Measuring and Improving the Viewing Experience of First-person Videos

Biao Ma
School of Electrical and Computer Engineering
Purdue University
ma336@purdue.edu

Amy R. Reibman
School of Electrical and Computer Engineering
Purdue University
reibman@purdue.edu

ABSTRACT

First-person videos (FPVs) captured directly from wearable cameras are usually too shaky for humans to watch comfortably. Existing video stabilization methods can solve the problem but remove important First-person motion information (FPMI) from the FPVs. FPMI contains both subjective and objective information: First-person feeling (FPF) and First-person motion range (FPMR), respectively. In this paper, we propose measurement of both the stability and FPMI of FPVs. To improve the viewing experience of FPVs, which includes both video stability and FPMI, we develop a video processing system based on these measurements. Objective experiments show that the measurement we propose is robust under time shift, angular estimation drifting and white noise. Our subjective tests show that (1) our measurement can correctly compare the stability and FPMI of a FPV across different versions of the same content, and (2) our video processing system can effectively improve the viewing experience of FPVs.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision*;

KEYWORDS

First-person videos; Video stabilization; Viewing experience

1 INTRODUCTION

First-person videos (FPVs) are captured using the wearable cameras which are becoming popular recently. Compared with traditional types of videos, they can be used to share experience from the First-person perspective, such as playing sports and more generally, the recorders' life-logs. Some producers even started to make First-person films. However, the original version of these FPVs are usually too shaky for humans to watch comfortably. As a result, additional tools such as gimbals are needed and the photographers are required to have specific training. Our target is to increase the viewing experience of FPVs based on computer vision techniques without any additional physical assistance.

The viewing experience of FPVs includes two parts. The most basic part is the stability of the FPVs. To enhance it, traditional video stabilization techniques are effective and have been well-developed. Normally, three steps are performed: motion estimation, motion smoothing and frame/video reforming. Based on the motion type they work with, the video stabilization techniques are classified into 2D [14, 24, 27, 32] and 3D solutions [21, 26, 33, 38]. In the 2D solutions, the motion on the image plane is estimated using either local features or pixel intensity information. Then the frame transformation is calculated based on the smoothed 2D motions. In contrast, 3D solutions estimate the camera motions in the 3D world. The estimation approaches mainly rely on the methods of either structure-from-motion (SfM) [6] or visual-based simultaneous localization and mapping (vSLAM) [37]. The advantage of 3D solutions is that they have a full understanding of the physical camera motions.

The second and less-often considered part of viewing experience of FPVs is the First-person motion information (FPMI). FPMI mainly consists of the recorder's motion intentions. For improving this part, the traditional stabilization techniques have some difficulties which are caused by their motion smoothing step. Traditional video stabilization techniques are primarily designed for hand-held videos which are preferred to be similar to cinematographic videos after processing[12, 13]. As a result, their motion smoothing approaches simply apply low-pass filters to the estimated motions [19, 25, 29, 36]. Some other works [12, 21, 26] modify this but only add constraints to the smoothing procedure such as minimizing the black area caused by the homographic transformation or minimizing the image mosaic errors. So either the low-pass filter approach or the constraint-based approach leads to the result of over-stabilizing the FPVs, which will remove almost all the FPMI. To fix this problem while also addressing the stabilization problem, we propose measurements for both the stability and FPMI based on a human perception model in order to carefully design the new camera motion.

Despite the traditional video stabilization techniques, there are also related works of processing FPVs [15, 21, 31, 35]. Their general goals are the same as ours: create watchable egocentric videos. However, they place different requirements on the resulting videos. To fix the problem of motion smoothing, their strategy is to find and remove the video parts that have the undesired motions, which means they allow the resulting video to be a reduced or fast-forwarded version of the original one. When constructing the reduced version of the video, they only choose the semantic segments, for example the segments that contain human faces. However, we believe that not only are the selected specific frames semantically meaningful but also there is semantic meaning in First-person motions. In short,

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-5416-5/17/10…$15.00
DOI: https://doi.org/10.1145/3126686.3126748

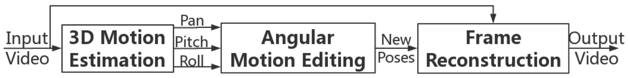


Figure 1: System Pipeline

Table 1: Concept Abbreviations

Abbreviations	Extend Names
FPVs	First-person videos
FPM	First-person motion
FPF	First-person feeling
FPMR	First-person motion range
FPMI	First-person motion information

these works ignore and discard the FPMI when processing FPVs. Only our prior work [28] takes a similar approach. However, they do not measure the viewing experience and their human perception model is inaccurate as we demonstrate in section 3.

In this work, we propose a system that can measure and improve the viewing experience of FPVs. Our strategy is to improve the stability of the FPVs while preserving an adequate amount of FPMI. And note that we do not discard any part of the video. In general, we follow the pipeline of 3D video stabilization techniques and replace the motion smoothing step with our approach, which can be illustrated by Fig. 1. In the next section, we introduce the geometric basis of our work, which is the 3D motion estimation. To form our measurement, we introduce a human perception model of FPVs in section 3. Then, in section 4, our proposed measurements are demonstrated. Based on the measurements, the detail of our motion editing method is shown in section 5. Objective and subjective experiments are illustrated in section 6 to demonstrate the robustness of our measurements and the whole system. Finally, we conclude our work in section 7.

Note that we define several new concepts in this work. For readers' convenience, we summarize them in Table. 1.

2 3D CAMERA MOTIONS

In [28], we demonstrated that to stabilize FPVs, only angular motions of the camera need to be estimated and stabilized. The translations known as head bobbing are necessary to provide a First-person feeling. The traditional SfM algorithm is simplified according to this statement, and the bundle adjustment is performed using the graph optimization tool provided by [5]. We adopt their algorithm for this part. Note that our motion estimation is not a real-time algorithm. Recent work in [10] introduced a robust real-time system. We do not focus on this topic in this paper.

Given the estimated camera poses, the Euler angles are used instead of quaternions to describe the rotation of a camera. This is because Euler angles have more physical meanings for humans. To further analyze and modify the angular motion, the rotation matrix is decomposed as:

$$R_{cam} = R_z(\theta_z)R_x(\theta_x)R_y(\theta_y), \quad (1)$$

where R_{cam} is the estimated camera motion. R_x, R_y and R_z are the rotation matrices for pitch, yaw and roll. This decomposition order aligns with the normal human motion order. The primary motion for human activities is yaw, which is performed to look around. Looking up and down has intermediate importance, and is performed by pitch. Roll is removed in the subsequent stabilization procedures since it is rarely performed on purpose.

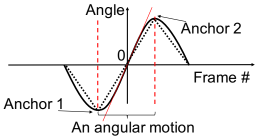


Figure 2: Example of a single angular motion

3 HUMAN PERCEPTION MODEL OF FPVS

In this section, we first demonstrate the difference between perception in real-life and while watching FPVs. Then we introduce the eye movement model proposed in [8] and the basic geometry of eye movement while watching FPVs. Then by considering a practical situation, we reform it into a more general eye movement model which forms the basis of measuring the viewing experience of FPVs.

3.1 Smooth pursuit and catch-up saccade

The experience of watching a FPV usually is not identical to what the recorder experienced. As demonstrated in [28], this is because these two situations are described using different human perception models. In real-life, the recorder performs the vestibulo-ocular reflex (VOR) to compensate for rotations of the head in order to maintain the image of the target on the fovea. The frequency of VOR can reach up to 100 Hz [1], which ensures the stabilization can be performed in real-time.

On the contrary, when watching a FPV, the spectator performs smooth pursuit eye movements (SPEM) to follow the target. Note that the SPEM only relies on visual clues and is less efficient than VOR. Before the spectator starts to pursue the target, a catch-up saccade needs to be performed to catch the target, which takes nearly 150 ms [7]. A catch-up saccade is also triggered when SPEM lag behind the target. During a catch-up saccade, visual information is not processed. This leads to an experience of instability.

3.2 Triggering condition of catch-up saccade

Fig. 2 shows an example of angular motion decomposed using the approach in section 2. A single motion starts from one local extreme and ends at the next one, defining two motion anchors. The motion changes rapidly around each motion anchor. Between them, there is a constant speed area in which the spectator can perform SPEM to follow the target.

In [28], three assumptions were made based on this motion model, which are not strictly precise:

(1) A catch-up saccade is always triggered at the beginning and the ending of a single motion;
(2) During the constant speed area, the spectator is guaranteed to perform SPEM without any catch-up saccade;
(3) As long as spectators' eye movement aligns with the estimated camera motion, they can smoothly pursue the target.

In this paper, we build a more precise model that is not restricted by these assumptions.

First, we introduce the triggering condition for a catch-up saccade using the model in [8]. In this model, a catch-up saccade is triggered based on the eye-crossing time (T_{XE}) that the human

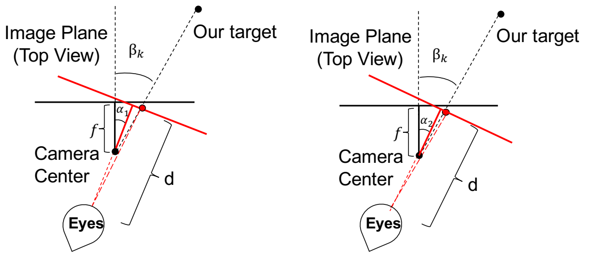


Figure 3: Camera motion (yaw/pitch) and object image

brain estimate. If the T_{XE} is between 40 ms and 180 ms, the catch-up saccade is not triggered. T_{XE} is defined as:

$$T_{XE} = \frac{-PE}{\omega_{rs}}, \tag{2}$$

where PE is the angular position error. ω_{rs} is the relative angular speed (target's image speed on fovea) or so-called retina slip. Intuitively, their ratio is the time that human eyes need to catch the target.

Based on this condition, we proposed a probabilistic model of SPEM for watching FPVs. We demonstrate the basic geometry in section 3.3. The probabilistic model and measuring approaches are shown in section 4.

3.3 Basic geometry of SPEM for FPVs

Fig. 3 shows the geometric relationship among the target object, camera and spectator across the time from the top view. Images from the left to right describe that the camera yaws from angle α_1 to angle α_2 and captures the object on the image plane. The spectator perceives this process through the images.

Note that for FPVs, the translation of the recorder can be ignored within a short period since it is relatively small. As a result, the relative position of the target with respect to the camera center is fixed. Also, in real life the depth of most objects is much larger than the focal length, so the height of the object in the image plane remains the same. Then it is reasonable and convenient to illustrate this geometry in 2D from the top view.

Suppose the target position with respect to the camera center is β_k at frame k while the camera focal length is f. The viewing distance of the spectator is d and the estimated camera position is θ. So the observation angle of the target for the spectator is:

$$\varphi_{obj}(k;\beta_k) = \arctan\left[\frac{f\tan\beta_k}{d}\right]. \tag{3}$$

At frame m, the observation angle changes to:

$$\varphi_{obj}(m;\beta_k) = \arctan\left[\frac{f\tan(\beta_k - \sum_{i=k}^{m-1}\alpha_i)}{d}\right], \tag{4}$$

$$\alpha_i = \theta(i+1) - \theta(i). \tag{5}$$

Assume that the frame rate is 30 and the spectator performs a SPEM from frame k to frame $(k+1)$, then the $PE(\beta_k)$ and $\omega_{rs}(\beta_k)$ at frame $(k+2)$ can be calculated as:

$$PE(\beta_k;k+2) = \varphi_{obj}(k+2;\beta_k) - 2\varphi_{obj}(k+1;\beta_k) + \varphi_{obj}(k;\beta_k), \tag{6}$$

$$\omega_{rs}(\beta_k;k+2) = 30\cdot PE(\beta_k;k+2). \tag{7}$$

The geometry of roll motion is different from yaw and pitch as shown in Fig. 4. Suppose the object is on the $x-z$ plane in the k^{th} frame. Then the position at frame m is:

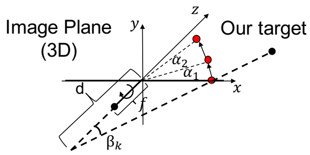


Figure 4: Geometry of roll motion

$$\varphi_{obj}(m;\beta_k) = \arctan\left[\frac{2r\sin(\sum_{i=k}^{n-1}\alpha_i/2)}{\sqrt{(f+d)^2+r^2}}\right], \tag{8}$$

$$r = f\tan\beta_k. \tag{9}$$

Directly applying the condition given in equation (2) indicates that the catch-up saccade is always being triggered, which is inconsistent with the actual situation. Several reasons cause this.

Firstly, the model in [8] treats the SPEM as an open-loop system. The position errors are generated by changing the target position abruptly. This weakens the predictive ability of SPEM, which is the key of the closed-loop characteristic of SPEM [22]. In addition, [8] used laser spots or circles as the tracking target to test the SPEM properties of human eyes, which will also underestimate the predictive ability of SPEM. According to [3, 4], the target shape can provide additional information for visual tracking. Moreover, [2, 22, 34] concluded that the predictive ability can be generated by scene understanding or the experiences of motion patterns. None of them are taken into account in [8]. Consequently, we relax the constraint by recognizing the spectator can utilize this information. For example, without the additional information, the gaze may lead the target in both position and velocity, which causes a negative value of T_{XE}. When the additional information is available, the eye movement can be de-accelerated before the next frame is shown, which makes the T_{XE} fall into the desired region.

Secondly, the sensitivity of the human visual system needs to be taken into consideration in practical situations. A position error less than the minimum angular resolution cannot be perceived by human eyes. Meanwhile, human eyes also have errors estimating the position error.

As a result, we relax the constraint in equation (2) by treating the SPEM as a closed-loop system. The condition required to maintain SPEM becomes:

$$0.04 \le \frac{|PE(\beta_n;n+2)| + b}{|\omega_{rs}(\beta_n;n+2)|} \le 0.18 \; or \; |PE(\beta_n;n+2)| < MAR, \tag{10}$$

where MAR is the minimum angular resolution of human eyes, and b is the bias of position error estimation which is set to MAR.

Note that equation (10) is a condition related to the angular position of the target β_n. By solving equation (10) for each frame, we can obtain an interval of β_n. Any object in the current frame that has an angular position within this interval can be tracked without having a catch-up saccade between the next two frames, if the SPEM has already been performed. We define this interval as $B(n)$ for future convenience.

4 VIEWING EXPERIENCE MEASUREMENTS

Given the camera motion of yaw, pitch or roll, we can find the corresponding object position interval $B(n)$. In this section, we first

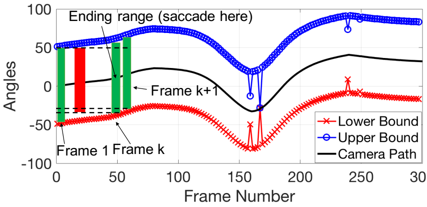


Figure 5: Example of $\tilde{B}(n)$

introduce a measurement of viewing experience (VE) using $B(n)$ based on a probabilistic model. Then, to speed up our algorithm, we propose a simpler measurement based on $B(n)$. However, it does not have physical meaning like the probabilistic model. After that, we introduce our approach to combine the measures of all 3 motions: yaw, pitch and roll.

4.1 Viewing experience score

Our viewing experience (VE) score is based on $B(n)$ and a probabilistic model. It measures the fraction of frames that can be viewed by the spectator using SPEM.

Recall that $B(n)$ is the object position interval of each frame. However, each frame has its own camera position θ with respect to the first frame. As a result, $\tilde{B}(n)$ is calculated where all frames share the same coordinate:

$$\tilde{B}(n) = B(n) + \theta(n). \tag{11}$$

In Fig. 5, the upper and lower boundaries show an example of $\tilde{B}(n)$. Unlike $B(n)$, $\tilde{B}(n)$ not only includes the object position interval of each frame but also has the spatial relationship between the intervals across the whole video.

To understand the procedure of computing VE, consider the example in Fig. 5. Suppose the spectator randomly chooses a target to track within the FOV from frame 1. Within a short period, the objects keep at the same location with respect to the camera in the first frame. So the trajectory of the target in this figure is a straight horizontal line. When this line intersects with the boundaries $\tilde{B}(n)$, the spectator loses this target. In this situation, one of two things happens. The spectator can perform a catch-up saccade to follow the previous target. Or he/she can perform a saccade eye movement to randomly retarget a new object within the FOV. Either of these two procedures takes nearly 6 frames[7].

The procedure of targeting and smooth pursuit is defined as a trail of tracking. A possible path of watching a video consists of several trails of tracking. By calculating the length of each possible path and their probability, we can find the expected value of the fraction of frames for which the spectator performs SPEM. As a result, a wider, more open pathway between the upper and lower bounds shown in Fig. 5 will produce a higher expected value, i.e. a higher VE score.

First, we compute the probability of a single tracking trail. Define $V_{i,j}$ to be the event:

$$V_{i,j} = \{\text{Target can be tracked from frame i to j}\}.$$

Define $T_{i,j}$ to be the event:

$$T_{i,j} = \{\text{Target is tracked from frame i and lost at frame j}\}.$$

Then we have:

$$\begin{aligned} Prob(T_{i,i+k+1}) &= Prob(\overline{V}_{i,i+k+1}, V_{i,i+k}) \\ &= Prob(V_{i,i+k}) - Prob(V_{i,i+k+1}), \end{aligned} \tag{12}$$

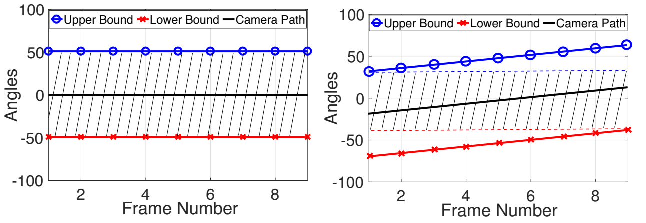


Figure 6: $\tilde{B}(n)$ of zero motion and motion with slope

where

$$Prob(V_{i,i+k}) = \frac{\max \bigcap_{n=i}^{i+k} B(n) - \min \bigcap_{n=i}^{i+k} B(n)}{FOV}. \tag{13}$$

Note that here we assume the objects are uniformly distributed within the FOV.

Then we need to compute the probability of a possible path. Suppose for an N-frame video, a possible path $path_m$ has n trails of tracking. Then the probability of $path_m$ is:

$$Prob(path_m) = \prod_{i=1}^{n-1} Prob(T_{l_m(i), l_m(i+1)-6}) \cdot Prob(V_{l_m(n), N}), \tag{14}$$

where $l_m(i)$ encodes the start frame index of each trail of tracking in $path_m$. And the length of this possible path is:

$$L(path_m) = N - 6(n-1). \tag{15}$$

So our VE score of a video from frame i to frame $(i+N)$ can be computed as:

$$VE(i; N, \theta) = \frac{1}{N+1} \sum_m Prob(path_m) \cdot L(path_m), \tag{16}$$

where θ is the camera motion.

The reason we calculate a VE score for $(N+1)$ frames instead of the whole video is that objects may only be visible for a short period. To compute the equation (16), we first identify all the possible paths for a length $(N+1)$ video. Then we check whether each path is feasible or not. A path is not feasible when any of its trails are not feasible, which is indicated when the probabilities in equation (12) and (14) are less than 0. Then we let:

$$Prob(path_m) = 0. \tag{17}$$

To reduce the computational complexity, we set $(N+1)$ to 10. As a result, for a K-frame video, $\tilde{B}(n)$ of yaw, pitch or roll is a $(K-2)$ by 2 vector, and their VE score has length $(K-12)$.

4.2 Structure Viewing experience score

The VE score proposed in the previous section has a physical meaning. It represents the fraction of frames that can be viewed using SPEM. However, we find that computing this VE score may be time consuming. So we propose a similar measurement that can be computed rapidly: the Structure Viewing experience (SVE) score.

We start with considering the mechanism of VE score. Fig. 6 shows the $\tilde{B}(n)$ of a zero motion and a motion with slope. The zero motion has the largest VE score since a target can be tracked across the whole video as long as it is within the FOV. However, the motion with slope yields the VE score that is smaller than 1. Intuitively, the shaded area in Fig. 6 changes with the motion slope. This illustrates that the VE score not only depends on the interval value $\tilde{B}(n)$ of each frame but also depends on the shape of $\tilde{B}(n)$. The more open the pathway of $\tilde{B}(n)$ is, the higher VE the video has.

Inspired by this, we proposed the SVE score. One approach would be to compute the shaded area in Fig. 6. However, it becomes more complex when the motion varies significantly. Instead, we use the following equations to compute SVE.

$$SVE(i; N, \theta) = 1 - \frac{\sum_{m=0}^{N-1} \left(\Delta\tilde{B}(i+m)\right)\left(N-m\right)}{N \cdot FOV}, \tag{18}$$

$$\Delta\tilde{B}(n) = \tilde{B}(n+1) - \tilde{B}(n). \tag{19}$$

We first compute the $\Delta\tilde{B}$ of each pair of adjacent frames using equation (19). $\Delta\tilde{B}(n)$ quantifies the number of objects that we lose tracking from frame n to $(n+1)$. Then we assign different weights to $\Delta\tilde{B}$ at different time instants in equation (18), because earlier time instants are more influential to the openness of the object position interval $\tilde{B}(n)$. As a result, SVE has the similar property with VE. The more complex the motion is, the smaller the SVE/VE is. By testing, when $N = 9$, computing VE needs 0.425 seconds for a 300-frame video while SVE only needs 0.155 seconds. More robustness experiments are shown in section 6.

4.3 Score of a video

Note that either the VE score or SVE score measures just a single motion. To obtain the viewing experience measurement of the whole video, we need to combine the scores of all three of its motions: yaw, pitch and roll. Suppose the measurement for these motions are M_y, M_x and M_z respectively. The measurement for the whole video is M_{all}. Then it is natural to have the following combination approach:

$$M_{all}(i; N, \theta) = \min_{j=x,y,z} M_j(i; N, \theta). \tag{20}$$

Equation (20) implies that the viewing experience measurement of the whole video is limited by the measurement of the most shaky motion.

5 MOTION EDITING

In this section, we introduce our motion editing method based on the VE/SVE score proposed in the previous section. Our target is to increase the stability of FPVs while preserving an adequate amount of First-person motion information (FPMI). To achieve this goal, we first systematically define the First-person motion (FPM) including its structure and properties. Then we illustrate the optimization procedure that helps us to re-design the camera path. After that, we supply details of how we speed up the optimization procedure.

5.1 First-person motion

First-person motion (FPM) is the camera motion estimated from FPVs. Although we introduce the perceptual geometry based on it in section 3.3, we have not demonstrated it systematically as an object that we are going to measure and edit. First of all, we clarify the intuitive parts of FPM: its properties, which are the parts the spectator directly perceives. Then we introduce our definition of the FPM structure, which actually controls its properties. After that, the logic of our measurement assignment of different parts of FPM properties would be more apparent.

5.1.1 Properties of FPM. FPM has two properties: stability and FPMI. We define the sum of them to be the viewing experience of a FPV. The stability describes the comfort extent of watching the FPV. A low stability makes it difficult for the spectator to perceive the content of the video and may even causes dizziness.

The FPMI is the information conveyed by FPM. It has two subparts: First-person motion range (FPMR) and First-person feeling (FPF). FPMR is the objective information, which is produced when the recorder looks around. The damage to FPMR when doing motion editing makes the spectator lose any chance to observe some particular objects. FPF is the subjective information, which is the sense of watching activities captured from the perspective of a person. It is produced by two conflicts. The first conflict is between the feeling of watching FPVs and traditional videos (such as cinematographic videos). The spectator finds that the current video (FPV) is not like the video he/she usually watches. By watching more FPVs, this conflict may be eliminated. The second conflict for the spectator is that the motion he/she perceives by watching FPVs is not consistent with what he/she perceives in his/her own First-person experience. The more obvious the conflict is, the more obvious the FPF is. Actually, in real life, spectator eyes do not perceive many large motions using SPEM (but using VOR).

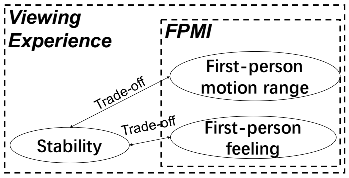


Figure 7: Properties of First-person motion

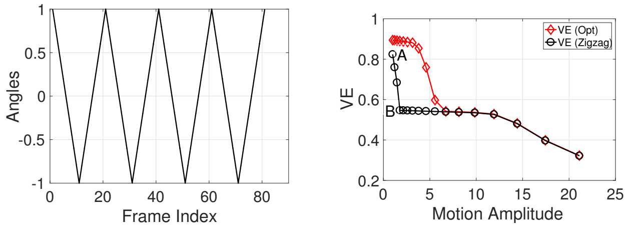
There is another potential conflict which is not included here. This third conflict happens in the vestibular system caused by mismatched motions: the visual motions make humans feel that their body is moving while the body has no physical motion. It causes disorder in the visual system, which is called vestibular illusion. However, this effect varies with the strength of visual cues as illustrated in [23]. Stronger visual clues make the spectator more confident about the mismatched motions. Compared with VR systems, this effect has limited influence for 2D screens where FPVs are played back.

Fig. 7 shows the relationship between the stability and FPMI. In general, there is a trade-off between them. When the FPM has large amplitude, the FPMI increases (in both FPMR and FPF) while the stability decreases. However, we show in section 5.2, it is still possible to increase the stability while preserving the FPMI, i.e. increasing the viewing experience of a FPV.

5.1.2 The structure of FPM. We model the FPM using the concepts of motion anchors and motion shape. The motion anchor is first defined in section 3 and Fig. 2. It is the core of FPM. Given fixed motion anchors for a FPV, the motion amplitude and motion frequency are fixed, which means the FPMI is fixed. The motion shape is the path between motion anchors. It only influences the stability of FPVs. Note that as the core of FPM, the motion anchors also influence the stability of FPVs. This is because, the motion anchors are the foundation of motion shapes. Motions have shapes after their amplitude and frequency are fixed.

So the strategy applied in here is a two-step iteration. First, we take a set of particular motion anchors to preserve the FPMI. Then we find the motion shapes that produce the highest stability based on this particular set of motion anchors. The iteration terminates when we find the desirable motions that preserve an adequate amount of FPMI and increase the stability.

However, we need a measurement for motion anchors. Considering the method proposed in section 4, we define the pure FPM as



(a) Example of pure motion **(b) VE scores**

Figure 8: VE scores based on pure motions

the zigzag path that connects all the motion anchors. An example of the pure FPM is shown in Fig. 2 using dotted lines. As long as the motion anchors are fixed, the pure FPM is determined, and vice versa. As a result, the pure FPM is independent of motion shape. So we can use the measurement in section 4 of the pure FPM as the descriptor of motion anchors. Since FPMI is only determined by the motion anchors, any measurement of the pure FPM is also a measurement of the FPMI.

To simplify the problem, in the current work, we fix the motion frequency of the FPM and only focus on the motion amplitudes and motion shape when we try to find the desirable motions.

5.1.3 Measurements assignment. So far, we introduced the properties and the structures of FPM. The properties are what we want to measure and the structures are what we can extract out of the video. The idea here is to assign different measures to the structures, which allows us to quantify the properties.

The VE/SVE score introduced in section 4 measures both FPMI and the stability of FPVs. Given a FPM, the stability is measured by the absolute VE/SVE score since their definitions align with each other. However, the FPMI is measured by the negative VE/SVE score of its pure FPM. This is because, firstly, as demonstrated in section 5.1.2, any measurement of the pure FPM can be seen as a measurement of the FPMI. Secondly, consider the second kind of conflict discussed in section 5.1.1. The FPMI increases when the conflict becomes more obvious. In this case, the FPV has more or large FPMs, which decreases the VE/SVE score.

Note that given a FPM, based on its pure FPM, we can find the motion shapes which combine with the motion anchors to give the highest VE/SVE score. This VE/SVE score measures the highest potential stability of this FPM based on its pure FPM.

To better illustrate the logic of measurement assignment, consider the Fig. 8. Fig. 8 (a) shows an example of pure motion: a zigzag path between motion anchors that have a constant frequency and amplitude. We use the easing function method (introduced in the next section) to find the optimal motions that give the highest VE scores. And we do this process by varying the amplitude of the zigzag motion from 1 to 20. Fig. 8 (b) shows the VE scores of both the zigzag motions and their corresponding optimal motions. The ratio of viewing distance with respect to the focal length is 6. Increasing (reducing) this ratio only shifts the curves to the right (left) while it does not change their shapes.

Fig. 8 shows that, as the motion amplitude increases, the VE scores of the optimal motions decrease since the motions become larger and shakier. Therefore, we assign VE/SVE score to be the measurement for motion stability. The VE scores of the zigzag motions are more interesting. There is a sharp decrease from point A to point B in Fig. 8 (b). As we proposed, the negative VE score of the zigzag motion measures the FPMI, where lower values indicate more FPMI. So the FPMI increases significantly from A to B. This is because point B is where the FPF becomes noticeable, which means the second kind of conflict we discussed in section 5.1.1 becomes obvious. As the motion amplitude increases beyond this point, the VE scores of the zigzag motion do not decrease as quickly after point B. The small decreases are due to the increasing angular variation corresponding to increasing FPMR.

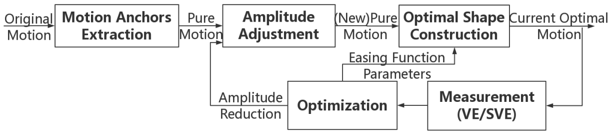


Figure 9: General procedure of motion editing

5.2 Optimization procedure and speeding up

In this section, we introduce the details of the optimization procedure of motion editing. First, we introduce its general procedure along with the parameters which need optimizing. Then the main objective function is discussed. After that, we demonstrate how we speed up the optimization procedure.

5.2.1 General procedure and optimization parameters. Fig. 9 shows the general procedure of our motion editing algorithm. Given the pure motion extracted from the original motion, the optimization module focuses on two tasks: adjusting the motion amplitude and finding the optimal shapes for the adjusted motion anchors. As a result, there are two kinds of parameters the optimization module must deal with: the motion amplitude reduction rates and the parameters describing the motion shape.

Suppose the frame index of the i^{th} motion anchor is $A(i)$, the original camera motion is θ and the new camera motion is $\tilde{\theta}$. Their pure motions are θ_p and $\tilde{\theta}_p$.Then the motion amplitude reduction rates of the i^{th} motion anchor is defined as $D(i)$:

$$D(i) = \frac{\left|\tilde{\theta}_p\big(A(i)\big) - \theta_p\big(A(i)\big)\right|}{\left|\theta_p\big(A(i)\big) - \theta_p\Big(A\Big(\arg\min_{i\neq j}\left|A(j) - A(i)\right|\Big)\Big)\right|}. \tag{21}$$

Easing function methods are popular in the computer graphics community [18, 30]. They are usually used to construct motion shapes. However, for a motion anchor pair, if we adopt the n-dimensional polynomial easing function, it will produce n unknown parameters. It is too time consuming to run the optimization since a normal FPV usually has over 30 motion anchors for every 300 frames. Considering the constraint that there is no local extreme between two adjacent motion anchors, we construct our own easing function in equation (22) which only requires 2 parameters for a motion anchor pair.

$$\tilde{\theta}(n; k_i, s_i) = \tilde{\theta}_p(n) + s_i \cdot \left[\frac{n - \big(A(i) + A(i+1)\big)/2}{\big(A(i+1) - A(i)\big)/2}\right]^{k_i} \Delta\tilde{\theta}_p(n), \tag{22}$$

$$\Delta\tilde{\theta}_p(n) = \tilde{\theta}_p\Big(\arg\min_{A}\big(A(i) - n,\, n - A(i+1)\big)\Big) - \tilde{\theta}_p(n), \tag{23}$$

where k is the parameter that controls the degree and s is the scalar.

5.2.2 Objective function and speeding up algorithm. Recall that our target is to increase the stability of FPVs while preserving an adequate amount of FPMI. Although we illustrate that the FPF is proportional to the degree of the second kind of conflict in section 5.1.1, the conflict is also harmful and may cause motion sickness. So the point B shown in Fig. 8 (b) is the optimal point we desire.

At point B, the FPF just becomes noticeable and the stability of the optimal motion we can get is at a high level. So the strategy is that we want the VE/SVE score of the optimal motion to be high (to have a high stability). Meanwhile, we want the VE/SVE score of the corresponding zigzag motion (pure motion) to be low (to have necessary amount of FPF).

In addition, we need to consider the FPMR. By decreasing the amplitude of motion anchors, we can increase the stability and preserve an adequate amount of FPF, but this damages the FPMR. To solve this problem, we make compensation for those motions which are larger than the FOV. Usually, motions smaller than FOV are just vibrations or unintentional head motions. As a result, our objective function is constructed as:

$$\min_{D,k,s}\left[1-\frac{\left\|M(i;N,\tilde{\theta})\right\|}{K-N-2}+\frac{\left\|M(i;N,\tilde{\theta}_p)\right\|}{K-N-2}\right]+\alpha_{FPMR}\cdot\Delta FPMR^T, \quad (24)$$

where K is the number of frames. $M(\cdot)$ is either $VE(\cdot)$ or $SVE(\cdot)$, which has dimension $(K-N-2)$ as shown in section 4.1. α_{FPMR} is the vector of weights for motions that are larger than FOV:

$$\alpha_{FPMR}(i)=\frac{\Delta\theta(i)\cdot\mathbb{1}_{\{x>FOV\}}\Delta\theta(i)}{\sum_j\Delta\theta(j)\cdot\mathbb{1}_{\{x>FOV\}}\Delta\theta(j)}, \quad (25)$$

$$\Delta\theta(i)=\theta\big(A(i)\big)-\theta\big(A(i-1)\big), \quad (26)$$

And $\Delta FPMR$ is the vector of distortions:

$$\Delta FPMR(i)=1-\frac{\tilde{\theta}\big(A(i)\big)}{\theta\big(A(i)\big)}. \quad (27)$$

The optimization is performed based on particle swarm. However, we notice that for a camera motion which has T motion anchors, the objective function has $T+2(T-1)$ variables. When T is large, more particles and longer convergence time are required. As a result, we construct a look-up table to reduce the number of variables. The periodic motion in Fig. 8 (a) is used. Since only $\|M(i;N,\tilde{\theta})\|$ contains parameters k and s, we can pre-find all the k and s using the objective function below for all possible motion amplitude:

$$\min_{k,s} 1-\frac{\|M(i;N,\tilde{\theta})\|}{K-N-2}. \quad (28)$$

This look-up table enables us to eliminate the parameter k and s in equation (24). However, note that this training process of roll motion needs to be taken separately from the yaw and pitch motion. This is because they have different equations of target position as shown in equation (5) and (9).

6 EXPERIMENTS

In this section, we conduct both objective and subjective tests to evaluate our measurements of viewing experience. The performance of our motion editing algorithm is also discussed.

6.1 Objective tests

Both of our measurements VE and SVE are subjective measurements. Although the subjective tests are usually applied to evaluate this kind of measurements, the objective tests are also useful to demonstrate its robustness.

Without subjective tests, we have no knowledge about the difference between two random motions. However, we can create similar enough motions that are expected to have nearly equivalent viewing experience for the spectator based on some simple human perception models. First, we use the sine-wave as our base motion. Then we modify the sine-wave motion by making small changes. The small changes are bounded by the minimum angular resolution (MAR) so that the spectator should not be able to perceive apparent difference between the base motion and the synthetic motions. After that, we compute the VE/SVE scores of the base motion and the synthetic motions. If the scores are close, then it is reasonable to conclude that the VE/SVE measurement is robust.

The synthetic motions are generated using 4 operators: flipping, shifting, adding Gaussian noise and adding slope. The flipping operator simply flips the sine-wave from left to right or up-side down. The shifting operator randomly chooses a frame index, replicates the corresponding position value to the next frame and all of the remaining position values are shifted to the right by 1 frame. The operator of adding noise adds normally distributed Gaussian noise using the following equation:

$$\theta_{syn}(n)-\theta_{syn}(n-1)=\theta(n)-\theta(n-1)+\mathcal{N}\Big(0,\big(\frac{MAR\cdot f}{3d}\big)^2\Big), \quad (29)$$

where f is the camera focal length and d is the viewing distance in pixels. The adopted distribution bounds the amplitude of the noise so that the noise added to the position error is smaller than MAR in human eyes with probability 99.73%. It is equivalent to:

$$\theta_{syn}=\theta+\mathcal{N}\Big(0,\big(\frac{MAR\cdot f}{12d}\big)^2\Big). \quad (30)$$

The operator of adding slope is the same with adding accumulated noise based on equation (30), which models the drifting in motion estimation. Then we are able to compute the measurement errors caused by the difference between the base motion and the synthetic motions using:

$$e=\frac{1}{r}\sum_{i=1}^{r}\left|\frac{\|M(i;N,\theta^i_{syn})\|-\|M(i;N,\theta)\|}{\|M(i;N,\theta)\|}\right|\cdot 100\%, \quad (31)$$

where r is the number of experiment trials. The sine-wave we use has amplitude 1, period 20 and 12 periods. We perform the experiment by increasing the amplitude of the sine-wave from 1 to 20. For each amplitude the number of trials n is set to 100 and MAR is set to be 0.02. We assume the resolution of the video of the synthetic motion is 1080p, the viewing distance is 3240 and focal length is 830. The error of VE and SVE scores are shown in Fig. 10.

From Fig. 10, we can see that our two measurements VE and SVE are both robust under different operators. There are three main observations. Firstly, the flipping operator does not influence the scores at all. Meanwhile, the shifting operator has the largest error since it changes motion shape more heavily than other operators. Secondly, VE and SVE have similar performance when the motion amplitude is large. Meanwhile, they are more robust at high amplitude than at low amplitude. This is simply because when amplitude is large, the noises are relatively small. Thirdly, in the low amplitude region, SVE is more stable than VE under all operators except flipping. However, we cannot assert that SVE is better than VE. On the contrary, larger changes under noise implies that VE is a more sensitive measure, especially in the low amplitude region is where the FPF increases dramatically according to Fig. 8 (b). We prefer a measurement that has a higher sensitivity. This is supported by the tests in the next section.

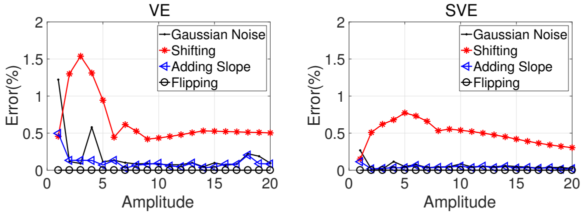


Figure 10: Error caused by synthetic motions

Table 2: Fraction of Correct Distance Conditions

	VE				
Conditions	Cdn 1	Cdn 2	Cdn 3	Cdn 4	Total
Stability	1	1	1	0.8	0.95
FPMI	0.8	1	1	0.8	0.9

	SVE				
Conditions	Cdn 1	Cdn 2	Cdn 3	Cdn 4	Total
Stability	1	1	1	0.6	0.9
FPMI	0.8	1	1	0	0.7

6.2 Subjective tests

Firstly, we treat our VE/SVE as quality estimators [17]. We use videos and subjective test results from [28] to test our quality estimators. As a result, this experiment is independent of our motion editing step and simply considers how effective the VE/SVE is to measure the viewing experience. Secondly, we conduct our own subjective test to evaluate our whole system.

6.2.1 Test for the measurements. The resulting videos in [28] include three versions: the original ones, results of their algorithm and the results from Microsoft Hyperlapse [20, 21]. Their subjective scores are computed based on the Bradley-Terry model [16]. To test the effective of measurements in the sense of subjective measurements, we apply the model proposed in [9].

The general idea of this model is to examine the consistency of the subjective scores and the quality estimator scores. As illustrated in [9], if the Bradley-Terry scores of three version of the videos have the relationship: $BT_1 < BT_2 < BT_3$, then the scores of a quality estimator should be: $QE_1 < QE_2 < QE_3$. Meanwhile, the distance between the subjective scores and the distance between the quality estimator scores should be similar. As a result, the model [9] yields the following conditions:

$$sign(BT_3 - BT_2) = sign(QE_3 - QE_2), \tag{32}$$

$$sign(BT_3 - BT_1) = sign(QE_3 - QE_1), \tag{33}$$

$$sign(BT_2 - BT_1) = sign(QE_2 - QE_1), \tag{34}$$

$$sign(BT_3 - 2BT_2 + BT_1) = sign(QE_3 - 2QE_2 + QE_1). \tag{35}$$

By calculating the number of conditions that are satisfied, we can use the fraction of correctness to evaluate the effectiveness of our measurements. The results are shown in Table. 2. We can see that for the first three conditions, VE and SVE have the same performance. It shows that they are both robust for the simple ranking tasks of stability or FPMI. However, for the condition 4, VE is much better than SVE, which can support the demonstration in the objective tests. VE is a more sensitive measure, which ensures the distance similarity between subjective scores and the quality estimator scores, especially for measuring the FPMI.

6.2.2 Test for the whole system. To evaluate our whole system, we conduct a similar subjective test to the one in [28]. We use the same source videos, which includes 5 different scenes shot by GoPro Hero Session 4 with 1080p. Three versions are prepared for each video: the original one, our resulting video and Microsoft's result. All the test parameters are the same: videos are played back on a 27-inch, 82 PPI screen and the ratio of viewing distance with respect to the equivalent focal length is 4.

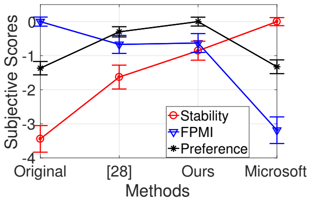


Figure 11: Result comparison of different version videos

Our test uses paired comparison, which includes 21 subjects. Subjects are asked the following questions after shown each comparison: (1) **Which video is more stable**; (2) **In which video you can recognize more First-person motion**; (3) **If your friend tries to share his/her First-person experience with you, which one do you prefer**. The subjective scores (with 95% confidence interval) calculated using the Bradley-Terry model [16] are shown in Fig. 11, which also includes the result from [28]. A higher score indicates higher stability, more FPMI or higher preference.

As expected, the original videos have the highest score for FPMI and Microsoft's videos have the highest stability. Our system is in the second place for these two scores. Meanwhile, the scores of our videos are closer to the desired videos: our FPMI score is closer to the original videos' while our stability score is closer to the Microsoft's. This illustrates that our system can effectively increase the stability while preserving an adequate amount of FPMI. [28] has a similar FPMI score to ours. However, their stability score is lower than ours. Moreover, the preference score of our videos is the highest. As a result, we can conclude that our system is more effective than the one in [28]. This is because our system is based on a more precise human perception model and has carefully designed measurements for the viewing experience.

7 CONCLUSIONS

In this paper, we proposed two measurements (VE and SVE) that can quantify both the stability and the First-person feeling of a FPV. Based on the measurements, we further proposed a system that can enhance the viewing experience of FPVs. To accomplish these two items, we described the human perception model, analyzed the perceptual geometry of watching FPVs and systematically defined the First-person motion and its properties. The objective tests show that our measurements are robust under different operators that can create visually equivalent motions. The subjective tests show that both measures of VE or SVE highly align with the subjective scores. Also, our system can effectively increase the stability of FPVs while preserving an adequate amount of FPMI.

Our work still has places can be improved. First, we will enhance and speed up 3D motion estimation algorithm based on a recent work [10], which can improve the robustness in a feature-less environment. Based on this, we further plan to remove the rolling shutter by incorporating [11]. We also consider including image stitching algorithms to remove the black area that appears in our videos.

REFERENCES

[1] ST Aw, GM Halmagyi, T. Haslwanter, IS Curthoys, RA Yavor, and MJ Todd. 1996. 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* 76, 6 (1996), 4021–4030.

[2] A Terry Bahill and Jack D McDonald. 1983. Smooth pursuit eye movements in response to predictable target motions. *Vision Research* 23, 12 (1983), 1573–1583.

[3] Brent R Beutter and Leland S Stone. 1998. Human motion perception and smooth eye movements slow similar directional biases for elongated apertures. *Vision Research* 38, 9 (1998), 1273–1286.

[4] Brent R Beutter and Leland S Stone. 2000. Motion coherence affects human perception and pursuit similarly. *Visual Neuroscience* 17, 01 (2000), 139–153.

[5] Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert. 2015. Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*. 4597–4604.

[6] Jonathan L. Carrivick, Mark W. Smith, and Duncan J. Quincey. 2016. Background to Structure from Motion. *Structure from Motion in the Geosciences* (2016), 37–59.

[7] Sophie de Brouwer, Marcus Missal, Graham Barnes, and Philippe Lefèvre. 2002. Quantitative analysis of catch-up saccades during sustained pursuit. *Journal of Neurophysiology* 87, 4 (2002), 1772–1780.

[8] Sophie De Brouwer, Demet Yuksel, Gunnar Blohm, Marcus Missal, and Philippe Lefèvre. 2002. What triggers catch-up saccades during visual tracking? *Journal of Neurophysiology* 87, 3 (2002), 1646–1650.

[9] Ali Murat Demirtas, Amy R Reibman, and Hamid Jafarkhani. 2014. Full-reference quality estimation for images with different spatial resolutions. *IEEE Transactions on Image Processing* 23, 5 (2014), 2069–2080.

[10] Jakob Engel, Vladlen Koltun, and Daniel Cremers. 2017. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[11] Per-Erik Forssén and Erik Ringaby. 2010. Rectifying rolling shutter video from hand-held devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 507–514.

[12] Michael L. Gleicher and Feng Liu. 2007. Re-cinematography: improving the camera dynamics of casual video. In *Proceedings of the 15th ACM international conference on Multimedia*. 27–36.

[13] Michael L. Gleicher and Feng Liu. 2008. Re-cinematography: Improving the camerawork of casual video. *ACM Transactions on Multimedia Computing, Communications, and Applications* 5, 1 (2008), 2.

[14] Matthias Grundmann, Vivek Kwatra, and Irfan Essa. 2011. Auto-directed video stabilization with robust L1 optimal camera paths. In *IEEE Conference on Computer Vision and Pattern Recognition*. 225–232.

[15] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European Conference on Computer Vision*. Springer, 505–520.

[16] John C. Handley. 2001. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In *PICS*, Vol. 1. 108–112.

[17] Sheila S Hemami and Amy R Reibman. 2010. No-reference image and video quality estimation: Applications and human-motivated design. *Signal processing: Image communication* 25, 7 (2010), 469–481.

[18] Łukasz Izdebski and Dariusz Sawicki. 2016. Easing Functions in the New Form Based on Bézier Curves. In *International Conference on Computer Vision and Graphics*. Springer, 37–48.

[19] Chao Jia and Brian L Evans. 2017. Online motion smoothing for video stabilization via constrained multiple-model estimation. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 25.

[20] Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F Cohen. 2015. Real-time hyperlapse creation via optimal frame selection. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 63.

[21] Johannes Kopf, Michael F. Cohen, and Richard Szeliski. 2014. First-person hyperlapse videos. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 78.

[22] Eileen Kowler. 2011. Eye movements: The past 25years. *Vision Research* 51, 13 (2011), 1457–1483.

[23] Steven M LaValle. 2016. Virtual reality. *Champaign (IL): University of Illinois* (2016).

[24] Ken-Yi Lee, Yung-Yu Chuang, Bing-Yu Chen, and Ming Ouhyoung. 2009. Video stabilization using robust feature trajectories. In *IEEE International Conference on Computer Vision*. 1397–1404.

[25] Andrey Litvin, Janusz Konrad, and William C Karl. 2003. Probabilistic video stabilization using Kalman filtering and mosaicing. In *Electronic Imaging 2003*. International Society for Optics and Photonics, 663–674.

[26] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. 2009. Content-preserving warps for 3D video stabilization. *ACM Transactions on Graphics* 28, 3 (2009), 44.

[27] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. 2011. Subspace video stabilization. *ACM Transactions on Graphics* 30, 1 (2011), 4.

[28] Biao Ma and Amy R. Reibman. 2017. Enhancing Viewability for First-person Videos based on a Human Perception Model. In *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*.

[29] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. 2006. Full-frame video stabilization with motion inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 7 (2006), 1150–1163.

[30] Robert Penner. 2002. Motion, tweening, and easing. *Programming Macromedia Flash MX* (2002), 191–240.

[31] Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg. 2015. Egosampling: Fast-forward and stereo for egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4768–4776.

[32] Hui Qu and Li Song. 2013. Video stabilization with L1–L2 optimization. In *IEEE International Conference on Image Processing*. 29–33.

[33] Erik Ringaby and Per-Erik Forssén. 2012. Efficient video rectification and stabilisation for cell-phones. *International Journal of Computer Vision* 96, 3 (2012), 335–352.

[34] Gerben Rotman, Nikolaus F Troje, Roland S Johansson, and J Randall Flanagan. 2006. Eye movements when observing predictable and unpredictable actions. *Journal of Neurophysiology* 96, 3 (2006), 1358–1369.

[35] Michel Melo Silva, Washington Luis Souza Ramos, Joao Pedro Klock Ferreira, Mario Fernando Montenegro Campos, and Erickson Rangel Nascimento. 2016. Towards Semantic Fast-Forward and Stabilized Egocentric Videos. In *European Conference on Computer Vision*. Springer, 557–571.

[36] Zhongqiang Wang and Hua Huang. 2016. Pixel-wise video stabilization. *Multimedia Tools and Applications* 75, 23 (2016), 15939–15954.

[37] Khalid Yousif, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. 2015. An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics. *Intelligent Industrial Systems* 1, 4 (2015), 289–311.

[38] Guofeng Zhang, Wei Hua, Xueying Qin, Yuanlong Shao, and Hujun Bao. 2009. Video stabilization based on a 3D perspective camera model. *The Visual Computer* 25, 11 (2009), 997–1008.