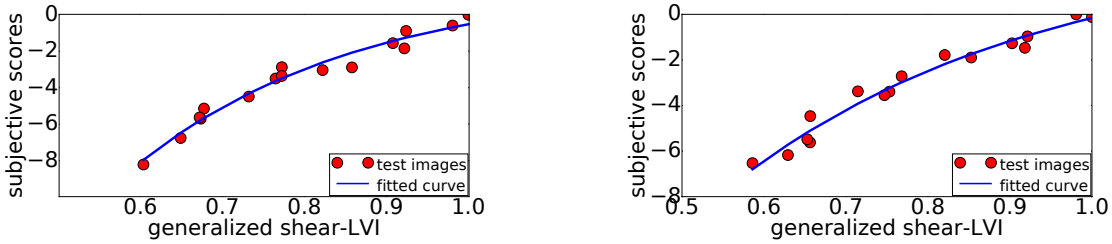
(a) (b)

Fig. 2.18.: Curve fitted with logistic function between subjective scores and rotation-LVI: (a) winter Hovde Hall (p=4.53) (b) bell tower (p=1.13)



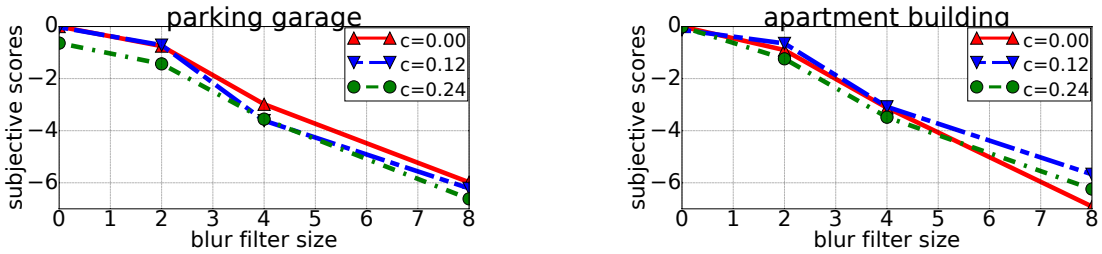
(a) (b)

Fig. 2.19.: Subjective test - shear and blur: (a) autumn Hovde Hall (b) parking lot



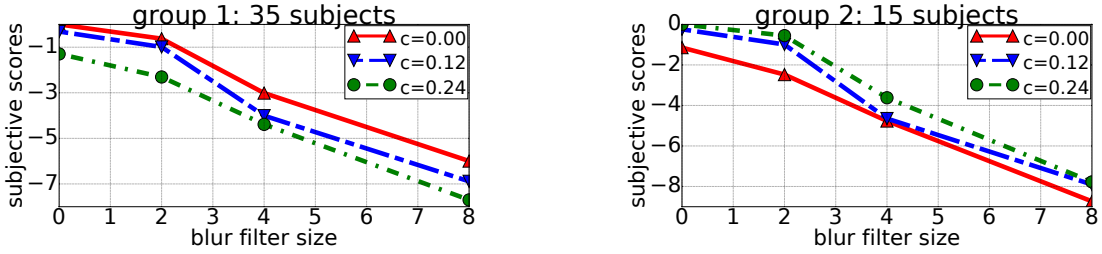
(a) (b)

Fig. 2.20.: Curve fitted with logistic function between subjective scores and generalized shear-LVI (g=4.07): (a) autumn Hovde Hall (b) parking lot



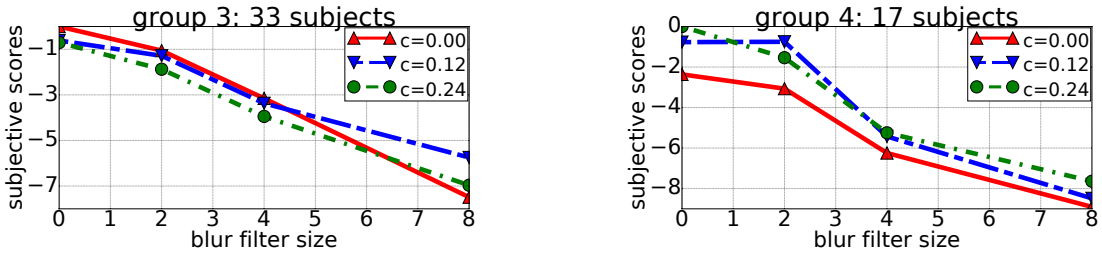
(a) (b)

Fig. 2.21.: Subjective test - fisheye and blur: (a) parking garage (b) apartment building



(a) (b)

Fig. 2.22.: Parking garage: group 1 prefers non-fisheye, group 2 prefers fisheye



(a) (b)

Fig. 2.23.: Apartment building: group 1 prefers non-fisheye, group 2 prefers fisheye

3. CONTROLLABLE ILLUMINATION ENHANCEMENT

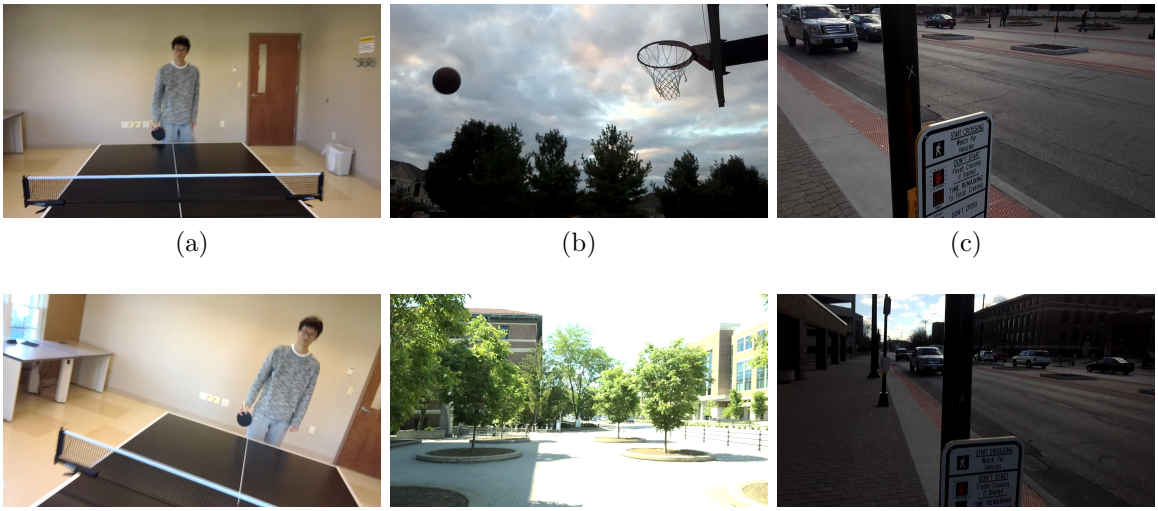
3.1 Introduction

Exposure distortions refers to that when the camera sensor is not exposed to proper amount of light so that the luminance histogram of the image fails to spread over the desired range [10]. The image is called well-exposed when there is no exposure distortions. The image is called over-exposed when the camera receives more light than its well-exposed version, otherwise the image is called under-exposed. Possibly, under-exposure and over-exposure occurs in specific regions of an image.

First-Person videos (FPVs) are often badly-exposed. Because FPVs are recorded under conditions that the wearer is not fully aware of the lighting condition for the camera, so the wearer has no intention to adjust the camera location so that the images or videos are often captured with exposure distortions.

The causes of exposure distortions in FPVs can be classified into motion-induced lighting variations, bad environmental lighting or a combination of both. Motion-induced lighting variations are caused by the change of lighting directions and motion itself. Since the wearer often has violent motion, the camera angle and location changes frequently during capture so that lighting conditions are very unstable. Bad environmental lighting refers to that the scene in front of the camera suffer from bad lighting condition during capture. Typical cases are blocked sunshine, dark indoor environment, nightfall, shadows, glare of the sun. In addition, the combination of both will introduce more temporal exposure change into the video. For example, the camera wearer moves his or her body from sunshine into a shadow, and then back to sunshine. Figure 3.1 shows the exposure distortions in the three cases.

The quality of FPVs can be improved by alleviating exposure distortions. One type of method is illumination enhancement. Illumination enhancement is to either



(a) (b) (c)

(d) (e) (f)

Fig. 3.1.: (a)(d) Motion induced lighting variation (b)(e) Bad environmental lighting (c)(f) Combination of both

increase or decrease the illumination locally or globally, so that the image has better perceived illumination. Some enhancement works do not use exposure, illumination, lighting, brightness or lightness consistently [24, 93, 94], we illustrate their definitions to clarify the concept of illumination enhancement. Exposure is the total amount of light received by the camera, affecting the whole image in most cases. Synthetic exposure change should be a replication of actual captured image with different exposures. Illumination or lighting refers to the use of light to achieve better practical effect in images. The modification of illumination can focus on specific regions of the image and often cannot be reproduced during actual capture. Brightness and lightness are terminology of perception.

This relationship between enhancement and image visual quality can be described as a concave function with a peak point. We consider the peak point as the optimal degree of enhancement, defined as the optimal point (OP). The concave relationship is produced by three aspects, contrast, exposure level and newly generated artifacts introduced by enhancement operations. First, image quality is a concave function of

contrast. According to the results from image quality database TID2013 [34], when the synthetic contrast manipulation is applied to an image, there exists a peak point of quality corresponding to its best contrast. Second, the exposure level change also has a concave relationship with image quality [10], and its best point corresponds to the exposure level at which the image is well-exposed, not either under-exposed or over-exposed. Third, enhancement operations often generate new artifacts, such as color shift or loss, noise amplification, structure modification or unnaturalness [95]. The combined visual effect of newly generated artifacts and contrast change can also be described as a concave function of image quality. The concave function of enhancement is content dependent, in that the OP varies for different content. One unsolved problem is how to define the OP for different images and characterize the concave function. Our solution is to enhance the image into different levels, and then characterize the concave curve including the OP using a content-independent over-enhancement measure.

We propose an controllable illumination enhancement method for which the degree of enhancement can be adjusted using a single parameter [22]. Many existing enhancement methods including histogram equalization [96, 97], retinex methods [98, 99] and others [100–103] have no clear relationship between their parameters and image quality. However, our single parameter has a concave relationship with image quality. In our method, we model under-exposure and over-exposure differently to assign under-exposed and over-exposed probabilities for each pixel. We then design a system that applies logarithmic mapping in the identified under-exposed pixels with boundary-artifact compensation. Our mapping uses the assigned under-exposed probabilities, the artifact compensation weights and the single adjustment parameter together to calculate mapping coefficients. We also propose an over-enhancement measure, Lightness Order Measure (LOM) to quantify the unnaturalness in the enhanced image. We consider the unnaturalness to be related to the inversion of relative lightness order between neighboring pixels, and which is influenced by both the proportion of inversions and the inversion magnitude.

In Section 3.3, we describe and illustrate the three major parts in our system of controllable illumination enhancement: under-exposed map and over-exposed map, boundary penalization, logarithmic mapping. We then illustrate our proposed over-enhancement measure, LOM, in Section 3.4. Section 3.5 implements a subjective test to explore the relationship between LOM and image subjective quality after enhancement, and demonstrates the effectiveness of LOM and our illumination enhancement method.

3.2 Related Works

In this section, two types of exposure distortions, over-exposure and under-exposure are are introduced. The assessment of over-exposure and under-exposure are related to two factors: pixel intensity and saturation. Second, there exists spatially-inconsistent exposure distortions within a FPV frame. Third, the over-enhancement should be measured and avoided.

3.2.1 Over-exposure and Under-exposure

Over-exposure and under-exposure are both highly correlated with pixel intensity. Over-exposure introduces a loss of details in bright areas. It occurs when the received light of the camera go out of its dynamic range. The resulting output over-exposed pixels clip at their maximum value. Under-exposure introduces a loss of details in dark areas. In low intensity regions, the threshold of just-noticeable-difference is larger than medium intensity region. Hence many details in dark regions are unable to be perceived because the contrast is not enough.

Another influence due to over-exposure and under-exposure is pixel saturation. When mid-tone colors are exposed as bright or dark colors, they often lose saturation. In [104], either low intensity or saturation would cause the perceived pixel color to be close to gray that are indistinguishable, otherwise the pixel can be considered as

a "true color" pixel. A well-exposed pixel should be a "true color" pixel that can be correctly perceived.

FPVs often contain spatially-varying exposure distortions within a frame. Because the lighting condition across the whole image is not consistent while the camera exposure length are consistent, one image contains different regions with different degree of over-exposure or under-exposure. [100, 101] both proposed over-exposure detection models that segment the image into over-exposed and under-exposed areas.

3.2.2 Existing Enhancement Methods

Existing video enhancement methods can be classified into self-enhancement and mutual-enhancement, that is called context-based fusion video enhancement in [105]. Self-enhancement methods mainly consists of three types: contrast-based [96, 98], HDR-based [106], transform-based [107, 108]. Contrast-based methods are widely used with computationally efficiency, usually using transformation function.

$$I'(x, y) = T[I(x, y)], \tag{3.1}$$

where (x, y) are pixel locations, I is the original image and I' is the enhanced image. $T(\cdot)$ is transformation function. $T(\cdot)$ is normally applied to spatial domain including histogram with certain constraints. Mutual-enhancement is to extract useful information for enhancement from multiple images, proposed in [109]. In addition, some works have been proposed to focus on specific cases: over-exposure correction [100, 101], low-light enhancement [102, 103], illumination editing [110].

3.2.3 Existing Enhancement Evaluation Metrics

The enhancement measures are proposed to evaluate the quality of enhanced image compared to the original. Commonly used Existing measures are Absolute Mean Brightness Error(AMBE) [111], Discrete Entropy (DE) [112], Image Enhancement Metric (IEM) [113], Measure of Enhancement (EME) [114] and RIQMC [115].

Over-enhancement refers to the artifacts introduced after enhancement. Typical over-enhancement artifacts include loss of edges, textures or unnaturalness. The measurement of over-enhancement is important in the design of enhancement algorithms so that the amount of enhancement can be constrained to avoid newly generated artifacts in enhanced images or videos. In [23], a Structure Measure Operator was proposed to detect structure change after enhancement. In [24], a Lightness Order Error was proposed to measure the unnaturalness of the enhanced image.

3.3 Controllable Illumination Enhancement

In this section, we propose a controllable illumination enhancement method that allows a single parameter to adjust the degree of enhancement. Our enhancement system has 3 major parts: under-exposure and over-exposure map, boundary penalization and logarithmic mapping. We separately model the under-exposed and over-exposed map based on an over-exposure model in [100]. Our logarithmic mapping takes into account the under-exposed map values and boundary-artifact compensation weights, and the single adjustment parameter β to assign mapping coefficients for each pixel.

Figure 3.2 shows the block diagram of our method. First, an under-exposed map and an over-exposed map are calculated for the input image. Then, the image is partitioned into either under-exposed or over-exposed regions. Third, a logarithmic mapping is applied to the under-exposed regions with penalization to compensate for the boundary artifacts. Finally, our proposed Lightness Order Measure (LOM) quantifies the unnaturalness of the output image, illustrated in Section 3. Details for each step are explained next.

Under-exposed map and Over-exposed map: We create an under-exposed map and an over-exposed map separately for an image considering both pixel saturation and intensity. Pixel saturation is affected similarly by both under-exposure and over-exposure, in that low saturation pixels are perceived to be close to gray, and therefore

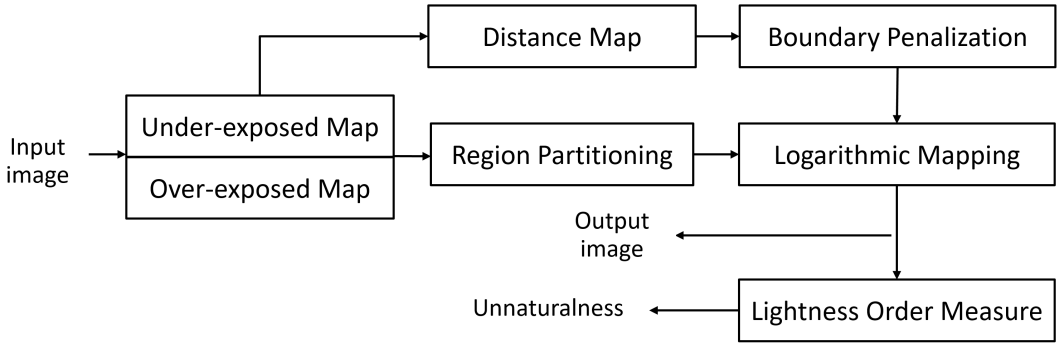


Fig. 3.2.: Illumination enhancement block diagram

are indistinguishable from each other [104]. A well-exposed pixel, on the other hand, has color that can be correctly perceived. However, pixel intensity is affected differently. Over-exposed pixels often have high intensity, while under-exposed pixels have low intensity.

Based on the over-exposure detection model proposed in [100], we model both the under-exposed map M_u and the over-exposed map M_o in $L^*a^*b^*$ space as

$$M_u = 0.5 \tanh\left(\delta(L_{ut} - (\sqrt{a^2 + b^2} + G(L))\right) + 0.5 \tag{3.2}$$

$$M_o = 0.5 \tanh\left(\delta(L_{ot} - (\sqrt{a^2 + b^2} - G(L))\right) + 0.5 \tag{3.3}$$

where L, a and b are rescaled pixel values (from 0 to 255) of L*, a* and b*. For a fixed L, when saturation drops, $\sqrt{a^2 + b^2}$ will decrease. $G(\cdot)$ is a 15×15 Gaussian filter with $\sigma = 3$. The range of M_u and M_o is from 0 to 1, corresponding to the probability of a pixel to be under-exposed or over-exposed, respectively. We set $L_{ut} = 255$ and $L_{ot} = 0$ so that M_u and M_o are both 0.5 when the pixel has intensity and saturation that are half of their entire range. δ controls how fast M_u and M_o increase or decrease with L or $\sqrt{a^2 + b^2}$, and is experimentally set to be 1/60. Figure 3.4 shows an example image with its under-exposed map and over-exposed map.

Boundary penalization: The image is partitioned into under-exposed regions R_u ($M_u > M_o$) and over-exposed regions R_o ($M_u < M_o$). To eliminate the artifacts near

the boundary of R_u and R_o after enhancement, we introduce a boundary penalization weighting function $\omega(x, y)$, where (x, y) is pixel location. We first compute the Euclidean distance between pixels to its closest partitioning edges between R_o and R_u, and normalize it to get distance map $\mathcal{D}$. Then $\omega(x, y)$ is calculated as

$$\omega(x, y) = \frac{\log(\mathcal{D}(x, y)(p - 1) + 1)}{\log(p)}, \tag{3.4}$$

where p is a constant parameter. To set the the value p, we experimentally test with p from 0.001 to 1000. When p decreases from 1000 to 10, the boundary artifacts is reduced, after p drops below 10, there is no apparent boundary artifacts. When p drops from 10 to 0.001, the boundary artifacts are eliminated, but we need to apply higher β to compensate the drop in average enhancement.

Logarithmic mapping: To enhance the illumination of the under-exposed regions, we use the logarithm mapping function

$$L'(x, y) = \frac{\log(L(x, y) * (\gamma(x, y) - 1) + 1)}{\log(\gamma(x, y))}, \tag{3.5}$$

where $L'(x, y)$ and $L(x, y)$ are luminance values in $L^*a^*b^*$ space for the enhanced image and the original image, respectively. $\gamma(x, y)$ is the mapping coefficient, calculated as

$$\gamma(x, y) = 1 + M_u(x, y) * \omega(x, y) * \beta, \tag{3.6}$$

where β is the control parameter that can adjust the amount of enhancement. We finally convert the image back to RGB space using the mapped luminance L' and original a^*, b^*. Figure 3.5 shows an example image, extracted from video "Alin, Day1" in [116], enhanced to 7 levels by adjusting β.

3.4 Over-enhancement Measure

In this section, we propose an over-enhancement measure, the Lightness Order Measure (LOM), to quantify the unnaturalness after enhancement, and we compare it with two existing metrics, SMO [23] and LOE [24].

(a)

(b)

(c)

Fig. 3.3.: (a) $p = 1000, \beta = 20$ (b) $p = 10, \beta = 50$ (c) $p = 0.1, \beta = 200$

The principle of our Lightness Order Measure (LOM) is to measure when the *relative* lightness order of pixels in the image is reversed. Relative Lightness order [24] refers to the pixel intensity order of the image, represented as $I(x_1, y_1) > I(x_2, y_2)$,

(a)

(b)

(c)

Fig. 3.4.: (a) original image (b) under-exposed map M_u (c) over-exposed map M_o

where (x_1, y_1) and (x_2, y_2) are two different pixel locations. The relative lightness order of an image should be preserved to keep its naturalness.

There are two existing over-enhancement measures, SMO and LOE. SMO measures the image structure change; it quantifies the difference of gradients, standard

deviation and entropy between the original image and the enhanced image. LOE measures the change of lightness order globally in the image; it compares every two pixels and calculates how many pairs are reversed. All three measures compare the original image to the enhanced image.

LOM shows advantages compared to SMO and LOE. Compared to SMO, LOM does not use content-dependent information, so it is subject to less influence from different contents. Compared to LOE, LOM considers the relative lightness order locally and quantifies the magnitude of the inversion; hence it improves the computational efficiency.

To compute LOM, let the original image be i_1 and the enhanced image be i_2 in luminance domain. First, the local mean filter is both applied to i_1 and i_2 with window size 31×31, and the filtered luminance images are f_1 and f_2, respectively. Second, we calculate the difference image $d_1 = f_1 - i_1$ and $d_2 = f_2 - i_2$. Third, we quantify the LOM as

$$LOM = \frac{1}{H \cdot W} \sum_x \sum_y \Big| (d_2(x,y) - d_1(x,y)) \cdot \frac{sign(d_2(x,y)) - sign(d_1(x,y))}{2} \Big|, \quad (3.7)$$

where H and W are image height and width. Larger values for LOM indicate greater unnaturalness. In Figure 3.6, three enhanced versions of the same image are shown with different LOM.

Enhancement upper limit refers to the upper bound that the enhancement method can achieve without introducing artifacts. In our case, the upper limit is the bound of enhancement that the image does not suffer from over-enhancement artifacts.

3.5 Experiments and Results

In this section, we implement a subjective test with two phases. The first phase explores subjective quality of enhanced images of different levels using our method. It also assesses the performance LOM, SMO and LOE to characterize the OP of the concave quality curve for different contents. The second phase evaluates the subjective

quality of images enhanced by prior existing methods and ours. Both phases are used to test the effectiveness of our enhancement method with over-enhancement measure, LOM.

Test sources are 6 video frames captured by a wearable camera Pivothead (1080p30fps), shown in Figure 3.7. The test images are ordered using the percentage of partitioned under-exposed regions P_u. Each test image is enhanced by our proposed method to 9 different levels, and by five existing enhancement methods, LDR [117], CVC [118], WAHE [97], SRIE [99], Low-light enhancement using camera response model (LL-CRM) [119]. Examples of enhanced images using five methods and ours are shown in Figure 3.8, where the original image is 6 in Figure 3.7.

The subjective test has two phases. The first phase evaluates 9 enhanced images using our method by adjusting β, and find the best β for each content. The second phase evaluates enhanced images using five existing methods and the best β image version obtained in the first phase.

Our test method is paired comparison. Each pair of enhanced images of the same content is presented side by side on a 4k monitor (DELL P2415Q), and the right-side image is horizontally flipped. The monitor resolution is 3840×2160. The image is symmetrically cropped to be 1900×1080. Each of the 20 subjects are asked to indicate *which image is perceptually better in terms of illumination, noise, naturalness, color and incorrect textures*. The subjective image quality is calculated from the paired comparison results using the Bradley-Terry Model [37]. The calculated subjective scores are all relative; the best quality score is set to be 0.

The results in Figure 3.9 show that each of LOM, SMO and LOE has a concave relationship with subjective image quality, and the concave curve varies for different contents. The comparison between Figure 3.9(a), 3.9(b) and 3.9(c) indicates that our LOM reduces content-dependency compared to SMO and LOE. The overlap between concave curves of different contents in Figure 3.9(a) is much greater than in Figure 3.9(b) and (c). For example in Figure 3.9(b), the best version of image 6 has an SMO of 5.5, but this is larger than the SMO of all versions of the other 5 images,

including their worst quality versions. In Figure 3.9(c), the comparisons between the best version of image 5 with LOE 440 and images 2, 3, 4, 6, and between the best version of image 6 with LOE 358 and images 2, 4 show the same situations as mentioned for Figure 3.9(b). This means SMO and LOE are unsuited for use to find the best degree of enhancement when applied to different contents. In contrast, Figure 3.9(a) shows a better set of concave curves; the LOM of all versions of one image are neither smaller or larger than the LOM value of the best version of another image.

Visual quality of an enhanced image is influenced by both illumination and naturalness. For example, image 6 has the highest P_u and its best version has the largest LOM compared to the other 5. The reason is that image 6 is heavily under-exposed, so the illumination improvement has a larger influence than unnaturalness.

Table 3.1 shows the results of subjective quality of images enhanced by the five enhancement methods and ours, and indicates that our method shows more balanced performance considering image quality and computational efficiency. The results of subjective scores show that the overall performance of the 6 methods can be listed from the best to the worst as SRIE, ours, WAHE, LDR, CVC, LLCRM. LLCRM is applied for low-light image enhancement, so it performs much worse when P_u is small for images 1 to 4 compared to other methods. LDR, CVC and WAHE focus on contrast enhancement, they all have relatively unbalanced performance compared to SRIE and ours. For example, their performance is worse for image 6 with $P_u = 0.82$ than images 1, 2, 3 with $P_u < 0.6$. SRIE and our method show the best or at least the 3rd performance for different contents. However, the processing time for SRIE is more than 50 times longer than the other five methods. Because SRIE uses an iterative optimization strategy, and the optimization time significantly depends on the content. Overall, the performance of our method is more balanced for contents that cover a range of P_u from 0.35 to 0.82.

We also apply our enhancement method into videos. For each signle frame in the video, we enhance it to the version with peak quality using LOM. In our experiment,

Table 3.1.: Negative subjective quality ("0" indicates the best) and average processing time of the 6 enhancement methods

image	LDR	CVC	WAHE	SRIE	LLCRM	ours
1	0.68	0.28	0.68	0	2.37	0.44
2	0.82	0.95	0	0.13	3.95	0.39
3	1.45	0.69	1.75	0	2.78	0.51
4	1.91	1.15	1.23	0.47	2.54	0
5	1.05	1.65	2.19	0	0.53	1.01
6	3.00	3.65	2.46	0.73	2.18	0
time(s)	0.42	4.88	0.41	89.84	1.74	1.81

the average LOM of the peak points in Figure 3.9, 0.127, is set to the threshold value T_{LOM} of LOM that corresponds to the optimal enhancement point. An image that has LOM greater than T_{LOM} is considered to be over-enhanced so that the amount of enhancement should be decreased. Then, Bisection search is used to find the enhancement parameter β that makes the image to have LOM to be T_{LOM}. The maximum β and the minimum β in the search process is defined to map the luminance value of 1 to half of maximum luminance and to luminance value of 2, respectively. The bisection range of β is then calculated to be $[1.2, 127.5]$. Figure 3.10 shows the example of enhanced video frames. Through our observation of the enhanced videos, the temporal consistency is well maintained despite the enhancement is individually applied to each of the frame.

3.6 Conclusions

In this chapter, we propose a controllable enhancement illumination method that allows the degree of enhancement to be adjusted using a single parameter. We then propose an over-enhancement measure, LOM, to evaluate the unnaturalness of en-

hanced images. Our results of subjective test indicates that our enhancement method has a balanced performance in terms of image quality and running time compared to existing enhancement methods. Our proposed over-enhancement measure, LOM, also shows its effectiveness in the subjective test. It reduces the content dependency and provides a score with interpretability compared to existing over-enhancement detection methods.

For future work, one issue is how to improve the illumination within over-exposed regions simultaneously. One difficulty is that over-exposed regions has a loss of details so that the enhancement algorithm should consider how to recover some of details based on the neighboring regions in one image or nearby frames. Another issue is how to design an objective measure for image or video quality after enhancement that provides a consistent evaluation for both different contents and enhancement methods [95].

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3.5.: (a) original image (b) $\beta = 2$ (c) $\beta = 4$ (d) $\beta = 8$ (e) $\beta = 12$ (f) $\beta = 16$ (g) $\beta = 20$ (h) $\beta = 24$

(a)

(b)

(c)

Fig. 3.6.: (a) $LOM = 0.07$ (b) $LOM = 0.10$ (c) $LOM = 0.13$

image 1

image 2

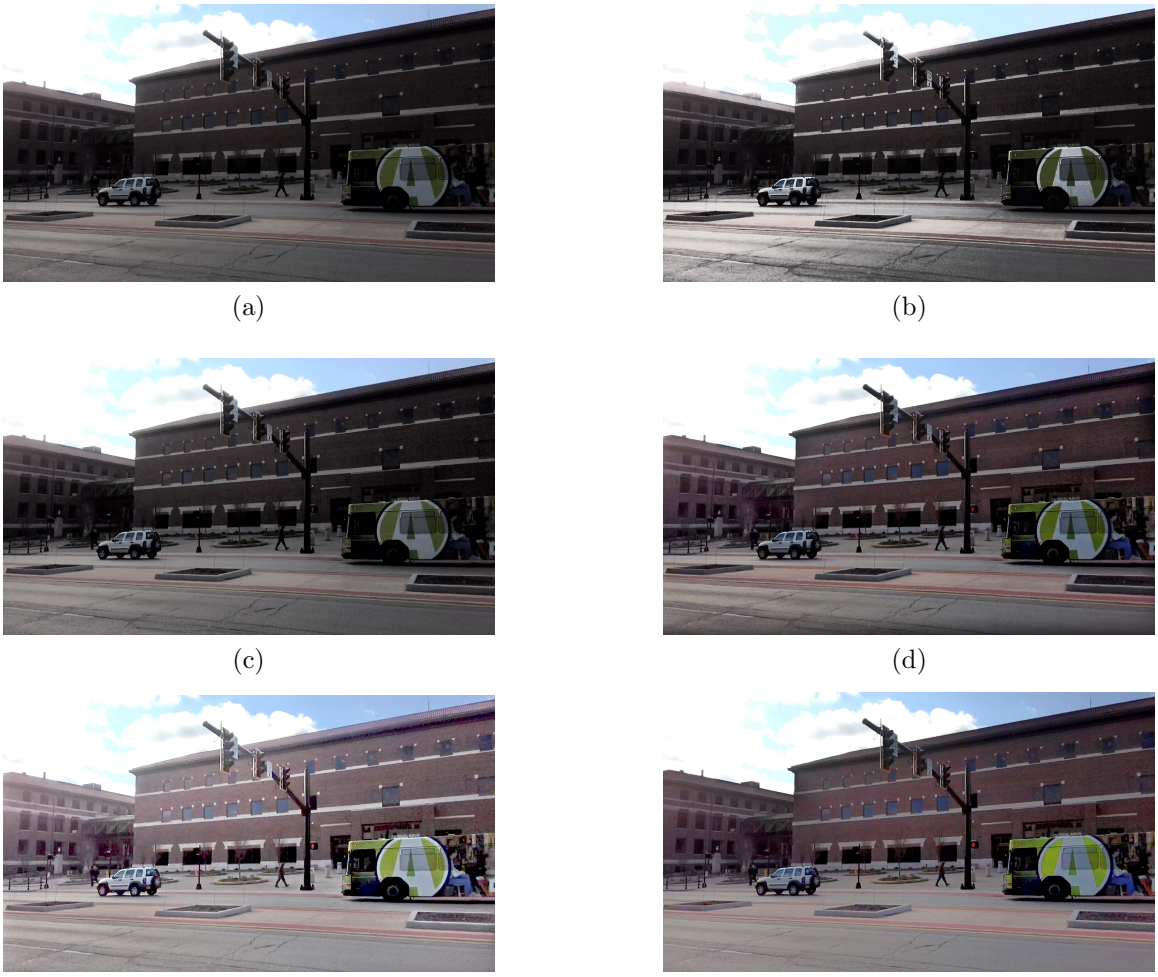
image 3

image 4

image 5

image 6

Fig. 3.7.: Test images: (1) $P_u = 0.35$ (2) $P_u = 0.57$ (3) $P_u = 0.58$ (4) $P_u = 0.76$ (5) $P_u = 0.76$ (6) $P_u = 0.82$



(a) (b) (c) (d) (e) (f)

Fig. 3.8.: Example enhanced images: (a) LDR (b) CVC (c) WAHE (d) SRIE (e) LLCRM (f) ours

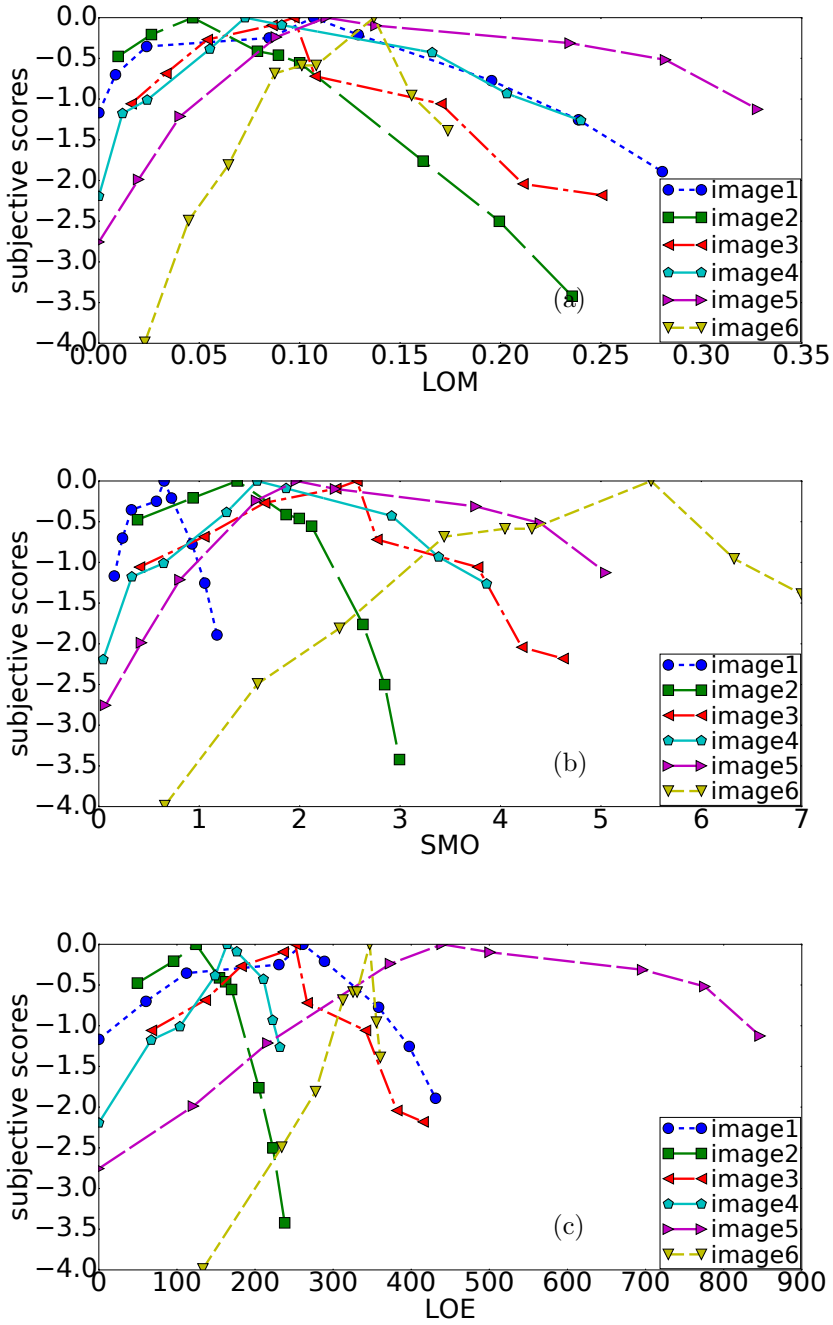


Fig. 3.9.: 9-level enhanced images: subjective quality with (a) LOM (b) SMO (c) LOE

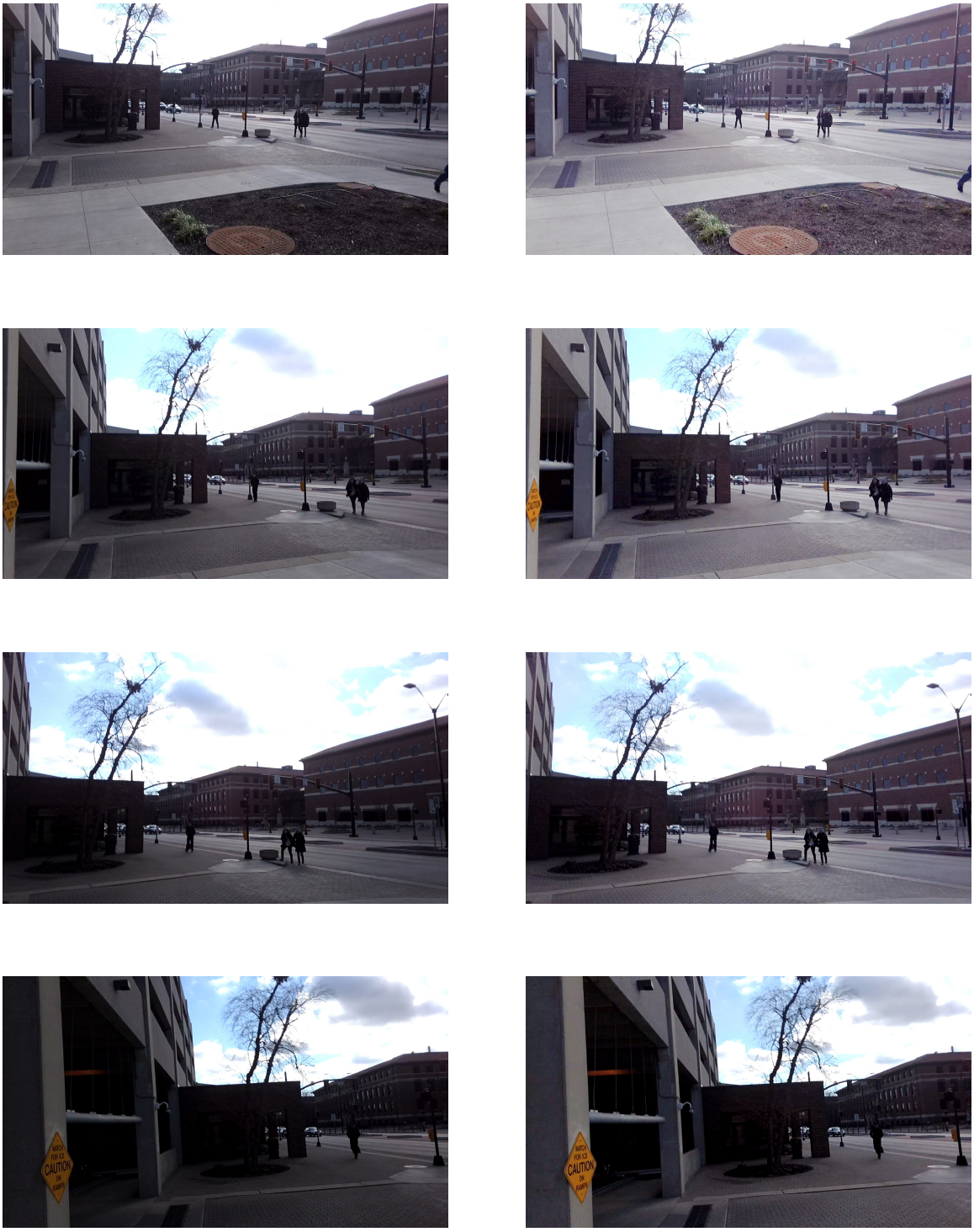


Fig. 3.10.: Video enhancement example: left frames are original, right frames are enhanced.

4. VISIBILITY-INSPIRED TEMPORAL POOLING WITH APPLICATION TO VIDEO STABILIZATION

4.1 Introduction

In the past decade, the development of capturing and recording devices has brought many new types of multimedia. The popularity of recording and sharing mobile videos [120], FPVs, and drone videos [121] has rapidly increased, and interactive live streaming [122] also arouses enormous interest from people. As capture of our daily lives becomes easier, high video quality becomes more important. However, those new types of videos often contain much larger motion than broadcast videos; we refer to them as large motion videos (LMVs). Note in this chapter, we focus on LMVs with FPVs to be a typical example.

LMVs are often low-quality due to camera motion during capture. Since LMVs can be captured using hand-held phones or body-mounted cameras, they are often quite shaky. In contrast, broadcast videos are recorded by stably-mounted cameras that are static or contain low-speed motion. Therefore, motion-induced quality degradations in LMVs are often much worse than broadcast videos [6].

Because of the low-quality nature of LMVs, the assessment of their video quality is important. Our goal is to design a LMV quality estimator (QE) to specifically consider the influence of motion on the perception of quality degradations, which differentiates LMVs from broadcast videos. When an image moves quickly, many details cannot be perceived so that the perceived amount of artifacts becomes smaller than a similar static version [123].

As we will show below, existing video quality estimators (VQEs) are not effective when applied to LMVs. Most VQEs [124] are proposed to measure multiple artifacts including blur, compression artifacts, noise, and they often extract temporal features

that do not measure the influence of large motion magnitude. In addition, most existing methods are only validated using videos that were captured by stably-mounted cameras [83, 125]. On the other hand, the development of still image quality estimators (IQEs) has matured, and they are effective for multiple type of distortions in different contents [19, 20, 126, 127].

To design a VQE, two strategies are commonly used. One strategy is to use an IQE that estimates the quality scores for each individual frame, followed by a temporal pooling mechanism to aggregate frame-level scores into a single video-level score. Another is to extract spatial features as IQEs do, and then combine them with temporal features to be mapped into a video quality score. Temporal pooling strategies are more suitable for the quality assessment of LMVs, because the incremental design allows it to leverage different accurate IQEs, thus avoiding the cost of metric redesign when the type of artifacts changes.

In this chapter, we propose a visibility-inspired temporal pooling (VTP) mechanism [128] for the quality estimation of LMVs, The VTP mechanism combines frame-based spatial quality scores into a video score by considering that the relative importance of an individual frame in the entire video depends on its visibility. We then apply our VTP mechanism to measure the relative perceptual blurriness before and after video stabilization. Existing methods to evaluate stabilized videos focus only on the motion stability [129–131]; however, perceptual blurriness is another important quality factor. The stabilized video is perceptually more blurry than its original for two reasons. First, the stabilization process applies a geometric transformation into each frame that adds spatially variant blur into the frame. The second more important reason is that the blur that already exists within frames, which was caused by camera motion, becomes more visible when there is less frame-to-frame motion. The estimation of quality drop due to increased perceived blurriness can be an important evaluation factor for stabilization algorithms.

Our major contributions in this chapter are:

(1) We propose a visibility measurement that estimates the perceivable proportion of

a frame under a given motion to consider the motion influence on perception.
(2) We consider the pooling weight of each frame to be a function of its visibility, which is measured by carefully constructed subjective experiments.
(3) We apply the VTP mechanism to the scenario of video stabilization, and demonstrate that it effectively measures the relative perceptual blurriness before and after stabilization.
(4) We design a method to synthesize shaky videos in a controlled manner, and we apply this method to create test videos in subjective tests.

Section 4.2 reviews the studies on the window of visibility and discusses existing temporal pooling methods and their weakness when applied to LMVs. Existing objective and subjective quality assessment strategies are also reviewed. In Section 4.3, we describe our proposed visibility measurement for individual frames under a given motion. Then we introduce the visibility-inspired temporal pooling (VTP) mechanism and a method to gather necessary subjective video quality data in Section 4.4. In Section 4.5, we implement a subjective test to gather subjective video quality scores, and validate our VTP mechanism by comparing it with existing temporal pooling mechanisms. In Section 4.6, the VTP mechanism is applied to measure the relative blurriness between videos before and after stabilization, and a systematically designed subjective test is implemented to successfully demonstrate its effectiveness.

4.2 Related works

4.2.1 The Window of Visibility

The visibility of quality degradations during motion has been studied in [25, 38], in which the theory of the window of visibility is proposed. The window represents human visual spatio-temporal contrast sensitivity function (STCSF). The perceivable contrast decides a boundary outside which the spatio-temporal content is not perceivable.

The window of visibility was derived based on the STCSF measured from [132]. [38] approximated the window shape based on isosensitivity coutours of contrast thresholds in STCSF, and indicated that the window of visibility is a simplified representation of spatial and temporal frequencies that are visible to human observers.

Figure 4.1 shows the window of visibility with an example of spatio-temporal content of a moving line, where the x-coordinate is spatial frequency (cycles/degree) and the y-coordinate is temporal frequency (Hz). The positive frequency part of the window is the triangle with three vertices, $(0, 0)$, $(u_0, 0)$ and $(0, w_0)$, where u_0 is spatial frequency limit and w_0 is temporal frequency limit. Consider the motion function of a line: $m(x, t) = \delta(x - rt)$, where x is the position, t is the time, and r is the speed. The transformed moving line in the spatio-temporal domain is determined by $f(u, w) = \delta(w + ru)$, where u and w are spatial and temporal frequency, respectively. $f(u, w)$ is shown as the red line in Figure 4.1, in which the dashed part of the line is the part of $f(u, w)$ outside the window of visibility that cannot be perceived.

The window limits u_0 and w_0 are determined by the display luminance I according to [38]. u_0 is saturated at around 50 $cycles/deg$ at $I = 7cd/m^2$, and w_0 has a linear relationship with display luminance $log_{10}(I)$ that can be approximated as $w_0 = 15 \cdot log_{10}(I) + 35$ Hz.

4.2.2 Temporal Pooling Methods

The process of temporal pooling maps frame-level quality to video-level quality. Average pooling is the simplest strategy. It assumes every frame contributes the same amount to the video quality, so the mean frame score is the video quality. However, human evaluation of video is influenced by the severity of quality degradations, the temporal variation of distortions [133], the temporal hysteresis effect [134], the motion influence and many other factors, so average pooling is not an accurate strategy.

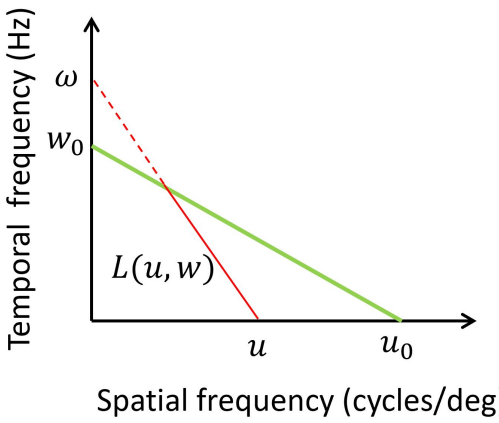


Fig. 4.1.: Green: the window of visibility (u_0, w_0) boundary. Red: spatio-temporal content of $\boldsymbol{u}$ in which the solid line is visible, and the dashed line is invisible

A typical method to assign unequal weights to frames based on the influence of frame quality scores is weighted average pooling:

$$Q = \frac{\sum_i \omega_i \cdot q_i}{\sum_i \omega_i}, \tag{4.1}$$

where q_i and ω_i are the frame quality score and pooling weight for i^{th} frame, respectively, and Q is the video quality score. There are two basic strategies to choose these weights; one considers only the value of frame quality scores, while the other accounts for other quality factors. The first strategy applies a mapping function to combine all frame scores. Typical methods are percentile pooling, Minkowski pooling and LocalMinimum pooling [135]. The percentile pooling method considers only the quality scores below a certain percentage threshold. Minkowski pooling emphasizes high quality frames, while LocalMinimum pooling emphasizes the worst part of the video. Some other temporal pooling also uses parametric functions such as SoftMax [136] or KMeans clustering algorithms [137]. The second strategy considers other quality influences, such as memory or motion. A hysteresis model in [134] emphasizes the memory effects for a human observer. In [133], the temporal pooling strategy measures the influence of the quality temporal variations.

Most current temporal pooling metrics do not consider the influence of motion, so they are not suitable for LMVs. The aforementioned second type of temporal pooling can be extended to our case. By emphasizing the motion influence in the computation of weights assigned to each frame, a temporal pooling mechanism can be applied to LMVs. One example is the human visual speed model proposed in [138].

4.2.3 Objective Quality Assessment

Existing image or video quality estimators are mostly designed to measure the perception of distortions and are built based on human visual system properties. While challenges remain, they have become quite mature in their ability to accurately estimate image quality. Typically, they can be classified into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. FR and RR methods [12, 13, 40, 41] need a high-quality reference image or video, while NR methods estimate the quality of a single image without relying on any reference [42]. A recently proposed new type of method, mutual reference (MR) [46], measures the relative quality between images with overlapping but not necessarily pixel-aligned content.

FR and RR methods are not appropriate to apply to realistic stabilization scenarios. When evaluating a distorted video, FR and RR methods need a corresponding reference that has pixel-aligned frames. Since stabilization operations introduce pixel-misalignment in frames, the original video can not be a reference for its stabilized version using either FR or RR methods.

4.2.4 Subjective Quality Assessment

Subjective quality assessment, in which human observers assess quality, is an effective tool to address two challenges. The first is the performance validation of newly proposed QEs. Many subjective video quality databases have been published for the evaluation of QEs in different applications involving compressed videos [28], camera

videos [29] and high frame-rate videos [30]. The second is obtaining data that can be used to help design QEs by using subjective data for parameter estimation in fitting a mapping model. A subjective test for the latter case should be designed to follow two principles: (1) human observers should be able to perceive the differences between images or videos, (2) extraneous quality factors that are not assessed should be equal. For example in [7], different amounts of rotation or shear have been synthetically injected into images while other types of distortions are maintained to be equal. [31] proposed a video stability estimator in which one parameter is estimated using gathered subjective data. Their test videos are synthetically created to have different motion but no other distortions.

4.3 Visibility Measurement

In this section, we propose a visibility measurement developed using the window of visibility described in Section 4.2.1. We measure the visibility to be the proportion of the overall power spectrum that is inside the window of visibility.

The visibility of frame i patch q, V_{iq}, is considered to be the perceivable proportion of its energy from all spatial frequencies. It is computed as the summation of the fraction of energy over all spatial frequencies, weighted by their visible proportion inside the window of visibility. V_{iq} is then spatially averaged to compute the visibility of frame i, V_i.

Specifically, given an image patch with speed $\boldsymbol{v}$ (where all bold font parameters indicate a vector variable), we have a fixed window of visibility represented as (u_0, w_0). Let $\boldsymbol{u}$ be one spatial frequency in the image patch. We consider only the part of $\boldsymbol{u}$ parallel to $\boldsymbol{v}$ that influences the visibility. Then the temporal frequency w for $\boldsymbol{u}$ is calculated as $w = \boldsymbol{u} \cdot \boldsymbol{v}$. This is illustrated in Figure 4.1, where $\|\boldsymbol{u}\| \cos\theta$, where $\|\boldsymbol{u}\|$ is the length of $\boldsymbol{u}$ and θ is the intersection angle between $\boldsymbol{u}$ and $\boldsymbol{v}$. We compute the fraction of energy, $P(\boldsymbol{u})$ for spatial frequency component $\boldsymbol{u}$, in the image patch to be

$$P(\boldsymbol{u}) = \frac{M(\boldsymbol{u})}{\int_{\boldsymbol{u}} M(\boldsymbol{u})}, \tag{4.2}$$

where $M(\boldsymbol{u})$ is the magnitude of the spatial power spectrum at $\boldsymbol{u}$ in the spatial power spectrum. Not all spatial frequencies $\boldsymbol{u}$ will be completely visible because some lie out of the window of visibility. The energy fraction $P(\boldsymbol{u})$ of spatio-temporal content at $\boldsymbol{u}$ is weighted by its visible proportion $\omega(\boldsymbol{u})$, calculated as

$$\omega(\boldsymbol{u}) = \frac{L(u,w)}{\sqrt{u^2+w^2}}. \tag{4.3}$$

Here, $L(u,w)$ is the length of the visible part shown in Figure 4.1, and $\sqrt{u^2+w^2}$ is the total length. The visibility of image patch q in frame i is then calculated as

$$V_{iq} = \int_{\boldsymbol{u}} \omega(\boldsymbol{u}) P(\boldsymbol{u}). \tag{4.4}$$

The visibility of frame i is spatially pooled from 31×31 patches overlapped by 15 pixels:

$$V_i = \frac{1}{N_q} \sum_q V_{iq}, \tag{4.5}$$

where q is the patch index, N_q is the total number of patches. The measured V_i is not very sensitive to the chosen patch size and the spatial pooling method.

From another point of view, we can also interpret Equation (4.4) by considering $\omega(\boldsymbol{u})$ to be the probability of spatial frequency content u and $P(\boldsymbol{u})$ to be the probability that u is perceivable given $\boldsymbol{v}$, u_0 and w_0. Then the visibility V_{iq} can be interpreted as the probability that the image patch is perceived.

In our actual implementation, the speed $\boldsymbol{v}$ refers to the viewing angular velocity $v_{angular}$ that depends on the viewing distance, and can be calculated using pixel speed v_{pixel}

$$v_{angular} = fps \cdot 2tan^{-1}(\frac{v_{pixel}}{D_{viewing}}), \tag{4.6}$$

where *fps* is the frame rate per second, and $D_{viewing}$ is the viewing distance. In addition, since the window limits u_0 and w_0 depend on the display luminance, we use the gamma correction display model in [139] to transform pixel values into display luminance with gamma value 2.2.

4.4 Visibility-Inspired Temporal Pooling Modelling

In this section, we propose the visibility-inspired temporal pooling (VTP) method that uses the weighted average pooling strategy of Equation (4.1) to combine frame scores to create a video quality score. The pooling weight assigned to each frame relies on the visibility estimated by the procedures described in Section 4.3. Then, we introduce a data gathering strategy to collect subjective data for the modelling and the validation of VTP.

4.4.1 Pooling Method

VTP employs the weighted average pooling strategy to combine frame spatial quality into a video quality. It considers that each frame is not equally important for the entire video, and their importance depends on the visibility. The pooling method is expressed as

$$Q = \frac{\sum_i \lambda(V_i) \cdot q_i}{\sum_i \lambda(V_i)}, \tag{4.7}$$

where q_i and V_i are the spatial quality and visibility of frame i, and Q is the video quality score. $\lambda(V_i)$ is the pooling weight, in which $\boldsymbol{\lambda}(\cdot)$ is a function that can be interpreted as the influence of the estimated visibility V_i on the relative importance of q_i.

4.4.2 Estimating the Function $\lambda(\cdot)$

The function $\lambda(\cdot)$ in Equation (4.7) can be measured using D (for $D > 1$) test video sequences that share the same visibility but have a different spatial quality in the temporal domain. To see this, we write Equation (4.7) to be

$$Q' = \boldsymbol{q}^T \lambda(\boldsymbol{V}), \tag{4.8}$$

where the scaled video quality $Q' = \sum_i \lambda(V_i) \cdot Q$. Let K be the number of frames. Then $\lambda(\boldsymbol{V})$ and $\boldsymbol{q}$ are both $K \times 1$ vectors that represent spatial quality and the

function $\lambda(\cdot)$ of the estimated visibility $\boldsymbol{V}$, respectively. To get the solution of $\lambda(\boldsymbol{V})$, we need to construct D different videos that all have the same visibility but different quality $Q'_1, Q'_2, \ldots, Q'_D$. Then we can apply least squares to find the solution $\lambda(\hat{\boldsymbol{V}})$ for $\lambda(\boldsymbol{V})$ in Equation (4.8) to be

$$\lambda(\hat{\boldsymbol{V}})^T = \begin{bmatrix} \boldsymbol{q_1}^T \\ \boldsymbol{q_2}^T \\ \vdots \\ \boldsymbol{q_D}^T \end{bmatrix}^{\dagger} \begin{bmatrix} Q'_1 \\ Q'_2 \\ \vdots \\ Q'_D \end{bmatrix}, \tag{4.9}$$

where † is the Moore-Penrose pseudo-inverse. The spatial quality values $\boldsymbol{q_1}, \boldsymbol{q_2}, ..., \boldsymbol{q_D}$ can be estimated using any existing IQE. After gathering the video quality $Q'_1, Q'_2, ..., Q'_D$, we can estimate $\lambda(\cdot)$.

4.4.3 Data Gathering Strategy

To estimate the $\lambda(\cdot)$ using procedures described above, we need subjective and objective measurements of the quality of D videos, all of which have same visibility but different spatial quality. However, using a camera to capture D such videos is challenging, because camera motion affects both the visibility and the spatial quality of individual frames. To decouple these two quantities, we choose to create D videos synthetically. Specifically, we take a large, high-quality still image to create a static video. First, we inject synthetic motion blur into the video to create the desired spatial quality. Then we crop the frame with a moving window to create a video with the desired motion.

Two approaches are possible to form the collection of D videos. The first is to have different levels of constant blur and adjust the motion to create equivalent visibility for the collection of videos. However, this would require large differences between the amounts of injected motion, which would create significant cognitive load during the subjective test. Therefore, we choose the second approach, which is to add a time-varying amount of blur to each video, and adjust the motion to achieve the

desired equivalent visibility. The same sinusoidal amount of blur is added into each video, but the blurs are temporally shifted from one video to the next. The resulting videos have similar amounts of motion, making subjective testing a simpler task for the subjects.

The objective quality of these D videos' frames can be obtained using any IQE. Clearly, the choice of IQE will affect the estimated $\lambda(\cdot)$. In this chapter, we use LVI [46] to estimate frame quality to estimate $\lambda(\cdot)$. LVI has been demonstrated to be effective at providing a consistent measure for blur [7].

Based on this strategy, the creation of a test video j needs the information of a motion profile A_j, a blur profile B_j, and a visibility profile P_j. The profiles describe the pixel shifts (A_j), the average filter kernel size (B_j) and the estimated visibility (P_j) temporally for each frame in a video. Assume a blur profile B_0 and a visibility profile P_0 where $P_0 \propto -B_0$. If B_0 is temporally shifted to a new blur profile B_0' while P_0 is maintained, there would be less masking effect for B_0'. Figure 4.2 shows the comparison between B_0, B_0' and P_0 in which $1 - P_0 = c \cdot B_0$ with c to be a constant parameter. When a temporal shift is introduced into B_0 to blur profile B_0', the overlap between B_0' and $1-P_0$ becomes smaller so that more blurry frames would have higher visibility. The intuition here is that the motion has a masking effect, as measured by the decreased visibility, on frame perceptual blurriness so that the perceived video quality increases. To inject the temporal shift to the video in a controlled manner, we shift the phase of the blur profile B_0 in the frequency domain. By shifting phase 0.125π, 0.25π, 0.375π, 0.5π, B_0 becomes new profiles B_1, B_2, B_3 and B_4, respectively. The A_0 is then edited to become A_j, for $j = 0, 1, 2, 3, 4$, in which the video visibility profile is maintained to be constant, P_0. A test set is then formed with videos created by (B_j, A_j), for $j = 0, 1, 2, 3, 4$.

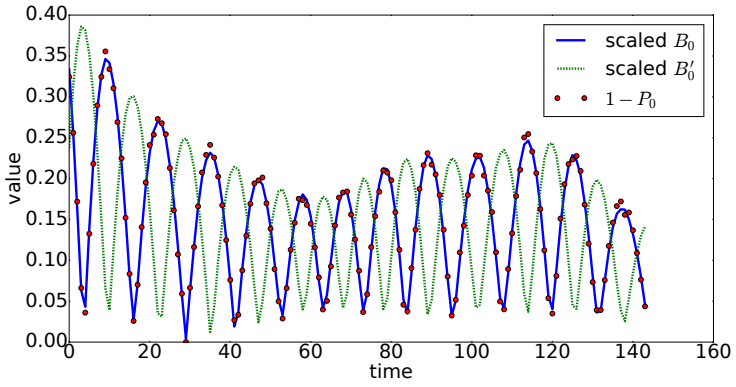


Fig. 4.2.: Comparison between Blur profile B_0, B_0' and visibility profile $1 - P_0$

4.5 Subjective test for visibility-inspired temporal pooling

In this section, we describe our subjective test using synthetic shaky videos. The test results from one source video are used to estimate the function $\lambda(\cdot)$ using method described in Section 4.4.2. Then, our pooling strategy is applied to three other source videos, and is demonstrated to perform the best when compared to different temporal pooling methods across a range of existing image quality estimators.

4.5.1 Test Video Sets

To create synthetic videos, we start with 4 high-resolution images corresponding to test sets Γ_j, where $j = 0, 1, 2, 3$. Videos are created by moving the cropping window in the original image using the strategy described in Section 4.4.3. The injected pattern of motion are extracted from one actual captured shaky video. Γ_1 and Γ_2 are created with frequency range between 1 and 2 Hz, and the frequency range for Γ_3 and Γ_4 is between 2 and 3 Hz. Each test set has five videos with blur phase shift $0, 0.125\pi$, 0.25π, 0.375π, 0.5π. All test videos with their corresponding reference videos and the video that is used to inject motion are available at [140].

To synthetically create videos in set Γ_j, we first obtain a motion profile A_j and then compute the corresponding blur profile B_j and visibility profile P_j. Assume we

want to have a motion profile A_j in the frequency range from a Hz to b Hz, where a and b are constants, we introduce a method to inject motion by extracting an A_j from a real shaky video:

1. Extract motion frequency spectrum F_M from one actual captured shaky video.

2. From a Hz to b Hz in F_M, find the peak frequency f_{peak}, and apply a Gaussian filter centered around f_{peak} to create a frequency spectrum $F_M(a, b)$ from a Hz to b Hz.

3. Transform $F_M(a, b)$ into motion in the time domain to create the motion profile A_j. At time t, the motion v_t can be

$$v_t = \sum_f \omega_f \sin(2\pi f t + \phi(f)) - \sum_f \omega_f \sin(2\pi f(t-1) + \phi(f)) \tag{4.10}$$

   where ω_f is motion magnitude corresponding to frequency component f in $F_M(a, b)$. $\phi(f)$ is the corresponding phase in $F_M(a, b)$. Note v_t can be either horizontal motion x_t or vertical motion y_t.

Next we compute blur B_j based on the motion A_j in which the window length of the average blur filter is proportional to the pixel displacement. P_j is computed based on B_j and is then used to edit A_j to get A'_j. Now, (B_j, A'_j) is the motion and blur information to synthetically create a video. Specifically,

1. Create motion profile A_j that at each time $A_j(t) = [x_t, y_t]$. $A_j(t)$ refers to angular velocity that requires the information of viewing distance and video frame rate.

2. Create corresponding blur profile B_j with filter window length

$$B_j(t) = [\max(1, x_t), \max(1, y_t)]. \tag{4.11}$$

3. Create visibility profile P_j so that

$$P_j(t) = \max(0, 1 - q \cdot \sqrt{x_t^2 + y_t^2}), \tag{4.12}$$

where q is a constant parameter. Note that P_j has the same temporal change as B_j so that the larger the amount of blur, the lower the visibility.

4. For frame at t, we search for the motion $A'_j(t) = [x'_t, y'_t]$ so that the frame with $A'_j(t)$ is measured to have visibility $P_j(t)$. Let $A'_j(t) = \alpha \cdot A_j(t)$, bisection search is applied to find the value of α that satisfy the condition for $A'_j(t)$.

4.5.2 Test Setup

Our subjective test method is paired comparison. All pairs to be compared are videos in the same set Γ_j. A pair of test videos is presented one after another on a monitor (DELL U2718Q) that has resolution 3840×2160. The video is presented at the center of the screen with resolution 1920×1080. The background is gray at 128. Each test video is 5 seconds with frame rate 30 frames/second. Since the calculation of the visibility relies on the viewing distance, it is fixed to be 3.2 height of the video. Each of the 20 test participants are asked to choose *in which video can you perceive more spatial details*.

4.5.3 Subjective Test Results

The relative subjective qualities are estimated using the Bradley-Terry Model [37]. The test results are shown in Figure 4.3 where the best quality is 0 for each test content.

The subjective results indicate that a larger phase difference between visibility and blur introduces more perceived quality degradations for a human observer in all four test contents. This demonstrates that the window of visibility does have a masking effect on the perception of blurriness; low quality frames have little influence when they have low visibility.

One additional comment about content differences is that content 1 and 2 show greater quality differences between videos with phase shift 0 and 0.5π than content

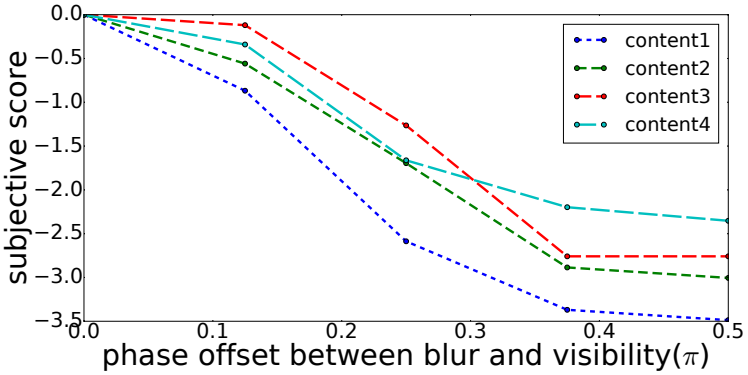


Fig. 4.3.: Subjective scores (0: best quality in each test set)

3 and 4. One reason is that content 1 and 2 have lower-frequency motion than do content 3 and 4. Content 1 has much greater quality difference between videos with phase shift 0 and 0.5π than other contents, because it contains a higher proportion of regions with high spatial frequencies that enable the differences to be more perceivable.

4.5.4 Estimating $\lambda(\cdot)$

We estimate the function $\lambda(\cdot)$ using the method illustrated in Section 4.4.2. We apply the subjective results from the four contents to estimate $\lambda(\cdot)$, and choose the estimated model using content 1 because it achieves the highest PLCC between $\boldsymbol{V}$ and $\lambda(\hat{\boldsymbol{V}})$ among the four contents.

The temporal weighting vector $\lambda(\hat{\boldsymbol{V}})$ is calculated by Equation (4.9), where $\boldsymbol{Q}$ is the subjective quality scores of the five test videos of content 1. Vector $\boldsymbol{q}$ is estimated by LVI [46]. Figure 4.4 shows the comparison between $\boldsymbol{V}$ and the estimated weighting vector $\lambda(\hat{\boldsymbol{V}})$. We fit function $\lambda(\cdot)$ using the logistic function.

$$f(x) = (t_0 - t_1)/(1 + \exp(-(x - t_2)/|t_3|)) + t_1 \tag{4.13}$$

Then we normalize the values after mapping, where the maximum value and minimum value for normalization is $f(1)$ and $f(0)$. The estimated $\lambda(\cdot)$ shown in Figure 4.5 maps measured visibility to pooling weight with fitted parameters $t_0 = 0.26, t_1 =$

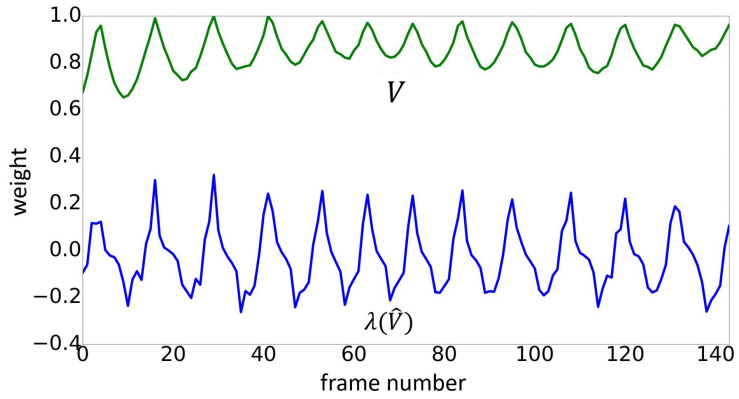


Fig. 4.4.: Comparison between $\boldsymbol{V}$ and estimated $\lambda(\hat{\boldsymbol{V}})$

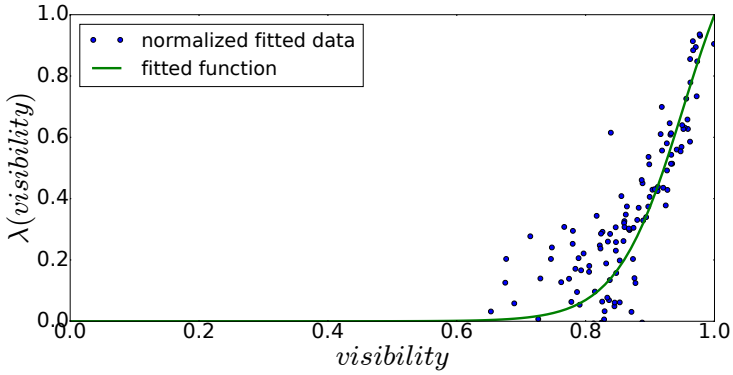


Fig. 4.5.: Function $\lambda(\cdot)$ in Equation 4.1: x-axis is measured visibility V_i, y-axis is $\lambda(V_i)$.

$-1.25, t_2 = 0.95, t_3 = -0.05$. Our measure for visibility is shown to have an nonlinear relationship with the pooling weight in Figure 4.5.

4.5.5 Evaluating the Overall Method

The function $\lambda(\cdot)$ is estimated using one test video content, while our VTP strategy is validated using the other three test video contents. We compare our method with existing pooling strategies: average pooling, percentile pooling (70th), Minkowski pooling (p=2), speed pooling [138], temporal variation pooling [133], and hysteresis

pooling [134]. In our implementation, the relative speed is zero in speed pooling, since all our test videos only contain global motion. In addition, we only consider the global temporal pooling method in [133] and set the distortion value to be the negative quality value plus the maximum quality value of the quality metric. To test the generality for different IQEs, we estimate the frame quality using four FR IQEs (SSIM [126], GSM [49], VSNR [127], VSI [141]), one mutual reference IQE, LVI [46], and two NR IQEs (BRISQUE [19], NIQE [20]). All quality scores are normalized to be between 0 to 1 using the minimum and maximum values in [44].

Table 4.1 shows the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SROCC) between the subjective video quality scores and the objective temporal pooling scores using different IQEs. For all three test contents, our pooling method shows the best overall performance.

Our VTP mechanism can achieve high PLCC and SROCC for two reasons. First, because of the limited number of test samples, PLCC and SROCC mainly measure whether the method correctly ranks the video quality. Second, the subjective test and our proposed method are both specifically designed for the masking effect on perceived blurriness due to motion.

The results also show that our method can generalize across different contents. Our method incorporates the influence of content since our estimation of visibility computes spatio-temporal information in a single frame. In addition, we model the relationship $\lambda(\cdot)$ between visibility and pooling weight based on gathered subjective data that has better cross-content performance than considering $\lambda(\cdot)$ to be linear.

Our VTP mechanism is not successful when pooling either BRISQUE and NIQE for content 2. BRISQUE and NIQE do not provide a consistent measure when the same amount of blur is added into pixel-shifted content. The test videos in content 2 are produced with greater frame-to-frame pixel shifts than content 3 and 4, so the BRISQUE and NIQE scores of content 2 are not as robust as in other contents.

Speed pooling has the second best performance among all. It computes temporal weights based on motion, but their model parameters are only evaluated on videos

with low-speed motion. The other 5 methods are not suitable for our situation of LMV. They pool the video quality using only frame scores. However, our videos are created to have similar frames scores with different visual qualities, so these methods cannot capture all the relevant information.

4.6 Measuring perceptual blurriness after video stabilization

In this section, we demonstrate the effectiveness of applying the VTP mechanism to measure the perceptual blurriness in stabilized videos using the gathered subjective data. As motion decreases, the blurriness becomes more visible. Therefore, the blurriness of a stabilized video is perceptually more severe than its original video. Instead of only improving the motion stability, the video quality drop due to increased blurriness should also be considered in the design of stabilization algorithms.

Section 4.6.1 illustrates the strategy of gathering subjective evaluation of shaky videos and their stabilized versions. We synthetically create shaky videos by adding synthetic motion and blur into high-quality stably-captured videos. Then, Section 4.6.2 introduces a motion-frequency method to inject real shaky motion to create test videos. In Section 4.6.3, the test methodology and setup are described. Section 4.6.4 and Section 4.6.5 show the subjective test results and validate the performance of the VTP mechanism, respectively.

4.6.1 Test Motivation and Strategy

To validate the effectiveness of our method in measuring the relative blurriness before and after stabilization, we need to gather the subjective evaluation of perceptual blurriness of shaky videos and their stabilized versions. In order to have videos independent of specific stabilization algorithms and their motion estimation strategies, we synthetically create shaky videos by adding ideal motion to a high-quality stable video instead of applying a stabilization algorithm to captured shaky videos.

The test videos need to follow the principles for subjective testing, in that they should have perceivable blur differences and equal quality factors that are not assessed. In order to evaluate only the blurriness, our test individually compares two sets of videos, where within each set the videos have the same motion. The first set Λ_u contains unstable (shaky) videos with different amounts of motion blur, the second set Λ_s contains stable videos that are stabilized versions of Λ_u. We gather the relative subjective scores, independently from both sets. Then, with one estimate of the relative quality between a single pair of videos in Λ_u and Λ_s, the relative quality between any pair from both sets can be computed.

The creation process of Λ_u and Λ_s is shown in Figure 4.6. Stable high-quality videos are our source videos. To create the Λ_u, we first add shaky motion into a source video to create the synthetic shaky video V_u^0; the motion transformation $\boldsymbol{T}$ is a set of geometric transformations for each single frame. Then, different amounts of blur are added into V_u^0 to obtain shaky test videos $V_u^0, V_u^1, V_u^2, V_u^3, V_u^4$. The set Λ_u consists of the shaky reference and the four shaky test videos, all with identical motion. To create Λ_s, we apply $\boldsymbol{T}^{-1}$, the inverse transformation of $\boldsymbol{T}$, to every video in Λ_u. Λ_s consists of the stable reference V_s^0 and the four stable test videos $V_s^1, V_s^2, V_s^3, V_s^4$. V_s^j and V_u^j form a stabilization pair, where $j = 0, 1, 2, 3, 4$. Note that $\boldsymbol{T}^{-1}$ is considered to be the stabilization process. By applying a known transformation $\boldsymbol{T}$, our stabilization using $\boldsymbol{T}^{-1}$ is well defined and identical for each video in Λ_s.

After gathering the subjective quality of Λ_u and Λ_s separately, we need to know the relative quality of an anchor stabilization pair to compute the quality differences of the other four pairs. V_u^0 and V_s^0 is chosen to be the anchor pair because of the convenience to compute their quality difference. Let the blurriness of V_u^0 to be $Q_s^0 = 1$ and V_s^0 have blurriness Q_b^0. Since the V_s^0 has little motion after stabilization, the impact of motion on visibility is negligible, so only its frame quality influences $Q_s^0 - Q_b^0$. By applying a specific IQE, we can measure the relative quality $Q_s^0 - Q_b^0$. Then the relative quality among all videos in Λ_u and Λ_s can be computed.

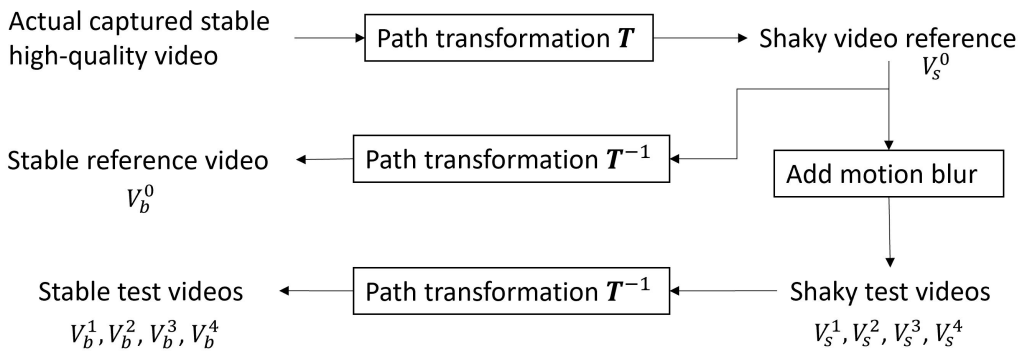


Fig. 4.6.: Diagram of creating test set: Λ_u and Λ_s

4.6.2 Synthetic creation of shaky videos

To synthetically create a set of shaky videos, Λ_u, that simulates the actual captured motion, we introduce a frequency-based method that injects the pattern of motion that is present in an actual captured shaky video into a stably-captured source video. Note this subsection only describes our method for injecting shaky motion into the video, not for also adding blur.

First, we extract motion-frequency information of three rotational components, yaw, pitch and raw, from a shaky video S that was captured using a real camera. Because of the low-quality nature of S, it is often not possible to accurately and directly estimate the three rotational components. Instead, we use 2D translation x_t and y_t at time t to approximate yaw and pitch at a given viewing distance and rotation θ_t to approximate roll. We estimate using the frame-to-frame homography decomposition in [6]. x_t and y_t is transformed to angular yaw velocity v_{yaw} and pitch velocity v_{pitch} by setting a standard viewing distance, 3.2 times the height of the image, using Equation (4.6), and angular roll velocity v_{roll} is directly computed as $fps \cdot \theta_t$. Specifically, v_t^{yaw} is the cumulative yaw rotation at time t. By applying discrete Fourier transform into yaw rotation in S to get its temporal spectrum f^{yaw}.

Then, peak frequencies every 1Hz frequency range is picked in f^{yaw} to compute v_t^{yaw} as

$$v_t^{yaw} = d \sum_{i=1}^{N} w_i(\sin \omega_i t + \phi_i), \tag{4.14}$$

where d is the magnitude scale factor in time domain, w_i and ϕ_i are the magnitude and phase of frequency range i in f^{yaw}. v_t^{yaw} represents the cumulative yaw rotation between frame 0 and frame t, v_t^{pitch} and v_t^{roll} can be derived using the same way as v_t^{yaw}.

Then, we want to create a motion transformation $\boldsymbol{T}$ that transforms the stably-captured source video with existing camera path D_t to have this estimated motion from S. From the estimated angular motion velocities in S, we can obtain the desired shaky camera path C_t. Let E_t be the transformation that warps the stable video frame to egocentric frame at time t, then we can compute E_t as follows:

$$C_t = E_t D_t \tag{4.15}$$

The relationship between frames in C_t and D_t can be modeled as

$$C_t = G_{t-1} C_{t-1}, \tag{4.16}$$

$$D_t = J_{t-1} D_{t-1}, \tag{4.17}$$

where G_t and J_t are frame-to-frame transformation for C_t and D_t. Then we can get

$$C_t = G_{t-1} G_{t-2} ... G_0 C_0 \tag{4.18}$$

$$D_t = J_{t-1} J_{t-2} ... J_0 D_0 \tag{4.19}$$

E_t can be calculated based on the relationship that $C_0 = D_0$,

$$E_t = C_t {D_t}^{-1} = (G_{t-1} G_{t-2} ... G_0)(F_{t-1} F_{t-2} ... F_0)^{-1}. \tag{4.20}$$

G_t can be decomposed as

$$G_t = K R_t [I | T_t] K^{-1}, \tag{4.21}$$

where K is the intrinsic matrix of the camera, T_t is kept as the same translation as original in D_t, R_t has three rotational components, yaw, pitch and roll. Then, the actual R_t is computed as

$$R_t = r_t {r_{t-1}}^{-1}, \tag{4.22}$$

where r_t is the 3D rotation matrix created by v_t^{yaw}, v_t^{pitch} and v_t^{roll}. E_t is applied to each frame in the stable source video to create Λ_u. With this E_t, the Λ_u will have motion-frequency characteristics similar to the desired motion in S. The inverse of E_t is then applied to videos in Λ_u to simulate the stabilization process to create videos in Λ_s.

It should be noted that Section 4.5.1 synthesizes the video using 2D motion for a cropping window in an image, while here we use 3D motion transformation applied to a video. Another difference is how motion-frequency information is extracted from an actual captured shaky video, Section 4.5.1 extracts motion within a frequency range such as 1 to 2 Hz, while here we extract from the full frequency range.

4.6.3 Test Description

Test video sources are high-quality stable videos captured using a GoPro6 (4k resolution, wide field of view, 30fps) mounted on a tripod. By smoothly moving the tripod, we can record stable videos with either forward or panning motion. Three high-quality stable videos captured on the Purdue University campus, a grocery store, and some apartment buildings are selected for creating test videos. For each of the three contents, Λ_u and Λ_s are synthetically created using the methods described in Sections 4.6.1 and 4.6.2, and all videos are cropped into resolution 1920 by 1080. It should be noted that the synthetic motion blur is added into the video using the method in Section 4.5.1. The test videos and the actual captured shaky videos that are used for injecting motion are available at [142].

The Double Stimulus Impairment Scale (DSIS) method is chosen for the experiment. Pairs of videos are displayed sequentially. 20 test participants are asked to

rate the blurriness of the second video compared to the first video. The rating scale is from 1 to 9 (9 is the best), and viewers are informed that the first video is a reference video that has quality score 9. The test environment is the same as described in Section 4.5.2. For the stabilized videos in Λ_s, V_s^0 is the reference video and V_s^j are the videos to be rated, where j refers to the five blur levels. Note that we also include the case where the reference videos appears as the test videos. We apply the same strategy to gather subjective data for the shaky videos in Λ_u.

4.6.4 Subjective Results

To apply statistical analysis on the subjective data, we normalize subjective ratings using the maximum and minimum scores of each participant. Then, an outlier detection is applied to remove subjective ratings that deviate more than two standard deviations from the mean (95% confidence interval). Two participants' ratings are considered to be invalid and removed. The Mean Opinion Score (MOS) is calculated as the mean of the subjective ratings:

$$MOS_{ij} = \frac{1}{N}\sum_k s_{ijk} \tag{4.23}$$

where i is the video index, j is the blur level index, k is the participant index, and N is the number of valid participants. s_{i0k} is the reference video subjective score for participant k. The 95% confidence interval of MOS is given by $[MOS_{ij} - \delta_{ij}, MOS_{ij} + \delta_{ij}]$, where δ_{ij} is calculated as

$$\delta_{ij} = 1.96\sqrt{\sum_k \frac{MOS_{ij} - s_{ijk}}{N(N-1)}} \tag{4.24}$$

Figure 4.7 shows the MOS of the three contents, with the video of blur level j defined to be V_s^j. By comparing the slopes of the two sets, we see that the shaky videos have smaller visual differences in adjacent blur levels than do the stable videos. For example in Figure 4.7 (a), the difference between V_u^2 and V_u^3 is smaller than that between V_s^2 and V_s^3. It can be interpreted as when the same amount of blur is

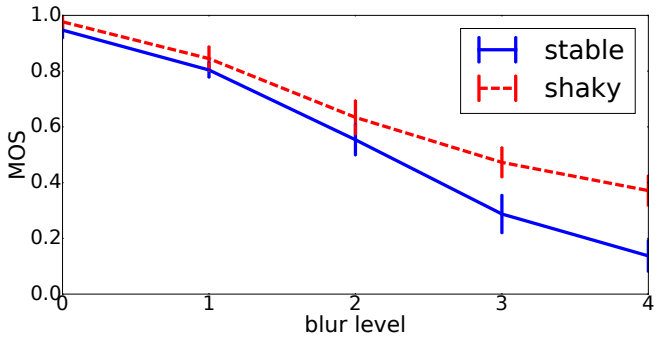
added, an unstabilized video has less visual blurriness increase than its stabilized version. Therefore, it can be concluded that stabilized videos are perceived to be more blurry than the unstabilized videos, when each have the same amount of blur. This corresponds to our assumption that the visibility of blur decreases when the video contains more motion. One exception is the blur level 3 and 4 in Figure 4.7(c); the score of the stable video, V_s^4, is 0.027, which almost reaches the worst possible quality of 0. The scale limitation restricts the slope.

The subjective differences between a reference and itself should always be zero. However, we see from Figure 4.7 that the reference video in the Λ_s is rated as having lower subjective quality than the reference video in the Λ_u. Therefore, we will eliminate this difference in our analysis. As our goal is to evaluate the perceptual blurriness between unstabilized and stabilized videos, we need to estimate the subjective differences between Λ_u and Λ_s. Our solution is to use objective quality measures. In actual implementation, V_u^0 is considered to be the reference video for both Λ_u and Λ_s, and objective quality metrics are used to estimate the visual blur differences between V_u^0 and V_s^0. Therefore, the subjective scores in Λ_u are maintained, while scores in Λ_s are adjusted using the estimated difference $(Q_{obj}(V_u^0) - Q_{obj}(V_s^0)))$, where $Q_{obj}(\cdot)$ is the objective video quality score. Note that all scores from $Q_{obj}(\cdot)$ should be normalized to have the same scale as the subjective scores.

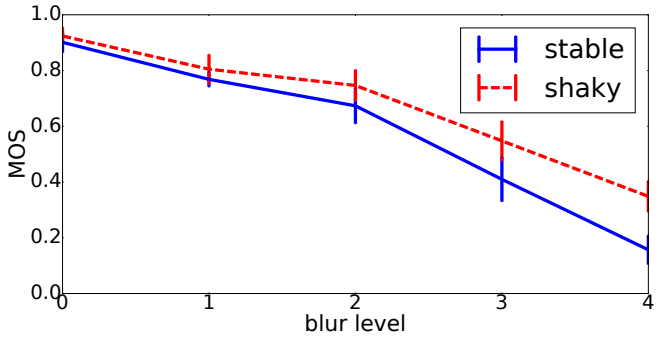
4.6.5 Method Validation

In this subsection, the effectiveness of the VTP mechanism in measuring the relative perceptual blurriness before and after video stabilization is validated using the subjective results from Section 4.6.4. We demonstrates that the VTP can estimate the perceptual blurriness of the combination set of shaky and stable videos.

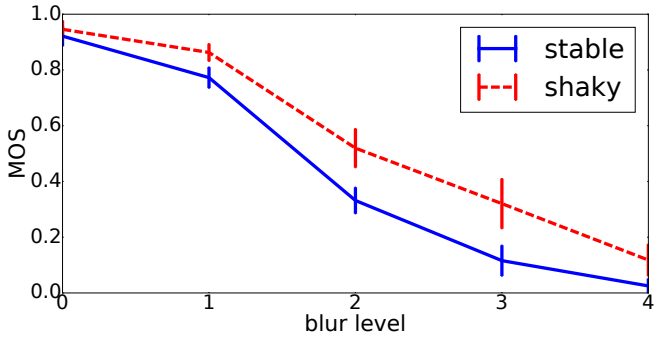
We apply our VTP to estimate the objective quality of test videos. Three NR IQE, BRISQUE, OG [143] and NIQE, and one mutual reference IQE, LVI, are used to estimate frame spatial quality in VTP. We also compare our method with two



(a)



(b)



(c)

Fig. 4.7.: (a) Campus (b) Grocery (c) Apartments (blur level j refers to V_s^j). The vertical bar is the corresponding confidence interval.

existing NR VQEs, V-BLIINDS [144] and VIIDEO [145]. Note that no FR methods are applied here since they cannot be used to compare videos before and after stabilization.

Table 4.2 shows the SROCC and PLCC between the MOS and the objective video quality scores. NIQE-VTP and LVI-VTP outperform all other video quality measures; both show high SROCC and PLCC for the three contents. OG-VTP and BRISQUE-VTP show good performance for contents Grocery and Apartment, however, neither are effective for content Campus. The content Campus has larger viewing angle change than the other two contents so that it contains larger frame-to-frame pixel shifts within a single test video. Since OG and BRISQUE have less consistency when measuring different contents, they do not perform as well on Campus.

NIQE-VTP and LVI-VTP demonstrate their ability to estimate the perceptual blurriness between shaky videos, between stable videos, and between pairs of them. Therefore, both VTP can be effectively applied to estimate the relative perceptual blurriness between a shaky video and its stabilized version. The estimated increase of blurriness after stabilization can be a quality evaluation factor in designing stabilization algorithms.

The VTP also shows its generalization ability using different IQEs. Since its pooling function is modelled by the IQE, LVI, that provides effective quality measure of blur. If VTP employs an IQE that has consistent measure as LVI does, the VTP is then shown to have similar and good performance, such as NIQE-VTP.

Two existing NR VQEs, V-BLIINDS and VIIDEO, are effective for assessing the perceptual blurriness of the stable videos, but not of the shaky videos or between a shaky video and its stabilized version, because the V-BLIINDS and VIIDEO are designed and tested only on stably-captured videos. Figure 4.8 that shows the comparison between LVI-VTP and V-BLIINDS. V-BLIINDS correctly ranks the set of stable videos Λ_s (circle points), while two pairs of videos are ranked falsely in the set of shaky videos Λ_u (triangle points). Because the inaccuracy in measuring the shaky videos, the difference between a shaky video and its stabilized version is not effectively estimated by V-BLIINDS. In contrast, LVI-VTP shows good performance in estimating the visual blur differences among shaky videos, and between a shaky video and its stabilized version.

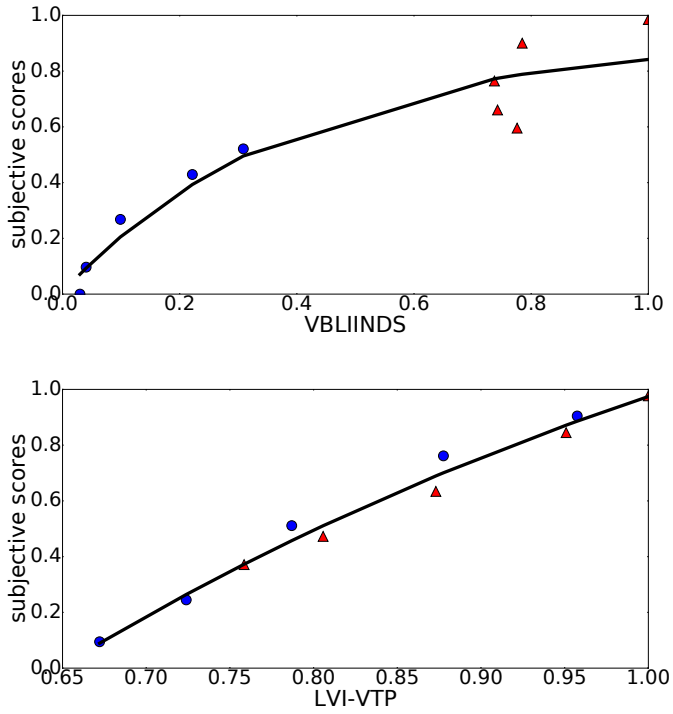


Fig. 4.8.: Objective scores versus subjective scores (circle points: stable videos Λ_s, triangle points: shaky videos Λ_u, black line: fitting curve)

4.7 Conclusions

In this chapter, we propose a visibility-inspired temporal pooling method (VTP) built on a measurement of visibility that is more effective at estimating the quality of LMVs than existing pooling strategies. The VTP is also demonstrated to be effective for the application of measuring the relative perceptual blurriness between videos both before and after stabilization.

The VTP method employs a weighted average pooling where the weight is calculated as a function of visibility. The measurement of visibility considers the fraction of visible details within a single frame under a given motion based on the window of visibility. A systematic subjective test is implemented to model the pooling function using our proposed visibility measure. The test results also indicate that our pooling

strategy is more suitable for LMVs and can be effectively applied to pool quality scores estimated by different types of image quality metrics.

The measurement of perceptual blurriness before and after video stabilization is validated by another subjective rating test, using synthetically created shaky and stable videos. The subjective results demonstrate that shaky videos have smaller visual blur differences than stable videos because the visibility increases as motion decreases. By applying the VTP to existing image quality metrics and comparing with existing VQEs, we demonstrate the effectiveness of measuring the blurriness among stable videos, among shaky videos, and between a shaky video and its stabilized version.

There also exists two additional potential applications for our VTP mechanism. First, we can compare the perceptual quality of videos that are simultaneously captured by multiple shaky cameras [146]. Second, the perceptual quality of a LMV after post-processing can be compared relatively to its original. One example is a LMV after illumination enhancement. The newly generated artifacts [22, 95] in its frames may be imperceivable due to the low visibility.

The future work is to investigate how to improve the design of the VTP mechanism. One potential improvement is to locally pool the image quality measure using the visibility as a spatial mask. Another potential improvement is to consider the influence of the temporal quality variation. In addition, more video content with different motion can be also be tested to further optimize our method.

Table 4.1.: PLCC (SROCC) between objective pooling scores and subjective scores.

Content 2

Pooling method	SSIM	GSM	VSNR	VSI	LVI	BRISQUE	NIQE
average	0.75(0.7)	0.7(0.6)	0.86(0.7)	0.48(0.6)	0.74(0.7)	0.49(0.3)	0.84(0.6)
Minkowski	0.7(0.6)	0.69(0.6)	0.84(0.7)	0.46(0.6)	0.78(0.7)	0.66(0.7)	0.82(0.5)
percentile	0.83(0.6)	0.77(0.5)	0.99(0.9)	0.24(-0.2)	0.86(0.6)	-0.91(-0.9)	0.71(0.4)
speed [138]	0.94(0.9)	0.92(0.9)	0.97(0.9)	0.90(0.9)	0.94(0.9)	0.76(0.8)	**0.95(0.9)**
hysteresis [134]	0.57(0.6)	0.55(0.6)	0.87(0.9)	0.70(0.6)	0.62(0.7)	**0.86(0.9)**	0.59(0.7)
variation [133]	0.76(0.6)	0.72(0.6)	0.86(0.7)	0.48(0.5)	0.74(0.7)	0.48(0.3)	0.86(0.6)
VTP	**0.99(1.0)**	**0.98(1.0)**	**0.98(1.0)**	**0.99(1.0)**	**0.99(1.0)**	0.64(0.6)	0.81(0.7)

Content 3

Pooling method	SSIM	GSM	VSNR	VSI	LVI	BRISQUE	NIQE
average	-0.46(0.1)	-0.52(0.1)	-0.74(-0.7)	-0.03(0.1)	0.04(0.1)	-0.58(-0.3)	-0.2(0.0)
Minkowski	-0.51(-0.4)	-0.52(0.1)	-0.96(-0.9)	-0.06(0.1)	0.1(0.1)	-0.55(-0.3)	-0.2(0.0)
percentile	-0.09(0.1)	-0.36(0.1)	0.72(0.7)	0.36(0.4)	-0.3(0.0)	-0.77(-0.9)	0.05(0.3)
speed [138]	0.80(1.0)	0.77(1.0)	0.51(0.7)	0.84(1.0)	0.65(0.6)	-0.27(-0.1)	0.41(0.2)
hysteresis [134]	0.82(0.7)	0.87(0.9)	-0.38(-0.3)	0.87(0.9)	0.37(0.3)	0.36(0.5)	0.61(0.5)
variation [133]	0.76(0.6)	0.72(0.6)	0.86(0.7)	0.48(0.5)	0.74(0.7)	0.48(0.3)	0.86(0.6)
VTP	**0.98(1.0)**	**0.98(1.0)**	**0.96(1.0)**	**0.98(1.0)**	**0.97(1.0)**	**0.99(1.0)**	**0.98(1.0)**

Content 4

Pooling method	SSIM	GSM	VSNR	VSI	LVI	BRISQUE	NIQE
average	0.22(0.0)	0.39(0.1)	0.29(0.0)	0.32(0.0)	0.69(0.5)	-0.03(-0.3)	-0.51(-0.4)
Minkowski	0.04(0.0)	0.38(0.0)	0.07(0.0)	0.28(0.0)	0.46(0.1)	0.09(-0.3)	-0.46(-0.3)
percentile	0.55(0.3)	0.72(0.8)	0.87(0.8)	0.73(0.9)	0.78(0.7)	-0.84(-0.9)	-0.72(-0.9)
speed [138]	0.85(0.9)	0.88(0.9)	0.72(0.6)	0.86(0.9)	0.89(0.9)	0.44(0.3)	0.49(0.3)
hysteresis [134]	0.79(0.6)	0.76(0.6)	0.50(0.1)	0.72(0.6)	0.59(0.3)	0.39(0.1)	0.70(0.4)
variation [133]	0.17(0.0)	0.38(0.0)	0.29(0.0)	0.33(0.0)	0.71(0.7)	-0.09(-0.4)	-0.64(-0.4)
VTP	**0.99(1.0)**	**0.99(1.0)**	**0.98(1.0)**	**0.99(1.0)**	**0.98(1.0)**	**0.97(0.9)**	**0.98(1.0)**

Table 4.2.: SROCC and PLCC between objective video quality scores and subjective scores.

SROCC

VQE	Campus	Grocery	Apartment
V-BLIINDS	0.951	0.806	0.648
VIIDEO	0.903	0.830	0.467
BRISQUE-VTP	0.794	0.951	0.976
OG-VTP	0.490	0.952	0.964
LVI-VTP	**0.988**	0.988	**1.000**
NIQE-VTP	0.964	**1.000**	0.976

PLCC

VQE	Campus	Grocery	Apartment
V-BLIINDS	0.931	0.798	0.756
VIIDEO	0.834	0.796	0.540
BRISQUE-VTP	0.850	0.920	0.990
OG-VTP	0.667	0.954	0.960
LVI-VTP	**0.978**	0.961	**0.983**
NIQE-VTP	0.914	**0.979**	0.974

5. CONCLUSIONS AND FUTURE WORK

5.1 Summary

1. 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. 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. We present different distortions in images of FPVs including motion blur, rolling shutter artifacts and rotation. Then we propose a measurement method for classification and quantification of these types of distortions. Our proposed algorithm provides information about how to design an image or video quality metric for FPVs.

2. We propose a controllable enhancement illumination method that allows the degree of enhancement to be adjusted using a single parameter. We then propose an over-enhancement measure, LOM, to evaluate the unnaturalness of enhanced images. Our results of subjective test indicate the effectiveness of our enhancement method and LOM. Remaining issues for future work are how to improve the illumination within over-exposed regions simultaneously and how to design an objective measure for image quality after enhancement that provides a consistent evaluation for both different contents and enhancement methods.

3. We propose a visibility-inspired temporal pooling method (VTP) built on a measurement of visibility that is more effective at estimating the quality of LMVs than existing pooling strategies. The VTP is also demonstrated to be effective

in the application of video stabilization, because it can measure the perceptual blurriness between videos before and after stabilization. VTP method employs the weighted average pooling that the weight is calculated as a function of visibility. The measurement of visibility considers the fraction of visible details within a single frame under a given motion based on the theory of the window of visibility. A systematic subjective test is implemented to model the pooling function using proposed visibility measure. The test results also indicate that our pooling strategy is more suitable for LMVs and can be effectively applied to pool quality scores estimated by different types of image quality metrics.

5.2 Future Work

Frame-quality Assessment: The first improvement is to remove the scaling constraint in our MR quality estimator so that the quality measure can be applied to images with different scales. Since the images with similar content in many cases have different scales, if our design can compare local corresponding patches in terms of their scale differences, the quality estimator can be applied into more different scenarios.

The second improvement is to develop a quality estimator between images that have no overlapping content and incorporate it into our present framework. Even we can partition the video into different near-sets and measure their quality differences, the quality comparison between different near-sets is still a problem. If we can build a NR quality estimator that can be combined with our MR quality estimator, we can use the best of information provided either from similar enough images and our knowledge of "good" images.

The third improvement is to incorporate measures of more varieties of quality degradations, such as illumination in another part of work. The quality measures

of different quality degradations can also be combined to a video-quality level score, which needs further investigation and can refer to the design of VTP strategy.

Illumination enhancement: The visual quality of an enhanced FPV is influenced by motion, video content and temporal illumination change. The illumination enhancement strategy can be further improved by considering these aforementioned FPV characteristics.

First, the motion in FPVs influences the visual quality of illumination enhancement. When the motion is small, the strategy to enhance a frame in a video is similar to enhance a single image. However, when the motion is large, frame content details cannot be perceived due to the masking effect of motion. In addition, motion-induced blur within frames also affects the "goodness" of enhancement.

Second, the content also influences the visual quality of illumination enhancement. There exist different salient regions within a FPV across time. During a specific time interval, a human observer may focus on a region that is originally well-exposed instead of a badly-exposed area. The enhanced badly-exposed region would be ignored so that the video quality is not enhanced as expected.

Third, the temporal illumination change across time influences how we design our enhancement strategy. The illumination difference between an original frame and its optimal enhanced version influences the amount of enhancement applied to it. For example, a very dark frame can have greater illumination enhancement than a frame that is close to well-exposedness. If we enhance all frames into its best achievable quality, the original illumination change within the FPV will be discarded. The remaining question is that whether the best enhanced strategy is to enhance every frame to its optimal point or to enhance each frame considering its illumination differences between neighboring frames.

VTP mechanism: One potential improvement for our VTP mechanism is to locally pool the image quality measure using the visibility as a spatial mask. Another potential improvement is to consider the temporal quality variation frequency influ-

ence after weighted average pooling. In addition, more video content with different structure of motion can be also be tested to further optimize our method.

The measurement of blurriness after stabilization can be applied to the design of stabilization algorithms. Since the motion stability is the key quality factor for stabilization, the increased blurriness is another aspect. The quality of a video after stabilization can expressed as the combination of the quality increase due to motion stabilization and the quality decrease caused by increased perceived blurriness. The question becomes how to design the quality metric that consider both quality factors for a video, and how to gather the subjective data to validate the metric.

Potentially, the visual quality of a video has a concave relationship with the degree of stability using a specific stabilization algorithm. Under this relationship, the design of stabilization algorithm can refer to the framework design of our illumination enhancement, which improve a video into the best achievable quality by controlling a knob. The quality metric to be designed can be used to adjust the parameter setting.

REFERENCES

REFERENCES

[1] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. 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] 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 (ECCV)*, 2016, pp. 557–571.

[3] B. Xiong and K. Grauman, "Detecting snap points in egocentric video with a web photo prior," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 282–298.

[4] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2714–2721.

[5] Y. Niu and F. Liu, "What makes a professional video? A computational aesthetics approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1037–1049, 2012.

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

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

[8] H. Jin, P. Favaro, and R. Cipolla, "Visual tracking in the presence of motion blur," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2005, pp. 18–25.

[9] S. Baker, E. Bennett, S. B. Kang, and R. Szeliski, "Removing rolling shutter wobble," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2392–2399.

[10] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, "A no reference metric for the quality assessment of videos affected by exposure distortion," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[11] Z. Tang, R. G. von Gioi, P. Monasse, and J.-M. Morel, "High-precision camera distortion measurements with a calibration harp," *Journal of the Optical Society of America A*, vol. 29, no. 10, 2012.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[14] D. M. Chandler and S. 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.

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

[16] W. Lin, L. Dong, and P. 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.

[17] H. Hu and G. De Haan, "Low cost robust blur estimator," in *IEEE International Conference on Image Processing (ICIP)*, 2006, pp. 617–620.

[18] N. D. Narvekar and L. 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.

[19] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

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

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

[22] C. Bai and A. R. Reibman, "Controllable illumination enhancement with an over-enhancement measure," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 385–389.

[23] H. Cheng and Y. Zhang, "Detecting of contrast over-enhancement," in *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 961–964.

[24] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[25] A. B. Watson, A. J. Ahumada, and J. E. Farrell, "Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays," *Journal of the Optical Society of America A*, vol. 3, no. 3, pp. 300–307, 1986.

[26] S. T. Hammett, M. A. Georgeson, and A. Gorea, "Motion blur and motion sharpening: temporal smear and local contrast non-linearity," *Vision Research*, vol. 38, no. 14, pp. 2099–2108, 1998.

[27] M. A. Georgeson and S. T. Hammett, "Seeing blur: motion sharpening without motion," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 269, no. 1499, pp. 1429–1434, 2002.

[28] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H.264/AVC video quality assessment dataset," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1509–1513.

[29] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014A database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[30] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3407–3411.

[31] B. Ma and A. R. Reibman, "Estimating the subjective video stability of first-person videos," *Human Vision and Electronic Imaging*, vol. 2018, no. 14, pp. 1–7, 2018.

[32] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[33] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 9, no. 1, March 2010.

[34] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[35] I.-T. RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[36] ——, "Methodology for the subjective assessment of video quality in multimedia applications," 2007.

[37] J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *Proc. IS&Ts Image Processing, Image Quality, Image Capture, Systems Conference*, 2001, pp. 108–112.

[38] A. B. Watson, "High frame rates and human vision: A view through the window of visibility," *Motion Imaging Journal*, vol. 122, no. 2, pp. 18–32, 2013.

[39] B. Ma and A. R. Reibman, "Measuring and improving the viewing experience of first-person videos," in *ACM Multimedia Thematic Workshops*, 2017, pp. 493–501.

[40] L. Zhang, L. Zhang, and X. Mou, "A comprehensive evaluation of full reference image quality assessment algorithms," in *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1477–1480.

[41] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3378–3389, 2012.

[42] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Processing: Image Communication*, Aug. 2010.

[43] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[44] H. Liu and A. R. Reibman, "Software to stress test image quality estimators," in *Quality of Multimedia Experience (QoMEX)*, 2016.

[45] C. Bai and A. R. Reibman, "Mutual reference frame-quality assessment for first-person videos," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 290–294.

[46] ——, "Image quality assessment in first-person videos," *Journal of Visual Communication and Image Representation, special issue on egocentric vision and lifelogging tools*, vol. 54, pp. 123–132, 2018.

[47] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[48] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[49] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[50] L. Zhang and H. Li, "Sr-sim: A fast and high performance iqa index based on spectral residual," in *IEEE International Conference on Image Processing*, 2012, pp. 1473–1476.

[51] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.

[52] X. Marichal, W.-Y. Ma, and H. Zhang, "Blur determination in the compressed domain using DCT information," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2. IEEE, 1999, pp. 386–390.

[53] P. Marziliano, F. Dufaux, and S. Winkler, "Perceptual blur and ringing metrics: application to JPEG2000," *Signal processing: Image communication*, vol. 19, no. 2, pp. 163–172, 2004.

[54] X. Wang, B. Tian, C. Liang, and D. Shi, "Blind image quality assessment for measuring image blur," in *Image and Signal Processing*, vol. 1, 2008, pp. 467–470.

[55] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *Human Vision and Electronic Imaging*, vol. 6492, 2007, p. 64920I.

[56] R. Hassen, Z. Wang, and M. M. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2798–2810, 2013.

[57] I. Setyawan, D. Delannay, B. M. Macq, and R. L. Lagendijk, "Perceptual quality evaluation of geometrically distorted images using relevant geometric transformation modeling," in *Security and Watermarking of Multimedia Contents*, vol. 5020, 2003, pp. 85–95.

[58] A. D'Angelo, M. Pacitto, and M. Barni, "A psychovisual experiment on the use of gibbs potential for the quality assessment of geometrically distorted images," in *Human Vision and Electronic Imaging*, vol. 6806, 2008, p. 680616.

[59] A. D. Angelo, L. Zhaoping, and M. Barni, "A full-reference quality metric for geometrically distorted images," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 867–881, 2010.

[60] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009.

[61] O. Barkol, H. Kogan, D. Shaked, and M. Fischer, "A robust similarity measure for automatic inspection," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2489–2492.

[62] P. Kellnhofer, T. Ritschel, K. Myszkowski, and H.-P. Seidel, "A transformation-aware perceptual image metric," in *Human Vision and Electronic Imaging*, vol. 9394, 2015, p. 939408.

[63] D. Michele A. Saad, M. Pinson *et al.*, "Image quality of experience: A subjective test targeting the consumers experience," in *Human Vision and Electronic Imaging*, 2016, pp. 1–6.

[64] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Forty-Sixth Annual Asilomar Conference on Signals, Systems, and Computers*, 2012, pp. 1693–1697.

[65] A. M. Demirtas, A. R. Reibman, and H. Jafarkhani, "Full-reference quality estimation for images with different spatial resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2069–2080, 2014.

[66] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[67] T. Stathaki, *Image fusion: algorithms and applications*. Academic Press, 2011.

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

[69] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, no. 2, pp. 156–160, 2008.

[70] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[71] M. A. Saad, M. H. Pinson, D. G. Nicholas, N. V. Kets, G. V. Wallendael, R. V. Jaladi, and P. J. Corriveau, "Impact of camera pixel count and monitor resolution perceptual image quality," in *Colour and Visual Computing Symposium*, 2015, pp. 1–6.

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

[73] A. Agarwal, C. V. Jawahar, and P. J. Narayanan, "A survey of planar homography estimation techniques," *Technical report, IIT-Hyderabad*, 2005.

[74] 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.

[75] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2537–2544.

[76] J. Mas and G. Fernandez, "Video shot boundary detection based on color histogram," *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST*, 2003.

[77] "Test video dataset for Mutual Reference Frame Quality Assessment of First-Person videos," *https://engineering.purdue.edu/VADL/resources/MRFQAFPV/test_videos.zip*.

[78] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[79] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to sift or surf," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[80] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[81] J. Košecká and W. Zhang, "Video compass," in *European Conference on Computer Vision*, 2002, pp. 476–490.

[82] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[83] ——, "A subjective study to evaluate video quality assessment algorithms," in *Human Vision and Electronic Imaging*, vol. 7527, 2010, p. 75270H.

[84] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in *IEEE International Conference on Image Processing (ICIP)*, 2012.

[85] J. W. Peirce, "PsychoPyPsychophysics software in python," *Journal of neuroscience methods*, vol. 162, no. 1, pp. 8–13, 2007.

[86] A. T1.801.03, "Digital transport of one-way video signals parameters for objective performance assessment," in *American National Standards Institute*, 1996.

[87] G. Boracchi and A. Foi, "Modeling the performance of image restoration from motion blur," *IEEE Transactions on Image Processing*, vol. 21, no. 8, 2012.

[88] W. Neale, D. Hessel, and T. Terpstra, "Photogrammetric measurement error associated with lens distortion," *SAE Technical Paper 2011-01-0286, 2011, doi:10.4271/2011-01- 0286.*

[89] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, 2010.

[90] N. Narvekar and L. J. Karam, "An improved no-reference sharpness metric based on the probability of blur detection," in *Workshop on Video Processing and Quality Metrics*, January 2010.

[91] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[92] P. Bex, "Sensitivity to spatial distortion in natural scenes," *Journal of Vision*, vol. 8, no. 6, pp. 688–688, 2008.

[93] D.-H. Lee, Y.-J. Yoon, S.-j. Kang, and S.-J. Ko, "Correction of the overexposed region in digital color image," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 2, pp. 173–178, 2014.

[94] K. Singh and R. Kapoor, "Image enhancement using exposure based sub image histogram equalization," *Pattern Recognition Letters*, vol. 36, pp. 10–14, 2014.

[95] M. A. Qureshi, A. Beghdadi, and M. Deriche, "Towards the design of a consistent image contrast enhancement evaluation measure," *Signal Processing: Image Communication*, vol. 58, pp. 212–227, 2017.

[96] Q. Wang and R. K. Ward, "Fast image/video contrast enhancement based on weighted thresholded histogram equalization," *IEEE transactions on Consumer Electronics*, vol. 53, no. 2, 2007.

[97] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 1921–1935, 2009.

[98] Z. U. Rahman, D. J. Jobson, and G. A. Woodell, "Retinex processing for automatic image enhancement," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 100–110, 2004.

[99] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2782–2790.

[100] D. Guo, Y. Cheng, S. Zhuo, and T. Sim, "Correcting over-exposure in photographs," in *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 515–521.

[101] Y.-J. Yoon, K.-Y. Byun, D.-H. Lee, S.-W. Jung, and S.-J. Ko, "A new human perception-based over-exposure detection method for color images," *Sensors*, vol. 14, no. 9, pp. 17 159–17 173, 2014.

[102] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, "Fast efficient algorithm for enhancement of low lighting video," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[103] K. Fotiadou, G. Tsagkatakis, and P. Tsakalides, "Low light image enhancement via sparse representations," in *International Conference Image Analysis and Recognition*, 2014, pp. 84–93.

[104] A. Vadivel, S. Sural, and A. K. Majumdar, "Human color perception in the hsv space and its application in histogram generation for image retrieval," in *Color Imaging X: Processing, Hardcopy, and Applications*, vol. 5667, 2004, pp. 598–609.

[105] Y. Rao and L. Chen, "A survey of video enhancement techniques," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 3, no. 1, pp. 71–99, 2012.

[106] A. Srikantha and D. Sidibé, "Ghost detection and removal for high dynamic range images: Recent advances," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 650–662, 2012.

[107] J. Tang, E. Peli, and S. Acton, "Image enhancement using a contrast measure in the compressed domain," *IEEE Signal Processing Letters*, vol. 10, no. 10, pp. 289–292, 2003.

[108] D. Liu, X. Sun, F. Wu, S. Li, and Y.-Q. Zhang, "Image compression with edge-based inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 10, pp. 1273–1287, 2007.

[109] M. H. Asmare, V. S. Asirvadam, L. Iznita, and A. F. M. Hani, "Image enhancement by fusion in contourlet transform," *International Journal on Electrical Engineering and Informatics*, vol. 2, no. 1, pp. 29–42, 2010.

[110] N. Kong and M. J. Black, "Intrinsic depth: Improving depth transfer with intrinsic images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3514–3522.

[111] S.-D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, 2003.

[112] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[113] V. Jaya and R. Gopikakumari, "IEM: a new image enhancement metric for contrast and sharpness measurements," *International Journal of Computer Applications*, vol. 79, no. 9, 2013.

[114] S. S. Agaian, K. Panetta, and A. M. Grigoryan, "Transform-based image enhancement algorithms with performance measure," *IEEE Transactions on Image Processing*, vol. 10, no. 3, pp. 367–382, 2001.

[115] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, "Subjective and objective quality assessment for images with contrast change," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 383–387.

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

[117] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5372–5384, 2013.

[118] T. Celik and T. Tjahjadi, "Contextual and variational contrast enhancement," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3431–3441, 2011.

[119] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3015–3022.

[120] D. Ghadiyaram and J. Pan, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, 2018.

[121] J. Harguess and M. Reese, "Aggregating motion cues and image quality metrics for video quality estimation," in *Geospatial Informatics, Motion Imagery, and Network Analytics VIII*, vol. 10645, 2018, p. 106450A.

[122] C. Zhang and J. Liu, "On crowdsourced interactive live streaming: a twitch.TV-based measurement study," in *ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2015, pp. 55–60.

[123] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., Sunderland, MA, 1995.

[124] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1–19, 2013.

[125] M. H. Pinson, "The consumer digital video library [best of the web]," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 172–174, 2013.

[126] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[127] D. M. Chandler and S. 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, Sept. 2007.

[128] C. Bai and A. R. Reibman, "Video quality temporal pooling using a visibility measure," in *IEEE International Conference on Multimedia and Expo (ICME) (to be appear)*.

[129] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-preserving warps for 3d video stabilization," in *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3. ACM, 2009, p. 44.

[130] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust l1 optimal camera paths," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 225–232.

[131] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, and S.-M. Hu, "Deep online video stabilization with multi-grid warping transformation learning," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2283–2292, 2019.

[132] J. G. Robson, "Spatial and temporal contrast-sensitivity functions of the visual system," *Journal of the Optical Society of America A*, vol. 56, no. 8, pp. 1141–1142, 1966.

[133] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Visual Media Quality Assessment*, vol. 3, no. 2, pp. 253–265, 2009.

[134] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1153–1156.

[135] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of temporal pooling method on the objective video quality evaluation," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2009, pp. 1–5.

[136] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning*, 2010, pp. 111–118.

[137] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610–620, 2013.

[138] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America A*, vol. 24, no. 12, pp. B61–B69, 2007.

[139] R. R. Hainich and O. Bimber, *Displays: Fundamentals & Applications*. CRC Press/A. K. Peters, 2011.

[140] "Visibility test video dataset," https://engineering.purdue.edu/VADL/resources/visibility_test/visibility_test.zip.

[141] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[142] "Stabilization blurriness test video dataset," https://engineering.purdue.edu/VADL/resources/visibility_test/stabilization_blurriness_test.zip.

[143] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Processing: Image Communication*, vol. 40, pp. 1–15, 2016.

[144] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[145] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[146] M. Elfeki, A. Sharghi, S. Karanam, Z. Wu, and A. Borji, "Multi-view egocentric video summarization," *arXiv:1812.00108*, 2018.

VITA

VITA

Chen Bai was born in Nanchang, China. He obtained his bachelor's degree from Huazhong University of Science and Technology, China, in 2014. He joined the Ph.D. program at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana in August 2014. He has worked as Research Assistant under the supervision of Professor Amy R. Reibman since 2015. His research interests are image and video quality assessment.

PUBLICATIONS

Journal Papers:

1. **Chen Bai**, Amy R. Reibman, "Visibility-inspired temporal pooling with application to video stabilization", in *IEEE Transactions on Circuits and Systems for Video Technology*, under review
2. **Chen Bai**, Amy R. Reibman, "Image quality assessment for first-person videos", in *Journal of Visual Communication and Image Representation, Special Issue on Egocentric Vision and Lifelogging Tools*, May, 2018

Conference Papers:

1. **Chen Bai**, Amy R. Reibman, "Video quality temporal pooling using a visibility measure", in *IEEE International Conference on Multimedia and Expo (ICME)*, July, 2019
2. **Chen Bai**, Amy R. Reibman, "Controllable image illumination enhancement with an over-enhancement measure", in *IEEE International Conference on Image Processing (ICIP)*, October, 2018
3. **Chen Bai**, Amy R. Reibman, "Mutual reference frame-quality assessment for first-person videos", in *IEEE International Conference on Image Processing (ICIP)*, September, 2017
4. **Chen Bai**, Amy R. Reibman, "Subjective evaluation of distortions in first-person videos", in *Human Vision and Electronic Imaging (HVEI)*, February, 2017
5. **Chen Bai**, Amy R. Reibman, "Characterizing distortions in first-person videos", in *IEEE International Conference on Image Processing (ICIP)*, September, 2016