

On Human Perception-Guided Video Stability Assessment

Marcos Roberto Souza  [Institute of Computing, University of Campinas | marcoosrs@gmail.com]

Luís Gustavo Lorgus Decker  [Institute of Computing, University of Campinas | luisgustavo.decker@gmail.com]

Jose Luis Flores Campana  [Institute of Computing, University of Campinas | jluisflores357@gmail.com]

Helena de Almeida Maia  [Institute of Computing, University of Campinas | helena.maia@ic.unicamp.br]

Helio Pedrini   [Institute of Computing, University of Campinas | helio@ic.unicamp.br]

 Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, Campinas, SP, 13083-852, Brazil

Received: 22 March 2026 • **Accepted:** 21 August 2026 • **Published:** 28 August 2026

Abstract User experience can be compromised if a video contains an in-capture distortion called camera instability, characterized by a typical visual effect caused by uncontrolled camera motion during recording. The process of removing this distortion is known as video stabilization. Despite recent advances in this field, few studies aim to define how video stability quality should be evaluated. In this work, we propose assessment metrics based on pixel profile derivatives and machine learning regressors. We compare them with existing measures on public datasets, correlating the results with human perception scores. Our findings indicate that a widely used measure, called Low-High Frequency Ratio (LHR), correlates poorly with human perception, achieving PLCC=0.388 on LIVE-Qualcomm and PLCC=0.538 on MIND-VQ. In contrast, the best result obtained in this work achieves PLCC=0.884 on MIND-VQ, improving LHR by 34.6 percentage points on that dataset.

Keywords: Video Stability, In-Capture Video Distortion, Assessment, Human Perception

1 Introduction

Amateur users have easy access to mobile devices with a camera. However, they often capture handheld videos without adhering to cinematographic principles, leading to camera motion instabilities.

Subjective studies indicate that video instability can be a highly annoying distortion in captured footage [Ghadiyaram *et al.*, 2017], negatively impacting the user experience. Video stabilization seeks to eliminate camera instability [Souza *et al.*, 2024b; Wang *et al.*, 2024a; Peng *et al.*, 2024; Wang *et al.*, 2024b; Souza *et al.*, 2022, 2023; Souza and Pedrini, 2018b,a], producing a stabilized video with smooth, non-oscillatory camera motion.

Evaluation metrics are crucial for comparing approaches and significantly influence the direction of advancements in the field [Souza and Pedrini, 2019; Souza *et al.*, 2018]. Despite the rapid progress in digital video stabilization methods, fewer works focus on stability assessment itself – particularly on how objective measures relate to subjective stability judgments under controlled user studies. As a result, it remains challenging to answer the following question based on current knowledge in the literature: “How closely are the available measures related to human perception of stability?”

Motivated by this question, we analyze the relationship between existing stability assessment measures and human perception stability scores from third-party user studies. We also propose and evaluate novel approaches based on pixel profiles, including experiments with traditional machine learning regressors.

In this work, we (i) derive stability measures from first-, second-, and third-order derivatives of optical-flow-based pixel profiles and (ii) study the prediction of subjective stability scores using conventional regressors trained on a multi-aspect feature set computed from pixel profiles and image similarity measures.

Compared to the conference version [Souza *et al.*, 2024a] – which was a condensed summary of the doctoral thesis and, therefore, included only a brief overview of the proposed metrics among many other contributions – this journal manuscript provides a deep dive into these metrics, expanding the methodological description and substantially extending the experimental analysis and discussion. In particular, this manuscript presents the detailed formulation and analysis of the pixel-profile-derivative measures ($VCP^2/ACP^2/JCP^2$) and the regression-based stability assessment, which were only briefly summarized in the conference version. The main contributions of this paper, compared to the conference paper and the existing literature, are:

1. **Analysis of existing evaluation metrics for video stability based on human perception:** We correlated the results of the most commonly used assessment measures in the literature with subject scores obtained from third-party user studies. This step is necessary for improving our understanding of how stability can be assessed, and developing metrics that better align with human visual perception.
2. **Proposal and analysis of pixel-profile-based stability measures and a regressor-based assessment approach:** We propose three stability measures based on

first-, second-, and third-order derivatives of optical-flow-based pixel profiles (VCP², ACP², and JCP²), and we study a learning-based approach that predicts stability mean opinion score (MOS) from a multi-aspect feature vector computed from pixel profiles. Our kinematic measures achieve stronger correlation with subjective stability than the commonly used measures evaluated in this work. In addition, in the intra-dataset setting, non-linear regressors trained on the proposed multi-aspect features achieve the highest correlations among the evaluated regressors.

This paper is organized as follows. Section 2 reviews the literature on topics relevant to this work. Section 3 describes the proposed measures and features. The results are presented and analyzed in Section 4. Finally, Section 5 includes some concluding remarks and directions for future work.

2 Literature Review

This section reviews three topics in the context of video stabilization and stability assessment: (i) non-referenced video stability assessment (Subsection 2.1), where we present competing measures in the literature and discuss learning-based strategies to predict subjective scores; (ii) feature trajectories and pixel profiles (Subsection 2.2), where the latter is used to compute the measures studied in this paper; and (iii) video quality assessment (Subsection 2.3). Some final remarks are presented in Subsection 2.4.

2.1 Non-Referenced Video Stability Assessment

Stability is a subjective concept related to human perception, and multiple approaches have been used to assess it. One of the first works to propose an objective metric for stability assessment was conducted by Morimoto and Chellappa [1997]. They introduced the Interframe Transformation Fidelity (ITF), which computes the average similarity between consecutive stabilized frames using the Peak Signal-to-Noise Ratio (PSNR). Some authors have used the Structural Similarity Index Measure (SSIM) [Wang et al., 2004] instead of PSNR, because the first considers elements such as luminance, contrast, and structure that are more relevant to visual perception when comparing images.

Liu et al. [2013] proposed a widely used stability measure, called Low-High Frequency Ratio (LHR). This measure assumes that the proportion of low-frequency energy increases with video stability. LHR is computed through frequency analysis by calculating the ratio between the low and high frequencies of the estimated two-dimensional camera motion.

Zhang et al. [2018a] introduced a measure that characterizes stability based on Intrinsic Geodesic Curvature (IGC). The final video smoothness is defined as the average of the weighted geodesic curvatures of each path on a 2D mesh.

Beyond handcrafted measures, a common strategy in the broader video quality assessment literature is to cast quality prediction as a supervised regression problem, where a model learns a mapping from objective features to subjective

scores (e.g., MOS) [Min et al., 2024; Streijl et al., 2016]. This paradigm is particularly appealing when multiple complementary cues are available, as regressors can combine them to better approximate human perception.

Despite the advances made by the literature studies [Kou et al., 2023; Rao et al., 2023; Liu et al., 2023; Ito and Izquierdo, 2019; Lee et al., 2021; Nie et al., 2017; James et al., 2023; Shi et al., 2022], a deeper understanding of the relationship between objective stability measures and human perception is still needed. Moreover, while supervised regression is a standard paradigm in no-reference video quality assessment for predicting MOS from objective features, several stability papers report objective stability measures without explicitly learning (and validating) a mapping to subjective stability MOS under controlled protocols. In contrast to methods found in the literature, our measures are based on derivatives of pixel profiles, and we correlate their outputs with third-party user studies on perceived stability. In addition, we investigate conventional regressors trained on multi-aspect features computed from pixel profiles to predict stability MOS values.

2.1.1 Stability Measures in Video Acceleration and First-Person Hyperlapse

Although our review so far focuses on works whose primary goal is to define stability assessment metrics, stability (or smoothness) criteria are also central in video acceleration and hyperlapse, particularly in first-person settings. In these problems, naive frame sub-sampling tends to amplify head-motion shake; thus, many approaches explicitly analyze camera motion and incorporate smoothness objectives (or stability/smoothness terms) when selecting frames or planning the output camera path.

For example, EgoSampling [Poleg et al., 2015] formulates adaptive frame sampling as an energy minimization problem to generate more stable fast-forward egocentric videos, relying on simple motion cues (e.g., epipole/FOE estimation) derived from optical flow rather than computationally demanding 3D reconstruction. An extended version [Halperin et al., 2017] further supports wide-view hyperlapse by mosaicing each output frame from multiple inputs; this repurposes shake as a cue to enlarge the field of view and enables joint processing of multiple egocentric videos.

In contrast, Kopf et al. [2014] reconstruct a 3D input camera path and proxy geometries, and then optimize an output path that is both smooth and renderable, incorporating a view-independent quality term to account for rendering limitations. Joshi et al. [2015] propose a real-time hyperlapse method based on optimal frame selection that targets a desired speed-up while explicitly optimizing for smooth camera motion, without relying on special sensor data. Semantic pipelines also incorporate smoothness: Lai et al. [2017] stabilize 360° video by smoothing inter-frame rotation and combine saliency/semantic cues with motion smoothness to produce NFOV hyperlapses (validated through a large-scale user study), while Silva et al. [2016] introduce a smoothness evaluation metric motivated by the observation that commonly used measures do not necessarily reflect viewers' preferences in this context.

These lines of work highlight that stability criteria are widely used beyond the stabilization-metric literature, often as objectives or constraints for producing watchable accelerated videos. By comparison, our focus is on evaluating stability assessment measures themselves by correlating objective outputs with per-video subjective stability scores (MOS) on public datasets, and on analyzing how different measurement principles relate to perceived instability.

2.2 Feature Trajectories and Pixel Profiles

Figure 1 presents representations of feature trajectories and pixel profiles. Trajectories are sparse in both space and time and change their position based on two-dimensional motion. On the other hand, pixel profiles are dense in both space and time, where each one has a fixed spatial position. In this way, the position of each feature in a trajectory is defined by real numbers, whereas the position indices of a pixel profile are integers.

For didactical purposes, we can refer to both features and pixels as points. Let us also consider that points have another attribute other than position: the accumulated motion vector. In both cases, this vector is computed based on the relative displacements between adjacent frames. However, as the position of the trajectory points changes due to these displacements, its accumulated vector and position are equal. Conversely, the accumulated motion vector for each point in the pixel profile is derived from displacements related to the same pixel, consequently differing from its fixed position.

We usually construct feature trajectories and pixel profiles using matched local features and optical flows, respectively. This is why the displacement vectors of pixel profiles are represented inversely.

Feature trajectories of a video with t frames at a resolution of $H \times W$ can be formally defined as a collection of n trajectories denoted by $\mathbf{FT} = \{\mathbf{ft}_1, \dots, \mathbf{ft}_n\}$. Each trajectory, indexed by i , can be described as $\mathbf{ft}_i = \{\mathbf{x}_s, \mathbf{x}_{s+1}, \dots, \mathbf{x}_e \mid 0 \leq s < e \leq t-1\}$, with a size of $e - s + 1$, where $\mathbf{x}_j$ is the local feature of the j -th frame in the video, corresponding to other local features within the same trajectory. The positional (and accumulated motion) vector of each feature can be defined as

$$\mathbf{x}_j = \begin{cases} (x_0, y_0), & \text{if } j = s \\ \mathbf{x}_{j-1} + (\Delta_j x, \Delta_j y), & \text{otherwise} \end{cases} \quad (1)$$

where (x_0, y_0) is the initial two-dimensional position of the path (first feature) and $(\Delta_j x, \Delta_j y)$ is the two-dimensional displacement between the features $\mathbf{x}_{j-1}$ and $\mathbf{x}_j$.

Similarly, pixel profiles [Liu et al., 2014] is the set $\mathbf{P} = \{\mathbf{pp}_0, \mathbf{pp}_1, \dots, \mathbf{pp}_{HW}\}$ of HW profiles of size t , where the j -th pixel $\mathbf{p}_j$ of $\mathbf{pp}_i$ refers to the i -th pixel of the j -th frame. Despite the fixed position, the accumulated motion vector of each pixel can be defined as

$$\mathbf{p}_j = \begin{cases} (m, n), & \text{if } j = 0 \\ \mathbf{p}_{j-1} + \mathbf{O}_j(m, n), & \text{otherwise} \end{cases} \quad (2)$$

where $0 \leq m < H$ and $0 \leq n < W$ are the pixel coordinates $\mathbf{p}_j$ and $\mathbf{O}_j(m, n)$ the optical flow vector in coordinates (m, n) between the frames $\mathbf{F}_{j-1}$ and $\mathbf{F}_j$.

The concept of a pixel profile is related to visual rhythms since both capture pixel variations throughout the video. Visual rhythms consists of a representation, proposed by Adelson and Bergen [1985] and Bolles and Baker [1985], which summarizes spatio-temporal information of a video into a single image or in a small set of images. This representation is built by the concatenation (typically along the x -axis of the rhythm image) of a predefined feature extracted from the video frames, so that it is possible to verify the behavior of that information over time. This representation can be seen as a slice in the video volume, where the third dimension corresponds to time.

Figure 2 illustrates examples of visual rhythms in different directions. In this work, we utilized the Horizontal Average and Vertical Average strategies for video visualization [Souza and Pedrini, 2020] due to their ability to provide information about the entire frame without sampling as severely as the simple Horizontal and Vertical methods [Souza et al., 2020; Souza and Pedrini, 2020].

The main difference between visual rhythms and pixel profiles is that, in the latter, we necessarily work with pre-computed motion information (optical flow). Besides, there is no reduction of the entire video information to a single image for pixel profiles. However, we relate the information from the same pixel over time in both cases. In this work, visual rhythms were used to represent video motion visually.

2.3 Video Quality Assessment

Video Quality Assessment [Min et al., 2024] can be classified into subjective and objective assessment categories. Subjective assessment involves human observers watching videos and providing their opinions for each video. This evaluation method is time-consuming and requires substantial human effort. On the other hand, objective assessment calculates a video's quality numerically using mathematical functions, which is advantageous for evaluating large datasets. However, to establish a connection between objective measures and human perception, mature subjective studies are necessary.

Mean Opinion Score (MOS) [Streijl et al., 2016] is the most commonly used measure to aggregate the overall quality of a stimulus. It is calculated as the arithmetic average of all valid user ratings. Most of the video quality assessment literature currently available is concerned with distortions obtained post-capture [Ghadiyaram et al., 2017], such as during transmission or compression of videos [Wu et al., 2022; Nuutinen et al., 2016; Kossi et al., 2022]. Regardless, some authors have studied in-capture distortions, such as stability.

Ghadiyaram et al. [2017] performed an unbiased and distortion-guided user-study with their own dataset, called LIVE-Qualcomm. It comprises 208 videos with six distinct in-capture distortions, all recorded at a resolution of 1920×1080 pixels and 30 FPS. Among these videos, 35 have a shaky camera and they computed a MOS for each of these videos according to the subjective assessment of 39 distinct subjects. They were captured using 8 mobile devices from different brands and models. Figure 3 shows examples of videos from this dataset. LIVE-Qualcomm videos exhibit multiple levels of stability in high-resolution frames, which align consistently

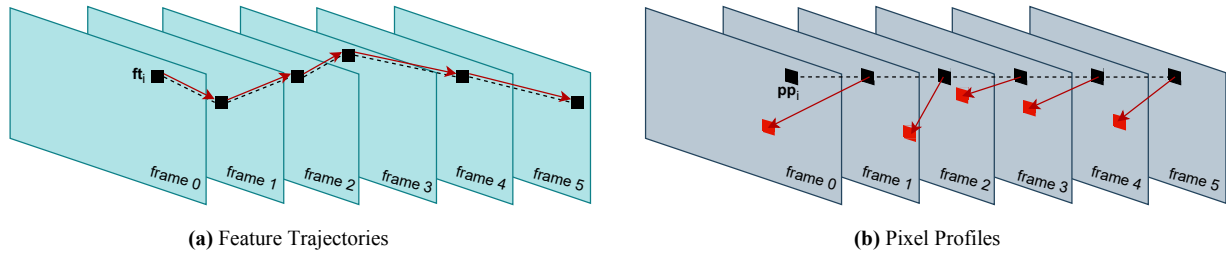


Figure 1. Representations for Feature Trajectories and Pixel Profiles. Dotted lines represent the spatial displacement of correspondent points. Red arrows represent the vectors used in the accumulated motion vector at each point. Adapted from Liu *et al.* [2014].

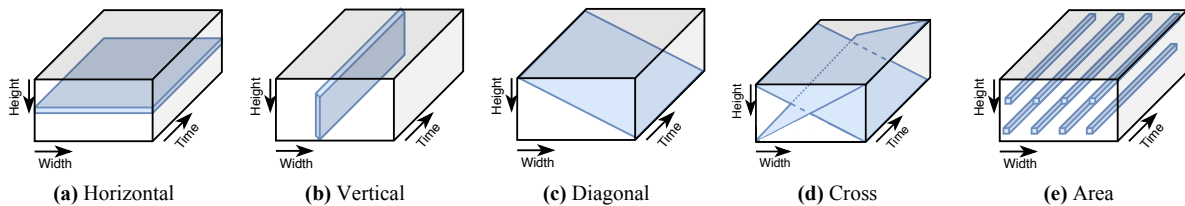


Figure 2. Different transverse directions in the construction of a visual rhythm [Souza *et al.*, 2020].

with MOS values.

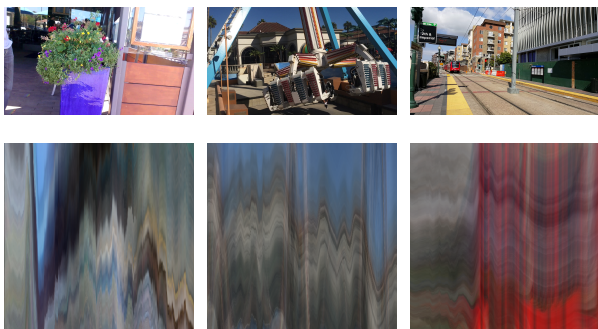


Figure 3. Videos from LIVE-Qualcomm [Ghadiyaram *et al.*, 2017]. We show a frame and the visual rhythm [Souza *et al.*, 2020] for each sample. From left to right, the MOS of these videos are 13.36, 46.46, and 64.91.

Guan *et al.* [2022] conducted a similar study on MIND-VQ dataset. The used dataset comprises 1181 videos featuring night scenes captured using various mobile devices by 21 users and evaluated by 435 participants. These users were encouraged to record videos with a wide range of diversity, both spatially and temporally. Each video was annotated in user studies involving a single stimulus, which considered five perceptual attributes, including stability. These subjective experiments were conducted in a controlled laboratory environment. Examples of videos from this dataset are presented in Figure 4.

MIND-VQ videos were recorded only at night but have a great variety in recording style and quality, which may make this dataset more realistic. The visual rhythms of these videos show differences in camera stability, which are related to three different MOS values. The reader may consult the following references on visual rhythms to better understand this visualization [Souza *et al.*, 2020; Souza and Pedrini, 2020].

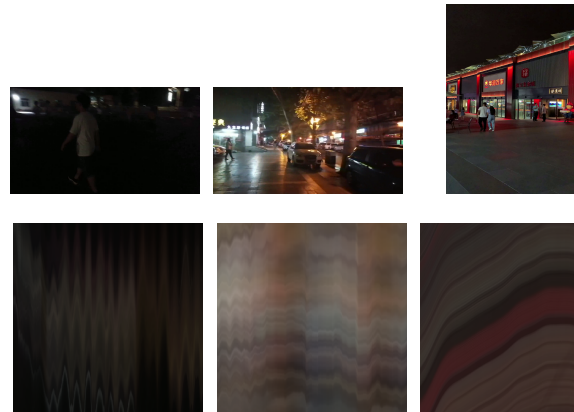


Figure 4. Visual frames and rhythms from MIND-VQ [Guan *et al.*, 2022] videos. The MOS, from left to right, are 54.42, 66.24, and 83.93.

2.4 Remarks

In this work, we propose an evaluation based on kinematic statistics extracted from the pixel profile. This approach was motivated by the superior performance of kinematics-based methods, as observed in an evaluation based on our own visual perception.

The utilization of pixel profiles instead of feature trajectories was driven by two primary reasons: (i) pixel profiles offer a dense representation that encompasses all pixels within each frame, and (ii) their implementation is more straightforward, as we do not need to track features across the entire video, avoiding complications such as temporal discontinuity and features leaving the image domain.

This strategy enabled us to achieve better results than the metrics available in the literature regarding correlation with the MOS of subjective studies conducted on the LIVE-Qualcomm and MIND-VQ datasets. Based on the premise that different approaches to assessing stability in the literature – image similarity, frequency analysis, geometric analysis, and kinematic metrics – can be complementary, we constructed features that encompass all of them and trained

regressors to predict MOS values. This further improved our results.

In the next section, we present the kinematic measurements extracted from the pixel profiles, along with our features based on multi-aspect. Subsequently, we report the experiments conducted on the LIVE-Qualcomm and MIND-VQ datasets.

3 Kinematic-based Measures and Features from Multiple Aspects

Figure 5 presents an overview of our assessment measures. The input for all the proposed metrics is a video frame sequence. In the preprocessing step, we compute the optical flow for each consecutive frame pair and the background/foreground segmentation mask for each frame. The optical flow is predicted using RAFT [Teed and Deng, 2020], whereas the segmentation masks are computed using BSUV-Net [Tezcan et al., 2020, 2021]. These masks are used to (i) restrict derivative-based kinematic measures to background regions (Alternative A) and (ii) separate static (background) and moving-object (foreground) regions when computing region-wise descriptors in Alternative B.

Throughout this section, let $\{\mathbf{I}_i\}_{i=0}^{T-1}$ denote a video with T frames defined over a pixel domain Ω , and let $\mathbf{p} \in \Omega$ be a pixel location. From the optical flow, we compute an optical-flow-based pixel profile $\mathbf{pp}_i(\mathbf{p})$ for each frame (Equation 2). We also compute a binary background mask $\tilde{\mathcal{M}}_i(\mathbf{p})$ for each frame. For Alternative A (derivative-based kinematic measures), we combine masks over the temporal neighborhood of each derivative so that pixels affected by foreground motion in any frame of the neighborhood are excluded from the kinematic pooling. For Alternative B (multi-aspect features), we use the same masks to define static (background) and moving-object (foreground) regions when computing region-wise descriptors (and full-frame summaries when applicable). Frame-level quantities are then aggregated over time to obtain a single score per video.

3.1 Pipeline and Aggregation Details

Figure 5 summarizes the overall pipeline. For completeness, we detail here how frame-level quantities are produced and aggregated into a single video score, as this aggregation is common to both alternatives.

Given consecutive frames $(\mathbf{I}_i, \mathbf{I}_{i+1})$, we estimate an optical flow field $\mathbf{O}_{i+1}(\mathbf{p})$ that maps each pixel location $\mathbf{p} \in \Omega$ from frame i to frame $i + 1$. From these fields, we construct an optical-flow-based pixel profile $\mathbf{pp}_i(\mathbf{p})$ recursively as in Equation 2, i.e., $\mathbf{pp}_0(\mathbf{p}) = \mathbf{p}$ and $\mathbf{pp}_i(\mathbf{p}) = \mathbf{pp}_{i-1}(\mathbf{p}) + \mathbf{O}_i(\mathbf{p})$ for $i \geq 1$.

For each frame i , we compute a binary mask $\tilde{\mathcal{M}}_i(\mathbf{p}) \in \{0, 1\}$ where $\tilde{\mathcal{M}}_i(\mathbf{p}) = 1$ denotes background and $\tilde{\mathcal{M}}_i(\mathbf{p}) = 0$ denotes foreground. For any temporal neighborhood $[i, j]$, we combine these masks as $\mathcal{M}_i^j = \tilde{\mathcal{M}}_i \wedge \tilde{\mathcal{M}}_{i+1} \wedge \dots \wedge \tilde{\mathcal{M}}_j$, so that a pixel is retained only if it is classified as background in all frames of the neighborhood.

Many descriptors considered in this work produce either (i) per-pixel values (indexed by $\mathbf{p}$, and possibly by time i), or (ii) per-frame scalars (indexed by i). For Alternative A,

the three kinematic descriptors produce per-pixel derivative vectors at each frame index, which are summarized into a frame-level scalar by averaging the squared magnitude over the valid background pixels given by $\mathcal{M}_i^j$. For Alternative B, several descriptors are summarized either region-wise (background/moving-object) or over the full frame, yielding one or more per-frame scalars.

All frame-level scalars are aggregated over time into a single video score. For any frame-level sequence $\{s_i\}$ defined for $i \in \mathcal{I}$ (e.g., $\mathcal{I} = \{0, \dots, T-2\}$ for pairwise descriptors, or $\mathcal{I} = \{0, \dots, T-4\}$ for third-order derivatives), we obtain a video-level quantity by temporal averaging:

$$\bar{s} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_i. \quad (3)$$

This convention is used for the kinematic measures (Alternative A) and also for intermediate per-frame quantities in the multi-aspect feature extraction (Alternative B).

3.2 Computational Considerations

Our assessment pipeline has two deep preprocessing stages and several lightweight post-processing stages. Given a video with T frames of resolution $H \times W$, optical flow estimation is performed for each consecutive frame pair, resulting in $(T-1)$ forward passes of RAFT. Background/foreground segmentation is computed per frame, resulting in T forward passes of BSUV-Net. These two steps dominate the computational footprint of the pipeline because they require running deep networks over full-resolution frames, whereas the subsequent computations (masking, finite differences on pixel profiles, and statistic aggregation) consist mainly of per-pixel arithmetic operations.

After preprocessing, Alternative A computes first-, second-, and third-order finite differences of the optical-flow-based pixel profiles and averages them over background pixels. This stage scales linearly with the number of pixels and temporal positions, i.e., $\mathcal{O}(HWT)$. Alternative B adds feature extraction across four aspects. In addition to the $\mathcal{O}(HWT)$ kinematic and geometric pooling, the frequency features require computing temporal spectra of pixel-profile signals; when implemented with FFT along the temporal axis, this scales as $\mathcal{O}(HW T \log T)$ for the horizontal and vertical components.

To make the computational footprint of each stage explicit, Table 1 summarizes an analytical breakdown of the main pipeline components, including the number of deep-model forward passes per video and the approximate scaling of the post-processing steps with video length and resolution.

Using the optical flow and segmentation masks, we explore two possible alternatives for obtaining the evaluation measure. Both alternatives rely on pixel profiles derived from optical flow and on background/foreground masks, but they use these masks differently: Alternative A restricts kinematic pooling to background pixels, whereas Alternative B uses the masks to compute region-wise descriptors (background vs. moving-object) and full-frame summaries, depending on the feature group. In Alternative A, presented in Subsection 3.3, we introduce novel kinematic measures using the first, second, and third derivatives of pixel profiles. To do this, we

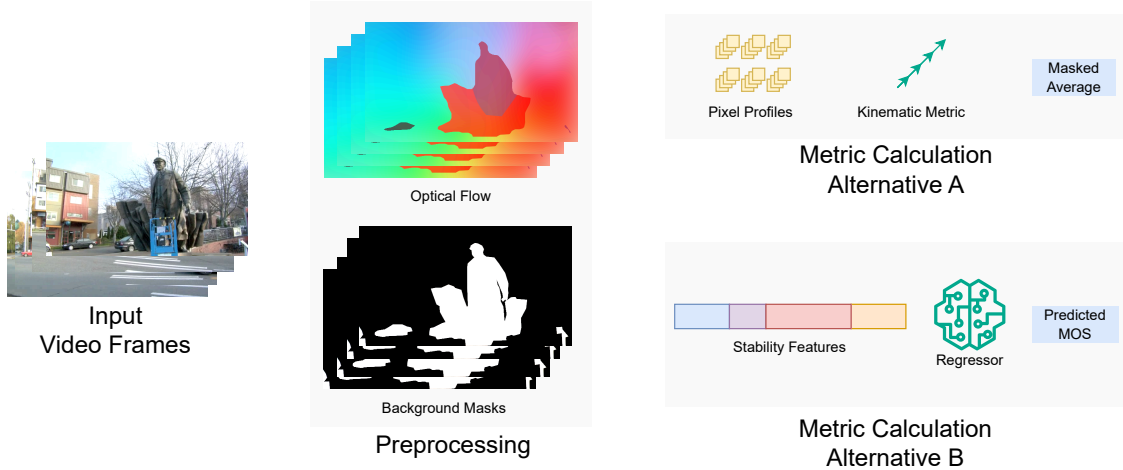


Figure 5. Overview of the computation of the two proposed alternatives for evaluating stability from a sequence of frames.

Table 1. Analytical breakdown of the main computational stages per video.

Stage	Calls per video	Scaling (approx.)
Optical flow (RAFT)	$T - 1$	Deep model over $H \times W$
Segmentation (BSUV-Net)	T	Deep model over $H \times W$
Kinematic derivatives + pooling	-	$\mathcal{O}(HWT)$
Frequency features (FFT over time)	-	$\mathcal{O}(HWT \log T)$

calculate the derivative of the pixel profile for each pixel. The value for each pair of frames is computed as the average of the pixel values in the background. The final video value is obtained by averaging the values of all frame pairs. The inspiration for investigating these measures originated from visual observations: we found that, in simple videos, techniques based on kinematic principles [Grundmann *et al.*, 2011] produce videos with better stability than those relying on high-frequency attenuation [Liu *et al.*, 2013].

Moreover, in Alternative B, we used these measures and other statistics to compose a feature vector and train a regressor to predict stability scores based on human perception, described in Subsection 3.4. The idea is to use different approaches to assess stability in the literature in a complementary way.

Comparing the two alternatives, the first is simpler to implement and does not depend on a training set. In contrast, the second alternative can achieve better results, showing a higher correlation with the MOS and greater similarity to it in terms of scale and numerical range. However, its success depends more on the data used and may be less robust in unseen scenarios.

3.3 Alternative A – Kinematic Measures

The diagram in Figure 6 illustrates the calculation of kinematic measurements. First, we compute pixel profiles from optical flows, as defined in Equation 2. Next, we generate background segmentation masks. Finally, we compute three

kinematic measures (first, second, and third derivatives) from the pixel profiles, which we respectively named Velocity of Camera Pixel Profiles (VCP²), Acceleration of Camera Pixel Profiles (ACP²) and Jerk of Camera Pixel Profiles (JCP²). To compute the average for each video, we consider only the derivatives of regions without foreground. This prevents the inclusion of vectors from moving objects, which do not represent camera motion.

These metrics were directly inspired by the stabilization approach developed by Grundmann *et al.* [2011], which minimizes the first, second, and third derivatives of unstable camera trajectories. A similar formulation has been used recently to generate stable videos on synthetic datasets [Souza *et al.*, 2024b]. Mathematically, the three frame-level measures can be defined as

$$\begin{aligned}
 \text{VCP}_i^2 &= \frac{\sum_{\mathbf{p}} \mathcal{M}_i^{i+1}(\mathbf{p}) \|\mathbf{pp}_{i+1}(\mathbf{p}) - \mathbf{pp}_i(\mathbf{p})\|^2}{\sum_{\mathbf{p}} \mathcal{M}_i^{i+1}(\mathbf{p}) + c}, \\
 \text{ACP}_i^2 &= \frac{\sum_{\mathbf{p}} \mathcal{M}_i^{i+2}(\mathbf{p}) \|\mathbf{pp}_{i+2}(\mathbf{p}) - 2\mathbf{pp}_{i+1}(\mathbf{p}) + \mathbf{pp}_i(\mathbf{p})\|^2}{\sum_{\mathbf{p}} \mathcal{M}_i^{i+2}(\mathbf{p}) + c}, \\
 \text{JCP}_i^2 &= \frac{\sum_{\mathbf{p}} \mathcal{M}_i^{i+3}(\mathbf{p}) \|\mathbf{pp}_{i+3}(\mathbf{p}) - 3\mathbf{pp}_{i+2}(\mathbf{p}) + 3\mathbf{pp}_{i+1}(\mathbf{p}) - \mathbf{pp}_i(\mathbf{p})\|^2}{\sum_{\mathbf{p}} \mathcal{M}_i^{i+3}(\mathbf{p}) + c},
 \end{aligned} \tag{4}$$

where VCP_i^2 is the average velocity of the camera pixel profiles for the i -th frame. The term $\mathbf{pp}_i(\mathbf{p})$ denotes the pixel profile at pixel $\mathbf{p}$ in the i -th frame. To compose the final value of VCP^2 , we compute the average of these values in each video frame. ACP^2 and JCP^2 were computed analogously. The constant $c = 10^{-6}$ was used for division stability.

We used a segmentation mask between background and foreground to ignore the latter, as shown in the representation in Figure 7. Mathematically, we define the mask to be used as

$$\mathcal{M}_i^j = \tilde{\mathcal{M}}_i \wedge \tilde{\mathcal{M}}_{i+1} \wedge \dots \wedge \tilde{\mathcal{M}}_j, \tag{5}$$

where $\tilde{\mathcal{M}}_k$ is the binary mask at the k -th frame where 1 denotes background and 0 denotes foreground. Consequently, the logical AND retains a pixel only if it is classified as background in all frames from i to j , i.e., if a pixel is labeled as foreground in at least one frame of the temporal neighborhood, then $\mathcal{M}_i^j$ becomes 0 for that pixel and it is ignored in the computation. For visualization purposes, we render

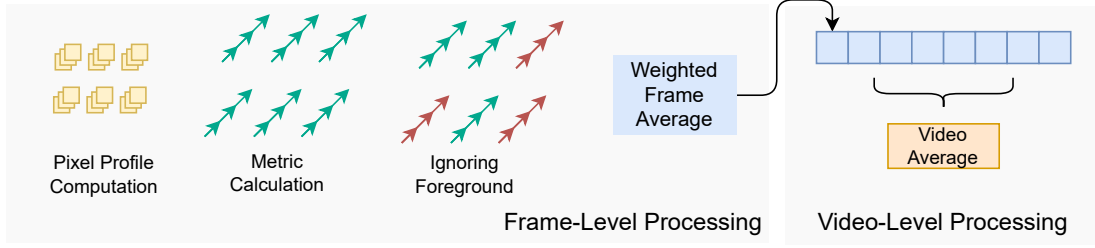


Figure 6. Calculation of kinematic metrics from pixel profiles at frame and video levels.

background pixels (value 1) in black and foreground pixels (value 0) in white in our figures.

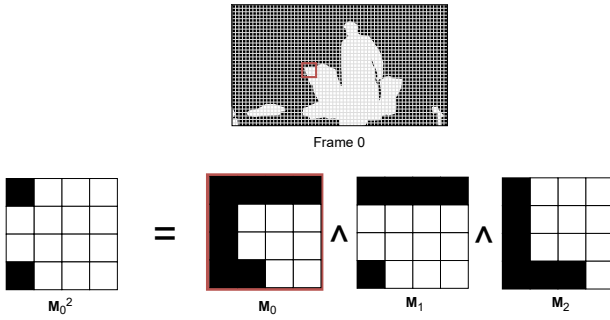


Figure 7. Representation of the background mask calculation for ACP2. The top image shows the background mask for Frame 0, where black pixels indicate background (i.e., $\mathcal{M}_0(\mathbf{p}) = 1$) and white pixels indicate foreground (i.e., $\mathcal{M}_0(\mathbf{p}) = 0$). In the bottom, we see a cropped region highlighted in red for Frame 0 and the following two, as well as the resulting mask (left).

From the representation, we note that a pixel is considered background only if all equivalent pixels within the temporal neighborhood are also background.

3.4 Alternative B – Stability Features from Multiple Aspects

As an alternative to the measure directly calculated from the pixel profile derivative, we also propose using a regressor to predict the MOS of a video for stability. We hypothesized that stability assessment should encompass multiple aspects. We defined the aspects based on existing methodologies in the literature: (i) image similarity, which measures the degree of similarity between two images and may be based on statistical, structural, or perceptual characteristics [Morimoto and Chellappa, 1997; Wang et al., 2004; Zhang et al., 2018b]; (ii) frequency analysis, which studies the components of a signal in different frequency bands [Liu et al., 2013]; (iii) geometry, which refers to spatial properties, including concepts such as distance and angles [Zhang et al., 2018a]; and (iv) the kinematic measures proposed in this work [Grundmann et al., 2011; Souza et al., 2024b]. We defined distinct measures for each aspect. Typically, these measures output a value per pixel or frame.

In this way, we used six statistical metrics to synthesize these values: average, standard deviation, median, interquartile range, kurtosis, and skewness. The following paragraphs show how we use these measures and statistical metrics to compose a vector with 204 features. Figure 8 shows a visual representation of these features.

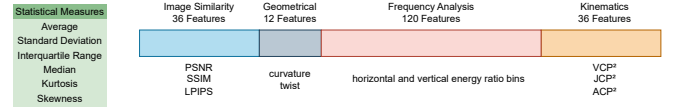


Figure 8. Representation of our feature vector.

Several descriptors in Alternative B yield either (i) a temporal sequence of scalars (one scalar per frame or per frame pair), or (ii) a pooled set of values defined over pixel locations and (optionally) time. To obtain fixed-length features, we summarize these outputs using the same set of six statistics: mean, standard deviation, median, interquartile range, kurtosis, and skewness.

For a temporal sequence $\{u_i\}_{i \in \mathcal{I}}$, we compute $\phi(\{u_i\}) = [\mu, \sigma, \text{median}, \text{IQR}, \text{kurtosis}, \text{skewness}]$, yielding six features. For a pooled set $\mathcal{U} = \{u(\mathbf{p}, i)\}$ collected over valid pixels (and time indices, when applicable), we compute the same six statistics over $\mathcal{U}$. This defines how the descriptors from different aspects are mapped into a fixed-length vector.

For image similarity, we consider ITF using PSNR, SSIM, and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018b] as similarity measures. Static and moving regions are defined using the background/foreground masks predicted by BSUV-Net. For a consecutive frame pair $(\mathbf{I}_i, \mathbf{I}_{i+1})$, we define a pairwise background mask as $\mathcal{M}_i^{i+1} = \mathcal{M}_i \wedge \mathcal{M}_{i+1}$, where $\mathcal{M}_k(\mathbf{p}) = 1$ denotes background and 0 denotes foreground. Thus, $\mathcal{M}_i^{i+1}(\mathbf{p}) = 1$ indicates static (background) pixels for that pair, whereas $1 - \mathcal{M}_i^{i+1}(\mathbf{p})$ indicates pixels affected by moving objects in at least one of the two frames. For PSNR, we compute two region-wise scalars per frame pair (static and moving-object). For SSIM, we compute three scalars per frame pair (static, moving-object, and full-frame). For LPIPS, we compute a full-frame scalar per frame pair. For region-wise PSNR, we compute the MSE restricted to the region and convert it to PSNR in dB. For SSIM, we first compute the per-pixel similarity map $\mathbf{S}_i(\mathbf{p}) = \text{SSIM}(\mathbf{I}_i, \mathbf{I}_{i+1})(\mathbf{p})$, and then summarize it into scalar values by averaging $\mathbf{S}_i$ over (i) static pixels ($\mathcal{M}_i^{i+1} = 1$), (ii) moving-object pixels ($1 - \mathcal{M}_i^{i+1} = 1$), and (iii) all pixels (full frame). Over the whole video, each scalar forms a temporal sequence, from which we extract the six statistics (mean, standard deviation, median, interquartile range, kurtosis, and skewness). Thus, PSNR, SSIM, and LPIPS have respective 12, 18, and 6 values, summing 36 features.

For a toy example with three frame pairs, suppose the mean SSIM over static pixels is $\{0.90, 0.86, 0.88\}$ and over moving-object pixels is $\{0.72, 0.70, 0.75\}$ (values shown only for illustration). Each set forms a temporal sequence, and apply-

ing $\phi(\cdot) = [\mu, \sigma, \text{median}, \text{IQR}, \text{kurtosis}, \text{skewness}]$ yields 6 features per sequence. Repeating this for (i) static, (ii) moving-object, and (iii) full-frame SSIM produces $3 \times 6 = 18$ SSIM-related features.

We calculated two measures related to the geometry of the pixel profiles: the curvature and the twist. Each of these measures outputs a new trajectory for each pixel profile. For each pixel location $\mathbf{p}$ and frame index j , we compute scalar values $\kappa_j(\mathbf{p})$ (curvature) and $\tau_j(\mathbf{p})$ (twist) using standard discrete (numerical) approximations. To reduce the influence of independently moving objects, we restrict the computation to static regions by ignoring samples where $\tilde{\mathcal{M}}_j(\mathbf{p}) = 0$. We then aggregate each descriptor spatiotemporally, i.e., we pool all valid values across all pixel locations $\mathbf{p}$ and all frame indices j in the video (flattening the space-time field into a single set of values) and compute six summary statistics (average, standard deviation, median, interquartile range, kurtosis, and skewness). This yields 6 features for curvature and 6 for twist, totaling 12 geometric features.

The frequency component is based on metrics such as LHR. However, instead of using a threshold and analyzing the global motion of the video, we compute the frequency spectrum of each pixel profile and separate the frequency energy by B bins. This yields a set of per-profile energies, which we then pool over pixel locations to obtain fixed-length frequency features. As we separately analyze the horizontal and vertical components, we have $2B$ values per pixel profile. We compute the six statistical measures over the pooled set of per-profile energies and obtained $12B$ values. We used $B = 10$, totaling 120 features.

Finally, the kinematic component comprises VCP^2 , ACP^2 , and JCP^2 . In this case, we compute the six statistical measures for each of the three kinematic metrics in pixels related to moving objects and static regions, totaling 36 features. Table 2 summarizes the composition of the final multi-aspect 204-dimensional feature vector used in our experiments.

Table 2. Composition of the 204-dimensional multi-aspect feature vector used in Alternative B.

Aspect	#Features
Image similarity (PSNR/SSIM/LPIPS + region-wise/full-frame summaries)	36
Geometry (curvature, twist + pooled summaries)	12
Frequency (per-profile spectrum with $B = 10$ bins; x/y components)	120
Kinematics ($\text{VCP}^2/\text{ACP}^2/\text{JCP}^2$ + region-wise summaries)	36
Total	204

4 Experiments

In this section, we present the experiments performed in this work. Subsection 4.1 describes the configurations used, while Subsection 4.2 presents and analyzes the obtained results.

4.1 Setup

We experimented with five regressors: Support Vector Machine Regressor (SVR), Random Forest (RF), Gradient Boosting Machine (GBM), and eXtreme Gradient Boosting (XGBoost), as well as a Linear Fit, where we fit a simple straight line on the features. We report the mean Pearson Linear Correlation Coefficient (PLCC) / Spearman Rank-Order Correlation Coefficient (SROCC) over the ten predefined training/testing splits used in the original evaluation.

For each regressor, we perform a 3-fold cross-validation for a hyperparameter grid search. For the SVR, we look for $C = \{2^{-5}, 2^0, 2^1, 2^3, 2^6, 2^{10}\}$ and $\gamma = \{2^{-15}, 2^{-10}, 2^{-5}, 2^0, 2^5\}$. For the RF, we look for $\text{max_features} = \{10, 15, 20, 25\}$ and $\text{n_estimators} = \{100, 200, 300, 400\}$. The hyperparameters fetched for the GBM were $\text{n_estimators} = \{30, 70, 100\}$, $\text{learning_rate} = \{0.1, 0.05\}$. Finally, we fetch the following hyperparameters for XGBoost: $\text{learning_rate} = \{0.01, 0.05, 0.1, 0.15, 0.2\}$ and $\text{max_depth} = \{3, 4, 5, 6, 7, 8, 9, 10\}$.

We performed ten executions with two non-overlapping subsets (training and testing) of the same size. We reported the averages of the correlation values across the 10 testing subsets (even for non-trainable measures). For each of the ten training/testing splits, we compute all non-trainable measures on the corresponding testing split only. Trainable regressors are fitted on the training split and evaluated on the same testing split. We then report the mean correlation across the ten testing splits, ensuring a fair comparison between trainable and non-trainable methods in Table 3.

Because the proposed feature vector mixes heterogeneous descriptors (e.g., PSNR in dB, SSIM/LPIPS in normalized ranges, spectral energies, geometric quantities, and kinematic derivatives), we normalize each feature dimension before training using `StandardScaler` from `scikit-learn` (z-score standardization: zero mean and unit variance). For each training/testing split, the scaler is fitted on the training split only and then applied to both training and testing features. This normalization is performed within the cross-validation/grid-search procedure to avoid data leakage.

4.2 Results

Table 3 presents the correlations between the human perception stability scores (LIVE-Qualcomm and MIND-VQ datasets), as well as (i) different methods of evaluating the stability from the literature, (ii) the kinematic measures, and (iii) different regression methods that use the previously described features as input (Figure 8).

For the regression-based approach, the reported results refer to MIND-VQ. Due to the small number of samples in LIVE-Qualcomm, we did not train the non-linear regressors on this dataset. For completeness, we report a Linear Fit baseline under the same split protocol, while the learning-based analysis focuses on MIND-VQ. To the best of our knowledge at the time of our experiments, no additional public datasets simultaneously provided (i) videos with varying levels of camera instability and (ii) per-video subjective stability scores (MOS) collected under controlled conditions with sufficient methodological detail for a comparable evaluation.

Therefore, we keep the main regression-based analysis in the intra-dataset setting on MIND-VQ. LIVE-Qualcomm is used for correlation analysis of non-trainable measures and as a testing set for inter-dataset generalization (Table 4), but we do not train regressors on LIVE-Qualcomm. Furthermore, running experiments with IGC in MIND-VQ was not feasible due to its high computational cost.

We report two correlation coefficients between an assessment output and subjective stability MOS: the Pearson Linear Correlation Coefficient (PLCC), which measures linear agreement, and the Spearman Rank-Order Correlation Coefficient (SROCC), which measures rank consistency.

Table 3. Correlation between human perception of stability scores and different strategies for assessing stability.

Measure	LIVE-Qualcomm		MIND-VQ	
	PLCC	SROCC	PLCC	SROCC
LHR	0.388	0.404	0.538	0.489
ITF (PSNR)	0.015	0.024	0.317	0.308
ITF (SSIM)	0.081	0.072	0.200	0.196
IGC	0.626	0.614	-	-
VCP ²	0.204	0.405	0.546	0.555
ACP ²	0.710	0.720	0.781	0.769
JCP ²	0.636	0.755	0.753	0.747
Linear Fit	0.142	0.181	0.706	0.767
SVR	-	-	0.853	0.815
RF	-	-	0.880	0.839
GBM	-	-	0.884	0.843
XGBoost	-	-	0.884	0.842

In most instances, ACP² exhibited the highest correlation values among the stability measures, except for SROCC in the LIVE-Qualcomm dataset. The commonly used metrics (LHR and ITF) achieved low correlations. These outcomes suggest that the quantitative results reported in the literature (typically relying on LHR) lack robust alignment with human perceptions of stability.

4.2.1 Non-trainable Measures vs. Subjective Stability

Table 3 contrasts widely used objective measures (LHR and ITF variants), the geometric IGC measure, and the proposed kinematic measures. Two observations follow directly from the table. First, on both datasets, LHR exhibits only moderate correlation with MOS, while ITF variants show substantially lower agreement. This suggests that, for the shaky videos considered here, interframe similarity alone is not a reliable proxy for perceived stability, and that the frequency-based assumption behind LHR does not fully capture the perceptual factors driving stability judgments. Second, among the non-trainable alternatives, the proposed kinematic measures (especially ACP² and JCP²) achieve the strongest correlations with MOS on both datasets, indicating that derivatives of pixel-profile motion provide informative cues for instability.

4.2.2 Interpreting the Kinematic Measures

The relative behavior of VCP², ACP², and JCP² can be qualitatively related to what each temporal derivative emphasizes.

VCP² captures first-order frame-to-frame changes and may respond to smooth camera motion (e.g., steady panning), which can be perceptually acceptable in many videos. Higher-order derivatives (ACP² and JCP²) emphasize abrupt temporal variations and are therefore more sensitive to oscillations and jitter commonly associated with perceived instability. This interpretation is consistent with the observed ranking in Table 3, where VCP² underperforms ACP²/JCP² on both datasets.

On LIVE-Qualcomm, ACP² attains higher PLCC while JCP² attains higher SROCC. Since PLCC measures linear agreement in scale and SROCC measures rank consistency, this split may reflect a more non-linear relationship between MOS and jerk-related cues on this dataset. In practice, ACP² may align better with the MOS scale under a near-linear mapping, while JCP² may better preserve relative ordering even when the mapping is non-linear.

4.2.3 Learning-Based Regressors on MIND-VQ (Intra-Dataset)

The bottom part of Table 3 reports learning-based results for MIND-VQ using the proposed 204-dimensional multi-aspect feature vector. The Linear Fit baseline already yields strong correlations (PLCC=0.706, SROCC=0.767), indicating that the features are informative even under a simple linear mapping. Non-linear regressors provide further gains, with boosting-based methods (GBM/XGBoost) achieving the highest correlations (PLCC=0.884, SROCC=0.843), which suggests that modeling non-linear interactions across heterogeneous feature groups (similarity, frequency, geometry, and kinematics) is beneficial.

4.2.4 Inter-dataset Generalization and Domain Shift

Table 4 evaluates generalization when regressors are trained on MIND-VQ and tested on LIVE-Qualcomm. The performance drop relative to the intra-dataset setting indicates that the learned mapping between features and MOS is sensitive to domain shift. In this work, such shift may arise from differences in capture conditions and content between the datasets (e.g., MIND-VQ consists of night scenes, while LIVE-Qualcomm contains a broader variety of in-capture distortions and capture settings). As a result, while learning-based models are effective in the intra-dataset setting, a non-trainable kinematic measure such as ACP² can be more robust when applied across datasets without adaptation.

For example, the best intra-dataset PLCC on MIND-VQ (0.884) drops to 0.355 when training on MIND-VQ and testing on LIVE-Qualcomm (GBM), highlighting the sensitivity to domain shift.

In the inter-dataset setting, the StandardScaler parameters are fitted on the MIND-VQ training data and applied unchanged when evaluating on LIVE-Qualcomm, ensuring an out-of-domain test without target-domain normalization.

4.2.5 Practical Implications

Taken together, the results indicate a clear trade-off between simplicity and data dependence. Alternative A provides a training-free assessment that achieves strong correlations with

MOS via higher-order kinematic derivatives computed from optical-flow-based pixel profiles and background masking. Alternative B can further improve correlation in the intra-dataset setting by learning a mapping from a multi-aspect feature set, but its performance degrades under domain shift, motivating future work on improving generalization and robustness.

Table 4. Correlation between stability scores (assigned by humans) and results of regression algorithms using our features with MIND-VQ training and LIVE-Qualcomm testing.

Measure	PLCC	SROCC
Linear Fit	0.240	0.216
SVR	0.238	0.307
RF	0.352	0.327
GBM	0.355	0.341
XGBoost	0.306	0.302

4.2.6 Computational Considerations

As discussed in Subsection 3.2, the main computational footprint of our pipeline comes from the preprocessing stages (optical flow and segmentation), which require running deep models for $(T - 1)$ frame pairs and T frames, respectively. In contrast, the proposed kinematic measures and the multi-aspect feature aggregation mainly involve per-pixel arithmetic and pooling operations. This breakdown is relevant when choosing between Alternative A (training-free, lightweight post-processing) and Alternative B (additional feature extraction and learning), and it motivates future work on reducing preprocessing cost (e.g., faster motion estimation or lower-resolution evaluation) without sacrificing correlation with subjective stability.

5 Conclusions

We presented an analysis of stability assessment measures, correlating them with user-provided stability scores from the LIVE-Qualcomm and MIND-VQ datasets. In addition to examining measures from the literature, we also proposed novel kinematics-based metrics. Our measures demonstrated a better performance, especially the one called ACP². We improved the results achieved by the most widely used measure (LHR) by 24 to 32 PLCC points.

Additionally, we introduced features based on multiple aspects: image similarity, geometry, frequency analysis, and kinematics. These features were employed to train a regressor aimed at predicting the scores. Experiments using the same dataset for training and testing yielded an enhancement by up to 10.3 additional PLCC points (mean over the ten testing splits, as seen in Table 3).

A practical limitation is the computational footprint of the preprocessing stages (optical flow and segmentation), as discussed in Subsection 3.2. While our post-processing is lightweight, reducing the cost of motion estimation remains an important direction for future work. In addition, the inter-dataset experiments indicate that regression-based mappings

are sensitive to domain shift, motivating future work on improving generalization.

Declarations

Authors' Contributions

MRS contributed to the conception of this study, the implementation of codes, the conduction of experiments, the evaluation of results, and the preparation of the manuscript. LGLD and JLFC contributed by editing and reviewing the manuscript, as well as analyzing the experimental results. HAM and HP contributed by editing and reviewing the manuscript, as well as supervising the research work. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the Laboratory of Visual Informatics (LIV) and the Laboratory of Image Data Science (LIDS) for providing the infrastructure necessary to develop this work.

Funding

The authors are grateful to the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq grant #304836/2022-2) for their financial support.

Availability of data and materials

Data can be made available upon request.

References

- Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal Energy Models for the Perception of Motion. *Journal of the Optical Society of America*, 2(2):284–299. DOI: 10.1364/josaa.2.000284.
- Bolles, R. and Baker, H. (1985). Epipolar-Plane Image Analysis: A Technique for Analyzing Motion Sequences. In *3th IEEE Workshop on Computer Vision, Representation, and Control*, pages 168–178. IEEE. DOI: 10.1016/B978-0-08-051581-6.50009-X.
- Ghadiyaram, D., Pan, J., Bovik, A. C., Moorthy, A. K., Panda, P., and Yang, K.-C. (2017). In-capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077. DOI: 10.1109/tcsvt.2017.2707479.
- Grundmann, M., Kwatra, V., and Essa, I. (2011). Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths. In *Conference on Computer Vision and Pattern Recognition*, pages 225–232. IEEE. DOI: 10.1109/cvpr.2011.5995525.

- Guan, X., Li, F., Huang, Z., and Liu, H. (2022). Study of Subjective and Objective Quality Assessment of Night-Time Videos. *Transactions on Circuits and Systems for Video Technology*, 32(10):6627–6641. DOI: 10.1109/tcsvt.2022.3177518.
- Halperin, T., Poley, Y., Arora, C., and Peleg, S. (2017). Egocentric Sampling: Wide View Hyperlapse From Egocentric Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5):1248–1259. DOI: 10.1109/tcsvt.2017.2651051.
- Ito, M. S. and Izquierdo, E. (2019). A Dataset and Evaluation Framework for Deep Learning Based Video Stabilization Systems. In *Visual Communications and Image Processing*, pages 1–4. IEEE. DOI: 10.1109/vcipc47243.2019.8966057.
- James, J. G., Jain, D., and Rajwade, A. (2023). GlobalFlowNet: Video Stabilization using Deep Distilled Global Motion Estimates. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5078–5087. DOI: 10.1109/wacv56688.2023.00505.
- Joshi, N., Kienzle, W., Toelle, M., Uyttendaele, M., and Cohen, M. F. (2015). Real-Time Hyperlapse Creation Via Optimal Frame Selection. *ACM Transactions on Graphics*, 34(4):1–9. DOI: 10.1145/2766954.
- Kopf, J., Cohen, M. F., and Szeliski, R. (2014). First-Person Hyper-Lapse Videos. *ACM Transactions on Graphics*, 33(4):1–10. DOI: 10.1145/2601097.2601195.
- Kossi, K., Coulombe, S., Desrosiers, C., and Gagnon, G. (2022). No-Reference Video Quality Assessment Using Distortion Learning and Temporal Attention. *IEEE Access*, 10:41010–41022. DOI: 10.1109/access.2022.3167446.
- Kou, T., Liu, X., Sun, W., Jia, J., Min, X., Zhai, G., and Liu, N. (2023). StableVQA: A Deep No-Reference Quality Assessment Model for Video Stability. In *ACM International Conference on Multimedia*, pages 1066–1076. DOI: 10.1145/3581783.3611860.
- Lai, W.-S., Huang, Y., Joshi, N., Buehler, C., Yang, M.-H., and Kang, S. B. (2017). Semantic-Driven Generation Of Hyperlapse From 360 Degree Video. *IEEE Transactions on Visualization and Computer Graphics*, 24(9):2610–2621. DOI: 10.1109/tvcg.2017.2750671.
- Lee, Y.-C., Tseng, K.-W., Chen, Y.-T., Chen, C.-C., Chen, C.-S., and Hung, Y.-P. (2021). 3D Video Stabilization with Depth Estimation by CNN-based Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10621–10630. DOI: 10.1109/cvpr46437.2021.01048.
- Liu, S., Yuan, L., Tan, P., and Sun, J. (2013). Bundled Camera Paths for Video Stabilization. *ACM Transactions on Graphics*, 32(4):1–10. DOI: 10.1145/2461912.2461995.
- Liu, S., Yuan, L., Tan, P., and Sun, J. (2014). Steadyflow: Spatially Smooth Optical Flow for Video Stabilization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4209–4216. DOI: 10.1109/cvpr.2014.536.
- Liu, X., Min, X., Sun, W., Zhang, Y., Zhang, K., Timofte, R., Zhai, G., Gao, Y., Cao, Y., Kou, T., et al. (2023). NTIRE 2023 Quality Assessment of Video Enhancement Challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1551–1569. DOI: 10.1109/cvprw59228.2023.00158.
- Min, X., Duan, H., Sun, W., Zhu, Y., and Zhai, G. (2024). Perceptual Video Quality Assessment: A Survey. *Science China Information Sciences*, 67(11):211301. DOI: 10.1007/s11432-024-4133-3.
- Morimoto, C. and Chellappa, R. (1997). Evaluation of Image Stabilization Algorithms. In *DARPA Image Understanding Workshop*, pages 295–302. DOI: 10.1109/icassp.1998.678102.
- Nie, Y., Su, T., Zhang, Z., Sun, H., and Li, G. (2017). Dynamic Video Stitching via Shakiness Removing. *IEEE Transactions on Image Processing*, 27(1):164–178. DOI: 10.1109/tip.2017.2736603.
- Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., and Häkkinen, J. (2016). CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086. DOI: 10.1109/tip.2016.2562513.
- Peng, Z., Ye, X., Zhao, W., Liu, T., Sun, H., Li, B., and Cao, Z. (2024). 3D Multi-Frame Fusion for Video Stabilization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7507–7516. DOI: 10.1109/cvpr52733.2024.00717.
- Poley, Y., Halperin, T., Arora, C., and Peleg, S. (2015). Egocentric Sampling: Fast-Forward And Stereo For Egocentric Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4768–4776. DOI: 10.1109/cvpr.2015.7299109.
- Rao, Q., Yu, X., Navasardyan, S., and Shi, H. (2023). Sim2RealVS: A New Benchmark for Video Stabilization With a Strong Baseline. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5406–5415. DOI: 10.1109/wacv56688.2023.00537.
- Shi, Z., Shi, F., Lai, W.-S., Liang, C.-K., and Liang, Y. (2022). Deep Online Fused Video Stabilization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1250–1258. DOI: 10.1109/wacv51458.2022.00094.
- Silva, M. M., Ramos, W. L. S., Ferreira, J. P. K., Campos, M. F. M., and Nascimento, E. R. (2016). Towards Semantic Fast-Forward And Stabilized Egocentric Videos. In *European Conference on Computer Vision*, pages 557–571. Springer. DOI: 10.1007/978-3-319-46604-0_40.
- Souza, M. R., Almeida, H. M., and Pedrini, H. (2024a). Digital Video Stabilization: Methods, Datasets, and Evaluation. In *37th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 42–48. SBC. DOI: 10.5753/sibgrapi.est.2024.31643.
- Souza, M. R., Almeida Maia, H., and Pedrini, H. (2023). Rethinking Two-Dimensional Camera Motion Estimation Assessment for Digital Video Stabilization: A Camera Motion Field-based Metric. *Neurocomputing*, 559:126768. DOI: 10.1016/j.neucom.2023.126768.
- Souza, M. R., Almeida Maia, H., Vieira, M. B., and Pedrini, H. (2020). Survey on Visual Rhythms: A Spatio-Temporal Representation for Video Sequences. *Neurocomputing*, 402:409–422. DOI: 10.1016/j.neucom.2020.04.035.
- Souza, M. R., da Fonseca, L. F. R., and Pedrini, H. (2018). Improvement of Global Motion Estimation in Two-Dimensional Digital Video Stabilisation Methods. *Image Processing*, 12(12):2204–2211. DOI: 10.1049/iet-

- ipr.2018.5445.
- Souza, M. R., Maia, H. A., and Pedrini, H. (2022). Survey on Digital Video Stabilization: Concepts, Methods, and Challenges. *ACM Computing Surveys*, 55(3):1–37. DOI: 10.1145/3494525.
- Souza, M. R., Maia, H. A., and Pedrini, H. (2024b). NAFT and SynthStab: A RAFT-based Network and a Synthetic Dataset for Digital Video Stabilization. *International Journal of Computer Vision*, pages 1–26. DOI: 10.1007/s11263-024-02264-8.
- Souza, M. R. and Pedrini, H. (2018a). Combination of Local Feature Detection Methods for Digital Video Stabilization. *Signal, Image and Video Processing*, 12(8):1513–1521. DOI: 10.1007/s11760-018-1307-8.
- Souza, M. R. and Pedrini, H. (2018b). Digital Video Stabilization based on Adaptive Camera Trajectory Smoothing. *EURASIP Journal on Image and Video Processing*, 2018(1):37. DOI: 10.1186/s13640-018-0277-7.
- Souza, M. R. and Pedrini, H. (2019). Motion Energy Image for Evaluation of Video Stabilization. *The Visual Computer*, 35(12):1769–1781. DOI: 10.1007/s00371-018-1572-0.
- Souza, M. R. and Pedrini, H. (2020). Visual Rhythms for Qualitative Evaluation of Video Stabilization. *EURASIP Journal on Image and Video Processing*, 2020:1–19. DOI: 10.1186/s13640-020-00508-4.
- Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems*, 22(2):213–227. DOI: 10.1007/s00530-014-0446-1.
- Teed, Z. and Deng, J. (2020). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*, pages 402–419. Springer. DOI: 10.1007/978-3-030-58536-5_24.
- Tezcan, M. O., Ishwar, P., and Konrad, J. (2021). BSUV-Net 2.0: Spatio-Temporal Data Augmentations for Video-Agnostic Supervised Background Subtraction. *IEEE Access*, 9:53849–53860. DOI: 10.1109/access.2021.3071163.
- Tezcan, O., Ishwar, P., and Konrad, J. (2020). BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2774–2783. DOI: 10.1109/wacv45572.2020.9093464.
- Wang, N., Zhou, C., Zhu, R., Zhang, B., Wang, Y., and Liu, H. (2024a). SOFT: Self-Supervised Sparse Optical Flow Transformer for Video Stabilization via Quaternion. *Engineering Applications of Artificial Intelligence*, 130:107725. DOI: 10.1016/j.engappai.2023.107725.
- Wang, Y., Huang, Q., Liu, J., Jiang, C., and Shang, M. (2024b). Adaptive Video Stabilization Based on Feature Point Detection and Full-Reference Stability Assessment. *Multimedia Tools and Applications*, 83(11):32497–32524. DOI: 10.1007/s11042-023-16607-z.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612. DOI: 10.1109/tip.2003.819861.
- Wu, H., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., and Lin, W. (2022). Disentangling Aesthetic and Technical Effects for Video Quality Assessment of User Generated Content. *arXiv preprint arXiv:2211.04894*, 1:1–8.
- Zhang, L., Zheng, Q., and Huang, H. (2018a). Intrinsic Motion Stability Assessment for Video Stabilization. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1681–1692. DOI: 10.1109/tvcg.2018.2817209.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018b). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595. DOI: 10.1109/cvpr.2018.00068.