Video Acuity Assessment in Mobile Devices

Eilwoo Baik, Amit Pande, Chris Stover and Prasant Mohapatra
Univerisity of California, Davis
Email: {ebaik, pande, cjstover, pmohapatra}@ucdavis.edu

***Abstract*—The quality of mobile videos is usually quantified through the Quality of Experience (QoE), which is usually based on network QoS measurements, user engagement, or post-view subjective scores. Such quantifications are not useful for real-time evaluation. As a result, they cannot provide on-line feedback for improvement of visual acuity, which represent the actual viewing experience of the end user. We present a visual acuity framework which makes fast online computations in a mobile device and provide an accurate estimate of mobile video QoE. We identify and study the three main causes that impact visual acuity in mobile videos: spatial distortions, types of buffering and resolution changes. Each of them can be accurately modeled using our framework. We use machine learning techniques to build a prediction model for visual acuity, which depicts more than 78% accuracy. We present an experimental implementation on iPhone 4 and 5s to show that the proposed visual acuity framework is feasible to deploy in mobile devices. Using a data corpus of over 2852 mobile video clips for the experiments, we validate the proposed framework.**

Index terms: Video Quality, Quality of Experience, Mobile Video

I. INTRODUCTION

The popularization of smartphones and tablets has led to mass content consumption as well as a flood of user generated content in both wireless and wired networks. Mobile data traffic will increase at a Compound Annual Growth Rate of 61%, approaching 15.9 exabytes per month by 2018. An increased proportion of this traffic is comprised of mobile videos. In 2013, mobile video generated around 53% of mobile data traffic and the trend will increase to 69% by 2018 [11]. This traffic includes various types of video services such as progressive download, real-time video streaming, HTTP-based streaming services and interactive telephony. In all these video content services, the major concern is to transmit videos to mobile devices in fluctuating network conditions yet ensuring high fidelity or acuity. (In our discussion, we call these videos 'Mobile Videos').

Our goal in this work is to propose a framework for online evaluation of visual acuity on smartphones or tablets. We define video acuity as the perceptual experience of watching the video itself [17] regardless of confounding factors such as content type or user expectations [5]. We distinguish acuity with user engagement [5] metrics such as viewing time which may be influenced by other factors such as popularity of video, video content and viewers' mood. Acuity is close to video distortion metrics, except that it is a statement of overall video experience of the user, not a specific distortion. Video acuity is best measured by subjective user assessments (such as Mean Opinion Scores (MOS) or Differential MOS (DMOS)) of perceptual video quality. However, MOS scores are post-view, time consuming and distracting to the end-user, and reduce the viewing experience itself. This necessitates the development of an objective framework which can predict video acuity in an online manner.

Some researchers have developed predictive models for QoE using user engagement as a predicted variable [5]. These models may not be helpful to network providers who want to fine-tune their networks to improve this visual acuity for the end user, irrespective of content type or users mood, which are subjective and beyond their control. We want to develop accurate video acuity models using online light-weight computations on mobile devices. Developing such a model is a challenging task because of following factors:

- Difficulty in analyzing video traffic in terms of spatio-temporal content. Most video quality metrics require original video or have huge computational cost, video itself has huge file sizes, leading to high latency and cost of computations.
- Understanding acuity requires subjective user evaluation. Network or server level traces can only infer about user engagement statistics (such as percentage of time viewed, number of hits) but not on the objective user opinion (unless explicitly collected).

Traditional approaches to video quality tend to map MOS to QoS or video distortion metrics. However, most existing QoS based models [15] or video distortion metrics [18] focus on a single distortion or scenario. Moreover, existing video metrics are designed to capture only spatial distortions, they cannot be applied to HTTP/ TCP-based internet videos. It is indeed very challenging, if not impossible, to derive a video acuity metric for mobile videos. However, we want to provide feedback of video acuity to service providers for online network optimization. We will show in later sections that how the proposed video acuity framework can be used by service providers to analyze the impact of visual distortions on the video acuity. A robust video acuity model can also be used by end users to set a value for the video service according to quality of reception instead of paying a fixed fee to cellular and Internet service providers.

In this paper, our goal is to develop a comprehensive framework that can objectively provide real-time quantification of video acuity on mobile devices. Our proposed model enable to support both UDP-based and HTTP-based video services as a single model. The method proposed to detect freezing effects and resolution-switch does not require any additional implementation/ frameworks in network-level. The framework involves identification of the factors impacting video acuity and the corresponding metrics that can provide accurate quantification of the impairments. Based on detailed insights, the framework derives an approximate quantification of the impact of distortions on visual acuity of mobile videos. We consider a heterogeneous pool of 2500+ HD and regular mobile videos in our experiments encompassing a range of network distortions, content genres and application-type. Further videos were generated using real network testbeds and streaming experiments, and subjective ratings were performed to recognize the impact of different distortions over visual acuity; overall analyzing more than 10 million frames of video data. Some of our key observations are as follows:

1) Existing video quality metrics have high correlation to QoE across a range of spatial distortions but not across content. For example, the same video distortion over different content is measured differently by existing metrics.
2) Apart from total buffering, our experiments reveal that intensity and position of buffering events are crucial for estimating video acuity.
3) Visual acuity increases linearly with an increase in video resolution. However, our experiments reveal that online-switches-in-resolution can have nonlinear or negative impacts on video quality.
4) A predictive model can be used to estimate video acuity. Generating separate models for spatial distortion, buffering and resolution-switching using machine learning technique gave us accuracy of 93%, 81% and 87% respectively. A generic combined model gives accuracy of 78%.
5) The proposed framework can run conveniently on smartphones.

The paper is organized as follows: Section 2 gives an overview of motivation and related works. The heterogeneous dataset used in the work is explained in Section 3. Section 4 gives an analysis of mobile video distortions to build a prediction model. Section 5 gives the combined model as well as smartphone app implementation. Even an old iPhone 4 model can conveniently run our app in the background with full HD streaming and other processes.

II. Motivation and Related Works

The increasing volume of mobile video traffic motivates us to obtain a metric for robust acuity assessment which accurately maps objective scores to subjective video acuity. The most precise method to measure acuity is to conduct subjective assessments such as Mean Opinion Score (MOS) ratings [3], but it is restricted to offline evaluation.

A. Requirements of a mobile video acuity metric

The increase in mobile video traffic necessitates the development of a framework for robust quality assessment in mobile videos. The desired qualities of a mobile video metric are explained next.

Robust to video content and distortions: Many kinds of distortions such as blocking, blurring, blackout, freezing or jerkiness can be introduced during video delivery. A desired metric should be robust to various distortions and video contents.

Simple and Light : Limited resources of mobile devices are not appropriate for heavy computational cost and complexity. It may be difficult to apply psycho-visual models for acuity assessment if they incur high computation cost.

Codec-Independent : With a plethora of codecs and containers available (such as H.264, MPEG and HEVC) and many of them being proprietary, it is desired that the metric is independent of codecs and universally applicable to all video content.

Accuracy in estimating end-user QoE : The most significant factor in acuity assessment is the accuracy in estimating user QoE. The most challenging problem in the existing metrics and schemes is that one metric might be able to detect and predict user QoE for certain conditions. However, they are not able to show consistent results when applied across various types of video properties such as resolution, frame rate, encoding rate, size of GOP (Group of Pictures) and motion speed.

No need for reference video : Original video is not available for 'online' evaluation scenarios, making it difficult to estimate video quality, particularly in the presence of multiple distortions. Most of the well established metrics such as PSNR, SSIM , VSNR, VQM [14] need a reference video, therefore they are not suitable for mobile videos.

B. Related Works

Most of existing works are restricted to detect designated distortions, and limited to video display distortions such as blocking, blurring or ringing. No Reference (NR) schemes [26], [4] show inconsistency in results across various types of video content. Full Reference (FR) schemes [24], [25] are not suitable to be used for online quality assessment due to non-availability of the source video. Reduced Reference-based (RR) metrics [25], [9] have a low overhead compared to FR metrics, but still require additional channels or systems to send a partial original video reference. RR based metrics may be useful in such a scenatio, although existing RR metrics are not robust to distortion, not real-time and not light-weight for mobile devices.

Chono et al. [10] use vector information of image features and transmit its Slepian-Wolfe syndrome using an LDPC encoder. However, it targets distributed source coding and has high computation and communication overhead. Rehman and Wang [22] proposed an RR metric using structural similarity (SSIM) index. The proposed RR-SSIM shows high to medium correlation across multiple image datasets. However, the computation cost is 11 seconds per image, which is unacceptable for mobile videos.

Mittal et al.[20], [19] proposed high accuracy performance NR metrics, Blind/referenceless Image Spatial Quality Evaluator (BRIQUE) and Natural Image Quality Evaluator (NIQE).

But their performance on videos hasnt been studied. BRISQUE gives a score on spatial quality while NIQE gives naturalness index.

Balachandran et al. [5] presented a data-driven video QoE metric, but it measured user engagement (watch time or number of hits) as an indicator of video quality. This metric may not be useful to network operators (to fine-tune their service) as user engagement is affected by many other confounding factors beside video acuity.

III. Dataset

To model video acuity, we took a data corpus of 2852 video clips in this work comprised of videos from Youtube, LIVE mobile dataset as well as locally captured videos transmitted over a wireless network. The duration of each video clip is approximately 30 seconds and analyzed on a per-frame basis (making more than 10 million data points). We used three different strategies to create the data corpus.

(a) `UDPStream`: First, we created a pool of 362 videos which represent various wireless network induced distortions (such as packet losses and delay) across 5 video content. We set up a single-hop wireless testbed and implemented middleware to introduce burst and uniform packet loss and delay effects at the IP-layer by using IPFirewall [13]. Packet loss is simulated in three ways: uniformly distributed loss, bursty loss or a combination of both. The video streaming sessions are established by VLC [23], FFmpeg [7] is used for video coding. Details on parameters are specified in the Table I. 50 HD videos from `UDPStream` were tested by 17 [1] subjects on a smartphone using ITU single-stimulus (SS) method [14].

(b) `LIVE`: 200 HD videos from the LIVE dataset [21] have MOS scores averaged over 50 subjects each using ITU single-stimulus method [14] with hidden references. The videos are compressed using H.264 scalable video codec (SVC) at four different compression rates/profiles (0.7 to 6 Mbps). For rate adaptation, four profiles are generated to vary the rate dynamically within a video stream between two compression rates. For temporal dynamics, the compression rate is varied between multiple compression rates with different rate-switching structures within a single video stream. Wireless channel packet-loss is introduced using trace-based simulation. Frame-freezes include live video freezes (due to packet losses)leading to loss of temporal continuity after freeze, and stored video freezes causing no loss of temporal continuity after freeze (delay). Stored video freeze is introduced in three profiles: 8 times for 1 sec each, 4 times for 2 secs each and 2 times for 4 secs each. The two datasets together form a wide range of content (video type or genres, with a set of 17 different representative videos), GOP size, motion speed (slow to fast action scene), duration, encoder, container, resolution and frames-per-second.

(c) `UTrailers`: We have collected 2,280 popular video HD trailers for the past 3 years (2011-2013) from Youtube. The videos include most genres: action, adventure, drama, sports, games, music videos, romance, thriller, animation, education, etc. The collected video trailers include 90% of popular movies released in the past 3 years. Each video trailer is reproduced for each resolution type from 240p to 1080p. This huge data set was evaluated by 183 people aged from 16 to 62 to measure MOS values depending on different resolutions and freezing effects in the display of mobile devices. Over 183 users participated in the study and evaluated over 500 clips each. The mobile devices used for the subjective ratings can support up to 720p (HD). The screen size of the mobile devices are from 3.7 to 4.1 inches. This data collection and processing task took over 120 days.

TABLE I
Properties of video dataset used

Compression/Rate distortions: LIVE [21]	
Type	Value
Compression (**R**)	4 different compression rates
Rate Adaptation (**S**)	3 rate-switching to highest quality
Temporal Dynamics (**T**)	5 profiles with multiple rate switches each (same resolution)
Freezing (**F**)	8 secs (4 variable profile)
Packet Loss (**W**)	Uniform 4 QAM at SNR (15db); plr$\leq$ 1.19% for each rate (4)
UDPStream : Network Distortions	
Packet loss (**A**)	*Uniform* $0.1 \sim 50\%$
Packet loss (**B**)	*Burst* 90%, $2 \sim 4$ secs
Freezing	*Delay* : $1 \sim 4$ secs
UTrailers (Youtube Trailes)	
Content Geners	All (30s playtime)
Duration	30, 60 secs
Resolution	Full HD(1080p) HD(720p), others(480, 360, 240)
Screen Size	$3.7 \sim 4.1$ inch
No. applicants	162 (Age : $18 \sim 60$; Gender : M/F)
Overall content settings in dataset	
Resolution	1080, 720, 480, 450, 360, 288, 240
GOP size	25, 15, 12
Frame per sec	50, 30, 24
Motion speed	$1(rel.slow) \sim 5(rel.fast)$
Duration	9, 30, 60 secs
Diversity of content	17
Encoder	mpeg4, mpeg2, H.264
Container	avi, mp4, mp2, (m)ts, yuv
Number of Videos	2852 : 210 (LIVE) + 362 (UDPStream) + 2280 (UTrailers)

We choose a heterogeneous data corpus to cover the broad range of mobile videos including different distortions, sizes and content. The properties are summarized in Table I.

IV. Insights

We identify the main factors that affect the video acuity in mobile videos: spatial distortions, freezing (buffering) and video resolution. Spatial distortions are only present in UDP-based services such as one-to-many video service, video chat applications and live streaming services, while most popular internet video services use TCP-based connections and have

[1] According to ITU-R BT.500-11 subjective assessment standard [14], 15 subjects would be enough for subjective quality evaluation

TABLE II

Accuracy performance test of Existing metrics in the four scenarios : SVSD, SVMD, MVSD and MVMD

Type	TVM	TVI	Blck.	Blur.	BRIS.	NIQE
SVSD	0.7129	0.8752	0.8085	0.7695	0.7259	0.2659
SVMD	0.8475	0.9145	0.8930	0.8425	0.8752	0.6321
MVSD	0.1359	0.6305	0.1077	0.1381	0.0814	0.1136
MVMD	0.0571	0.7395	0.0974	0.0403	0.0113	0.0876

TABLE III

Proposed Model gives improvement in performance improvement when comparing videos on 'Per distortion' basis

Metrics / Dist. Type	R	S	T	W+A+B
WSNR	0.785	0.603	0.130	0.726
SNR	0.683	0.500	0.217	0.609
VSNR	0.812	0.583	0.324	0.790
VIF	0.867	0.629	0.077	0.876
MOVIE	0.819	0.742	0.157	0.802
TVI	0.806	0.699	0.305	0.768
SS-SSIM	0.667	0.517	0.237	0.660
MS-SSIM	0.734	0.637	0.310	0.738
NQM	0.822	0.545	0.242	0.852
UQI	0.498	0.269	0.279	0.656
VQM	0.712	0.540	0.340	0.763
Proposed	**0.926**	**0.817**	**0.471**	**0.934**

no spatial distortions. Only freezing and resolution-switching are experienced by the end users.

A. Insight on Distortions

Video connections over UDP are always vulnerable to packet loss and delay/jitter. These network impairments definitely cause video distortions like blocking, blurring, partial/whole blackout and color loss/change. However, existing video quality metrics are designed to focus on a single distortion. Therefore, they easily lose their accuracy when they are exposed to different display impairments that they are not designed to detect. By following experiments in the given conditions, we test accuracy of the recent representative spatial and temporal video metrics with various video distortion sets.

We configure four cases of video dataset depending on the video content and type(/degree) of video distortions. Single Video-Single Distortion (SVSD) is the case that different degrees of a single type of distortion to a single video content. In Single Video-Multiple Distortions (SVMD), we consider the case when multiple distortions are applied across the same video content. Multiple Videos-Single Distortion (MVSD) is the case that different degrees of a single distortion are applied across different videos. Multiple Videos- Multiple Distortions (MVMD) is for a scenario where both the distortions and the videos are variable.

The existing metrics work well with a single content or single video, but the performance degrades when we vary the content, shown in Table II. Temporal information-based metrics (TVM or TVI) gives relatively higher correlation than other metrics in most scenarios, but it is also not able to shows consistent accuracy in all cases.

$*$ *Machine Learning model:* Due to heavy computation and high complexity, most spatial-based metrics and temporal metrics based on motion-vector are not suitable for mobile devices. Temporal information-based metrics cannot tell specific type of video distortions, but it is suitable for capturing overall video distortions, as our experiments have shown in Table II. Thus, we use frame-based temporal information by difference of two consecutive frames [2] to capture temporal information, called as Ti.

$$\begin{aligned} d_f &= \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left(F_n(i,j)-F_{n-1}(i,j)\right)^2 \\ Ti(n) &= 10\sum_{i=1}^{fps}\log_{10}\left(\frac{c^2}{d_f(n)}\right) \end{aligned}$$

where resolution size is represented with HW, fps for frame per second and c is for color depth. The Ti_s values obtained from original video are compared with the Ti_d values of reference video. The difference of two Ti values indicates the video distortions, called as DTi. The DTi values are accumulated, reported and compared with DTi of original video per second. This difference will indicate the current degree of user video quality experience during the video service time. We call it $TIPS$ (Temporal information score Per Second).

$$\begin{aligned} DTi(t) &= \frac{|Ti_s(t)-Ti_d(t)|}{Ti_s(t)} \\ TIPS_i(t) &= \frac{1}{t}\sum_{t=1}^{T_{pt}}\{DTi_i(t)\} \\ DC(d_f) &= 0, \text{ if } d_f < 10^{-2} \end{aligned}$$

where T_{pt} for total playtime. No motion or freezing in a video yields a value 0, called as discontinuity (DC), in calculating differences (d_f) in two continuous frames. For example, if an original video has 5 DC and the stream video has 7 DC, the difference of 2 ($7-5$) DC indicates that freezing distortions has occurred due to network delay or packet-losses. DC enable to detect where and how long each freezing happens by simply counting it (more specified in section of freezing). Introduction of $TIPS$ and DC does not require any additional overheads or implementations in network layer like tracking sequence number and subtraction of timestamps in RTP packets of UDP stream and overheads for calculations of packet arrival/delay/recover times in TCP-based connections.

Feature Vectors: To model distortions using a machine learning model, we first select feature vectors:

- Ti_s : Ti of source video.
- Ti_d : Ti of destination video.
- DTi : Difference of DTi_s and DTi_d
- $Tips$: Ti Score per second
- DC : Number of discontinuity
- MOS : Subjective score

Feature selection was done using Information Gain [16] criterion and computational analysis, details of which are skipped for brevity.

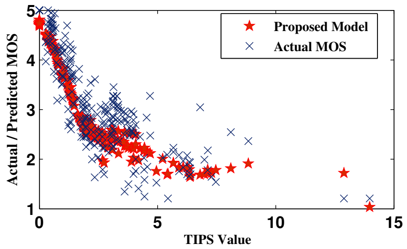


Fig. 1. Actual MOS and predicted values by Proposed Model

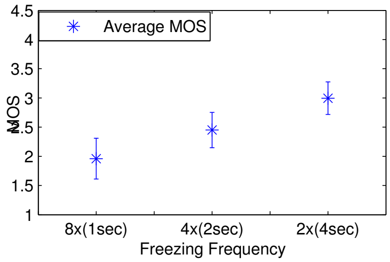


Fig. 2. Plot of MOS ratings for freezing events with different frequency

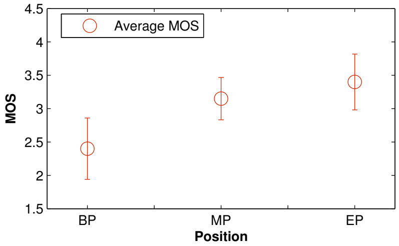


Fig. 3. Plot of MOS ratings for freezing events with different position

Regression Tree: Our model is built using Bootstraping aggregating (Bagging) [8], ensemble technique with reduced-error pruning decision tree as the underlying regression model to estimate MOS. The bagging ensemble technique is presented here because it was superior to models generated using other techniques (e.g., multilayer perceptron, support vector machines, linear regression, naive Bayes, reduced-error-pruning decision trees and nave Bayes) in terms of predictive accuracy. The bagging technique is an ensemble meta-algorithm to improve the stability and accuracy in statistical regression obtained by decision tree. The decision tree is based on Information-theoretic criterion for selecting the nodes. Once the tree is built, reduced error pruning is used, where each node, beginning with the leaves, is replaced with its most popular class. We divide the data for the model into n = 10 folds, where, n-1 folds are for supervised learning and one fold is used to test the model for errors. The errors obtained in a fold are added to the weights of nodes of next fold in the training set. Ten-fold cross validation was used to evaluate the model in order to ensure that the model was tested on data that it had not seen while training, to minimize chance for over-fitting. Statistical analysis was performed using Weka 3.6.10 and Matlab R2013a (Ver 8.1.0.604) software.

The proposed model is compared with different video quality metrics in each distortion categories, shown in Table III. The abbreviations are explained in the Table I. This analysis helps us to see that the proposed model has good expressive power in terms of capturing the effects caused by different distortions. Particularly, we observe a steep improvement in the case of (W+A+B) distortions where correlation shows over 93% accuracy. Since packet-loss happens normally in bursty ways, the lost video frames means missing information and finally this degrades accuracy of a metric. We obtain similar improvements on a 'per video' - SVSD or SVMD, but the results are omitted of brevity. Figure 1 shows the plot for 250 HD videos with predicted values of our proposed model and actual MOS values.

B. Insight on Freezing

The dominating HTTP-based internet videos like Netflix, Youtube, Hulu, etc, do not experience the spatial distortions such as blocking, blurring and color damages due to TCP recovery schemes for packet losses and errors. Freezing is the main distortion which can happen in the HTTP-based video content services. Prior studies only use total freezing time as a metric, not accounting for the type of freezing events. Burst or uniform freezing as well as position of freezing event (beginning or end of video) will impact acuity. This motivates us to introduce two new factors in freezing distortion - $frequency$ (how it happens) and $position$ (where it happens).

Frequency: The first factor we consider is frequency of freezing. For example - the same freezing event of total length 8 seconds can occur in two ways : 4 time×2sec-freezing or 2 time×4sec-freezing. In both cases, the fraction of video freezing (length of freezing / total length of video) is same. However, researchers have come across frequency effect in word recognition tasks [6] which shows that repetition skews the impression or perception of individual.

Position: The position of freezing can also be another factor that affects visual acuity. Freezing can result in differences on users' evaluation depending on the position where freezing happened. We gain this inspiration from psychological studies: 'primacy effect' - cognitive bias that primacy information presented in the beginning is likely to be recalled than information presented later on, or 'recency effect' - earlier information is easier to be recalled than later ones [12].

Setup / Observations: For $frequency$ experiments, different freezing effects are distributed over the entire video playback time with the same total freezing time (∼8sec, max 25%) but different combinations of intensity and frequency. We consider three scenarios here : A (8×, 1sec), B (4×, 2sec) and C (2×, 4sec).

In Figure 2, the results show that users are more sensitive and uncomfortable to $frequent-but-shorter$ freezing events rather than $less-but-longer$ freezing events. The interesting thing is that the visual acuity difference between [8×1sec, 2×4sec] is enough to change degree of user satisfaction. This supports our hypothesis that frequency tends to bring higher influence than intensity in visual degradation.

In the experiments for $position$ factor in freezing, we split each video into three granularities - BP (Beginning Position), EP (End Position) and MP (Middle Position, not in BP and EP). According to the plot diagram [1], first impression and recent impression in a video are normally decided first 20% and last 20% of video play timeline respectively. Thus, we set BP for the first 20% of video playtime, EP for last 20% and

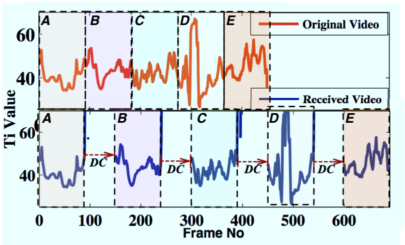


Fig. 4. Appearance and detection of DC at network delay events (2sec) happen 4 times

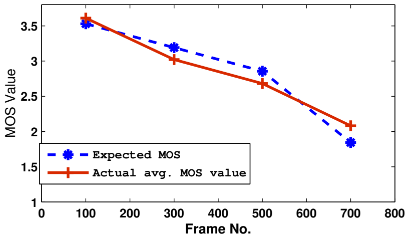


Fig. 5. Comparison of average actual MOS of users with expected MOS of TIPS scoring in each freezing distortions used in Figure 4; (Video - Oldwoman)

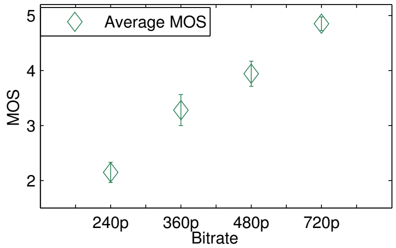


Fig. 6. Plot of actual MOS ratings for each resolutions (no resolution switches)

MP for the rest. A freezing effect is introduced in only one of the three positions. The total freezing duration is 3sec out of 30sec video. Figure 3 reports that freezing in start position (BP) results in worse visual acuity than later-introduced ones. This result is another example that supports the psychological study of primacy effect [12]. Abandonment (or leaving due to just changing-mind and content-disatisfaction) is not considered in this experiment.

∗ *Machine Learning model:* Freezing is detected and measured by DC. Each freezing time (T_f) and total freezing time (T_{tf}) can be derived as below:

$$\begin{aligned} T_f(i) &= \frac{|DC_s(d_{f_i}) - DC_d(d_{f_i})|}{\{fps(i)|\overline{fps}\}} \\ T_{tf} &= \sum_{i=1}^{T_{pt}} T_f(i) \end{aligned}$$

where DC_s is for the number of DC in a source video, DC_d for a destination video, $fps(i)$(dynamic) or $\overline{fps}$(fixed) for frame per second and T_{pt} for playtime. DC is measured and calculated every second so that it can track real timely user's video QoE depending on freezing effects. In Figure 4, the graph in the bottom (received video) has 4 DCs when network delays (2sec - 4 times) are given. The DC moves the graph as long as the freezing motions are displayed due to the given network delay conditions. This analysis enables to detect exactly where and how long freezing events happens. **Features:** From the above observations, we derive following features for a machine learning model.

- Total freezing time : $\bar{T_{tf}}$
- Each freezing time : $\bar{T_f}$
- Frequency : $\dot{F_f}$
- Position : BP($\vec{P_b}$), EP($\vec{P_e}$) and MP($\vec{P_m}$)
- Freezing ratio : $\tilde{F_r} \leftarrow \frac{T_f}{T_{pt}}$
- Subjective score : MOS

We combine all freezing datasets that are configured of different combinations of freezing time [1s 1.5s 2s 3s 4s 8s], frequency [1, 2, 4, 8 time(s)] and position [BP, MP, EP]. Total 643 subjective ratings were used to build our modeling.

Linear Regression : Following freezing model can be obtained as output of the linear regression. The results are also based on 10-fold cross validation. Let $\widehat{F_{AC}}$ give the estimated value of acuity with freezing effects, using the following regression modeling:

TABLE IV
ACCURACY OF FREEZING EFFECT MODELING : REALTIME-TRACKING USER QOE IN FREEZING EVENTS

Realtime-Tracking predicted MOS in Freezing	
Correlation Coefficient	**0.8175**
Mean absolute error	0.2565

$$\widehat{F_{AC}} = \alpha \cdot \dot{F_f} + \beta \cdot \bar{T_f} + \gamma \cdot \vec{P_b} + \delta \cdot \vec{P_e} + \eta \cdot \tilde{F_r} + \lambda \qquad (1)$$

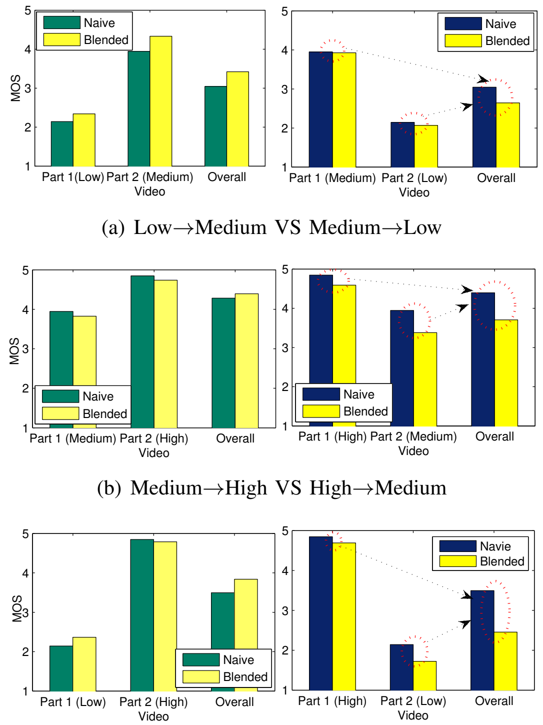
for constants α(-0.2333), β(0.0598), γ(-0.8636), δ(0.1897) η(1.5559) and λ(3.0551), showing that the position of MP is irrelevant of freezing effect. We also test non-linear models such as decision tree and naive Bayes, but they do not give recognizable differences in performance. So, we choose a linear regression technique. Table IV reports the freezing model gives us over 81% accuracy in estimating visual acuity. Figure 5 shows how users QoE degrades as network delay conditions of Figure 4 are given. Our model enables to track how users QoE reach to the final MOS value in the middle of service, and this is validated by actual values obtained users at each freezing event.

C. Insight on Resolution

Internet video offers multiple resolutions for each video content depending on the network conditions between service provider and end users. Resolution may impact visual acuity. Previous work [5] has shown non-monotonic relationship between average bitrate and user engagement. Bitrate is usually proportioned to video resolution, however, many factors such as switches in resolution and video popularity, etc may affect user engagement. In our work, we focus on video acuity and conduct separate experiments to study the impact of (a) higher resolution of video acuity and (b) online-resolution-switches in video acuity. Thus, it is the first effort to study the impact of resolution-change on user's video watching experience.

Resolution switch: Assessment based on averaging resolutions can not say how resolutions have been changed during video playback time. The traditional method of 'post-view' is not suitable to predict real-timely visual acuity during playback. This leads to missing interdependency/interaction between resolution-switching.

For example, let us consider the following experiments where we have a video played with resolution x for the



(a) Low→Medium VS Medium→Low

(b) Medium→High VS High→Medium

(c) Low→High VS High→Low

Fig. 7. Compare User QoE in opposite cases of resolution-switches

first 15 seconds and resolution y for next 15 seconds. In the given condition, we figure out how users react on resolution-switching during video playback time. We select three types of resolution mainly serviced in internet videos for mobile devices - L (Low, 240p), M (Medium, 480p) and H (High, 720p). The applicants were requested to evaluate three MOS values during playback time in the videos containing two different resolutions. The first score is for the first part, the second is for next part of video and last score is for overall video quality (not average of the first and second one). In order to prevent changing mind or abandonment due to video content itself, applicants were requested not to measure videos which are out of their interests. Applicants in the experiments experience and evaluate 5 types of resolution video clips respectively. The resolution 1080p is excluded in this experiments because there are rare number of phones which support screen size to display 1080p, and users do not recognize between 720p and 1080p in mobile display.

In the case of no resolution-switches from start and end, it is natural for users to have more satisfactions in higher bitrates when no bitrate-switchings happen, shown in Figure 6. We observed an increasing trend between MOS values and screen resolution in the Figure 6. However, following observations show that users' satisfaction vary considerably depending on how resolution have changed during the playtime.

Obervations Our results are plotted in Figure 7. Naive refers to the MOS score by the users when there was no resolution switch. For example - in Figure 7 (a) (left), part 1 of video clips, Naive refers to MOS score when a full L video was played for audience, and part 2 of video clips, Naive refer to MOS score when a full M video was played for audience. One may expect that a video with resolution-switch from ($L \rightarrow M$) will have MOS score which is average of the two. This is depicted as Naive in 'overall' legend. In another set of experiments, the users gave us three MOS scores, one for first half, one for second half and one for overall video. The score for these blended videos are also shown in Figure 7 for all cases.

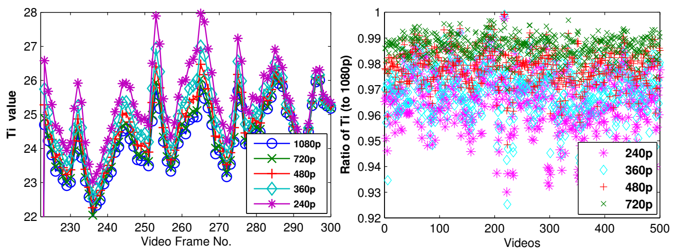


Fig. 8. (Left) Ti values and (Right) Ti ratio depending on each resolution; Need for Speed 2014

TABLE V
CORRELATION BETWEEN USER MOS AND PROPOSED MODEL IN RESOLUTION-CHANGES

Realtime-Tracking predicted MOS in Resolution-switch	
Corr. Coef.	0.871
Mean Abs.Err.	0.399

We make some key observations. Unlike the cases of freezing, where we observed primacy effect, here we experience that users tend to retain the last experience while watching videos. In all the cases, increasing the resolution in second half was helpful to boost MOS value, while a decrease in resolution in second half leads to rapid degradation in MOS than expected average value (Naive).

$*$ *Machine Learning model:* **Features** From the above observations, we derive following features for a machine learning model :

- $R\{r_1|r_2\}$: set of two resolutions
- $O_r\{r_1 \rightarrow r_2\}$: order of two resolutions
- $T\{r_1|r_2\}$: played-time at each resolution
- Ti and DTi at $R\{r_1|r_2\}$: temporal information of each resolution
- MOS : overall and instant subjective scores

For real application of a model in real world, it is significant to detect resolution resolution-switch during playback time. Figure 8 (a) draws Ti values of each resolution frame by frame in the video (Need for Speed 2014), and (b) shows distribution of Ti ratio values of videos. Ti values show the consistent distance between resolutions, shown in Figure 8 (a). The small distances between resolutions become clearer when it is presented by the proportion of each resolution to 1080p(max quality), shown in Figure 8 (b), which enable by using classification algorithm to detect which resolution is in service during the playback time.

Modeling We apply bagging regression tree to find out accurate modeling for predicting better visual acuity in

TABLE VI
PERFORMANCE COMPARISON OF PROPOSED MODEL ACROSS ALL VIDEOS WITH ALL DISTORTIONS (GLOBAL; R+S+T+W+A+B)

Model	WSNR	SNR	VSNR	VIF	MOVIE	TVI	SS-SSIM	MS-SSIM	NQM	UQI	VQM	Proposed
Corr.Coef.	0.649	0.537	0.702	0.421	0.749	0.739	0.577	0.665	0.740	0.494	0.659	**0.892**
RMSE	0.677	0.751	0.628	0.562	0.589	0.571	0.719	0.662	0.597	0.751	0.682	**0.412**

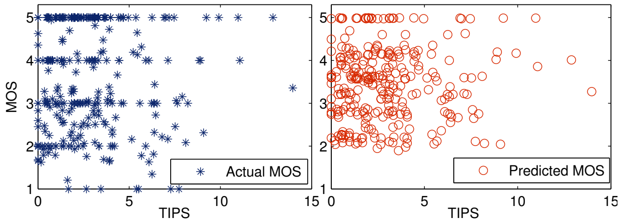


Fig. 9. Actual and Predicted MOS values with TIPS

resolution-switching by the same way of our distortion modeling. Table V reports that the proposed model for resolution shows over 87% accuracy. Naive Bayes model does not show better performance than decisionTree-based algorithm because they do not find interdependent information between inputs. It is also not suitable for linear regression to capture non-lineary properties.

V. MOBILE VIDEO ACUITY

A real practical scenario in mobile video services necessitates a robust and feasible model that can measure concurrent complex distortions. We propose a comprehensive model with the features derived by each model in the previous sections.

A. Acuity Modeling

In this section, we combine all factors causing various visual impairments with newly derived features in order to make a single framework that enables us to measure video acuity in real-time. The vectors (V) from each model are denoted as below: (duplicate feature vectors removed)

- Distortion (V) ← $\{Ti_s, Ti_d, DTi, Tips, DC_s, DC_d\}$
- Freezing (V) ← $\{\widehat{F_{AC}}, \bar{T}_{tf}, \bar{T}_f, \dot{F}_f, \vec{P}_f$ and $\vec{P}_r\}$
- Res. (V) ← $\{R\{r_1|r_2\}, O_r\{r_1 \rightarrow r_2\}, T\{r_1|r_2\}\}$
- Format (V) ← {color depth (c), $\{fps(i)|\overline{fps}\}$, $T_{pt}\}$
- Subjective Score ←{MOS}

All types of distortions can happens discretely or simultaneously in practical scenario. It is required for a framework to show the robustness in predicting user experience when exposed to diverse distortions. Thus, we configure all combinations of datasets which are represented and specified by all factors mentioned in the each modeling. First, we consider all combined spatial distortions with multiple videos : MVMD $(R + S + T + W$ and $A + B)$ except events of freezing and resolution changes. Table VI shows comparison of the accuracy performances of video metrics and that our proposed model shows more robust accuracy than the other representative metrics (based on FR/ RR/ NR). Second, we compare whole sets of distortions including freezing and resolution-swtich. Since the other video quality metrics are not able to capture freezing and resolution changes, we compare our model built on bagging regression tree with other decision tree algorithm and regression models such as Multi-Layer-Perceptron (MLP), SMO, etc. We can identify how our feature vectors behaves in different modelings and the accuracy of the selected modeling. Table VII reports the accuracy of compared models, and the proposed model is dominating with 78% correlation which is higher than other schemes. Decision-based algorithm, DS (decision stump) shows lower accuracy than Linear Regression (LR) because simple decision-based rules are not able to capture priorities obtained from interdependent relationships of input factors. LR also could not show high performance in predicting MOS when combining all features. Figure 9 draws the example of actual and the predicted acuity by our proposed model.

TABLE VII
COMPARING MODELINGS WITH ALL COMBINED-FACTORS OF DISTORTIONS

Model	Proposed	MLP	LR	DecisionStump
CorrCoef	**0.7877***	0.6929	0.5905	0.4476
Mean Abs.Err.	0.5631	0.6738	0.8231	0.899

MLP (MultiLayer Perceptron) LR (Linear Regression)

B. Acuity Application

In this work, we have proposed measuring visual acuity (due to distortions, freezings and screen resolution). These observations help network/ service providers to learn the impact of their streaming strategy (combined with wireless channel effects) on the visual acuity to end user, and make necessary changes. However, it is important that such computations are feasible on the end devices.

We implement a middleware of $TIPS$-scoring in a server side and application of Ti and DC in a client side, shown in the Figure 10. The testbed is configured with a single wireless hop between a content server and a mobile device. WLAN configuration for the testbed is IEEE 802.11n 2.4GHz and minimum end-to-end delay is 0.604 ms. FFmpeg [7] is used for coding/decoding and videoLAN [23] is used for video streaming connection. A Free BSD-based iMAC server (8GB, 3.2GHz Intel Core i3) was used for streaming to iPhone 4 (1 GHz Cortex-A8) and iPhone 5s mobile device.

We test different types of video content to calculate $TIPS$ computational time in this environment. We configure and run the testbed under the scenario detailed in Figure 10. (1) The implemented application (app) in iPhone 4/5 obtains video information (fps, bitrates, etc) from the server when a video streaming session is established. (2) With the frame-per-second (fps) offered by the sever, the app can decide the capture-rate to obtain video frame information loaded in memory. (3) The

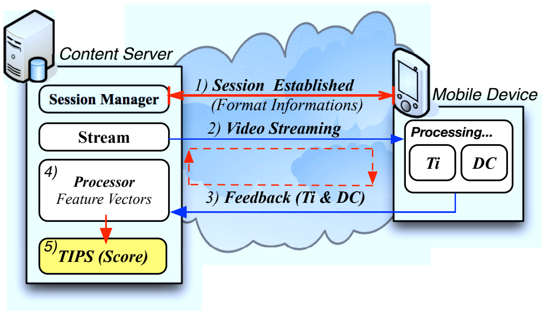


Fig. 10. Scenario and Testbed configuration

source and destination Ti and DC values are calculated and sent to a server. (2-3) steps are repeatedly processed during the playtime. (4) The server calculates feature vectors to be applied in the proposed model. (5) Finally, it reports $TIPS$ score and predicted user QoE per second. The computation time and memory usage includes loading the received video frame to memory and pixel-wise computations to obtain Ti and DC values. Table VIII reports computation time and memory usage in iPhone 4 and 5s. The computational complexity and memory usage are directly proportional to the resolution of video. However, even a full HD video (1920×1080 pixels) can be processed in real time. It can be seen that our app runs on iPhone4 conveniently in parallel with video streaming application and can provide real-time feedback to the remote streaming server.

TABLE VIII
COMPUTATION TIME FOR PROCESSING ONE SECOND OF VIDEO AT 30 FPS AND MEMORY FOR RUNNING MVM IN MOBILE DEVICE (IPHONE 4 AND 5S)

Resolution	Proc. Time(sec)		Memory
	iPhone 4	iPhone 5s	Usage(Kb)
352*288	0.114	0.0789	2148
960*640	0.339	0.187	4101
1920*1080	0.592	0.386	6738

VI. CONCLUSIONS

In this work, we proposed a framework for end-user side measurements which allows us to detect visual acuity. We analyzed different mobile video distortions and extracted important features. From the observations, we showed that user QoE varies depending on where and how long freezing happens and how resolutions switch. We have provided features on how to detect freezing and resolution changes without additional implementations or overheads, which enable to detect realtimely display conditions to track the variation of user QoE. We showed the feasibility of such computation on mobile devices and proposed a bagged regression tree based model for obtaining visual acuity.

REFERENCES

[1] Elements of fiction. http://learn.lexiconic.net/elementsoffiction.htm.

[2] *Subjective video quality assessment methods for multimedia applications*, volume ITU-T P.910. 4 2008.

[3] I. Assembly. *Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union, 2003.

[4] D. Bailey, M. Carli, M. Farias, and S. Mitra. Quality assessment for block-based compressed images and videos with regard to blockiness artifacts. In *Tyrrhenian International Workshop on Digital Communications, Capri, Italy*, 2002.

[5] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. In *Proceedings of the ACM SIGCOMM 2013*, pages 339–350. ACM, 2013.

[6] D. A. Balota and D. H. Spieler. Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128(1):32, 1999.

[7] F. Bellard, M. Niedermayer, et al. Ffmpeg. *Internet: http://www. ffmpeg. org,[Dec. 27, 2012]*, 2007.

[8] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[9] M. Carnec, P. Le Callet, and D. Barba. An image quality assessment method based on perception of structural information. In *IEEE International Conference on Image Processing, 2003.*, volume 3, pages III–185, 2003.

[10] K. Chono, Y.-C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod. Reduced-reference image quality assessment using distributed source coding. In *IEEE International Conference on Multimedia and Expo, 2008*, pages 609–612, 2008.

[11] CiscoSystem. Cisco visual networking index: Global mobile data traffic forecast update, 2013 - 2018, 2011.

[12] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Number 3. Teachers college, Columbia university, 1913.

[13] F. Group. Freebsd ipfirewall, 2012.

[14] ITU. *Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union: Recommendation ITU-R BT.500-11, 2002.

[15] A. Khan, L. Sun, and E. Ifeachor. QoE prediction model and its application in video quality adaptation over UMTS networks. *IEEE Transactions on Multimedia,*, 14(2):431–442, 2012.

[16] S. Kullback. *Information theory and statistics*. Courier Dover Publications, 2012.

[17] W. Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[18] X. Marichal, W.-Y. Ma, and H. Zhang. Blur determination in the compressed domain using DCT information. In *IEEE Proceedings International Conference on Image Processing*, volume 2, pages 386–390, 1999.

[19] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing,*, 21(12):4695–4708, 2012.

[20] A. Mittal, R. Soundararajan, and A. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2012.

[21] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing,*, 6(6):652–671, 2012.

[22] A. Rehman and Z. Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on Image Processing,*, 21(8):3378–3389, 2012.

[23] V. S. Solutions. VLC media player, 2006.

[24] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters,*, 9(3):81–84, 2002.

[25] S. Wolf and M. H. Pinson. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. In *International Society for Optics and Photonics*, pages 266–277, 1999.

[26] A. K. Wörner. A real time single ended algorithm for objective quality monitoring of compressed video signals. In *Conferences of Society of Motion Picture and Television Engineers*, pages 1–8, 2002.