https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-510

Estimating the Subjective Video Stability of First-Person Videos

Biao Ma and Amy R. Reibman;
Purdue University, West Lafayette, IN, USA

Abstract

First-Person Videos (FPVs) captured by body-mounted cameras are usually too shaky to watch comfortably. Many approaches, either software-based or hardware-based, are proposed for stabilization. Most of them are designed to maximize stability of videos. However, according to our previous work [1], FPVs need to be carefully stabilized to maintain their First-Person Motion information (FPMI). To stabilize FPVs appropriately, we propose a new video stability estimator Viewing Experience under "Central bias + Uniform" model (VECU) for FPVs on the basis of [1]. We first discuss stability estimators and their role in applications. Based on the discussion and our application target, we design a subjective test using real scene videos with synthetic camera motions to help us to improve the human perception model proposed in [1]. The proposed estimator VECU measures the absolute stability and the experimental results show that it has a good interval scale and outperforms existing stability estimators in predicting subjective scores.

Introduction

As wearable cameras, such as GoPro or Pivothead, become popular, many people start to capture and share their own First-Person Videos (FPVs). FPVs allow viewers to revisit recorders' life-logs in the First-Person Perspective. The interesting parts of FPVs are not only the objects recorded in the video, but also the recorder's reactions reflected by their First-Person Motions. However, there is a problem with FPVs: most of them are too shaky for humans to watch comfortably.

To solve this problem, both software-based and hardware-based approaches can be applied. Hardware-based approaches are mainly based on gimbals. Software-based approaches [2–10] estimate the camera motion in either 2D or 3D and then use filters to smooth the computed camera motion and re-project frames to create a stabilized sequence.

Nevertheless, these approaches treat the FPVs as normal shaky videos and try to stabilize them as much as possible, which damages the First-Person Motion Information (FPMI) with high probability [1]. When the FPMI is damaged, the viewing experience is limited, which degrades FPVs to normal videos. To appropriately stabilize a FPV, it is necessary to have an accurate estimator of its stability.

Quality estimators (QEs) characterize the quality of images or videos, which is useful for many applications. For a stability estimator, two relevant applications are algorithm optimization and product benchmarking [11]. Usually, QEs are classified into No-Reference, Reduced-Reference and Full-Reference estimators, but they can also be categorized by their purpose [11]. Using this principle, we partition stability estimators in 3 ways, which are: (1) Whether the estimator measures the relative or absolute stability of the video, (2) Whether the estimator is based on an objective model or a perceptually-motivated model, (3) Whether the estimator measures the ordinal scale or interval scale of stabilities of different videos.

Relative or Absolute: Some applications are only interested in comparing relative video stability or relative stabilization performance of different algorithms. In this situation, the relative measurement is sufficient. The Motion-Vector-based Mean Squared Error (MV-MSE) [12] is an example of such an estimator. MV-MSE measures the MSE between the intentional motion and the original motion. The intentional camera motion can be computed in many ways. For example, in [12], the original motion is transformed into the frequency domain and low-pass filtered to generate the intentional motion. However, no matter how accurate the computed intentional motion is, MV-MSE can only measure the instability of the original motion with respect to the intentional motion. The results in [1] indicate that the subjective stability of intentional motions strictly depends on their amplitude and frequency characteristics, which means the intentional motions themselves are seldom fully stable. As a result, it is hard for MV-MSE to measure the absolute stability of videos.

Some applications want to know the absolute stability of videos, for example the video stabilization systems. Based on their criteria of "stable", they can stabilize the video to the degree they want. Several works [6–10] for video stabilization techniques implicitly generate stability estimators as side-products. When creating the smooth camera path, they perform optimization processes which minimize the sum of squared values of the second derivative of motion curves. This objective function can be thought of as a stability estimator. Smaller values indicate higher stability. Some other objective functions are also adopted. For example, in [7], the mean square value of the third derivative of motion curves is also used. In [6], the difference of simplified affine motion parameters between adjacent frames is used.

Objective or Perceptually-motivated: Applications that need fast computation or fair comparison prefer to use objective estimators. These estimators mainly use objective information of the video. As a result, these estimators are not favorably biased toward any stabilization techniques. However, the scores of these estimators may not have high correlation with subjective scores. A good example is the Inter-frame Transformation Fidelity (ITF), which was proposed in [13] and used by many works [14–17] for comparing performance of different stabilization methods. First, the Peak Signal-to-Noise Ratio (PSNR) between adjacent frames are computed. The mean value of PSNR of all adjacent pairs is the ITF of the video, which can be computed easily and quickly. It is believed that the PSNR does not have a high correlation with subjective quality scores [18, 19], which means ITF is objective.

For applications that are interested in specific problems of video stability, perceptually-motivated estimators are more useful. In our previous works [1, 20], we focused on stabilizing FPVs

while preserving their First-Person Motion Information (FPMI). In [20], we first explored the stabilization technique by considering the human perception model. Then in [1], we carefully modeled the perception model and proposed a video stability estimator, the Viewing Experience (VE) score. By modeling the smooth pursuit eye movements of a viewer, the VE score estimates the fraction of the FPV that the viewer can watch comfortably. [1] showed that VE scores correctly characterized the rank order of 3 versions of stabilized videos for several scenes. Also, the stabilization algorithm based on VE scores has similar stabilization performance with Microsoft Hyperlapse [8, 21] and preserves much more FPMI.

The estimators from stabilization techniques [6–10] are somewhere between these two categorizations. They consider that viewers prefer smooth motion but do not explore the perception model. As a result, they do not have high correlation with subjective scores and are also not fair enough for comparing different algorithms.

Ordinal scale or Interval scale: The results in [1] show that VE scores based on this model well reflect human subjective scores. However, they only have similar rank to subjective scores, which means they have a good ordinal scale [22]. For example, if the subjective scores of three version of the videos have the relationship: $SJ_1 < SJ_2 < SJ_3$, then our VE score is: $VE_1 < VE_2 < VE_3$. It is hard to use VE_i to accurately predict SJ_i, which means VE_i and SJ_i do not have similar interval scales [22].

An estimator that has a good interval scale is more valuable. For example, watching videos with First-Person Motion may cause dizziness or motion sickness. It is useful to measure the impact of these side-effects, which is possible if the estimator can precisely predict the subjective video stability.

In this paper we aim to find a such video stability estimator than can help us predict the subjective video stability and guide us to appropriately stabilize FPVs. To achieve this target, the estimator needs to measure the **absolute stability** and be based on a **perceptually-motivated model**, which is satisfied by our precious VE score. For this reason, we only need to design a subjective test to improve the VE score to have **a good interval scale** by considering saliency models and refining the basic human perception model. In the next section, we briefly review the human perception model in our VE score and point out the parts need improving. In section 3, we introduce the subjective test including the video sources and test design. The experimental results are shown in section 4. We conclude our work and summarize interesting findings in section 5.

Viewing experience score

Our proposed video stability measurement [1] is called the viewing experience (VE) score, which measures the fraction of frames of the FPV that viewers can watch comfortably. In this section, we first review the motivation and the mathematical model of our VE score. After that we discuss the current limitation of our VE score, which inspires this work.

Human perception on stability

The motivation of our VE score is that we perform different eye movements in real life and when we are watching videos. Our proposed VE score measures the video stability based on our mathematical model of eye movements when we are watching videos.

In real life, we perform vestibulo-ocular reflex/eye movement [23] to physically compensate the unintentional motion of our body to retain our visual target fix at the center of the retina. This eye movement uses the information from our internal-ears and the reaction time is around 0.01 second [23]. However, the eye movement we are using is called smooth pursuit eye movement (SPEM) and saccade [24]. The saccade helps us to catch our visual target and the SPEM uses and only uses the visual clues to assist us to smoothly track the catched target. The time our visual system needs for perform a saccade is 0.15 to 0.2 second [25], which is much longer than the vestibulo-ocular reflex. And during such long period, there is no visual information can be perceived by our visual system, which leads to the unstable feeling of the video.

Our VE score estimates the duration of the total saccade period when we are watching the video to measure the video stability.

Previous work

Our previous VE score models the eye movements of viewers as a random process [1] to estimate the fraction of the FPV that the viewer can watch comfortably (without saccade). The model assumes viewers perform SPEM and saccade [26]. The process is described as follow. Viewers choose a random target in frame n and perform SPEM to track the target until frame $(n+1)$. If the target suddenly moves out of the area predicted by the human visual system in frame $(n+2)$, the viewers may fail to track it and would need to perform a saccade to catch the target or retarget another target.

If $PE(\beta_n; n+2)$ is the position error between eye position and target position at frame $(n+2)$, then according to the model in [1], the predicted area is where PE satisfies condition (1):

$$0.04 \leq \frac{|PE(\beta_n; n+2)| + b}{|\omega_{rs}(\beta_n; n+2)|} \leq 0.18 \; or \; |PE(\beta_n; n+2)| < MAR, \quad (1)$$

where MAR is the minimum angular resolution of human eyes, b is the bias of position error estimation which is set to MAR, and β_n is the target angular position with respect to the camera at the n^{th} frame.

By solving condition (1) for each frame, we can compute the target position interval β_n. Any target within this interval can be tracked using SPEM between the next two frames without a catch-up saccade or retargeting. Note that the catch-up saccade or retargeting process spends nearly 150 ms [26]. During this period, visual information is not processed, which leads to the experience of instability. If we assume the chance of choosing a target is uniformly equal across the whole frame, then based on the interval sequence $\{\beta_n\}$, the VE score can be computed as the fraction of a video that viewers can perform SPEM.

Existing limitations

Two potential weaknesses in the previous model restrict the previous VE score to only have a good ordinal scale instead of interval scale. First, we assumed that all targets have the same chance to be selected, which is not accurate. For example, people prefer to choose targets at the center of the frame [27], which is called the central bias. Based on this fact, we explore here including a saliency models into our human perception model. There

(a) Lobby

(b) Market

(c) University Street (US)

(d) Park Walkway (PW)

Figure 1: Example frames of source videos

are many saliency models [28–30] that have high performances available online. We discuss our adoption criteria and final choice in the next section.

Second, the bias parameter in condition (1) may not be accurate enough. The bias b was directly set to MAR without any experiment and discussion. As a result, the interval sequence $\{\beta_n\}$ computed from this condition may not be accurate enough, which consequently would degrade the precision of the VE score. To address this problem, in this paper, we design a subjective test and use the human subjective score as ground truth to find an appropriate value of b.

Subjective test

Our target is to use the subjective test to improve the design of our VE score. We want to determine which of 4 saliency models is the most effective, and find an accurate bias parameter for our system. In this section, we first introduce the process by which we create the source videos. After that, we show how we design the subjective test.

Our subjective test is based on paired comparison, which includes 19 subjects and is operated according to the recommendations in [31]. For each of the 4 scenes, there are 9 versions of videos, each having a different combination of motion amplitude in both yaw and pitch. Each video is 5 seconds long, since for repeated motion patterns, viewers are able to learn and predict well within 5 seconds [32]. Tested videos are played back on a 27-inch, 82 PPI screen at full resolution. The ratio of viewing distance with respect to the equivalent focal length is 4. During the test, for each comparison, only one question is asked: **Which video is more stable?**

Video source

To create the videos for our subjective test, we add synthesized motions into four different scenes that were captured using a 360° camera-set on a wheeled tripod. Then we slowly and smoothly move the tripod forward and record the scene. The obtained six videos are stitched using Kolor Autopano Video 3 [33] with D-warp, color normalization, video stabilization and other options on. Each frame of the resulting video is a equirectangular image. Based on the synthesized motion of each frame, we centralize the camera view at a particular part of the equirectangular image to create a perspective image. Collecting all resulting perspective images, we generate a video with synthesized motions. The camera-set consists of 6 GoPro Hero Session 4 cameras. The scenes are shown in Fig. 1. The resolution of all videos is 1080p and the frame rate is 30 fps.

The synthesized motion of a video consists of two different kinds of First-Person Motions: yaw and pitch, which are the motion of looking side to side and upside down. According to the motion model used in First-Person video games [34], yaw and pitch motions are synthesized using sinewaves while the rate of pitch is twice as that of yaw. All the synthesized videos used in this paper are designed to mimic First-Person running videos. If the normal running speed is R meters per second, the step size is S meters and the frame rate of recorded video is F_{rate}, then the period of pitch motion T is $\frac{F_{rate}}{R} \cdot S$. Suppose the original video records M meters' moving in N frames, then the fast-forward multiple of the original video is $\frac{N \cdot R}{F_{rate} \cdot M}$.

There are two main reasons for using this synthesizing process. First, it enables easy and accurate access to the camera motion since we actually generate it. Second, it enables us to create videos that do not contain several potential distortions. For example, since we move the tripod smoothly and slowly, videos are free from rolling shutter. And since we use the 360° videos, all the test videos are free from black areas.

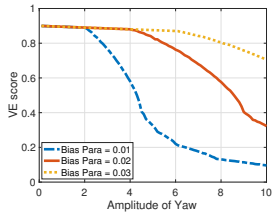
Design of synthesized motions

To generate different versions of videos, we can vary the amplitude of yaw and pitch. However, a question is raised: what value of amplitude we should choose? Paired-comparison-based subjective tests are time consuming, especially for videos. So we must answer this question wisely.

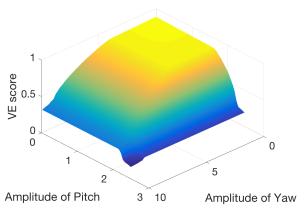
First, there are some constraints on the amplitude of yaw and pitch. By examining the estimated motions of recorded videos obtained in our previous work [20], we find that the minimum ratio between the amplitudes of yaw and pitch is around 4. Also, the maximum amplitude of yaw is around 10°. To make synthesized videos more realistic, we follow these constraints when choosing the amplitude of yaw and pitch.

Second, to guide us toward an effective selection of videos, we explore the impact of the bias parameter. An optimal bias parameter makes the VE score have a good interval scale; different bias parameters results in different VE curves. Fig. 2 (a) shows the VE curves of different bias parameters when we vary the amplitude of yaw and fix the amplitude of pitch. To learn about the perceptual impact of the bias parameter, the videos chosen to compare must have different relative VE scores for all values of the bias parameters. Correspondingly, the curve shapes in Fig. 2 (a) should be distinct. Otherwise, our videos will not provide information to identify the optimal bias parameter. For example, we want to avoid a yaw amplitude below 2, since for all bias parameters, the relative VE scores are the same. The reason to have different **relative** scores is that the subjective results are expressed using Bradley-Terry scores, which measure the relative subjective quality.

We aim to learn about an optimal bias parameter by calculating N different VE scores for each of N bias parameters, and correlating the subjective results to each set of VE scores. We know that the true value of the bias parameter is around 0.02, which was used in [1] and performed well. So we vary the bias parameter from 0.01 to 0.03 with interval 0.001 and vary the amplitude of yaw from 0.1 to 10 with interval 0.1. And the amplitude of pitch is varied from 0.025 to 2.5 with interval 0.025. For the bias parameter b, we obtain a 100 by 100 matrix P_b filled up with VE



(a) VE curves



(b) VE surface

Figure 2: VE scores under different conditions: (a) VE curves of different bias parameters when the amplitude of pitch = 0.025 (b) VE surface when bias parameter = 0.02

scores. Fig. 2 (b) shows the example of when the bias parameter is 0.02. $P_b(x,y,z)$ is the VE score of the corresponding yaw (x axis), pitch (y axis) and bias parameter (z axis) settings.

Our target here is to choose motion settings that result in videos that have distinctly different relative VE scores for different bias parameters. Suppose we adopt the corresponding motion settings of entries $(x-3n:n:x, y-3n:n:y)$, where n is the step size. Then we use the vector $D_{(x,y,b,n)}$, which consists of the diagonal entries of $P_b(x-3n:x, y-3n:y)$, as the representative VE scores of test videos under bias parameter b. Each entry of $D_{x,y,b,n}$ can be thought as a random variable. We compute the covariance matrix $T_{(x,y,n)}$ of $D_{x,y,b,n}$ and obtain the set of the eigenvalues of $T_{(x,y,n)}$ (denoted as $\{\lambda(T_{(x,y,n)})\}$). The VE scores of preferred motion settings should be at the (x,y) position of the 100 by 100 matrices with step size n, which satisfy the following condition:

$$\min_{x,y,n} J = \frac{Std\left(\lambda(T_{(x,y,n)})\right)}{Mean\left(\lambda(T_{(x,y,n)})\right)}. \tag{2}$$

This makes the vectors $\{D_{(x,y,b,n)}\}$, which are the representative VE scores of test videos under all possible bias parameter, well and equally separated in the space. Note that when computing the VE score, we do not apply any saliency model.

Saliency models

The saliency model is important to our perception model. To choose the optimal model among the ones available online, we follow the suggestions from [35]. In particular, since we do not want to miss any target in the frame and do not want to falsely include targets, we sort the saliency models listed in the MIT Saliency Benchmark [36] using the metric Normalized Scanpath Saliency (NSS), which is equally affected by false positive and false negative.

Before the test, we do not know how important the saliency model is for our human perception model. Thus, we want to test with saliency models that have different categories of performance. Four saliency models are tested. The saliency model that has the best performance among the ones are available online is [28]. We also test the model proposed in [29], an uniform model and a central bias model [27]. Note that [29] ranks at the middle of the list of [36] under the metric NSS.

Results and Discussion

Based on the discussion of test settings, we finalize the motion amplitude settings as shown in Fig. 3, where the settings marked are used to generate videos. All these videos are used in the subjective test and their Bradley-Terry (BT) scores [37] are computed together.

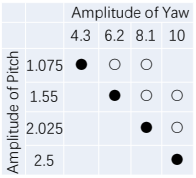


Figure 3: Motion amplitude settings

Table 1: Bradley-Terry scores of all videos

			Yaw Motion Amplitude			
			4.3	6.2	8.1	10
Lobby	Pitch Motion Amplitude	1.075	0 (±0.345)	-3.057 (±0.422)	-5.105(±0.550)	
		1.55		-4.027 (±0.533)	-5.987 (±0.622)	-7.813 (±0.809)
		2.025			-6.464 (±0.652)	-9.021 (±0.956)
		2.5				-9.857 (±1.095)
Market		1.075	0 (±0.387)	-1.262 (±0.431)	-2.184 (±0.495)	
		1.55		-1.268 (±0.468)	-2.683 (±0.555)	-4.445 (±0.720)
		2.025			-3.199 (±0.580)	-4.906 (±0.799)
		2.5				-5.298 (±0.870)
US		1.075	0 (±0.370)	-3.057 (±0.463)	-5.041 (±0.588)	
		1.55		-3.734 (±0.551)	-5.761 (±0.671)	-7.087 (±0.796)
		2.025			-6.948 (±0.763)	-7.981 (±0.897)
		2.5				-8.641 (±1.006)
PW		1.075	0 (±0.434)	-1.885 (±0.468)	-3.202 (±0.515)	
		1.55		-2.083 (±0.455)	-3.792 (±0.566)	-5.101 (±0.658)
		2.025			-4.530 (±0.624)	-5.950 (±0.729)
		2.5				-6.516 (±0.827)

Our study includes two parts: training and testing. The videos with settings on the diagonal (marked by solid circles) are used as training set. The remaining videos construct the testing set. The training process is used to find the bias parameter as we discuss in the previous section, which is based on the linear regression. First, the subjective test provides the subjective score of each of the source videos. Given a saliency model and a bias parameter, VE scores of training videos are computed. The training outputs are the bias parameter that results in the highest correlation coefficient.

The testing process is used to find which of the 4 saliency models is the most effective and to verify the accuracy of the bias parameter we obtain. The testing process examines the training outputs (bias parameter and regression models) by computing the predictive error of the VE scores. The effectiveness of saliency model is discussed by comparing the slopes of the regression models of different video scenes. When an accurate saliency model is applied, the regression model should be independent of video scenes.

Subjective test result

The subjective scores with 95% confidence interval (in parentheses) are shown in Table 1. Our videos include 4 different scenes, which can be classified into 2 kinds of scene structure. The "Lobby" and "University Street" are scenes that have open views while the "Market" and "Park Walkway" have narrow corridor-like views. Several subjects report that it is easier to distinguish the stability of videos with open view scenes than that of videos with narrow view scenes. This aligns with the BT scores shown in Table 1 where BT scores decrease more in open view scenes when the motion amplitudes increase.

There is another interesting phenomenon in our subjective test, which is the masking issue between different motions. Our VE model introduced in [1] did consider that when the yaw motion is much more shakier than the pitch, no matter how we change the pitch motion, the stability of the video is only decided by the yaw, or vice versa. The data in Table 1 provide some clue.

Based on the BT scores, we see that it becomes more difficult for viewers to distinguish the stability of videos with different pitch motions when the amplitude of the yaw increases. Also, this becomes more obvious in open view scenes. For example, in the "Lobby" scene, when the amplitude of yaw is 6.2, the BT scores of different pitch motions are do not overlap, which is different when the amplitude of yaw is increased to 8.1. To explore it more explicitly, we may need more subjects and to test with more motion amplitude values.

Effectiveness of saliency models

The training process helps us to find the optimal bias parameter. Suppose the bias parameter is b, BT scores of i^{th} scene are stored in vector BT_i and the corresponding VE scores under a saliency model is VE_i. Then we use following equation to find the optimal bias parameter for each saliency model:

$$\max_b J = \sqrt{\sum_i (corr(BT_i, VE_i))^2}, \tag{3}$$

where $corr(\cdot)$ computes the Pearson linear correlation coefficient (PLCC). As discussed in previous sections, we first apply 4 different saliency models: SAM, iSEEL, a uniform model and a central bias model [27]. The resulting bias parameters and PLCCs of 4 scenes under the corresponding parameters are shown in Table 2. For each saliency model, the PLCCs are shown in order: "Lobby", "Market", "University street" and "Park walkway". The corresponding linear models are shown in Fig. 4 (a)-(d).

For a good no-reference quality estimator, the relationship between its objective scores and subjective test scores should be similar **across different scenes**. If our VE scores and BT scores have similar relationship across all scenes, then our no-reference estimator works well. However, as can be seen from Fig. 4 (a)-(d), no matter which of the 4 saliency models is used, the line slopes of the narrow view scenes are quite different from that of the open view scenes. This only indicates our VE model has some weakness because **for the scenes of the same structure, the line slopes are quite similar**.

The four saliency models we adopt have good performance on predicting human saliency of images. The failure in Fig. 4 (a)-(c) only indicates that the saliency models applied do not work on video sequences since these models never consider the influence of camera motions on the saliency.

As a result, we propose a hypothesis for the saliency model: **for open view scenes, viewers are uniformly likely to look at the parts in the frames while for narrow view scenes, they prefer to look at the center of the frames**. This hypothesis is inspired by the statement of some subjects after the test. They report that for egocentric videos with forward motions, they are more willing to look across the entire frame for open view scenes since they are curious about the environment. And for narrow view scenes, the forward motion suggests there is an important target ahead, so they prefer to look at the center of the frames.

To verify this, we apply the uniform saliency model to the "Lobby" and "University street" and apply the central bias model [27] to the "Market" and "Park walkway". Its training outputs are shown in Fig. 4 (e) and Table 2. Under this saliency model setting, the line slopes of both narrow view scenes and open view scenes are similar.

To accurately show the difference of line slopes, the coefficient of variance (cv%) is computed for all saliency models and shown in Table 2. The "uniform + central bias" model outperforms all other saliency models with the lowest cv%.

Table 2: Results of training and testing

Measurement	Bias	PLCC	MSE	Average MSE	cv% of slopes
VE (SAM)	0.0180	0.9767	0.5932	0.4997	31.1
		0.9942	0.5539		
		0.9604	0.4149		
		0.9895	0.4370		
VE (iSEEL)	0.0181	0.9633	0.5922	0.4851	28.1
		0.9959	0.5811		
		0.9634	0.3782		
		0.9948	0.3887		
VE (Uniform)	0.0183	0.9744	0.4886	**0.4091**	29
		0.9967	0.4614		
		0.9687	0.3291		
		0.9952	0.3435		
VE (Central Bias)	0.0196	0.9656	0.8473	0.6318	31.6
		0.9392	0.7487		
		0.9945	0.562		
		0.9671	0.3693		
VE (Central Bias + Uniform)	0.0183	0.9744	0.4886	0.4927	**7.9**
		0.923	0.7962		
		0.9687	0.3291		
		0.9513	0.3570		
MV-MSE		0.9657	1.2538	1.0571	28.6
		0.9644	0.864		
		0.9374	1.3048		
		0.9496	0.8056		
ITF		0.9960	0.7539	0.9086	12.3
		0.9526	1.1421		
		0.9985	0.9		
		0.9940	0.8383		

Predictive performance

After we trained the models using BT scores and VE scores, we need to test their predictive performance since our target is to create a video stability estimator that has a good interval scale. We use the models we obtained to predict the BT scores of test videos. The predictive errors are shown in Table 2. To show the power of our model, we also train the models for MV-MSE [12] and ITF [13]. Their training and testing results are shown in Fig. 4 (f)-(g) and Table 2.

The VE under uniform saliency model has the smallest average mean square error (MSE). However, this model does not unify line slopes. The VE under "Central bias + Uniform" (VECU) model has the lowest cv% and the 3^{rd} smallest average MSE. Although ITF has low cv%, it has a much larger average MSE and does not compete with the VECU model. Also, although ITF is a no-reference estimator, its scores do not have physical meanings while the VE based models do. The scores VECU indicate the fraction of frames that the viewer can track without saccade and so that can be watched comfortably.

The performance of the VECU model shows that given a video, it is possible to accurately predict its BT score with respect to another video that records the same scene. However, we do not have the subjective data to explore whether the VECU model can compare the stability of videos that record different scenes. For that, we would have needed to include fully stable videos of each scene in our subjective test. Without these "anchors," the relationship between line models of different scenes cannot be established.

We anticipate that the VECU does have a highly potential to achieve it since line models for both narrow view scenes and open view scenes have quite similar slopes. Also, for each kind of scene construction, line models of different scene content align

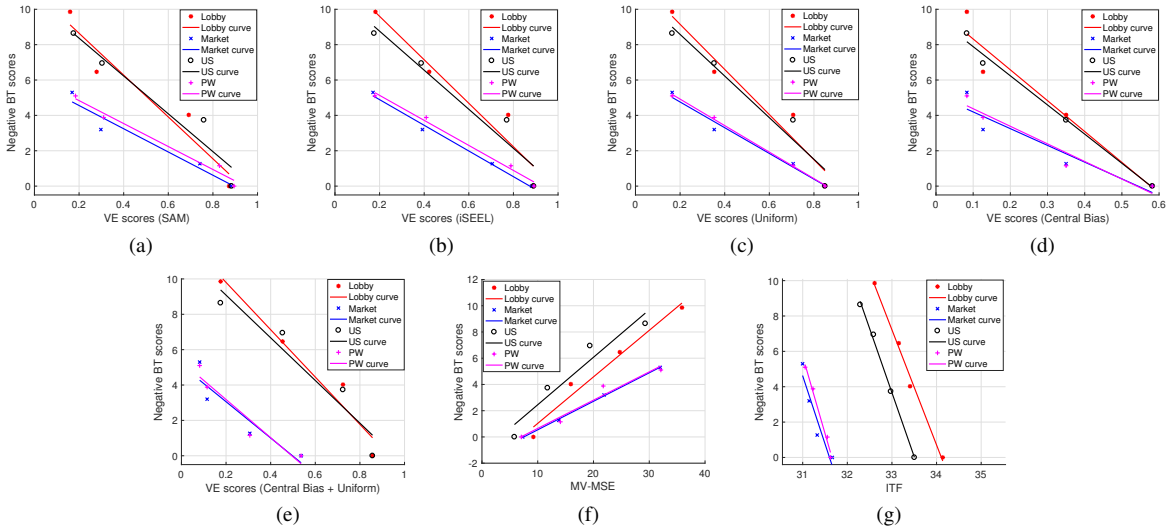


Figure 4: Fitting results between negative BT scores and objective scores

well with each other, which indicates the distance between the line models of open view scenes and narrow view scenes may be caused by the scene structure and not sensitive to the scene content. In addition, VECU scores have a physical meaning which is unchanged by the recorded content.

Conclusion

In this paper, we have 3 main contributions. First, we review existing stability estimators and classify them in 3 different ways. Second, we improve our previous video stability estimator. The obtained video stability estimator can measure the absolute video stability and has a good interval scale. It can be used to predict subjective scores and has high potential ability to compare the stability of videos that record different scenes. Third, we shows that image-based saliency models may not be effective for FPVs. Then we propose and simply verify the hypothesis that viewers have different viewing preference for narrow view scenes and open view scenes.

The obtained FPV stability estimator is valuable for many applications. For example, it can be related to measuring the motion sickness of VR videos since VR videos are similar with FPVs. It can be applied for motion design in First-Person games. Since it can accurately predict the subjective stability quality, game designers can use it to control the amount of First-Person Motions in order to create obvious but comfortable First-Person feelings [1].

References

[1] B. Ma and A. R. Reibman, "Measuring and Improving the Viewing Experience of First-person Videos," in *ACM Multimedia Thematic Workshops 2017*. ACM, 2017.

[2] F. Liu *et al.*, "Subspace video stabilization," *ACM Transactions on Graphics*, vol. 30, no. 1, p. 4, 2011.

[3] C. Jia and B. L. Evans, "Online motion smoothing for video stabilization via constrained multiple-model estimation," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 25, 2017.

[4] F. Liu *et al.*, "Content-preserving warps for 3D video stabilization," *ACM Transactions on Graphics*, vol. 28, no. 3, p. 44, 2009.

[5] Z. Wang and H. Huang, "Pixel-wise video stabilization," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 15 939–15 954, 2016.

[6] H.-C. Chang *et al.*, "A robust and efficient video stabilization algorithm," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 29–32.

[7] 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*, 2011, pp. 225–232.

[8] J. Kopf *et al.*, "First-person hyper-lapse videos," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 78, 2014.

[9] K.-Y. Lee *et al.*, "Video stabilization using robust feature trajectories," in *IEEE International Conference on Computer Vision*, 2009, pp. 1397–1404.

[10] H. Qu and L. Song, "Video stabilization with L1–L2 optimization," in *IEEE International Conference on Image Processing*, 2013, pp. 29–33.

[11] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal processing: Image communication*, vol. 25, no. 7, pp. 469–481, 2010.

[12] M. Tanakian, M. Rezaei, and F. Mohanna, "Camera motion modeling for video stabilization performance assessment," in *Machine Vision and Image Processing (MVIP), 2011 7th Iranian*. IEEE, 2011, pp. 1–4.

[13] L. Marcenaro, G. Vernazza, and C. S. Regazzoni, "Image stabilization algorithms for video-surveillance applications," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1. IEEE, 2001, pp. 349–352.

[14] D.-Y. Koh *et al.*, "Bioinspired image stabilization control using the adaptive gain adjustment scheme of vestibulo-ocular reflex," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 922–930, 2016.

[15] M. Favorskaya and V. Buryachenko, "Fuzzy-based digital video stabilization in static scenes," in *Intelligent Interactive Multimedia Systems and Services in Practice*. Springer, 2015, pp. 63–83.

[16] W. Hong, D. Wei, and A. U. Batur, "Video stabilization and rolling shutter distortion reduction," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 3501–3504.

[17] K. L. Veon, M. H. Mahoor, and R. M. Voyles, "Video stabilization using SIFT-ME features and fuzzy clustering," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 2377–2382.

[18] Z. Wang, H. R. Sheikh, A. C. Bovik *et al.*, "Objective video quality assessment," *The handbook of video databases: design and applications*,

vol. 41, pp. 1041–1078, 2003.

[19] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communications*. Prentice Hall Upper Saddle River, 2002, vol. 5.

[20] B. Ma and A. R. Reibman, "Enhancing Viewability for First-person Videos based on a Human Perception Model," in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 2017.

[21] N. Joshi *et al.*, "Real-time Hyperlapse creation via optimal frame selection," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 63, 2015.

[22] A. Leontaris, P. C. Cosman, and A. R. Reibman, "Quality evaluation of motion-compensated edge artifacts in compressed video," *IEEE transactions on image processing*, vol. 16, no. 4, pp. 943–956, 2007.

[23] S. Aw, G. Halmagyi, T. Haslwanter, I. Curthoys, R. Yavor, and M. Todd, "Three-dimensional vector analysis of the human vestibuloocular reflex in response to high-acceleration head rotations. II. Responses in subjects with unilateral vestibular loss and selective semicircular canal occlusion," *Journal of Neurophysiology*, vol. 76, no. 6, pp. 4021–4030, 1996.

[24] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. Williams, "Types of eye movements and their functions," *Neuroscience*, pp. 361–390, 2001.

[25] S. de Brouwer, M. Missal, G. Barnes, and P. Lefèvre, "Quantitative analysis of catch-up saccades during sustained pursuit," *Journal of Neurophysiology*, vol. 87, no. 4, pp. 1772–1780, 2002.

[26] S. De Brouwer *et al.*, "What triggers catch-up saccades during visual tracking?" *Journal of Neurophysiology*, vol. 87, no. 3, pp. 1646–1650, 2002.

[27] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of vision*, vol. 7, no. 14, pp. 4–4, 2007.

[28] M. Cornia *et al.*, "Predicting human eye fixations via an LSTM-based saliency attentive model," *arXiv preprint arXiv:1611.09571*, 2016.

[29] H. R. Tavakoli *et al.*, "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, pp. 10–18, 2017.

[30] M. Cornia *et al.*, "A deep multi-level network for saliency prediction," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 3488–3493.

[31] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 629–632.

[32] A. T. Bahill and J. D. McDonald, "Smooth pursuit eye movements in response to predictable target motions," *Vision research*, vol. 23, no. 12, pp. 1573–1583, 1983.

[33] Kolor, "Autopano video." [Online]. Available: http://www.kolor.com

[34] A. Thorn and M. S. arer, *Pro Unity Game Development with C#*. Springer, 2014.

[35] Z. Bylinskii *et al.*, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[36] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark."

[37] J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *PICS*, vol. 1, 2001, pp. 108–112.

Author Biography

Biao Ma received his BS in Automatic control from the Beijing Institute of Technology (2014) and now is a PhD student in Electrical and Computer Engineering at Purdue University. He is interested in viewing experience of videos and video compression.