Wu XQ, Li HS, Cao J *et al.* Geometry of motion for video shakiness detection. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(3): 475–486 May 2018. DOI 10.1007/s11390-018-1832-5

Geometry of Motion for Video Shakiness Detection

Xiao-Qun Wu, *Member*, *CCF*, Hai-Sheng Li, *Member*, *CCF*, *Member*, *IEEE*, Jian Cao, *Member*, *CCF* and Qiang Cai, *Senior Member*, *CCF*, *Member*, *IEEE*

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China

Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University Beijing 100048, China

E-mail: xiaoqunwu@gmail.com; {lihsh, caojian, caiq}@th.btbu.edu.cn

Received January 4, 2018; revised March 23, 2018.

Abstract This paper presents a novel algorithm for automatically detecting global shakiness in casual videos. Per-frame amplitude is computed by the geometry of motion, based on the kinematic model defined by inter-frame geometric transformations. Inspired by motion perception, we investigate the just-noticeable amplitude of shaky motion perceived by the human visual system. Then, we use the thresholding contrast strategy on the statistics of per-frame amplitudes to determine the occurrence of perceived shakiness. For testing the detection accuracy, a dataset of video clips is constructed with manual shakiness label as the ground truth. The experiments demonstrate that our algorithm can obtain good detection accuracy that is in concordance with subjective judgement on the videos in the dataset.

Keywords video shakiness, kinematic model, motion perception

1 Introduction

With the proliferation of inexpensive video recording devices, there has been a dramatic increase in the amount of video content. But image distortions frequently appear in video frames due to the involuntary vibration of cameras in the recording[1], especially for those casual videos recorded by the amateurs using hand-held video cameras. Video shakiness and motion blur are two frequent distorting effects related to this scenario, which often degrade the visual quality of videos by hiding vital information. Therefore, it needs to eliminate these distortions prior to assisting some intelligent video processing and multimedia applications[2], e.g., video conferencing, video surveillance, and multimedia communication. Many image/video deblurring and stabilization algorithms have been developed in the past decade[3-4], where the automatic detection of motion blur and shakiness becomes necessary in the pre-processing stage. While the detection of motion blur has been extensively studied and is relatively mature[5-7], video shakiness detection is less explored in the field of video stabilization, especially designed for casual shaky videos. In this paper, we concern on the phenomenon of video shakiness, caused by camera vibration, and seek for an automatic shakiness detection algorithm oriented to casual videos.

Video shakiness, also known as video jitter or instability, is caused by the disturbance of steady camera movement in the recording process, which generates the sense that the scene is oscillating through frames. It is a very important preprocessing step for a variety of video processing tasks. Unfortunately, a few existing algorithms for shakiness detection do not suit casual videos[8-9]. The challenge to this problem mainly lies in that the sense of shakiness essentially relates to subjective response on motion conditions like frequency and amplitude. But there is less appropriate formulation yet in accordance with perceptual interpretation that

Regular Paper

Special Section of CVM 2018

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61602015, the Open Funding Project of State Key Laboratory of Virtual Reality Technology and Systems at Beihang University under Grant No. BUAAVR-16KF-06, Beijing Natural Science Foundation under Grant No. 4162019, and the Research Foundation for Young Scholars of Beijing Technology and Business University.

is able to faithfully characterize these conditions. Besides, unlike motion blur that can be detected just by the spatial appearance degradation even from a single image[7], video shakiness involves motion disturbance within an ambiguous timeline window in the temporal domain. Thus it seems to be very difficult to identify video shakiness from just one frame. Considering the above issues, the purpose of this paper is to propose an appropriate motion model that is able to faithfully identify shakiness of casual videos.

The main contribution of our work is a new kinematic model to identify the video shakiness. The geometry of motion defined by inter-frame transformations enables per-frame shakiness amplitude computation that conforms to motion perception. We also build a dataset with manual labels on frames to annotate their shaky or stable attributes, which can be used as the ground truth to evaluate the accuracy of the shakiness detection algorithm. The experiments on the dataset demonstrate the efficiency and effectiveness of our algorithm for detecting shakiness, especially for casual videos.

2 Related Work

Video shakiness is ubiquitous and gets involved in a variety of video processing tasks, like video encoding[10], quality assessment[11], retargeting[12], stabilization[4,13-14], editing[15-16], and hyperlapse creation[17-18]. Although a wealth of methods for motion blur detection exist, like using domain adaption[5-6,19] or fusion[20-21], there has been much less effort towards video shakiness detection. Unlike motion blur, the detection of video shakiness is impossible from just a single image, while it requires temporal analysis on the continuous change of frame appearance. However, some existing work has drawn on the idea of domain adaption from blur detection for shakiness detection, which is relevant to our purpose.

Visentini-Scarzanella and Dragotti[8] proposed a video jitter analysis tool in the scenario of video forensics. They computed the high energy components of feature trajectories in the frequency domain, by using 2-level wavelet decomposition. Then, they compared the high energy components with some trained video dataset to identify the jitter. This method is good at detecting small global jitter in re-capturing videos, but does not work well for detecting shakiness in casual videos with non-static background. Sibiryakov[9] defined the video jitter as the global projection-based error between two adjacent frames on the intensity domain, and used it as a sort of descriptors for matching videos in a dataset. This method is designed for video content identification, but does not give any criterion for judging the shakiness of a single video itself. For home videos, Yan and Kankanhalli[22] defined video shakiness as the repeated forward and backward movement along same directions. They used the difference of direction and amplitude of relative displacement vectors between adjacent frames for video shakiness identification, but this method can only detect translational motion shake.

Video stabilization is highly related to video shakiness detection. Video stabilization usually contains three steps: motion estimation, motion smoothing, and motion compensation, which resorts to optimizing on the motion representation for visual smoothness[13,23]. Recently, paper [14] proposes a global approach which is claimed flexible and efficient by solving a quadratic minimization problem defined for image warps. And a novel formulation of video stabilization in the space of geometric transformation is presented in [24]. The optimized smooth path is cast as the geodesics on the Lie group embedded in transformation space. However, all these approaches focus on the stabilization without detecting whether the video is shaky or not.

Considering the above minor progress on shakiness detection for casual videos, we attempt to step forward by making comprehensive investigation on the perception of video shakiness and then propose a shaky motion model for automatic shakiness detection in the sequel.

3 Kinematic Model Based on Inter-Frame Geometric Transformation

To describe the shakiness characterization, we need to establish a descriptive motion model to represent the smooth or shaky motion. Furthermore, the sensation of shaky motion is also a perceptual attribute, while the motion model should be adapted to accommodate the motion perception.

3.1 Kinematics-Based Motion Model

Kinematics describes the object motion from the geometry of the system without considering the physical conditions of the object, like mass or force that causes the motion. It is suggested that trajectories for human reaching motions from one point to another are chosen so that they minimize the integral

of the square norm of jerk (i.e., the derivative of motion acceleration)[25]. Thus, an ideal smooth motion described by a path $\boldsymbol{A}(t)$ from the perspective of kinematics is assumed to minimize the following energy functional[26]:

$$\mathcal{J} = \frac{1}{2}\int_{t_0}^{t_1} < \nabla_V \nabla_V V, \nabla_V \nabla_V V > \mathrm{d}t, \qquad (1)$$

where $V = \frac{\mathrm{d}\boldsymbol{A}(t)}{\mathrm{d}t}$ is the velocity of the motion, ∇ is the operator of affine connection defined in the tangent space at each instant position along the path, and $< \cdot, \cdot >$ is the inner product. Actually, (1) gives a kinematic description on the smoothest motion to bring the hand from the initial position to the final position in a given time interval $[t_0, t_1]$. Physically, this equation describes the change of motion accelerations along the path.

Unfortunately, (1) cannot be solved analytically, but it can be proved that with homogeneous boundary conditions in velocities and accelerations, the minimum jerk curve trajectories follow the same path as the geodesics[26]. Therefore, we can use the geodesic path to approximate the optimal smooth path defined by (1). Mathematically, the geodesics between the initial position at the time t_0 and the final position at t_1 is defined as the path optimizing the following integral energy functional:

$$\mathcal{G}(\boldsymbol{A}(t)) = \int_{t_0}^{t_1} < \frac{\mathrm{d}\boldsymbol{A}(t)}{\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{A}(t)}{\mathrm{d}t} > \mathrm{d}t, \qquad (2)$$

which minimizes the change of motion velocities along the path. Here, we assume the geodesic solution of (2) is $\boldsymbol{P}(t)$. In mathematics, solving $\boldsymbol{P}(t)$ is a mature problem from the perspective of differential geometry[27], which can be easily computed once specifying the form of camera motion $\boldsymbol{A}(t)$. Specially, such a geodesic path even has a closed-form solution as demonstrated in Section 4.

For shaky motion, its path deviates a potential geodesic path as described above. Then, the amplitude of shakiness can be formulated as the instant deviation between $\boldsymbol{A}(t)$ and $\boldsymbol{P}(t)$, denoted by $s_t = \boldsymbol{P}(t) \ominus \boldsymbol{A}(t)$, where the symbol $\ominus$ means the relative difference between the two paths with respect to a given metric.

3.2 Inter-Frame Geometric Transformation Motion

The camera motion typically has two kinds of representations: parametric and non-parametric. We can apply the motion pattern analysis as in [28-29] to obtain fine-scaled non-parametric motion model, but it is computationally intensive and does not suit analytic calculation in (1). Therefore we resort to the parametric motion model of geometric transformations, which converts to a series of mapping functions between corresponding points of two frames. Generally, the mapping function is parameterized by a matrix $\boldsymbol{M}_t \in \mathbb{R}^{n\times n}$ for representing the camera motion, which can be chosen from a set of transformations, e.g., rigid, similarity, affine and projective transformations.

In reality, the ideal camera movement follows a three-dimensional (3D) path defined by rigid transformations in the Euclidean space, which involves 3-axis shifts (forward/backward, left/right, and up/down) and 3-axis rotations (yaw, pitch, and roll). But in the case only given the two-dimensional (2D) video frames, recovering the 3D path by using the methods like structure-from-motion (SfM) is expensive and brittle in practice[23]. And thus previous work prefers to inferring 2D movement directly from adjacent frames to approximate the camera movement.

When the camera rotates by θ_p, θ_t, θ_r in pan, tilt, and roll axes, respectively, an arbitrary point $\boldsymbol{p} = (x, y)$ of the image is moved to the following point $\boldsymbol{p}' = (x', y')$[30]:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta_r & -\sin\theta_r \\ \sin\theta_r & \cos\theta_r \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} d_p \\ d_t \end{pmatrix}, \qquad (3)$$

where $d_t = L \times \tan\theta_t \approx L \times \theta_t$, $d_p = L \times \tan\theta_p \approx L \times \theta_p$, and L is the camera-to-object distance. Therefore considering the parallel shifts with respect to the image plane, i.e., left/right and up/down shifts, we can use 2D rigid transformation (including the components of rotation and translation) to approximate the 2D camera movement, which derives the following parametric expression for the transformation $\boldsymbol{M}_t$:

$$\bar{\boldsymbol{p}'} = \boldsymbol{M}_t(\bar{\boldsymbol{p}}) = \begin{pmatrix} \boldsymbol{R}_t & \boldsymbol{d}_t \\ \boldsymbol{0} & 1 \end{pmatrix} \bar{\boldsymbol{p}}, \qquad (4)$$

where $\boldsymbol{R}_t \in \mathbb{R}^{2\times 2}$ is the planar rotation transformation matrix, $\boldsymbol{d}_t \in \mathbb{R}^2$ is the planar translation transformation, and $\bar{\boldsymbol{p}'} = [\boldsymbol{p}', 1]^{\mathrm{T}}$ and $\bar{\boldsymbol{p}} = [\boldsymbol{p}, 1]^{\mathrm{T}}$ are the homogeneous coordinates of the corresponding points. Actually, all the rotation and translation transformations form the special Euclidean group (SE(2))①, which is essentially a Lie group, i.e., an algebraic group with

① The special Euclidean group is $\mathrm{SE}(2) = \left\{ \begin{pmatrix} \boldsymbol{R} & \boldsymbol{d} \\ \boldsymbol{0} & 1 \end{pmatrix}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \boldsymbol{Id}, \det \boldsymbol{R} = 1, \boldsymbol{d} \in \mathbb{R}^2 \right\}$, where $\boldsymbol{Id}$ is the identity matrix.

the structure of differentiable manifold[27]. Then, the kinematic smooth motion $\boldsymbol{P}_t$ can be formulated as the geodesics on SE(2). We will take the above motion model as an example, and demonstrate how to engage the above perceptual and kinematic characterization in shakiness detection on SE(2).

Remark. The global parametric motion model of (4) has been a general and mandatory step in many video processing methods like [22, 24, 30] and software like VirtualDub Deshaker②. However, the novelty of our work is to take it as the element in the Lie group, by which we can derive the shakiness amplitude in the sense of kinematics based on the manifold metric. This will assist the shakiness detection as elaborated in Section 4.

3.3 Shaky Motion Perception

Generally, human perceive motion from the displacement of retinal images, together with the persistence of vision to form the temporal variation[31]. Thus, besides the kinematic motion model, we need to investigate the perceptual cues about shakiness, by which the established motion model is able to correctly interpret the perception on motion shake.

Typically, the motion perception of shakiness arises from the perceived deviation from the pursuit on the illusory smooth motion[32]. Thus, it is closely related to two vibration conditions, amplitude and frequency, with respect to an intended smooth motion trend. To sense the video shakiness, a necessary condition is that the displacement between consecutive frames is visible by stimulating retinal neuron, or saying it is beyond the visual acuity of human eyes. It has been demonstrated that visual acuity is about 1 arcmin (minute of arc) on average, i.e., the human eye can resolve around $\frac{1}{60}$ of a degree[33]. Considering the preferred viewing distance between 20 and 40 inches (about 50 cm and 100 cm) for viewing computer screen with retinal quality resolution, it can be derived that the minimal displacement that human eyes can sense is about 0.5 pixels in general. Hence, the amplitude of shakiness is assumed to be at least 0.5 pixels for visible identification on the shaky motion.

As for the condition of frequency, it is found that human eyes themselves enable voluntary adaption on target vibration in the range of 0.5 Hz∼2 Hz[34], due to the vestibular system (organ of balance) functions through vestibulo-ocular reflex and vestibulo-spinal reflex[32]. Actually, this body resonance mechanism imposes a constant attenuation on shakiness from outside vibration. Moreover, the persistence of vision is between 0.1 seconds and 0.4 seconds around. Therefore, a necessary condition for sensing shaky motion is that the duration of displacement caused by retinal imaging changes is more than 0.1 seconds. While common video frame rate is about 25∼30 frames per second, it is about 0.1 seconds for three frames. Hence, the shakiness inbetween three frames can be sensed by human visual system, which means the shakiness frequency is at least 3 fps in sensing video shaky motion.

Consequently, inspired by the above motion perception cues, shakiness can be characterized by the relative displacement through three consecutive frames with amplitude more than 0.5 pixels, termed as just-noticeable shakiness (JNS) conditions for sensing shaky motion.

4 Video Shakiness Detection on SE(2)

The kinematic characterization of shakiness in Section 3 relies on the geodesics as the ideal motion path to define the amplitude. Fortunately, we have a closed-form solution for computing the geodesics on SE(2). In this case, the geodesic solution $\boldsymbol{P}_t$ of (2) between the initial position at t_0 and the final position at t_1 can be explicitly computed based on the rotational and translational components as follows:

$$\tilde{\boldsymbol{R}}_t = \boldsymbol{R}_{t_0} \exp(\Omega_0 \tilde{t}), \quad \tilde{\boldsymbol{d}}_t = \tilde{t}(\boldsymbol{d}_{t_1} - \boldsymbol{d}_{t_0}) + \boldsymbol{d}_{t_0}, \tag{5}$$

where $\tilde{t} = (t - t_0)/(t_1 - t_0)$, $\Omega_0 = \log(\boldsymbol{R}_{t_0}^{\mathrm{T}} \cdot \boldsymbol{R}_{t_1})$, and $\exp(\cdot)$ and $\log(\cdot)$ are the matrix exponential map③ and logarithm④ respectively[26].

Mathematically, the Lie group of SE(2) is equivalent to the product group of rotation transformations and translation transformations, which can be denoted by SO(2) and $\mathbb{T}(2)$ respectively, i.e., $\mathrm{SE}(2) = \mathrm{SO}(2) \otimes \mathbb{T}(2)$, where $\otimes$ is the direct product of groups. Besides, we can parameterize the rotation transformation of SO(2) with a 1D rotational angle $\theta_t \in \mathbb{R}$ as $\boldsymbol{R}_t(\theta_t)$, and the translation transformation of $\mathbb{T}(2)$ with 2D vector as $\boldsymbol{d}_t = (d_t^x, d_t^y)^{\mathrm{T}}$. Hence, we have the following parametric expression of the transformation motion model of

②http://www.guthspot.se/video/deshaker.htm, Mar. 2018

③The exponential of a matrix $\boldsymbol{A}$ is defined by $\exp(\boldsymbol{A}) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$.

④A matrix $\boldsymbol{B}$ is a logarithm of $\boldsymbol{A}$ if $\exp(\boldsymbol{B}) = \boldsymbol{A}$.

(4) as

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} \cos\theta_t & -\sin\theta_t & d_t^x \\ \sin\theta_t & \cos\theta_t & d_t^y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \quad (6)$$

Substituting the above transformation into (5), we obtain the geodesic solution represented by

$$\boldsymbol{R}_t(\theta_{\tilde{t}}) = \begin{pmatrix} \cos\theta_{\tilde{t}} & -\sin\theta_{\tilde{t}} \\ \sin\theta_{\tilde{t}} & \cos\theta_{\tilde{t}} \end{pmatrix}, \quad \boldsymbol{d}_t(d_t^x, d_t^y) = \tilde{t}\boldsymbol{d}_{t_0,t_1}, \quad (7)$$

where $\theta_{\tilde{t}} = \tilde{t}\theta_{t_0,t_1}$ and θ_{t_0,t_1} is the rotational angle from the starting to the end in the interval $[t_0, t_1]$, and $\boldsymbol{d}_{t_0,t_1} = \boldsymbol{d}_{t_1} - \boldsymbol{d}_{t_0}$ is the translational vector from the starting to the end in the interval $[t_0, t_1]$ (see Fig.1).

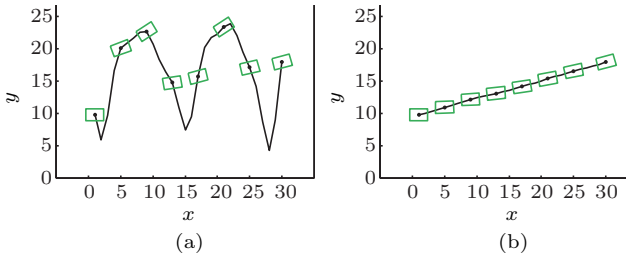


Fig.1. The geodesics on SE(2) is the linear interpolation of rotational angles and translational vectors between the starting and end points (blue dots) in the interval. (a) Original path. (b) Geodesic path. The green rectangles illustrate the rotation transformations of some sampled frames along the motion paths (black lines). The x-axis is the frame index, and the y-axis records the x-coordinate of the trajectory of the frame center.

As the shaky motion perception indicates the just-noticeable shakiness amplitude among three frames (see Subsection 3.3), we need to compute the geodesics going through consecutive three frames. Assuming the frames I_{k-1}, I_k and I_{k+1} and according to (7), we can derive the following geodesics through the three frames as

$$\begin{cases} \tilde{\boldsymbol{R}}_{k-1} = \boldsymbol{Id}, \\ \tilde{\boldsymbol{d}}_{k-1} = (0,0)^{\mathrm{T}}, \\ \tilde{\boldsymbol{R}}_k = \boldsymbol{R}(\tilde{\theta}_k = \dfrac{\theta_{k-1,k+1}}{2}), \\ \tilde{\boldsymbol{d}}_k = (\dfrac{d_{k-1,k+1}^x}{2}, \dfrac{d_{k-1,k+1}^y}{2})^{\mathrm{T}}, \\ \tilde{\boldsymbol{R}}_{k+1} = \boldsymbol{R}(\theta_{k-1,k+1}), \\ \tilde{\boldsymbol{d}}_{k+1} = (d_{k-1,k+1}^x, d_{k-1,k+1}^y)^{\mathrm{T}}, \end{cases} \quad (8)$$

where $\theta_{k-1,k+1}$ is the rotational angle from I_{k-1} to I_{k+1}, and $\boldsymbol{d}_{k-1,k+1}$ is the translational vector from I_{k-1} and I_{k+1}. It should be noted that the rotational angle here is directional such that the anti-clockwise angle has a positive value and the clockwise angle has a negative value. Intuitively, the geodesics through the three frames boils down to the linear interpolation of the rotational angle on SO(2) and translational vector on $\mathbb{T}(2)$ separately, which finally reconciles the composition of rotation and translation transformations as the smooth path through the three frames.

Then for the frame I_k, the deviation from the smooth path defined by the consecutive three frames is

$$\begin{aligned} \delta_k^{\boldsymbol{d}} &= \boldsymbol{d}_{k-1,k} - \tilde{\boldsymbol{d}}_k = \boldsymbol{d}_{k-1,k} - \boldsymbol{d}_{k-1,k+1}/2, \\ \delta_k^{\theta} &= \theta_{k-1,k} - \tilde{\theta}_k = \theta_{k-1,k} - \theta_{k-1,k+1}/2, \end{aligned} \quad (9)$$

which includes the components of both translation and rotation transformations (see Fig.2).

However, the deviation defined by (9) concerns the aspect of motion direction or magnitude. To obtain the real amplitude corresponding to motion shakiness, we must exclude the deviations of rotational angle and translational vectors without direction changes through the three frames (e.g., Fig.2(a)), and confine it to only the shakiness amplitude arising with directional changes (e.g., Figs.2(b) and 2(c)). Consequently, we define the shakiness amplitude with respect to the translation and the rotation transformations as:

$$\begin{aligned} s_k^{\boldsymbol{d}} &= \Xi\left(\frac{\boldsymbol{d}_{k-1,k} \cdot \boldsymbol{d}_{k,k+1}}{\|\boldsymbol{d}_{k-1,k}\|_2 \|\boldsymbol{d}_{k,k+1}\|_2}\right) \|\delta_k^{\boldsymbol{d}}\|_2, \\ s_k^{\theta} &= \Xi(\theta_{k-1,k}\theta_{k,k+1}) \|\delta_k^{\theta}\|_2, \end{aligned} \quad (10)$$

where $\|\cdot\|_2$ is the L_2-norm, $\Xi(\cdot)$ is the truncation function that satisfies $\Xi(x < 0) = 1$ and $\Xi(x \geqslant 0) = 0$. Thus, we obtain the amplitude as the pair of scalars of directional deviation, i.e., $s_k = \boldsymbol{P}(t) \ominus \boldsymbol{A}(t) = (s_k^{\boldsymbol{d}}, s_k^{\theta})$.

Actually, the amplitude defined in (10) also entails the degree of shakiness, i.e., a larger amplitude implies more severe shake. We use the scheme of threshold contrast (TC) to determine the shaky or stable attributes of each frame. TC is the minimum contrast at which the target can be distinguished from its surroundings. Formally, we set two parameters α and β as the criteria for translational and rotational shakiness thresholds, respectively, whereupon the frame with $s_k^{\boldsymbol{d}} > \alpha$ or $s_k^{\theta} > \beta$ is identified as shaky; otherwise stable. Thus, we can obtain the shakiness detection results for the input video frames. Fig.3 shows two examples of using our algorithm to detect the shakiness in the frames, with the shaky motion of either moderate amplitude (Fig.3(a)) or large amplitude (Fig.3(b)).

Setting of Threshold Values. Obviously, the shakiness detection result depends on the setting of the two

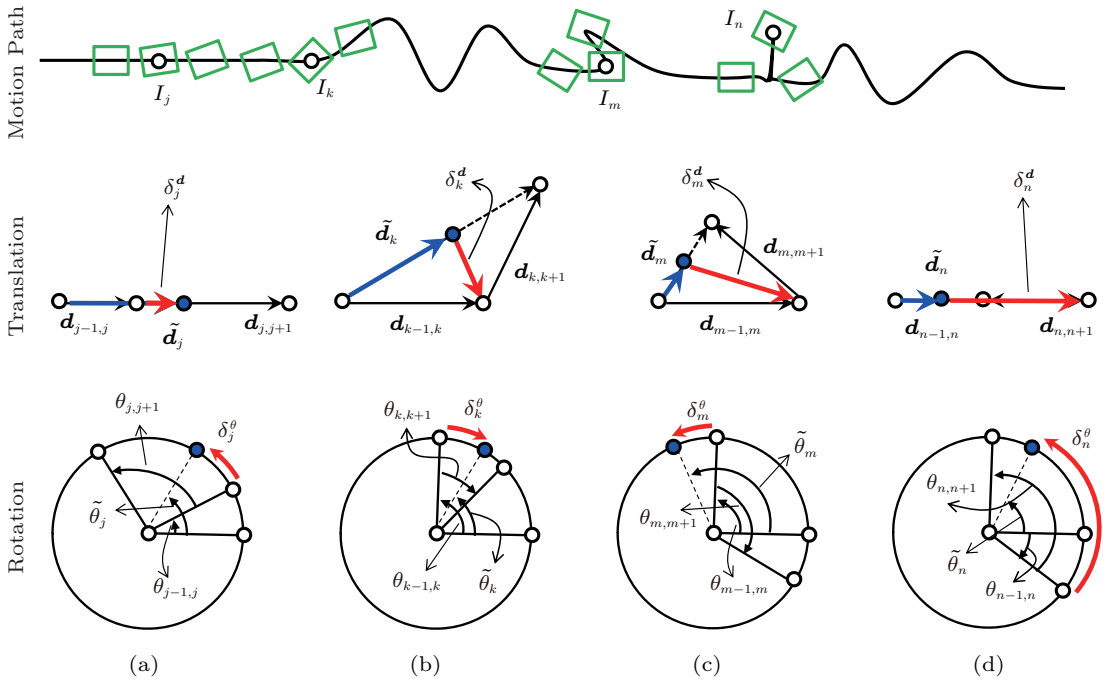


Fig.2. Shakiness amplitude corresponding to four instances of the frames: (a) I_j, (b) I_k, (c) I_m, and (d) I_n. From top to bottom: motion path, translational amplitude and rotational amplitude (as shown by the red arrows). The green rectangles in the motion path illustrate the rotation transformations between adjacent frames.

threshold values. The minimal amplitude for shakiness sensation in Subsection 3.3 suggests the threshold values to be $\alpha = 0.5$ pixels for the translational component. As for the rotational component, it is noticed that human eyes are sensitive to a rotation angle as small as about 3 arcmin[35], which suggests a default value for β. We call these two values to be the setting of the just-noticeable shakiness, because they are from the limited frequency response and lower bound of amplitude to induce the shaky motion perception, respectively. But in reality the values might be influenced by the imaging noise, and the computational model of shakiness amplitude also depends on the accuracy of feature points matching. We will give a full investigation on the setting of the threshold values in the experiments of Section 5.

Implementation Details. To compute the rigid transformation between two frames in (4), we adopt the pyramidal Lucas-Kanade[36] for feature points detection and matching among consecutive three frames. We also use the standard RANdom SAmple Consensus (RANSAC) method to filter out the outlier points that mostly locate in the moving foreground objects. Then, the remaining feature points of the two frames ($\{\boldsymbol{p}_k\}$ and $\{\boldsymbol{p}'_k\}$) are used to fit the best rotation transformation $\boldsymbol{R}_t$ and the best translational vector $\boldsymbol{d}_t$. Concretely, the translation is the vector between the geometric centers of the matching points, i.e., $\boldsymbol{d}_t = \boldsymbol{p}'_o - \boldsymbol{p}_o$, where $\boldsymbol{p}_o = \sum_k \boldsymbol{p}_k/N$, $\boldsymbol{p}'_o = \sum_k \boldsymbol{p}'_k/N$, and N is the number of feature points. Then, the rotation transformation is $\boldsymbol{R}_t = \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}$[26], where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the orthogonal matrices in the singular value decomposition (SVD), i.e., $[\boldsymbol{p}_k - \boldsymbol{p}_o]_{2\times N}[\boldsymbol{p}'_k - \boldsymbol{p}'_o]^{\mathrm{T}}_{2\times N} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, and $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values.

5 Experiments

We implement our video shakiness detection algorithm by C++ programming language, and also build a dataset of videos with different motion types as the benchmark for testing shakiness detection. Next, we will elaborate the details of the dataset and evaluate the performance of our shakiness detection algorithm by running it on the video dataset.

5.1 Dataset of Videos with Shakiness Label

For evaluating the performance of shakiness detection algorithm, we establish a dataset of video clips and label ground truth shakiness as the benchmark.

Dataset. The collected video clips are classified into two categories: professional videos and casual videos. The professional videos refer to ideally steady

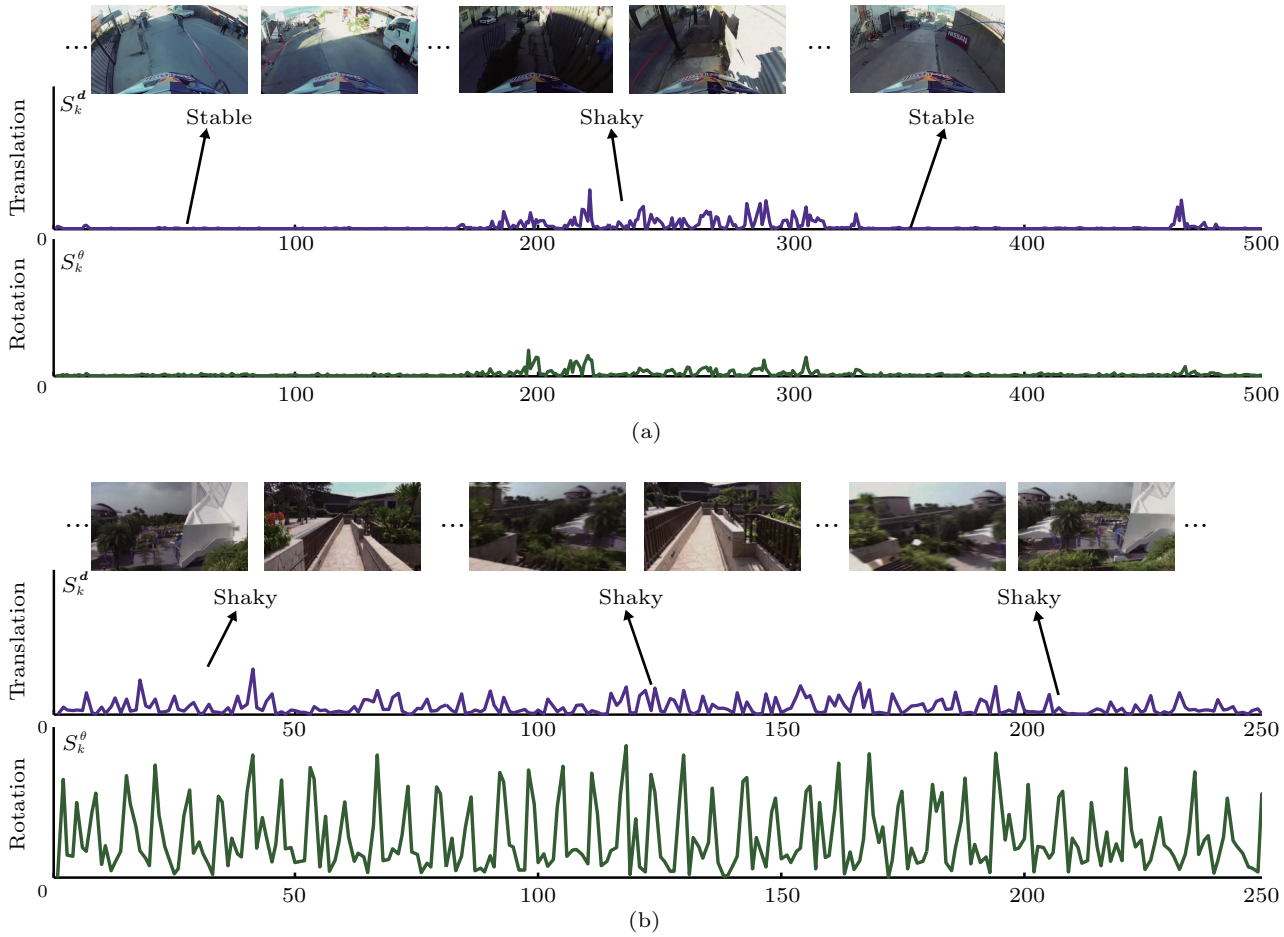


Fig.3. (a) and (b) are two examples of video shakiness detection according to the amplitudes of translational and rotational components respectively. The x-axis is the frame index and the y-axis is the amplitude.

videos of high quality, which consist of clip footages of movies and some other videos recorded by professional cameras equipped with hardware stabilizers. Concretely, we use the tool of tripod for recording steady videos by position-fixed camera, Dji Osmo[⑤] to obtain steady videos when walking on the ground, and drone equipped with steadicam for the steady aerial videos (see Fig.4(a)). Consequently, there are four sources of professional videos obtained from movie, tripod camera, Osmo camera and drone camera (see Table 1 and Fig.4(b)). It should be noted that the ideal steadiness of these professional videos is guaranteed by the auxiliary hardware and also confirmed by human subjective judgement in selecting them.

The casual videos refer to the ones that are recorded in the natural usage of the camera without any stabilizer, e.g., walking, running, riding, and driving. Generally, these videos possess the most common shaky motions in daily lives. We collect these videos from the Internet repository or record by ourselves using the amateur video cameras (see Table 1 and Fig.4(c)). For example, the riding videos are recorded by the camera fixed on a bicycle, and the driving videos are recorded by the dashcam installed in the automobile. Such casual videos usually interlace the steady and shaky motions in the frame sequence, which can be used to test shakiness detection algorithms.

Finally, our dataset has 60 video clips in total, of which there are 30 clips in the category of professional videos, and 30 clips in the category of casual videos, and the resolution is 1 280 × 720. These videos present a variety of motion types. The average length of the video clips in the dataset is about 17.2 seconds. Table 1 shows the statistics on the videos in the dataset, and Figs.4(b) and 4(c) show some sampled frames of video clips in the dataset. All the videos are with the frame-based shak-

⑤http://www.dji.com/product/osmo, Mar. 2018.

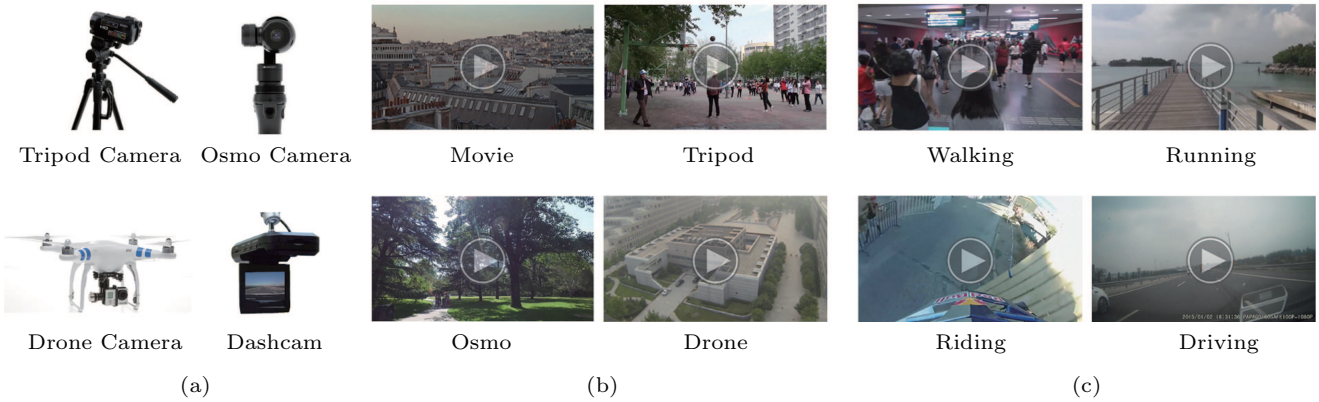


Fig.4. Dataset. (a) Some devices used in recording videos for our dataset. (b) Sampled frames of professional videos in our dataset. (c) Sampled frames of casual videos in our dataset.

Table 1. Statistics on the Video Clips in Our Dataset

Professional Video			Casual Video		
Source	#Videos	Avg. Length (s)	Source	#Videos	Avg. Length (s)
Movie	6	9.8	Walking	9	21.9
Tripod camera	9	21.8	Running	7	19.2
Osmo camera	7	20.3	Riding	8	18.3
Drone camera	8	9.7	Driving	6	17.8

Note: #Videos means the number of video clips recorded from different sources, and avg. length means their average length in seconds.

iness labeling results as the ground truth, whereupon this dataset can further be used as the test data for video shakiness detection algorithms.

Labeling Shakiness. To assess the shakiness detection accuracy, we need a definite shakiness label for each frame as the ground truth. Because video shakiness is a perceptual understanding, its criterion is obviously derived from human judgement on viewing the video content. Therefore, we recruit 20 people to label the shaky frames manually and watch the videos one by one, whilst they stuck the labels of shaky or stable on video frames to complete the ground truth shakiness label assignment. When watching the videos, we encourage people to sense the shaky motion based on the deviation from their experience on watching normal movie or TV videos, because these videos are ideally stable videos that are recorded by a static or moving camera mounted on the tripod or dolly. Besides, people are instructed to pay more attention to the motion shake located on the background, which actually delivers the shakiness caused by the camera vibration rather than the dynamic foreground objects.

The professional videos in the dataset are assumed to be ideally stable, and are all labeled as stable for the frames of each video by default. On the contrary, labeling the casual videos is an expensive task due to the complexity of shaky motion, which needs more efforts to assign the correct frame-based labels. It should be noted that the extreme fineness of labeling shakiness by checking each frame is meaningless and impossible in reality. Here, we just encourage people to isolate the sequential frames that they thought are shaky or stable as short as possible. Then, we average the beginning and the ending of the corresponding segments and label their frames as shaky or stable.

To complete this task, we design a custom interface to assist manually labeling on the video frames. This interface has a window to display the video, and the user can mark the beginning and the ending of a segment by clicking the window to add time stamps on the timeline. Thus if shakiness is viewed, the user clicks once as the starting of the shaky segment. If the shakiness disappears, the user clicks again as the end of the shaky segment. Our interface allows playback and fast-forward to help the user correct the beginning and the ending time. Then, we record the shakiness labels of the segments by shaky or stable for each video in the categories of casual videos. After the manually labeling process by all the people on the same video, the interface releases the average beginning and the average ending time as the time stamp for the initial position and the final position of the segments, respectively. In the final stage, it is allowed to further refine the beginning and the ending time by examining sequential

segments based on the averaged time stamps.

Consequently, we obtain a dataset that contains professional and casual videos with the shakiness labels assigned on their frames as the ground truth. This dataset involves common shaky motions that frequently appear when recording casual videos, which are assigned manual labels for shakiness identification. Next, we employ the videos in the dataset to evaluate the performance of our video shakiness detection algorithm.

5.2 Shakiness Detection Results

Given a video in the dataset, we run our algorithm and make statistics on the shakiness identification results. Then, we can obtain the shakiness label shaky or stable for each frame according to the decision made by thresholding based on (10). We did the experiments on both professional videos and casual videos in the dataset. All the experiments were done on a PC machine with 3.1 GHz Intel® Dual Core CPU and 8 GB RAM.

We report the performance of our algorithm in this subsection. The performance focuses on the precision and recall rate of the shakiness detection by comparing our resultant labels with the ground truth ones. It should be noted that the computation of precision and recall rate is based on the manually labeled segments in Subsection 5.1, i.e., if the segment which the detected shaky frame falls in is also labeled shaky as the ground truth, we consider the segment as the correct identification by our algorithm. Concretely, the precision P of the shakiness detection is defined as:

$$P = \frac{\#(\{\text{detected shaky segments}\} \cap \{\text{shaky segments}\})}{\#\{\text{detected shaky segments}\}}, \tag{11}$$

and the recall rate R in the shakiness detection is defined as

$$R = \frac{\#(\{\text{detected shaky segments}\} \cap \{\text{shaky segments}\})}{\#\{\text{shaky segments}\}}, \tag{12}$$

where $\#(\{A\})$ denotes the number of set A, {shaky segments} is the set of all the shaky segments that are manually identified, and {detected shaky segments} is the set of automatically detected shaky segments. We mainly use these two indices for evaluating the detection performance, and Fig.5 and Fig.6 show some statistics of the results by applying our algorithm on the dataset.

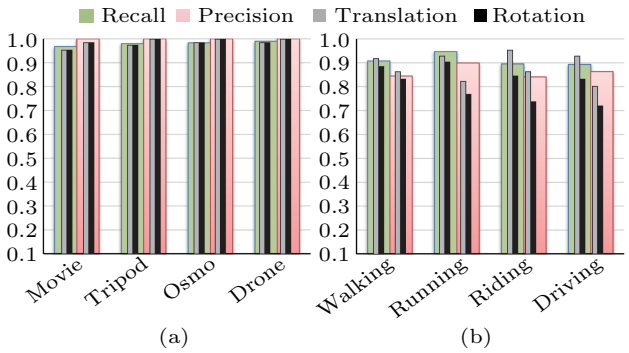


Fig.5. Statistics on precision and recall rate of shakiness detection by using the just-noticeable shakiness (JNS) threshold values $\alpha = 0.5$ pixels and $\beta = 3$ arcmin. (a) Professional videos. (b) Casual videos. The gray and dark bars show the separate influence of the translation and rotation on precision and recall rate.

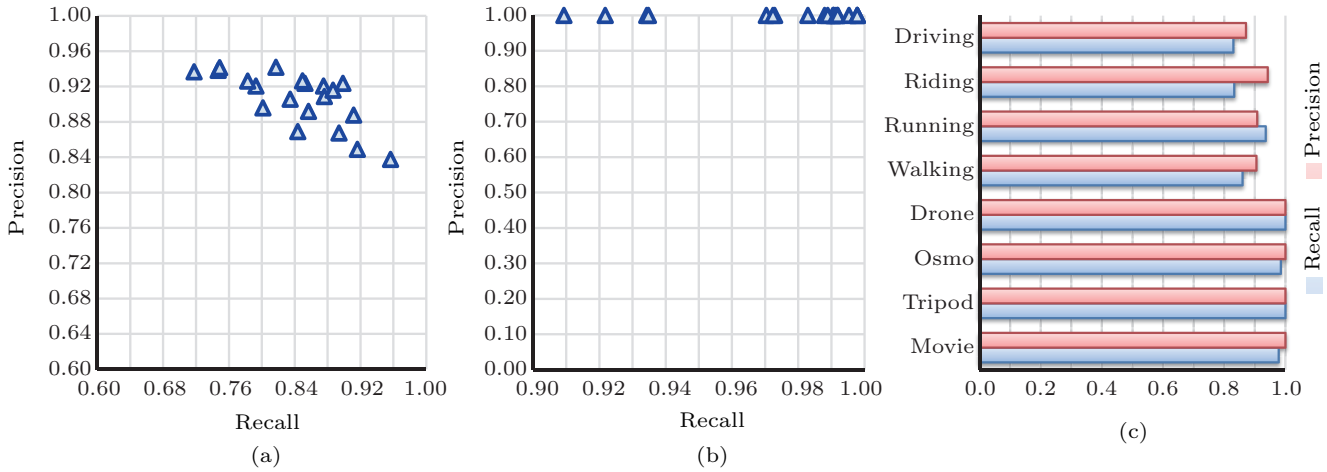


Fig.6. Precision-recall curves by setting different threshold values for (a) causal videos and (b) professional videos, where the purple diamonds denote some sampling threshold values. (c) Statistics on precision and recall rate by using the optimal threshold values $\alpha = 0.8$ pixels and $\beta = 2.2$ arcmin.

The running time of our algorithm is very short, which enables 82 fps for the video of $1\,280 \times 720$ resolution, due to the simplicity of shakiness amplitude computation. The memory cost of our algorithm is also low in the implementation because only three successive frames are involved in the computation. More importantly, our algorithm gains significant performance on the accuracy of video shakiness detection by setting the appropriate threshold values suggested as follows.

Influence of Translation and Rotation. To investigate the influence of translational and rotational components, we compute the precision and recall rate by shakiness detection based on just one of the two components (see Fig.5). It can be seen that both the translation and the rotation contribute to the overall motion shake, while the translational shakiness usually dominates the detection performance. Besides, the influence is also related to the types of scenes, where the use of one component might induce bad shakiness detection, e.g., for the videos of riding and driving in the scene of the street.

Threshold Values Setting. A key issue in shakiness detection is the setting of threshold values α and β for making shaky or stable decision. We have given the ideal just-noticeable shakiness conditions with $\alpha = 0.5$ pixels and $\beta = 3$ arcmin based on the motion perception analysis in Subsection 3.3. Fig.5 shows the shakiness detection results of the videos in the dataset by using these two JNS threshold values, where the statistics on the detection precision and the recall rate are illustrated. It should be noted that the precision and the recall for casual videos are computed based on the detected shaky frames with respect to the ground truth shaky labels. For professional videos, as all the frames are labeled stable as the ground truth, they are computed with respect to the stable frames, i.e., calculating the accuracy of frames labeled as stable by our shakiness detection algorithm. This is the reason why the precision values are all 1 in Fig.5(a). Overall, it can be seen that setting these two values gains a commendable performance on the precision and recall rate for the casual videos, 86.6% and 90.9% on average respectively.

To understand the influence of the threshold values to the shakiness detection results, we further do experiments on the precision and recall by gradually changing the values of α and β in a certain range, e.g., $\alpha \in [0.5, 2]$ and $\beta \in [0.1, 5]$. Then, we make statistics on the precision and recall rate based on the corresponding threshold values. Figs.6(a) and 6(b) show the precision-recall curve of the experiment on the casual videos and the professional videos respectively, where the precisions are all 1 in Fig.6(b) by detecting the stable frames instead of shaky frames in the professional videos. Based on the above statistics, we find the optimal setting of the threshold values is $\alpha = 0.8$ pixels and $\beta = 2.2$ armin in the application, which gains the precision and recall rate about 91.7% and 87.1% respectively. Table 2 shows the average precision and the average recall rate for the videos in each category.

5.3 Comparisons

We also compare our algorithm with other shakiness detection methods. Though there is a large body of studies on video motion model, as far as we know, the systematic approach to shakiness detection is still the minority in the research, especially for casual videos. Here, we choose the method of [22] as the representative for comparison, which is also based on a global inter-frame motion model. Table 2 shows the quantitative comparison on the precision and recall rate of shakiness detection by different methods, where our algorithm gains better performance on the precision and recall rate of the shakiness detection. Actually, the method of [22] mainly deals with translational shakiness, thereby it generates bad detection for the complicated shaky motions like walking and running in the casual videos. Contrarily, our algorithm involves the shakiness in both translational and rotational components, which enables better shakiness detection.

5.4 Limitations and Discussions

Although our algorithm enables commendable performance on shakiness detection, it is not without limi-

Table 2. Comparison of Shakiness Detection Precision and Recall Rate (P/R) by Our Algorithm and the Method of Yan and Kankanhalli[22]

Method	Professional Video				Casual Video			
	Movie	Tripod	Osmo	Drone	Walking	Running	Riding	Driving
Ours	**1.00/0.978**	**1.00/1.00**	**1.00/0.985**	**1.00/1.00**	**0.905/0.860**	**0.908/0.936**	**0.942**/0.833	**0.871/0.830**
Yan and Kankanhalli[22]	0.904/0.904	1.00/1.00	1.00/0.961	0.926/0.926	0.563/0.842	0.691/0.719	0.692/**0.909**	0.734/0.809

Note: The bold numbers indicate the performance of the better method.

tations. Firstly, our algorithm uses the rigid transformations between adjacent frames as the camera motion model. Therefore it cannot well deal with the shakiness generated by quick zooming in/out and out-of-plane rotation. Secondly, the computation of inter-frame transformation relies on feature points detection and matching. Therefore it might generate erroneous estimation on the translation and rotation for the frames with large homogeneous regions, which influences the amplitude computation and shakiness detection. Thirdly, our algorithm takes the spatial content of the entire frame for shakiness amplitude computation, which ignores some other factors, like visual attention, spatial resolution and frequency that also influence the sense of motion stability.

Actually, the principle of our kinematics-based shakiness detection can be generalized by using a more complex shaky motion model, e.g., similarity, affine or homography transformations for modeling the camera motion. The key step is to compute the corresponding geodesics on the Lie group of the corresponding transformations, e.g., the similarity group Sim(2), general affine group GA(2) or projective group PG(2)[27]. Once we have the geodesics as the intended smooth motion path for amplitude computation, we can obtain the shakiness metric for detection. Besides, we can adopt content or perception analysis on the videos to improve the fidelity of shakiness detection. For example, we can use the distribution of region of interest (ROI), visual saliency or spatial frequency to weight the influence of shaky motions in different regions. Overall, our algorithm suggests a novel way to automatic video shakiness detection, which would facilitate the application of video processing like stabilization, in dealing with the casual videos.

6 Conclusions

We presented a perception-inspired and kinematics-based algorithm for automatically detecting video shakiness caused by camera vibration. The mechanism of motion perception is adopted in order to analyze and characterize the just-noticeable shakiness conditions (amplitude and frequency), and the corresponding shaky motion is modeled by the deviation from kinematic transformation motion path. Specially, we gave a concrete solution for shakiness detection on SE(2) with its explicit geodesic for amplitude computation. A benchmark dataset is also established to evaluate the shakiness detection accuracy by our algorithm.

As the future work, we plan to investigate more perceptual cues, like ROI and visual saliency in our shakiness detection algorithm. Besides, it is promising to combine motion blur detection in our framework. Because shaky videos often incur blurred appearance in the frames, the combination of the two detection schemes will improve the quality enhancement by video deblurring and stabilization.

References

[1] Abdollahian G, Taskiran C M, Pizlo Z, Delp E J. Camera motion-based analysis of user generated video. *IEEE Trans. Multimedia*, 2010, 12(1): 28-41.

[2] Hu S M, Chen T, Xu K, Cheng M M, Martin R R. Internet visual media processing: A survey with graphics and vision applications. *The Visual Computer*, 2013, 29(5): 393-405.

[3] Zhang L, Zhou L, Huang H. Bundled kernels for nonuniform blind video deblurring. *IEEE Trans. Circuits and Systems for Video Technology*, 2017, 27(9): 1882-1894.

[4] Yan F, Iliyasu A M, Yang H M, Hirota K. Strategy for quantum image stabilization. *Science China Information Sciences*, 2016, 59(5): 052102.

[5] Kakar P, Sudha N, Ser W. Exposing digital image forgeries by detecting discrepancies in motion blur. *IEEE Trans. Multimedia*, 2011, 13(3): 443-452.

[6] Su B L, Lu S J, Tan C L. Blurred image region detection and classification. In *Proc. the 19th ACM Int. Conf. Multimedia*, November 2011, pp.1397-1400.

[7] Yu X, Xu F, Zhang S L, Zhang L. Efficient patch-wise non-uniform deblurring for a single image. *IEEE Trans. Multimedia*, 2014, 16(6): 1510-1524.

[8] Visentini-Scarzanella M, Dragotti P L. Video jitter analysis for automatic bootleg detection. In *Proc. the 14th Int. Workshop on Multi-Media Signal Processing*, September 2012, pp.101-106.

[9] Sibiryakov A. Hand jitter descriptor for mobile video identification. In *Proc. Int. Conf. Consumer Electronics*, January 2011, pp.77-78.

[10] Chen H H, Liang C K, Peng Y C, Chang H A. Integration of digital stabilizer with video codec for digital video cameras. *IEEE Trans. Circuits and Systems for Video Technology*, 2007, 17(7): 801-813.

[11] Xue Y Y, Erkin B, Wang Y. A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing. *IEEE Trans. Multimedia*, 2015, 17(1): 134-139.

[12] Yan B, Yuan B H, Yang B. Effective video retargeting with jittery assessment. *IEEE Trans. Multimedia*, 2014, 16(1): 272-277.

[13] Zhang F L, Wang J, Zhao H, Martin R R, Hu S M. Simultaneous camera path optimization and distraction removal for improving amateur video. *IEEE Trans. Image Processing*, 2015, 24(12): 5982-5994.

[14] Zhang L, Xu Q K, Huang H. A global approach to fast video stabilization. *IEEE Trans. Circuits and Systems for Video Technology*, 2017, 27(2): 225-235.

[15] Huang H Z, Fang X N, Ye Y F, Zhang S H, Rosin P L. Practical automatic background substitution for live video. *Computational Visual Media*, 2017, 3(3): 273-284.

[16] Hasegawa K, Saito H. Synthesis of a stroboscopic image from a hand-held camera sequence for a sports analysis. *Computational Visual Media*, 2016, 2(3): 277-289.

[17] Joshi N, Kienzle W, Toelle M, Uyttendaele M, Cohen M F. Real-time hyperlapse creation via optimal frame selection. *ACM Trans. Graphics*, 2015, 34(4): Article No. 63.

[18] Wang M, Liang J B, Zhang S H, Lu S P, Shamir A, Hu S M. Hyper-lapse from multiple spatially-overlapping videos. *IEEE Trans. Image Processing*, 2018, 27(4): 1735-1747.

[19] Tong H H, Li M J, Zhang H J, Zhang C S. Blur detection for digital images using wavelet transform. In *Proc. Int. Conf. Multimedia and Expo.*, June 2004, pp.17-20.

[20] Tico M, Trimeche M, Vehvilainen M. Motion blur identification based on differently exposed images. In *Proc. Int. Conf. Image Processing*, October 2006, pp.2021-2024.

[21] Liu R T, Li Z R, Jia J Y. Image partial blur detection and classification. In *Proc. Conf. Computer Vision and Pattern Recognition*, June 2008.

[22] Yan W Q, Kankanhalli M S. Detection and removal of lighting & shaking artifacts in home videos. In *Proc. ACM Int. Conf. Multimedia*, December 2002, pp.107-116.

[23] Liu F, Gleicher M, Jin H L, Agarwala A. Content-preserving warps for 3D video stabilization. *ACM Trans. Graphics*, 2009, 28(3): Article No. 44.

[24] Zhang L, Chen X Q, Kong X Y, Huang H. Geodesic video stabilization in transformation space. *IEEE Trans. Image Processing*, 2017, 26(5): 2219-2229.

[25] Wolpert D M, Ghahramani Z. Computational principles of movement neuroscience. *Nature Neuroscience*, 2000, 3(Suppl): 1212-1217.

[26] Murray R M, Li Z X, Sastry S S. A Mathematical Introduction to Robotic Manipulation. CRC Press, 1994.

[27] Zacur E, Bossa M, Olmos S. Left-invariant Riemannian geodesics on spatial transformation groups. *SIAM Journal on Imaging Sciences*, 2014, 7(3): 1503-1557.

[28] Duan L Y, Jin J S, Tian Q, Xu C S. Nonparametric motion characterization for robust classification of camera motion patterns. *IEEE Trans. Multimedia*, 2006, 8(2): 323-340.

[29] Afonso M V, Nascimento J C, Marques J S. Automatic estimation of multiple motion fields from video sequences using a region matching based approach. *IEEE Trans. Multimedia*, 2013, 16(1): 1-14.

[30] Nishi K, Onda T. Evaluation system for camera shake and image stabilizers. In *Proc. Int. Conf. Multimedia and Expo.*, July 2010, pp.926-931.

[31] Albright T D, Stoner G R. Visual motion perception. *Proceedings the National Academy of Sciences of the United States of America*, 1995, 92(7): 2433-2440.

[32] Peli E, García-Pérez M A. Motion perception during involuntary eye vibration. *Experimental Brain Research*, 2003, 149(4): 431-438.

[33] Healey C G, Sawant A P. On the limits of resolution and visual angle in visualization. *ACM Trans. Applied Perception*, 2012, 9(4): Article No. 20.

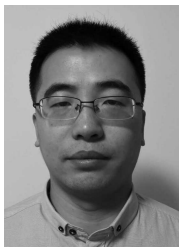
[34] Martins A J, Kowler E, Palmer C. Smooth pursuit of small-amplitude sinusoidal motion. *Journal of the Optical Society of America A*, 1985, 2(2): 234-242.

[35] He K M, Chang H W, Sun J. Content-aware rotation. In *Proc. Int. Conf. Computer Vision*, December 2013, pp.553-560.

[36] Shi J B, Tomasi C. Good features to track. In *Proc. Computer Society Conf. Computer Vision and Pattern Recognition*, June 1994, pp.593-600.

Xiao-Qun Wu is now a lecturer in the School of Computer and Information Engineering, Beijing Technology and Business University, Beijing. She received her B.S. and M.S. degrees in mathematics from Zhejiang University, Hangzhou, in 2007 and 2009, respectively, and her Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. Her research focuses on computer graphics, multimedia processing and applications.

Hai-Sheng Li is a professor in the School of Computer and Information Engineering, Beijing Technology and Business University, Beijing. He received his Ph.D. degree from Beihang University, Beijing, in 2002. His current research interests include computer graphics, scientific visualization, etc.



Jian Cao is now an associate professor in the School of Computer and Information Engineering, Beijing Technology and Business University, Beijing. He received his B.S. and Ph.D. degrees from Beijing Institute of Technology, Beijing, in 2004 and 2010, respectively. His research interests include image processing, pattern recognition, and machine learning.

Qiang Cai is a professor in the School of Computer and Information Engineering, Beijing Technology and Business University, Beijing. He is the dean of School of Computer and Information Engineering, Beijing Technology and Business University, Beijing. His research interests are in the areas of computer aided design, computational geometry, scientific visualization and knowledge map for food safety.