No-reference Video Shakiness Quality Assessment

Zhaoxiong Cui, Tingting Jiang

National Eng. Lab. for Video Technology, Cooperative Medianet Innovation Center, School of EECS, Peking University, Beijing, 100871, China

Abstract. Video shakiness is a common problem for videos captured by hand-hold devices. How to evaluate the influence of video shakiness on human perception and design an objective quality assessment model is a challenging problem. In this work, we first conduct subjective experiments and construct a data-set with human scores. Then we extract a set of motion features related to video shakiness based on frequency analysis. Feature selection is applied on the extracted features and an objective model is learned based on the data-set. The experimental results show that the proposed model predicts video shakiness consistently with human perception and it can be applied to evaluating the existing video stabilization methods.

1 Introduction

With the development of digital video capture devices, such as smart phones or wearable devices, more and more people are able to take videos in daily life and upload these videos to the social media. Compared to traditional broadcast videos, these handy videos usually are not perfect because most of them are taken by amateurs. For example, due to the lack of tripods, many videos encounter the problem of shakiness. If the shakiness is severe, it will influence the video quality perceived by people. Therefore, understanding the subjective perception of human to video shakiness is important for many video applications, *e.g.*, video editing, bootleg detection. Furthermore, how to design an objective assessment model for video shakiness which is consistent with subjective perception is a challenging problem. That is, given an input video, it is expected to output a shakiness score which is consistent with human perception.

Video shakiness has been extensively studied by many researchers from different perspectives. Some works [1–4] take the amount of camera motion into account. The underlying assumption is that the larger the camera motion is, the more shaky the video is. However, this assumption is not always true. For example, if the camera moves constantly, even if the motion is large, it would not affect the video quality that much. On the other hand, if the camera moves up and down frequently, even if the motion is small, it will be annoying for the audience. Therefore, there are several methods proposed based on the frequency analysis [5–8]. They apply different filters on the motion signals and design frequency-based models.

In this paper, we first conduct subjective experiments and construct a data-set which can provide ground truth for the design of object assessment models. Second, based on this subjective data-set, we propose a frequency-based model in order to objectively evaluate the video quality with respect to shakiness. Specifically, we extract motion signals (including translation, rotation and scaling) from videos and then apply frequency band decomposition on each signal. Later these frequency-related features from videos are selected by a genetic algorithm and an objective video shakiness assessment model is learned by support vector regression method (SVR). The experimental results show that our objective assessment model can predict the video shakiness score more consistently with subjective scores than previous work.

Besides the above subjective experiments and objective model design, another contribution of our work is that we apply the proposed video shakiness assessment model on evaluating video stabilization methods. By comparing the shakiness scores given by the proposed model before and after the video stabilization, we can objectively compare the improvements of different stabilization methods, while this comparison was usually performed by human eyes subjectively before. This demonstrates one application of our method.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 explains the subjective experiments and Section 4 shows the feature extraction for video shakiness. The objective model learning and experimental results are shown in Section 5. Section 6 demonstrates its application of evaluating performance of video stabilization algorithms. Finally conclusion is given in Section 7.

2 Related Work

2.1 Video Quality Assessment

According to the availability of reference videos, video quality assessment (VQA) can be classified as three kinds : full-reference (FR) VQA, reduced-reference(RR) VQA and no-reference(NR) VQA. Among these works, NR-VQA is most challenging because no reference video information can be used. To solve this problem, many methods have been proposed. For example, Bovik *et al.* [9, 10] extract video features and apply machine learning methods in order to design a general-purpose VQA model. Our work also belongs to NR-VQA, but we are specifically interested in video quality regarding shakiness.

2.2 Video Shakiness Analysis

Most previous work on video shakiness analysis can be classified as two categories in general: one is based on camera motion without filtering and the other is based on frequency analysis.

As for the former category, the underlying assumption is that the degree of video shakiness depends on the amount of camera motion only. For home

video editing, Girgensohn *et al.* [1] compute a numerical "unsuitability score" based on a weighted average of horizontal and vertical pan. According to the unsuitability score, videos can be classified as four categories. Besides pan, Mei *et al.*[2] represent the camera motion as three independent components(pan, tilt and zoom) and proposes a "jerkiness factor" for each frame as follows:

$$S_i = \max\{(\omega_p P + \omega_T T)/(\omega_p + \omega_T), Z\} \tag{1}$$

where S_i denotes the jerkiness factor for i-th frame, P is pan, T is tilt, Z is zoom. P, T, Z are all normalized to $[0, 1]$, ω_p and ω_q are weighting factors. A video's jerkiness is defined as the average of frame-level factors. These two works are cited by Xia *et al.* [3] and used as a component as a general video quality assessment system for web videos with weighting parameters as $\omega_p = 1, \omega_q = 0.75$. Similarly, Hoshen *et al.* [4] defines the shakiness of t-th frame $Q_{stab}(t)$, as the average square displacement of all feature points between adjacent frames, *i.e.*,

$$Q_{stab}(t) = \sqrt{(dx(t))^2 + (dy(t))^2} \tag{2}$$

where $dx(t)$ and $dy(t)$ denote the horizontal and vertical movement of this frame. The above methods all take the amount of camera motion between frames as the indicator of video shakiness, but ignore that different frequency components contained by the camera motion have different influences on human perception.

To address this issue, the latter category of previous work applies frequency analysis on motion signals from videos. For example, Shrestha *et al.* [5] and Campanella *et al.* [6] apply a FIR filter on the translation of video frames and then take the difference between the filtered signal and original signal, which corresponds to the high-frequency component, as the amount of shakiness. Alam *et al.* [7] and Saini *et al.* [8] take similar approaches but median filter is used. Although these works realize the importance of frequency decomposition, they only exploit the high-frequency component and discard other frequency components. In addition, the influence of frame rates on the filtering is ignored.

Besides these two categories, there are some previous work using other methods. For example, Yan *et al.* [11] compare the movement vectors between adjacent frames. If the angle between the two vectors is larger than $\pi/2$, they think this frame contains shakiness. In order to detect bootleg automatically, Visentini-Scarzanella *et al.* [12] retrieve the inter-frame motion trajectories with feature tracking techniques and then compute a normalized cross-correlation matrix based on the similarities between the high-frequency components of the tracked features' trajectories. Bootleg classification is based on the comparison between the correlation distribution and the trained models. However, these two works do not give quantitative metrics for video shakiness evaluation.

It is worth noting that all the above works do not consider video watching conditions, such as the screen size and watching distance. And these models are not verified by subjective experiments devoted to video shakiness.

3 Subjective VQA Experiment

3.1 Test sequences

We selected 4 queries, "scenery", "animal", "vehicle" and "sport", designed to retrieve the top ranked, high-definition real-world videos in four respective categories from youku.com. In November 2015, we issued the four queries to video search engine, soku.com, and collected all retrieved videos. All original videos we collected are encoded by H264/AVC codec, with target bit-rate 1600kbps, all in .flv format. For the sake of compatibility with our test platform, we converted the videos into .webm format encoded by VP8 codec. We used FFmpeg libvpx library for trans-coding, and set the quality parameters of output videos good enough ($\mathrm{crf} = 4, \mathrm{targetbitrate} = 2\mathrm{Mbps}$) to guarantee the fidelity. Then we cropped the videos into 512 sequences as our data-set. Each sequence lasts 10 seconds, and most ($> 99\%$) of the sequences are cropped within one shot to avoid the influence of scene switching between shots. Sequences with other severe distortions, like blurring and color distortion, were eliminated to avoid the masking effects. Numbers of the sequences in each categories are listed below:

Category	Number of Sequences
scenery	35
animal	134
vehicle	297
sport	46
Total	512

Table 1. Size of our data-set

As recommended in ITU-T Recommendation P.910 [13], we calculated the Spatial Information(SI) and Temporal Information(TI) of video sequences. SI and TI metrics quantify spatial and temporal perceptual information content of a given sequence. As shown in Fig. 1, the sequences span a large portion of spatial-temporal information plane, which implied a good variety of our data-set.

Among 512 video sequences in the data-set, 2 sequences falling at the extremes of the shakiness quality scale (one for the best quality, the other for the worst) were chosen for anchoring. Anchoring sequences were displayed with shakiness quality labeled to indicate the range boundaries of shakiness intensity. For the purpose of training, another 10 sequences were randomly selected as dummy (or stabilizing) presentations. Dummy presentations were adopted to familiarize the participants with the experiment process and to stabilize their opinion. The remaining 500 sequences were used as real presentations.

For the session division, the real presentations were divided into 4 parts (125 for each part). The session division, display orders of the dummy presentations, and display orders of the real presentations, were randomized for each observer to avoid the influence by the order of presentations.

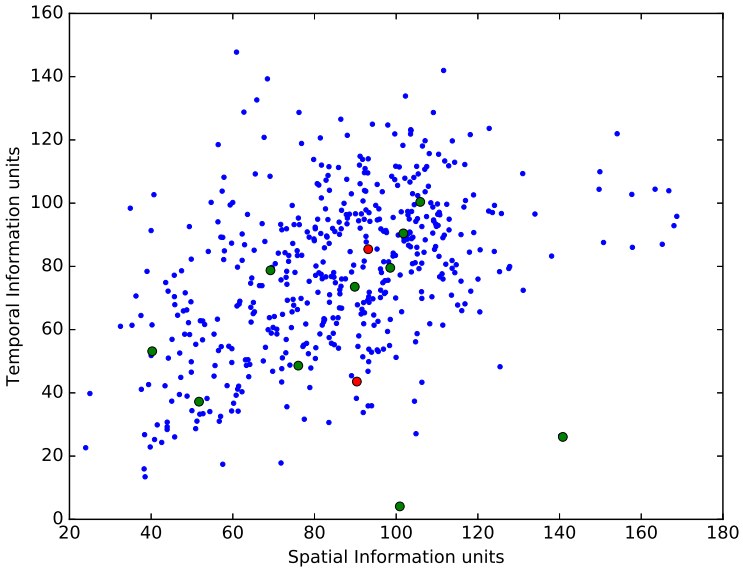

Fig. 1. Spatial-temporal plot for our test data-set: red circles for 2 anchoring sequences; green circles for 10 dummy presentations; blue dots for 500 real presentations.

3.2 Test Protocol

Test Environment The test sequences were displayed on a Dell UltraSharp U2414H 23.8-inch light-emitting diode liquid crystal display (LED-LCD) monitor (1920×1080 at 60Hz). At default factory settings, the U2414H was set to 75% brightness, which we measured at $254cd/m^2$. The contrast ratio was 853:1 and the viewing angle reached 178-degree. Other room illumination was low. A mini-DisplayPort video signal output from a HP folio 9470m laptop computer was adopted as signal source. The distance between the observer and the monitor was held at about 85cm which is about three times the height of the monitor.

Observers Twenty adults, including 9 female and 11 male, aged between 19 and 22, took part in the experiment. All of them were undergraduate college students, 13 majored in Computer Science, 5 in Electronics Engineering, 1 in Physics and 1 in Maths. 4 observers were practitioners in related fields (Computer Vision, Computer Graphics, or Image Processing), and the remaining 16 observers had no related expertise. No observers had experience with video quality assessment study, and no observers were, or had been, directly involved in this study. All observers reported normal visual acuity and normal color vision.

Voting Method The Single Stimulus (SS) non-categorical judgement method, with numerical scaling, was adopted for this experiment. In our SS method, an observer is presented with a video sequence, and then asked to evaluate the

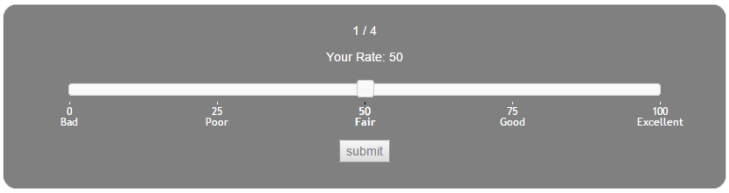

Fig. 2. The voting panel

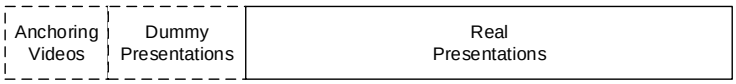

Fig. 3. A session

shakiness of the sequence by drawing a slider on a numerical scale from 0 to 100. 0 means "bad" in quality, or shaking violently, and 100 represents "excellent" in quality, or shaking unnoticeably. We labeled 5 ITU-R semantics of quality [14] with respective scores at two ends of scale and three intermediate points ("Bad" at 0, "Poor" at 25, "Fair" at 50, "Good" at 75 and "Excellent" at 100) for reference of more specific quality levels (see Fig. 2).

3.3 Experiment Procedure

An experiment contains 4 sessions in total. At the beginning of each session, anchoring sequences were presented first, followed by dummy presentations, then real presentations. Breaks were allowed between three phases. Although assessment trials in real and dummy presentations are just the same, subjective assessment data (voting scores) issued from real presentations were saved and collected after experiment, but results for dummy presentations were not processed.

In an assessment trial, a 10-second sequence faded in, presented and faded out. After that, voting panel faded in, the observer was asked to evaluate the video. Then voting panel faded out when evaluation submitted. The rating time was given at least 5 seconds, assuring that the observer voted carefully and adjacent stimuli were well isolated. The duration of a fade-in or a fade-out was set to 500 milliseconds, which provided comfortable transitions between tasks.

Observers were carefully introduced to the voting method, the grading scale, the sequence and timing at the beginning of experiment. A session lasted about

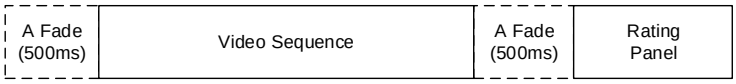

Fig. 4. A trial

half an hour, which meets the requirement prescribed by [14]. Observers were allowed to rest for a while between sessions. Usually an observer completed all 4 sessions at one time as suggested. Observers who didn't complete at one time were introduced again before their next session.

3.4 Results

A total number of 10,000 (500 sequences by 20 observers) voting scores were processed. As recommended in [14], the outlier detection for observers were imposed, but no outlier was detected.

The Mean Opinion Score (MOS) value of i-th sequence is defined as

$$MOS_i = \frac{1}{K} \sum_k S_{ki}, \tag{3}$$

where S_{ki} is score of sequence i voted by observer k, K is the number of observers. A higher MOS indicates better subjective shakiness quality of a sequence.

4 Feature Extraction

In this section, we design a no-reference video quality metric to predict perceived shakiness quality of web videos. Firstly, the global motion, namely the motion between adjacent frames, is extracted from the video sequence. Next, we transform the translation into the *deflection angle*, directly relating to the signal perceived by human visual system (HVS). Thereafter, the motion signals are decomposed into sub-bands, which contain frequency components of different levels. In the end, the statistics of each sub-band of the motion signals are calculated, as the features we designed for the video shakiness quality.

4.1 Global Motion Estimation

Global motion is defined as the geometrical transformation between adjacent video frames. It also indicates the motion of camera. Here we describe the global motion with a similarity transformation model, with four parameters $[d_x, d_y, \theta, \rho]$, corresponding to pan, tilt, rotation and isotropic scaling. Assuming (x_1, y_1) is the coordinates (with respect to the center of the frame) of a point in current frame F_t, and (x_2, y_2) is the coordinates of the corresponding point in next frame F_{t+1}, global motion can be illustrated by

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \rho \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix}. \tag{4}$$

Generally, there are two types of global motion estimation (GME) approaches: feature-based methods and featureless methods. Feature-based methods (e.g. [15–18]) utilize geometric features extracted in frames, such as Harris corners (see [19]), and then estimate the motion by matching the corresponding features

between adjacent frames. Featureless approaches directly estimate the global motion from all pixels on each frame. Usually, feature-based approaches are fast and accurate, however fragile. On the contrary, though time-consuming, featureless approaches are usually robust. In our task, web videos contained complex and intensive motion. So robustness is necessary. We adopted an FFT-based featureless approach in [20–22], measuring translation (d_x, d_y), rotation θ and scaling ρ directly from the spectrum correlation between frames. During our test on web videos, this FFT-based approach reaches a satisfying robustness and accuracy, with an acceptable time cost.

4.2 Perceptual Modeling

To properly measure the influence of global motion perceived by the viewer, we need to model the global motion signal in a physical meaning, and consider its impact on human visual system (HVS).

In previous section, the translation signals (d_x, d_y) between adjacent frames are estimated, in pixel unit. However, we need physics quantities directly related to the stimulus received by HVS. Considering the viewing condition, including viewing distance and display size, translation (d_x, d_y) shall be transformed into *deflection angle* (α_x, α_y), *i.e.*,

$$\alpha_{x,y}(t) = \arctan\left(\frac{L_d d_{x,y}(t)}{Zs}\right) \approx \frac{L_d d_{x,y}(t)}{Zs} \tag{5}$$

where $d_{x,y}(t)$ is translation at frame t in pixel unit, L_d is the diagonal length of the display monitor, Z is the viewing distance, and $s = \sqrt{h^2 + w^2}$, where h, w is the height and the width of the video frame in pixel unit, respectively. Deflection angle indicates the *shift* of viewing angle caused by the translation between the two frames (see Fig. 5).

It is noticed in [23] that the subjective sensation of motion is proportional to the logarithm of the stimulus intensity, *i.e.*, velocity (Weber-Fechner law [24]). So we take the logarithm of $\alpha_{x,y}(t)$, called *logarithm of deflection angle*, as

$$l_{x,y}(t) = \log \alpha_{x,y}(t). \tag{6}$$

Rotation signal $\theta(t)$ and scaling signal $\rho(t)$ are used directly, This is because HVS perception of the rotation and scaling signals are not directly influenced by viewing condition.

4.3 Sub-band Decomposition

There is evidence that different frequency compositions have different impact on HVS perception [25], more specifically on shakiness perception. So we decompose the signals into three different frequency sub-bands: low band for $(0, 3\text{Hz})$, mid band for $(3\text{Hz}, 6\text{Hz})$ and high band for $(6\text{Hz}, 9\text{Hz})$. Decomposition is done by filtering the original signal by three respective filters, *i.e.*,

$$S_{\mathrm{l,m,h}}(t) = S(t) * h_{\mathrm{l,m,h}}(t) \tag{7}$$

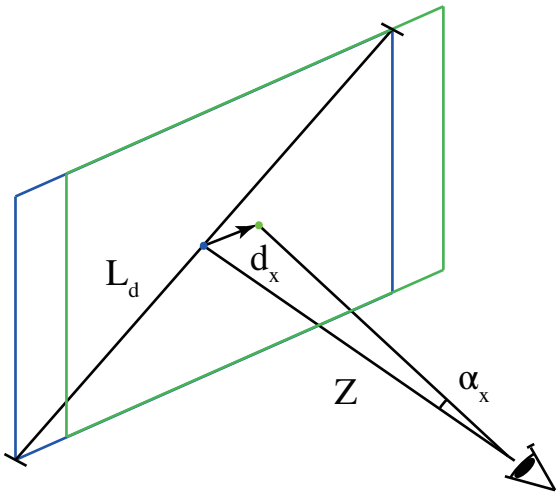

Fig. 5. Illustration of the deflection angle α_x.

original	$\alpha_x(t)$	$\alpha_y(t)$	$l_x(t)$	$l_y(t)$	$\theta(t)$	$\rho(t)$
low band	$\alpha_{x,\mathrm{l}}(t)$	$\alpha_{y,\mathrm{l}}(t)$	$l_{x,\mathrm{l}}(t)$	$l_{y,\mathrm{l}}(t)$	$\theta_{\mathrm{l}}(t)$	$\rho_{\mathrm{l}}(t)$
mid band	$\alpha_{x,\mathrm{m}}(t)$	$\alpha_{y,\mathrm{m}}(t)$	$l_{x,\mathrm{m}}(t)$	$l_{y,\mathrm{m}}(t)$	$\theta_{\mathrm{m}}(t)$	$\rho_{\mathrm{m}}(t)$
high band	$\alpha_{x,\mathrm{h}}(t)$	$\alpha_{y,\mathrm{h}}(t)$	$l_{x,\mathrm{h}}(t)$	$l_{y,\mathrm{h}}(t)$	$\theta_{\mathrm{h}}(t)$	$\rho_{\mathrm{h}}(t)$

Table 2. Sub-bands of motion signals

where $S(t)$ is the original signal, $S_{\mathrm{l,m,h}}(t)$ are filtered low-, mid- and high-band signals, and $h_{\mathrm{l,m,h}}(t)$ represent the corresponding impulse response functions of the three filters, and $*$ denotes the convolution operation. An illustration of the band decomposition is shown in Fig. 6.

Ideal filters are adopted in this decomposition work. The ideal filters keep frequency components only in an interval of frequency. The frequency response $H(f)$ of ideal filters are

$$H(f) = \begin{cases} 1 & f_t < f \leq f_h \\ 0 & \text{else} \end{cases} \tag{8}$$

where $H(f)$ is the Fourier transform of impulse response $h(t)$.

The six motion signals $\alpha_x(t), \alpha_y(t), l_x(t), l_y(t), \theta(t), \rho(t)$ are extracted from every video sequence. Decompositions are done for each of them. As a result, the six signals are decomposed into 18 sub-bands (see Table 2).

4.4 Statistics

Finally, statistical features that capture the impact of motion signals on HVS are extracted from each sub-band. Suppose that $s_i(t)$ is one of the sub-band

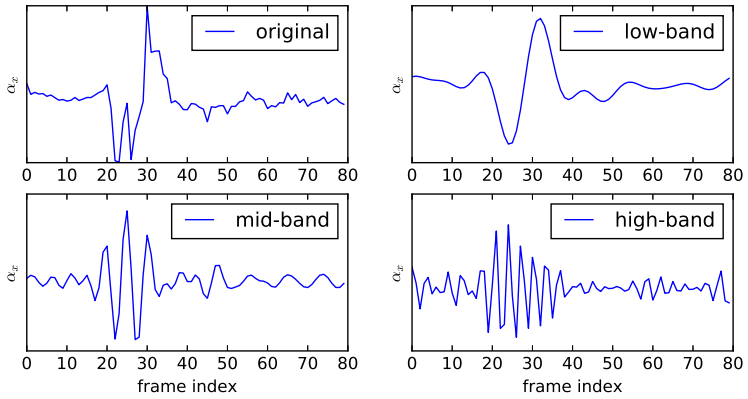

Fig. 6. Band decomposition result of $\alpha_x(t)$ signal of a video sequence.

signals, and T is the total number of frames of the video sequence, we estimate the first to the fourth central or standardized moments of $s_i(t)$, *i.e.*,

$$\begin{aligned}
\text{mean } s_i^1 &= \textstyle\sum_t s(t)/T \\
\text{variance } s_i^2 &= \textstyle\sum_t [s(t) - s_i^1]^2/T \\
\text{skewness } s_i^3 &= \textstyle\sum_t [s(t) - s_i^1]^3/[T(s_i^2)^{3/2}] \\
\text{flatness } s_i^4 &= \textstyle\sum_t [s(t) - s_i^1]^4/[T(s_i^2)^2]
\end{aligned}$$

In summary, for each video sequence, six motion signals are extracted, and decomposed into 18 sub-bands. In the next step, four statistics are estimated from each sub-band. In total, 72 ($6 \times 3 \times 4$) feature values are calculated from each video sequence.

5 Objective Experiment

In this section, we validate the performance of the extracted features, and obtain an objective no-reference video shakiness metric. We run the cross-validation test on the features, and validate the performance of the features by SROCC[26]. This cross-validation process is used for feature selection, and an optimal subset of features is obtained. We also compare our approach with other related works.

5.1 Cross Validation

We use a hold-out cross validation to evaluate the performance of the features. In each iteration, the data-set is randomly split into two parts, training set (90%) and validation set (10%). On the training set, a SVR model is trained, and then tested by the validation set. Then the performance of features is validated by

calculating SROCC between the subjective MOS and the output of the SVR model on the validation set.

LIBSVM [27] is used for SVR training. We adopt a ν-SVR with RBF kernel to get the optimal result of training. Given a set of features, this train-test process is repeated on the data-set 1000 times randomly. The median SROCC is calculated as the final performance of the set of features.

5.2 Feature Selection

In the previous section, 72 feature values are calculated from one video sequence. To get the best performance, an optimal subset of features must be chosen where SVR performs the best.

We resort to a wrapper model feature selection using a genetic algorithm (GA) [28]. Each subset of features is regarded as a genome, represented by a 72-bit number x. $x(i) = 1$ denotes the feature i is chosen and $x(i) = 0$ denotes the feature is not chosen. The fitness of genome x is determined by the median-SROCC of a 1000-times cross validation with the corresponding feature subset:

$$\text{fitness(x)} = \begin{cases} \text{SROCC}_{1000} - \text{P} & \text{SROCC}_{1000} \geq \text{P} \\ 0 & \text{SROCC}_{1000} < \text{P} \end{cases} \tag{9}$$

where P denotes the pressure of the evolution. During each generation, genomes with larger fitness are more likely to be selected to breed a new generation. More specifically, genomes with 0 fitness would never be chosen. So P determines the minimum fitness allowed in the evolution. The population of the next generation is generated by both crossover and mutation of the selected genomes (see [28]).

We adopt a two-step solution to find the optimal feature subset. In the first step, initialize the genomes by randomly choosing $x(i)$ for each i, setting $P = 0.8$, and run the genetic algorithm for 100 generations. In the second step, initialize the genomes by the genomes of the last generation in the first step , setting $P = 0.85$, and run the genetic algorithm again. The first run picks out a group of genomes with high fitness (SROCC). The second run imposes a more strict restriction, and purifies the genomes to be optimal. Finally the genome with the highest fitness in the last 5 generations during the second GA run is selected to be the optimal subset of the features. From the feature selection result, we find the translation is more important than rotation and zooming, and the low-rank moments of middle- and high-bands are more significant.

5.3 Results

The GA finally chooses an optimal set of 32 features from the 72 features. This optimal feature set performs a good result in cross-validation, with median-SROCC reaching 0.8767 (90% data for training, 10% for testing). Fig. 7 shows the scatter plot of MOS versus objective scores.

By adjusting the portion of training data, the relationship between the algorithm's performance and the amount of training data can be investigated.

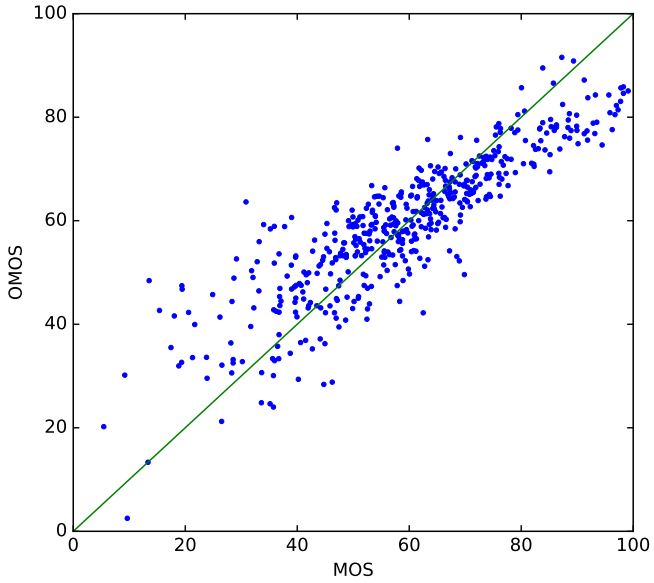

Fig. 7. MOS and objective scores (OMOS)

Starting with 1%, we gradually increased the portion of training data, and got a curve as shown in Fig. 8. With training portion exceeding 20% (train with only 100 video sequences), median-SROCC reaches 0.8. When training portion exceeds 40%, the SROCC will become stable. It shows that our approach can reach a good performance with small amount of training data. The generalization ability of our algorithm is excellent.

We compare our approach with related works [4, 3, 6], as well as only SI and TI features (trained with SVR). See Table 3. Note that the authors of the related works have not yet shared the source code, so we implement their works and test on our data-set by ourselves. The result shows that, our approach outperformed all the state-of-the-art methods.

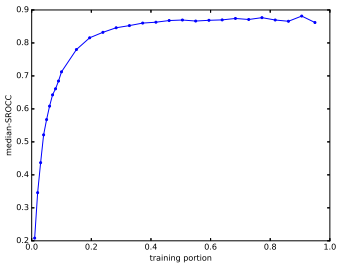

Fig. 8. The relationship between the performance and the training portion.

Method	SROCC
SI + TI (SVR)	0.4945
[4]	0.6506
[3]	0.6534
[6]	0.7669
Our Approach	**0.8767**

Table 3. Results

6 Benchmark for Stabilization Algorithms

Stabilization algorithms are designed to eliminate shakiness artifacts in videos. However, objective benchmark to test the performance of stabilization algorithms do not exist. As an application of our shakiness VQA approach, we propose a method to evaluate stabilization algorithms, by means of the NR-VQA model we learned.

Suppose V_i to be i-th video with shakiness artifact, $O(V_i)$ to be the original score given by our shakiness NR-VQA model, and suppose V_i^k to be the video stabilized by k-th stabilization algorithm, $O(V_i^k)$ to be the shakiness score of the stabilized video. Then $O(V_i^k) - O(V_i)$ is called the *enhancement* E_i^k of the stabilization algorithm k on the video V_i.

It is supposed that, the shakiness score will increase after stabilization, *i.e.*, $E_i^k > 0$. Unfortunately, E_i^k may also decrease after stabilization. For instance, if a video without shakiness artifact is stabilized, it is possible that stabilization algorithm unwillingly introduces a motion artifact to the video. In such cases, E_i^k will be less than zero, and we call the video quality is *degenerated*.

To evaluate the performance of a certain stabilization algorithm k, we define the following two indexes:

1. Average Enhancement E^k: the enhancement of stabilization algorithm k on the given data-set.

$$E^k = \frac{1}{N} \sum_i \big(O(V_i^k) - O(V_i)\big). \tag{10}$$

2. Degeneration Frequency P_d^k: the frequency of degeneration in videos stabilized by algorithm k on the given data-set.

$$P_d^k = \frac{1}{N} \sum_i I(O(V_i^k) < O(V_i)). \tag{11}$$

N is the amount of videos in the data-set. I is the indicator function: $I(A) = 1$ when A is true, otherwise $I(A) = 0$.

We stabilize all videos in our data-set, by three popular stabilization tools: Microsoft Project Oxford Video API [29], proDAD Mercalli 2.0 [30] and Adobe After Effect CC 2015 (VX deformation stabilizer) [31]. Then, we score original videos and stabilization videos by our NR-VQA model. We plot the scores of stabilization videos of three algorithms, in reference of the original scores, see Fig. 9. As shown in the figure, after stabilization the scores of videos increase generally. This shows the effect of stabilization algorithms. It is also observed that the enhancement of low-quality videos is more significant than that of high-quality videos. Moreover, indeed some videos degenerate after stabilization, and high-quality videos degenerate more frequently, exactly as expected.

Therefore, we calculate E^k and P_d^k for each algorithm separately in videos of three different quality levels: high-quality level (videos with the highest 100 $O(V_i)$), low-quality level (videos with the lowest 100 $O(V_i)$), and mid-quality level (the other 300 videos) . From the following table, it can be seen that proDAD Mercalli 2.0 performs best.

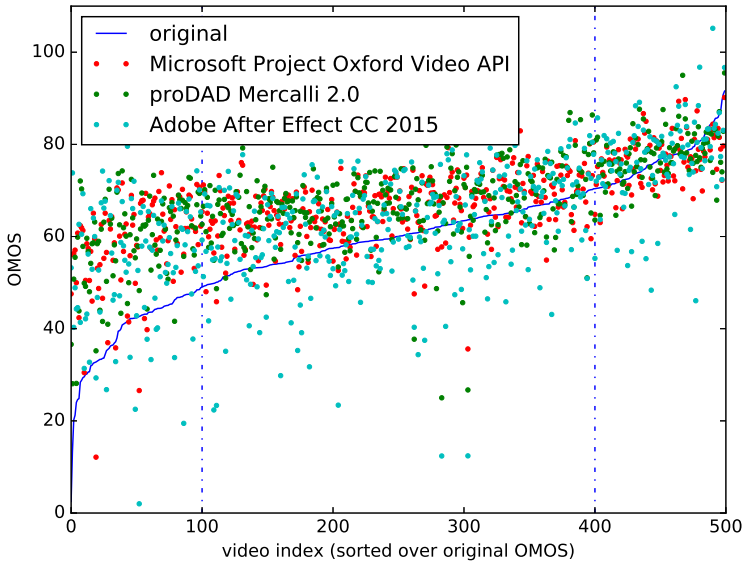

Fig. 9. Objective scores of stabilized and original videos. Dash lines indicate the boundary of video quality levels.

Stabilization Algorithm	Microsoft Project Oxford Video API		proDAD Mercalli		Adobe After Effects CC 2015	
Quality Level	P_d	E	P_d	E	P_d	E
Low	0.060	18.997	0.010	19.607	0.140	14.075
Mid	0.160	6.151	0.170	6.857	0.367	2.246
High	0.530	-0.137	0.440	0.860	0.450	-0.166
Overall	0.214	7.459	0.192	8.208	0.338	4.129

Table 4. P_d and E indexes of three stabilization algorithms

7 Conclusion

We propose a new method for video shakiness quality assessment. First, we construct a data-set based on subjective experiments. Second, based on this data-set we extract video features and learn an objective model to predict video quality in terms of shakiness. The proposed model has been validated on the constructed data-set and used to evaluate the performance of existing video stabilization methods.

Acknowledgment. This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and NSFC under contracts 61572042, 61390514, 61421062, 61210005, 61527084, as well as the grant from Microsoft Research-Asia.

References

1. Girgensohn, A., Boreczky, J., Chiu, P., Doherty, J., Foote, J., Golovchinsky, G., Uchihashi, S., Wilcox, L.: A semi-automatic approach to home video editing. In: ACM Symposium on User Interface Software and Technology. (2000) 81–89
2. Mei, T., Hua, X.S., Zhu, C.Z., Zhou, H.Q., Li, S.: Home video visual quality assessment with spatiotemporal factors. IEEE Transactions on Circuits and Systems for Video Technology **17** (2007) 699–706
3. Xia, T., Mei, T., Hua, G., Zhang, Y.D., Hua, X.S.: Visual quality assessment for web videos. Journal of Visual Communication and Image Representation **21** (2010) 826–837
4. Hoshen, Y., Ben-Artzi, G., Peleg, S.: Wisdom of the crowd in egocentric video curation. In: Computer Vision and Pattern Recognition Workshops. (2014) 587–593
5. Shrestha, P., Weda, H., Barbieri, M., With, P.H.N.D.: Video Quality Analysis for Concert Video Mashup Generation. Springer Berlin Heidelberg (2010)
6. Campanella, M., Barbieri, M.: Edit while watching: home video editing made easy. In: Electronic Imaging 2007. Volume 6506. (2007) 65060L–1–65060L–10
7. Alam, K.M., Saini, M., Ahmed, D.T., Saddik, A.E.: Vedi: A vehicular crowd-sourced video social network for vanets. In: IEEE Conference on Local Computer Networks Workshops (LCN Workshops). (2014) 738–745
8. Saini, M.K., Gadde, R., Yan, S., Wei, T.O.: Movimash: Online mobile video mashup. In: ACM International Conference on Multimedia. (2012) 139–148
9. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Transactions on Image Processing **25** (2016) 289–300
10. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Transactions on Image Processing **23** (2014) 1352–1365
11. Yan, W.Q., Kankanhalli, M.S.: Detection and removal of lighting and shaking artifacts in home videos. In: Proc. ACM Multimedia. (2002) 107–116
12. Visentini-Scarzanella, M., Dragotti, P.L.: Video jitter analysis for automatic bootleg detection. In: IEEE International Workshop on Multimedia Signal Processing. (2012) 101–106
13. ITU-T, R.: P.910: Subjective video quality assessment methods for multimedia applications. (1999)
14. ITU-R, R.: BT. 500-13, methodology for the subjective assessment of the quality of television pictures. International Telecommunications Union, Technical report (2012)
15. Perez, P., Garcia, N.: Robust and accurate registration of images with unknown relative orientation and exposure. In: International Conference on Image Processing. Volume 3., IEEE (2005) 1104–1107
16. Torr, P.H., Zisserman, A.: Feature based methods for structure and motion estimation. In: Vision Algorithms: Theory and Practice. Springer (1999) 278–294
17. Huang, J.C., Hsieh, W.S.: Automatic feature-based global motion estimation in video sequences. Consumer Electronics, IEEE Transactions on **50** (2004) 911–915
18. Ryu, Y.G., Chung, M.J.: Robust online digital image stabilization based on point-feature trajectory without accumulative global motion estimation. Signal Processing Letters, IEEE **19** (2012) 223–226
19. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference. Volume 15. (1988) 50

20. Kuglin, C.D.: The phase correlation image alignment method. Proc. Intl Conf. cybernetics and Society (1975) 163–165
21. Reddy, B.S., Chatterji, B.N.: An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Transactions on Image Processing **5** (1996) 1266–1271
22. Wolberg, G., Zokai, S.: Robust image registration using log-polar transform. In: International Conference on Image Processing. Volume 1., IEEE (2000) 493–496
23. Wang, Z., Li, Q.: Video quality assessment using a statistical model of human visual speed perception. J. Opt. Soc. Am. A **24** (2007) B61–B69
24. Hecht, S.: The visual discrimination of intensity and the weber-fechner law. Journal of General Physiology **7** (1924) 235–67
25. Winkler, S.: Issues in vision modeling for perceptual video quality assessment. Signal Processing **78** (1999) 231–252
26. Wang, Z., Sheikh, H.R., Bovik, A.C.: Objective video quality assessment. The handbook of video databases: design and applications (2003) 1041–1078
27. Chang, ChihChung, Lin, ChihJen: Libsvm: A library for support vector machines. Acm Transactions on Intelligent Systems and Technology **2** (2011) 389–396
28. Yang, J., Honavar, V.: Feature subset selection using a genetic algorithm. Intelligent Systems and Their Applications IEEE **13** (1998) 44–49
29. Microsoft: Cognitive services - video API. (https://www.microsoft.com/cognitive-services/en-us/video-api)
30. proDAD: Mercalli v2. (http://www.prodad.com/Home-29756,l-us.html)
31. Adobe: After effects CC. (http://www.adobe.com/products/aftereffects.html)