Towards a comprehensive model for predicting the quality of individual visual experience

Yi Zhu^{la}, Ingrid Heynderickx^b, Alan Hanjalic^a, Judith A. Redi^a a Delft University of Technology, Mekelweg 4, Delft, The Netherlands, 2628 CD ^bEindhoven University of Technology, Postbus 513, Eindhoven, The Netherlands, 5600 MB

ABSTRACT

Recently, a lot of effort has been devoted to estimating the *Quality of Visual Experience* (QoVE) in order to optimize video delivery to the user. For many decades, existing objective metrics mainly focused on estimating the perceived quality of a video, i.e., the extent to which artifacts due to e.g. compression disrupt the appearance of the video. Other aspects of the visual experience, such as enjoyment of the video content, were, however, neglected. In addition, typically *Mean Opinion Scores* were targeted, deeming the prediction of *individual* quality preferences too hard of a problem. In this paper, we propose a paradigm shift, and evaluate the opportunity of predicting individual QoVE preferences, in terms of video enjoyment as well as perceived quality. To do so, we explore the potential of features of different nature to be predictive for a user's specific experience with a video. We consider thus not only features related to the perceptual characteristics of a video, but also to its affective content. Furthermore, we also integrate in our framework the information about the user and use context. The results show that effective feature combinations can be identified to estimate the QoVE from the perspective of both the enjoyment and perceived quality.

Keywords: Quality of Viewing Experience, Bitrate, Perceptual Quality, Enjoyment Prediction

1. INTRODUCTION

With the explosion of streamed video data, maintaining a satisfactory quality in video delivery is challenging for multimedia providers. Due to technological limitations, visual artifacts such as blockiness, blur or packet loss can occur at any stage of the video delivery chain [1, 2]. This severely degrades the user's satisfaction, and users are known to be less willing to pay for (or even to quit) the service if video quality cannot meet their expectations [3]. As a result, video service providers are eager to implement quality control mechanisms to steer improvements when needed. To do so, it is useful to have an objective model that can automatically estimate the satisfaction of the user at any point of the video delivery chain.

User satisfaction can be quantified by estimating the *Quality of Visual Experience* (QoVE), i.e., the level of delight or annoyance of a user with a video application or service [4]. Up to now, objective QoVE estimation mostly targeted the prediction of the perceived quality of a video. That is, satisfaction has been considered to depend fully on the visibility and annoyance of visual artifacts [6]. Assessing the visibility of artifacts by means of e.g. models of the Human Visual System has been therefore the dominant approach in the objective QoVE assessment, with some remarkable results [5]. Since recently, though, QoVE is increasingly considered as a multifaceted concept, not necessarily limited to the perceived quality of a video, but also including e.g. enjoyment and endurability aspects [4,8]. Enjoyment of a video experience, specifically, has been shown in [8] to be only moderately correlated with perceived quality. Therefore, to be able to assess the QoVE in a more reliable way, video quality metrics also need to include the information related to factors other than the pure visibility of video artifacts.

The first efforts towards broadening the scope of QoVE identified three classes of factors (i.e., system, human and context) to potentially influence QoVE [4]. System factors relate to technical properties of a video service (e.g., bitrate, bandwidth, or genre), and they can typically be inferred from the video itself or from the knowledge of the delivery system properties. Human factors refer to the individual characteristics of a user (e.g., age, gender, interest or

Human Vision and Electronic Imaging XX, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Huib de Ridder, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9394, 93940A · © 2015 SPIE-IS&T CCC code: 0277-786X/15/\$18 · doi: 10.1117/12.2085002

 ¹ x^3 *Y.Zhu-1@tudelft.nl; phone: + 31 (0)15 2782549

personality), and as such are independent from the video itself. This is the case also for the contextual factors, which refer to the environment in which the video is experienced (e.g., presence of co-viewers, environmental noise, lighting conditions etc.). The latter two are typically not taken into account by objective quality metrics. We mention here further also some later studies supporting the hypothesis that QoVE is significantly influenced by affective aspects [7] and social context [8].

In addition to the above, quality assessment mostly targeted the prediction of perceived quality as experienced by an average user. It, therefore, became common practice in visual quality research to calculate the average over the opinion scores gathered from subjective tests (also known as Mean Opinion Score, MOS) [9] to quantify perceived video quality. This MOS indicator was widely accepted as the measure used for the comparison of the performance of different objective quality metrics [10], such as MOVIE [11], VQM [12] or VSSIM [13]. None of these metrics, however, are able to take into account the dependency of QoVE on user-specific factors, such as expectations, memory effects or current mood [14]. In fact, prediction of individual experiences has often been considered too difficult of a problem, also due to the difficulty of gathering user-specific information.

In summary, the existing objective QoVE metrics mainly have two disadvantages: 1) they fail to capture dimensions of QoVE other than perceived quality [8], and 2) they ignore individual uniqueness, by targeting only Mean Opinion Scores. To the best of our knowledge, no quality metric exists that overcomes both disadvantages. Nevertheless, more and more knowledge is being generated on the multiple facets of QoVE and the factors that influence it [4, 15]. Furthermore, through social networks and social media, it is becoming easier to estimate user factors, such as personality and preferences [16, 17]. Thus, the time has come to start approaching the prediction of individual preferences of QoVE.

In this paper, we explore opportunities for predicting the QoVE for an individual user, making use of information not only about the video, but also about the user who is judging the video and about the context in which the video experience happens. We set the first steps towards predicting two aspects of QoVE, i.e., the well-known perceived quality and video enjoyment. We evaluate the added value of a number of features describing the perceptual content of the video as well as the characteristics of the user and context, and assess their contribution to the goodness of our prediction. Finally, we formulate recommendations to setup further work on fitting QoE prediction to an individual user.

2. PROPOSED FRAMEWORK

Our proposed QoVE prediction framework is presented in Figure 1. It is based on the concepts developed in the Qualinet White Paper [4] and further deepened in [15]. QoVE depends on video characteristics, but also on the context around the video experience [4] and on the specific characteristics of the user him/herself. As a consequence, we propose an overall prediction framework that processes not only video features, but also information on context of experience and user characteristics, eventually reaching a better prediction of QoVE. All features feed a prediction module, which is in charge of combining them into an estimate of one or more aspects of QoVE. In this work, we consider the *Perceived Quality* of the video as well as the *Enjoyment* during the video watching as targets for the prediction. In practice, though, also other aspects such as endurability [8], or willingness to pay [3] could be set as targets. The prediction module can be implemented in multiple ways, starting with a linear combination of the input features (as attempted in this paper), to more complex learning methods. Depending on the learning methods, which may be more or less prone to the curse of dimensionality (and therefore, to overfitting), it may be wise to envision a feature selection step before proceeding with the prediction [18]. Furthermore, some features may be more beneficial to the prediction of the specific QoVE aspect than others. Thus, to make our framework as general as possible, we include this step in our overall framework.

In the following, we further detail the type of features we include in this preliminary study. The list of features we consider is not to be intended as exhaustive, and we acknowledge that more appropriate system, context and user descriptors may be useful for future models. Nevertheless, as a proof of concept of the feasibility of our approach, we make some initial choices on what could be potentially relevant information to be processed by a framework aimed at predicting an individual's QoVE.

Figure 1: Overview of our proposed framework for individual QoVE prediction. Video, Context and User information is evaluated to predict QoVE per user whose characteristics are given as input.

2.1 Detailed overview of the features of the experience

An overview of all extracted features is given in Table 1. To describe the perceived video quality, we use the information related to *perceptual characteristics* of artifacts. Perceived quality has been extensively investigated in the past and there is no doubt that the perceptual characteristics of a video have a great impact on QoVE. Features listed under *affective content* describe the intensity of emotion expected to be elicited in the user when watching the video, and are expected to inform the assessment related to enjoyment.

Regarding the *context* factors, there is so far very limited knowledge on which contextual elements impact QoVE. It is known, for example, that users evaluating perceived quality are more critical when doing so in a controlled lab condition [19]. With respect to enjoyment, the knowledge on context factors is even more scattered. In our previous study [8], though, we were able to prove that social context (i.e., the presence/absence of co-viewers) influences different aspects of QoVE. Specifically, the presence of co-viewers increases the video enjoyment. Thus, in our framework, we encompass this information too, expressed as a binary value, where '1' indicates that the viewing experience happened in a group situation, whereas '0' indicates that the user watched the video alone.

Finally, we also incorporate features in our framework that are related to *user characteristics* and that, in addition to the use context described above, serve to tune the QoVE even further towards the individual viewing situation of the user.

Table 1: An overview of all features used in this study. The number of features are given between brackets.

In total, 68 features were included in this study: 44 describing perceived quality, 4 describing the video affective charge, one context feature and 19 describing user characteristics. These features were all normalized between 0 and 1 and are explained in more detail in the remainder of this section.

2.1.1 Perceptual Characteristics.

For this preliminary experiment, we used features from a well-known no-reference quality metric [20,21] to describe the perceptual characteristics of the video. We chose to use off-the-shelf features rather than designing new ones because the focus of this work was to unveil the role of context and user features in predicting QoVE, rather than to improve existing descriptors of perceived quality. Thus, we picked readily available features which could predict video quality without the reference signal [20].

We included 44 features related to the perceptual characteristics (represented as *PQ* features) in our study. Thirty-six of them, based on the spatial domain natural scene statistics (NSS) model [21], were calculated at each frame k ($k = 1$, ...K). We represent them here as $PQ_n(k)$, where $n = 1,2,...36$. These features were computed as follows. First, each frame *k* was partitioned into squared patches of $n \times n$ pixels. The sharpness of each patch - as specified in [21], was computed and only those patches whose sharpness value was higher than a certain threshold (i.e., 0.75 in this work) were selected. Within the selected patches at frame *k*, intensity values were transformed by applying mean removal and divisive normalization. The transformed values were then used to fit a Generalized Gaussian Distribution (GGD), which is known to represent the distribution of such values in unimpaired images:

$$
f_k(\mathbf{x}; \alpha, \beta) = \frac{\alpha}{2\beta \Gamma(\frac{1}{\alpha})} \exp\left(-\left(\frac{|\mathbf{x}|}{\beta}\right)^{\alpha}\right), \quad \text{with} \quad \Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt \quad a > 0 \tag{1}
$$

To evaluate the impact of distortions at each frame, the products of adjacent transformed values (in the horizontal, vertical, and diagonal orientations) were used to fit an Asymmetric Generalized Gaussian Distribution, described by the parameters γ, $β_l$, $β_r$:

$$
f(x; \gamma, \beta_l, \beta_r) = \begin{cases} \frac{\gamma}{(\beta_l + \beta_r) \Gamma(\frac{1}{\gamma})} \exp\left(-\left(\frac{-x}{\beta_l}\right)^{\gamma}\right) & \forall x \le 0\\ \frac{\gamma}{(\beta_l + \beta_r) \Gamma(\frac{1}{\gamma})} \exp\left(-\left(\frac{-x}{\beta_r}\right)^{\gamma}\right) & \forall x \ge 0 \end{cases}
$$
(2)

Finally, the mean of the distribution was computed as well:

$$
\eta = (\beta_r - \beta_l) \frac{r(\frac{2}{\gamma})}{r(\frac{1}{\gamma})}
$$
\n(3)

The parameters α , β , γ , β_l , β_r , η are known to capture the differences between lossless and distorted images [21]. In order to capture multi-scale behavior, this set of 18 parameters (γ, $β_l$, $β_r$, $η$ being calculated for 4 orientations, plus , $β$) was computed for two patch sizes (i.e., 96x96 and 48x48 in this case). Thus, eventually, 36 NSS features were computed for each frame. In our study, we used the average of these features across all K video frames as input. A $37th$ feature, being the temporal variation across frames of the frames mean DC coefficients, representing sudden local changes which may arise from various temporal distortions in the video, was added. Six statistical DCT features (represented as *PQ38* to *PQ43*) were also computed from each frame difference using the spatio-temporal model described in [11]. These features were suggested to reflect perceptual differences between distorted and pristine videos. Finally, feature *PQ44* relates to motion estimation, and was computed based on the coherence in strength and direction of motion vectors (within windows of predefined width m , in this paper set to 10). Specifically, the motion coherence feature C_s as proposed in [20] was computed at each frame and then averaged across all frames. This feature denotes variations in local motion due to temporal distortions [20].

2.1.2 Affective content.

As shown in Table 1, four features, being *motion activity, sound energy, hue ratio* and *shot cuts* (all based on [22]) were extracted from the video to describe the intensity of the affective response elicited in the viewer. Also here, we relied on existing work, as the design of features describing a video's affective charge was not the goal of this paper.

The *Motion activity* feature *M* was estimated from the overall magnitude of all (*N*) motion vectors \vec{v}_k computed between two adjacent frames *k* and $k+1$, normalized by the maximum length of \vec{v}_k . Such value, representing the motion activity at frame *k*, was then averaged across all *K* frames to yield the Motion activity value:

$$
M_k = \frac{1}{\kappa} \sum_{k=1}^{K} \frac{100}{N \left| \text{MAX}_{\overrightarrow{V_k}} \right|} \left(\sum_{i=1}^{N} \left| \frac{\rightarrow}{v_i} (k) \right| \right) \tag{4}
$$

Prof. of SPIE-IS&T Vol. 9394 93940A-4

Such feature is suggested to be able to evoke individual emotional response and is extensively investigated in affective content analysis research [22, 23].

Evidence has shown that also color features extracted from a video can elicit emotions in viewers [23]. We therefore include *Hue Ratio* as a feature*,* computed as:

$$
H = \sum_{k=1}^{K} \frac{G_k}{N_k} / K \tag{5}
$$

where N_k is the total number of pixels in frame k while G_k is the number of green pixels in frame k, K is the total number of frames in the video [24].

The *shot cut* rate represents the proportion of sudden shot changes in a video, and is considered to have an impact on video arousal [22]. The shot cut rate *S* in a video is defined as

$$
S = \frac{\sum_{k=1}^{K} 100e^{\lambda}((1 - (n(k) - p(k)))/\delta)\%}{K}
$$
(6)

where $n(k)$ and $p(k)$ are the frame indexes of the two closest shot boundaries to the left and right of frame *k* and δ is a constant value, set as recommended in [22]. Finally, we extracted the feature *Sound energy* [22] from the audio track of the videos. The sound energy E_k at frame k is defined as the sum of the power spectrum of the A_k audio samples corresponding to frame k . The overall sound energy feature E is then computed as the mean of E_k across all frames. This energy feature is commonly used to represent video arousal [22].

2.1.3 User Characteristics

In principle, to feed a fully automatic model of QoVE, it would be desirable to measure user characteristics automatically e.g., by crawling data from social media. For this preliminary study, which is intended to be a proof of concept that user characteristics are essential to the prediction of QoVE, we made use of data provided directly from the users that judged the video experiences. Specifically, we use user information collected during a previous experiment [8], during which users evaluated QoVE of videos from different genres and in presence/absence of co-viewers. An overview of the experimental details is given in Section 3.1; below we specify which type of user information we collected and how.

To describe the user judging the visual experience, we made use of 19 features, summarized in table 2. *Personal interest* in different video content is proved to influence user's QoVE judgment [25]. Moreover, it is shown that users tend to rate a video with the same bitrate as higher in QoVE when they are more interested in the content of the video [26]. Thus, we included in our feature set values quantifying the level of prior interest into the (genre of) the video that the user was about to watch and judge. In our specific case, we quantified interest on a 7-point likert scale (with 7 being the highest possible level of interest in the video content). Similarly, we included values that quantified the user's *Immersive Tendency* (IT), which reflects how easily a user gets involved in a particular task [27]. Evidence shows that a high level of involvement may result in high satisfaction [28]. In our experiments, we quantified immersive tendency via the immersive tendency questionnaire [27], returning an IT score on a scale ranging from 18 to 126 (the higher the score, the higher the immersive tendency).

Personality is another important aspect of user characteristics, and is claimed to potentially influence QoVE [4]. Evidence shows that personality significantly influences user's performance on different tasks [29]. Therefore, we included a set of fifteen personality features in our study. These features were used to represent the "the big five" personality traits i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism [30]. These personality traits were quantified via the 10-items TIPI questionnaire, where each item was assessed on a 7-point likert scale, and each trait was measured by a pair of opposite items [31]. Participants were requested to specify to which extent a specific item applied to them. For instance, the trait "*Agreeableness*" was measured by a positive item *Sympathetic & Warm* and a negative item *Critical & Quarrelsome*. The inversed score from the negative item and the score from the positive item were finally summed to represent the corresponding personality trait. The ten items and five generated personality traits were all considered as user features in this study.

Finally, we included in our feature set some descriptors of a user's demographics i.e., the user's *gender* and *cultural background.* Gender has been shown to have an impact on emotional experiences as well as visual perception [15]. The cultural background was defined based on the user's nationality. Users with different nationality may have a different understanding of experience, and thus may perform differently towards the same task [32]. For example, we found that Asian users rated their QoVE higher than Western users in our previous experiments [8]. Thus, we included this trait as

well, discretized into a binary item, discriminating only between Asian and Western participants to represent their cultural background.

User characteristic	Type	Possible values
Interest in sports/comedy/education videos	Discrete (likert scale)	1 (min) to 7 (max)
Gender (Sex)	Binary	Male/Female
Cultural background	Binary	Asian/Western
Personality	Discrete (likert scale)	1 (min) to 14 (max)
Immersive tendency	Discrete (likert scale)	18 (min) to 126 (max)

Table 2: An overview of features related to user characteristics

3. DATASET USED

As mentioned above, for proving the concept of using video, context and user information to predict an individual's QoVE, we resort to an empirical dataset from a previous study [8]. In that study, we used six different videos covering three genres (i.e., comedy, sports, education) and these videos were further encoded at two bitrate levels (i.e., 600kbps and 2000kbps) to enforce different perceived quality levels. Participants were split into two disjoint groups: the 30 participants of one group watched all videos by themselves, whereas the 30 participants in the second group watched all videos with two friends (so, in groups of three; interaction among them was allowed). In total, sixty participants were involved in the experiment. Before starting the experiment, participants were asked to fill in some questionnaires investigating personal information, such as the level of interest they had (a priori) in the video genres they were about to see, their IT, and their personality. After filling in the personal data questionnaire, participants watched one of the two versions (i.e., 600 or 2000 kbps) of each video, for all videos. Bitrate levels as well as video order were counterbalanced across participants to guarantee that each participant would witness a similar range of perceived quality, and to avoid fatigue and learning effects.

For each video, participants then scored the QoVE through a questionnaire which consisted of 18 items [8], addressing enjoyment, endurability, satisfaction, involvement and perceived visual quality. Eventually, after discarding the data of one unreliable participant, we obtained 354 individual scores (59 participants * 6 videos) for each QoVE aspect. All aspects (excluding Perceived Quality) were quantified by means of 4 questions to be answered on a 7-point likert scale. The higher the score, the higher the appreciation for the experience. Since the statistical analysis performed in [8] indicated high inter-item consistency, the answers to these questions were summed into an overall score finally ranging from 4 to 28. To quantify Perceived Quality, we used the scores given on a 5-point ACR scale (i.e., from poor to excellent [9]), answering the question "What did you think about the video quality?".

The distributions of user ratings for Perceived Quality (representing opinion score, OS) and Enjoyment as obtained in our previous study are reported in Figure 2. In general, most Perceived Quality ratings were above 4, in fact only in 63 cases the Perceived Quality of a video was rated lower than 4. As mentioned earlier, only two bitrate levels were used in our previous study; 600kbps of bitrate is considered as a reasonable video quality level that users are willing to pay for [33], while 2000kbps of bitrate is considered as the entry level of HD quality. As a result, participants from our previous study thought that the quality of the videos was reasonable in most cases. The Enjoyment ratings, on the other hand, were distributed on the whole scale (from 4 to 28), suggesting that participants had different opinions regarding the enjoyment of the video experiences.

Figure 2: The distribution in ratings for Perceived Quality (OS) and Enjoyment as obtained from our previous study [8]

4. VALIDATION OF THE QOVE PREDICTION FRAMEWORK

In evaluating our proposed framework, we were are aiming at delivering a fully functional model for QoVE prediction, but rather at proving that including context and user information in QoVE prediction is beneficial. In the following, we describe the experimental setup that we used to evaluate our approach, and then test it to answer the following questions:

- 1. Is the use of user, context and video information together beneficial to the prediction of Enjoyment and Perceived Quality? With respect to Enjoyment one may think that this is the case, though for Perceived Quality user and context information may be superfluous. Here, we test whether using system, user and contex information contextually is more beneficial than using them separately. ve
s:
nd
ty
xt
- 2. Are there specific subsets of features that can optimize the prediction of individual Enjoyment and Perceived Quality, or are all features equally relevant in the prediction? Because of the curse of dimensionality and t Quality, or are all features equally relevant in the prediction? Because of the curse of dimensionality and the limited extent of our dataset, we would expect a large number of features to increase the bias of our framework We perform therefore feature selection to identify a minimal set of features that maximizes prediction accuracy, and related to (1) we verify whether information related to system, context and user is maintained in these minimal sets.

4.1 Target de efinition

For this preliminary study, we decided to binarize QoVE judgments into two classes, i.e. acceptable and not acceptable QoVE. To do so, we thresholded the individual QoVE scores for both aspects. For Perceived Quality, the threshold was set to 4: videos scored higher than 4 were defined as having Excellent Quality (EQ) for a specific participant, and scores lower than 4 (including 4) indicated that the video was perceived to have Average Quality (AQ) by a given participant Eventually, we had 173 samples of EQ and 181 samples of AQ. The reason to use such setting is that we aim to identify excellent Perceived Quality of videos. As mentioned earlier, only two quality levels were used in our previous study (177 samples per quality level): 600kbps of bitrate as a reasonable video quality level and 2000kbps as the entry level of HD quality. In our dataset, 118 samples of the 600kbps videos had EQ while 122 samples of the 2000kbps videos had AQ In line with Perceived Quality, we aim to predict high levels of Enjoyment. In psychological literature, aspects concerned with contentment are considered to be high when assigned a score corresponding to at least 75% of the maximum rating score [34]. Following this practice, we set the threshold of Enjoyment to 21: thus, participants who scored their enjoyment equal to or higher than 21 were considered to have had an enjoyable video experience, and were defined as Happy (H), whereas participants scoring lower than 21 were defined as Not Happy (NH) with the video. Eventually, we had 212 samples of NH and 143 samples of H. ee
eeus
t. ify
77
ID
Q. dge ir us verseres),

4.2 Setup of the Framewo ork

For this proof of concept, we were primarily interested in establishing the contribution that video, context and user information might give in the prediction of personalized QoVE. Thus, we chose to implement the predictor with a simple linear classifier at this stage, as its functioning is straightforward to interpret, and due to its simplicity (low variance) it is less prone to overfitting. Three classifiers were considered in this study, i.e., Linear Discriminant Classifier (LDC)

Logistic Classifier (LC) and Least Squares Classifier (LSC). These three classifiers calculated the linear combination of features in order to minimize a certain criterion, which in this case was the number of misclassified observations. Predictors were trained independently for each QoVE aspect. That is, a classifier was trained to predict either Perceived Quality or Enjoyment, and not both simultaneously. This setup allowed us later to select features that were more informative for these two different aspects, and to evaluate whether the selected feature sets overlapped (indicating thereby similarity across the QoVE aspects). Feature selection was indeed performed as anticipated in Section 2, to minimize the input entering the prediction, and therefore the risk of overfitting. In this paper, we adopted the Sequential Forward Feature Selection (SFS), which retained the minimum set of features that minimized the number of misclassifications. Features were selected starting from an empty feature pool, to which they then were sequentially added until there was no improvement in minimizing the number of misclassified observations.

A video-content-sensitive cross validation was performed to estimate the predictor accuracy in a robust way, especially considering the low number of video contents evaluated in the dataset. Thus, the dataset was split in 6 folds, each containing all entries related to one and only one of the 6 video contents. Then, six runs were performed, where the instances of one video were used as test data, while the rest of the videos were used for training. Each run returned a misclassification rate (*MisRate*) on the test data. The final accuracy of the model was defined as:

$$
Accuracy = 1 - \sum_{k=1}^{N} M isRate_k / R
$$

where R was the number of runs.

4.3 Contribution of different type of information to QoVE prediction

The prediction was first performed using all the 68 features presented in section 2.1. The resulting accuracy is reported in Table 3. In general, the performance is low, and borderline random. With regard to Perceived Quality prediction, the accuracy of the three classifiers is more or less the same (around 55%). With respect to Enjoyment prediction, the accuracy of LC and LSC is extremely poor, that is even lower than 50%; the accuracy of LDC, on the other hand, is 65.82%.

In order to better understand the contribution of each of the different types of information we use to predict QoVE, we also performed the classification using as input only the features of a specific group (i.e., user, affective content and perceptual characteristics information; context was omitted as it would imply the prediction to be based on a single, binary variable). The accuracy of these "information-specific" classifiers is reported in Table 3 as well. For Perceived Quality prediction, features related to perceptual characteristics are more informative than features from the other two groups, giving an accuracy of the three classifiers around 60%. For Enjoyment prediction, features describing user characteristics are shown to be better predictors than features from the other two groups, leading to an accuracy of around 64%. Features related to affective video content have a similar accuracy in predicting both Enjoyment and Perceived Quality (around 55%), thus seem to be not fully relevant on their own. In general, it can be noticed that the "information-specific" predictors achieve better performance than the predictors using all 68 features. Overfitting may be a reason for this discrepancy. Indeed, when using all features we may have fed too many input features to the framework, thereby increasing the number of parameters to be set. Given the relatively low number of observations in our dataset, this may have led to overspecialization of the parameters on the training data. Thus, a feature selection may be beneficial

prior to classification in order to reduce the dimensionality of the feature pool.

4.4 Feature Selection

The feature selection step had a two-fold rationale: (1) to reduce the feature pool and consequently the complexity of our model, and (2) to identify which features, and from which type of information, were important to predict Enjoyment and Perceived Quality. Should the feature selection have provided two (partially) overlapping sets of features, we had identified a set of features generally good for the prediction of any aspect of QoVE. But given the observation in [8] that enjoyment and perceived quality are two clearly distinct aspects of QoVE, we were expecting these selected feature sets to be largely disjoint.

The entire dataset was used in the feature selection process, without distinguishing training and test data. Table 4 reports the selected features per category and classifier, ordered according to decreasing importance. With regard to Enjoyment prediction, *Interest*, *Hue ratio* and *Extraversion* were selected as the most relevant features for the three classifiers. Different features related to perceptual characteristics were also selected for each classifier, suggesting that perceptual quality still influenced the enjoyment of a video experience. In addition, the social context feature was selected as a relevant one for the LDC and LC functions, as expected from [8].

PQ5, *PQ44* and *PQ26*, on the other hand, were selected as the three most relevant features for estimating perceived quality, independent of the classifier used. It is interesting to note that personality features - albeit not always the same ones - appeared in each prediction model. This finding suggests that personality is a potential factor in personalized QoVE estimation.

4.5 Model accuracy based on specialized feature sets

Based on these newly selected and specialized feature sets, we trained again the classifiers. The related performances on Enjoyment and Perceived Quality are reported as confusion matrices in Figure 3. Note that, each fold in the crossvalidation returned a partial confusion matrix; these six matrices were further added up to compose the overall confusion matrices.

In general, the performance of Perceived Quality prediction is better than that of Enjoyment. For predicting Enjoyment, LSC achieved the best accuracy (68.93%), with a large improvement with respect to the model using all features

Figure 3: Confusion matrices and accuracy of the three classifiers for Enjoyment and Perceived Quality. "P" stands for predicted samples, whereas "TPR" stands for the false positive rate.

(48.02%, see Table 3). Also with respect to the classifier using only features related to user factors, this is an improvement of over 5%. The accuracy of LDC is much lower than that of the other two classifiers with only 60.45%.

With respect to Perceived Quality prediction, the three classifiers have more or less the same accuracy. Compared to the classifier fed with all features, the accuracy of this classifier is increased by around 14%, showing the added value of the feature selection. Also compared to the accuracy obtained by only considering features related to perceptual characteristics (60.17%), the feature selection brings an improvement, indicating that personality information can be beneficial to Perceived Quality prediction too.

5. DISCUSSION AND CONCLUSION

In this study, we investigated how to include in QoVE prediction not only the features related to perceptual characteristics of the video, but also the features related to the video's affective content, user characteristics and social context. In this way, we not only aimed at expanding the QoVE prediction model to include enjoyment in addition to the perceptual quality, but also at making the prediction better fit the context and characteristics of an individual user. We identified the related features and showed that combining them improves the prediction accuracy of Enjoyment and Perceived Quality as compared to considering only features of one category.

As expected, we found three features related to perceptual characteristics (i.e., *PQ5*, *PQ44* and *PQ26*) to be crucial in estimating Perceived Quality. As described in section 2.1, *PQ5* and *PQ26* describe spatial characteristics of the distortions [21], while *PQ44* describes the effect of temporal distortions on motion [20]. This selection suggests that both spatial and temporal information from videos is important in Perceived Quality prediction. Features related to perceptual characteristics were also found to be relevant in estimating Enjoyment, independent of the classifier used. With respect to features representing affective content of the video, *Hue ratio* was found as the most important feature in Enjoyment prediction. Since the human visual system is much more sensitive to variations in brightness than color, Perceived Quality estimation usually only considers the luminance component of a video, neglecting chromaticity information [35]. However, color, especially the color green, has been proven to be able to evoke positive emotions [24]. Thus, features related to color may be worthwhile to be considered in Enjoyment prediction. Although in our study we did not find evidence of affective content features to be relevant for the prediction of Perceived Quality, we can't exclude that a relationship exists between the two. We encourage future research to further explore this opportunity, by evaluating other affective features in relation with a larger selection of video content.

Features related to user characteristics were found to play a primary role in both OS and Enjoyment prediction, proving the intrinsic appropriateness of our proposed framework. User *interest* was a key component in estimating Enjoyment, as also expected from the subjective studies in [8, 26]. *Extraversion*, which is a personality trait describing enthusiastic and talkative individuals [30], was also key in Enjoyment prediction, possibly because, *extraversion* is proven to have a significantly positive effect on user's perceived enjoyment [36]. For Perceived Quality prediction, especially the personality trait *Agreeableness* had some impact. This trait, selected by the linear classifier, describes kind, sympathetic individuals. *Critical & Quarrelsome,* being the inverse item of agreeableness, was selected to predict Perceived Quality using LSC and LDC. Critical people usually have higher standards, which may have an effect on the way they judge perceived quality. In practice, the addition of user *Agreeableness* information allowed to improve the accuracy of Perceived Quality prediction by 9% with respect to when only perceptual characteristics features were considered. Summarizing, collecting (or estimating, e.g via social media information) information of user characteristics is indeed of major importance to predict an individual's enjoyment for the video experience.

In general, we can conclude that predicting perceived quality and enjoyment requires different sets of features in our study. No single feature was found to be relevant in both Perceived Quality and Enjoyment prediction. The accuracy achieved for the prediction of Enjoyment and Perceived Quality was comparable. This finding suggests that the user's level of enjoyment with video is also predictable and encourages us to further investigate the other aspects of QoVE (e.g. involvement, edurability, etc.).

On the other hand, we acknowledge that there is large room for improvement in terms of prediction accuracy of our models. Also, admittedly, this is an exploratory, initial study towards a prediction of individual QoVE. The set of features we considered is very limited, and diverse video, user and context factors may need to be tested in future studies (e.g., other personality traits, environmental features, Quality of Service parameters). Furthermore, our results are based on a relatively small dataset, including only six different video contents and two bitrate level. Therefore, we cannot guarantee the generalizability of the results to larger video collections or different bitrate levels. Future experiments should be based on larger datasets, including more diversity in video genre and possibly also in artifacts (e.g. packet loss or blur). Finally, our framework was (on purpose) implemented with very simple, linear classifiers. This was done to ensure the interpretability of the results and of the role played by each feature in the prediction. Nevertheless, more sophisticated classifiers (e.g., SVM), may prove more effective in terms of classification accuracy.

Reference

- [1] Pérez, P., Macías, J., Ruiz, J.J., García, N., "Effect of packet loss in video quality of experience," Bell Labs Technical Journal 16, 91-104 (2011).
- [2] Wang, D., Speranza, F., Vincent, A., Martin, T., Blanchfield, P., "Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," Visual Communications and Image Processing 2003. International Society for Optics and Photonics, pp. 198-209 (2003).
- [3] K. Yamori, Y. Tanaka, "Relation between willingness to pay and guaranteed minimum bandwidth in multiplepriority service," in Proc. Joint Conf. 10th Asia–Pacific Conf. Commun. and 5th Int. Symp. Multi-Dimensional Mobile Commun. vol. 1, 113–117 (2004).
- [4] Le Callet, P., S. Möller, A. Perkis., "Qualinet white paper on definitions of quality of experience," (2012).
- [5] S. S. Hemami, A. R. Reibman., "No-reference image and video quality estimation: Applications and humanmotivated design," Signal Processing: Image Communication, vol. 25, no. 7, pp. 469-481 (2010).
- [6] Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L.J., "Objective video quality assessment methods: A classification, review, and performance comparison" IEEE Transactions on Broadcasting 57, 165-182 (2011).
- [7] De Moor, K., Mazza, F., Hupont, I., Quintero, M. R., Mäki, T., & amp; Varela, M., "Chamber QoE: a multiinstrumental approach to explore affective aspects in relation to quality of experience," IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, (2014).
- [8] Yi, Zhu, Ingrid Heynderickx, and Judith A. Redi., "Alone or together: measuring users' viewing experience in different social contexts." IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, (2014).
- [9] ITU-T RECOMMENDATION, P. "Subjective video quality assessment methods for multimedia applications," pp: 34-35 (1999).
- [10] Streijl, Robert C., Stefan Winkler, and David S. Hands. "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives." Multimedia Systems, pp:1-15 (2014).
- [11] K. Seshadrinathan, A. C. Bovik, "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos," IEEE Transactions on Image Processing, vol. 19, no. 2, pp. 335-350 (2010).
- [12] M. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. Broadcast., vol. 50, no. 3, pp.312-322 (2004).
- [13] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," Signal Process. Image Commun., vol. 19, no. 2, pp. 121-132 (2004).
- [14] Hobfeld, T., Raimund Schatz, and Sebastian Egger, "SOS: The MOS is not enough!." Quality of Multimedia Experience (QoMEX), Third International Workshop on. IEEE (2011).
- [15] Judith, A. Redi, Yi Zhu, Huib de Ridder, Ingrid Heynderickx. "How Passive Image Viewers Became Active Multimedia Users." In Visual Signal Quality Assessment, pp. 31-72 (2014).
- [16] Hughes, D. J., Rowe, M., Batey, M., & amp; Lee, A. "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage." Computers in Human Behavior 28.2, pp: 561-569 (2012).
- [17] Correa, Teresa, Amber Willard Hinsley, and Homero Gil De Zuniga. "Who interacts on the Web?: The intersection of users' personality and social media use." Computers in Human Behavior 26.2, pp: 247-253 (2010).
- [18] Gastaldo, P, Zunino, R, Redi, J, "Supporting Visual Quality Assessment with Machine Learning," EURASIP Journal on Image and Video Processing, accepted, in press (2013).
- [19] Pinson, M. H., Janowski, L., Pépion, R., Huynh-Thu, Q., Schmidmer, C., Corriveau, P., ... & Ingram, W. "The influence of subjects and environment on audiovisual subjective tests: An international study." Selected Topics in Signal Processing, IEEE Journal of 6.6, pp: 640-651 (2012).
- [20] Saad, Michele A., Alan C. Bovik, and Christophe Charrier, "Blind prediction of natural video quality." IEEE Transactions on Image Processing 23.3, pp: 1352-1365 (2014).
- [21] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer," IEEE Signal processing Letters, vol. 22, no. 3, pp. 209-212 (2013).
- [22] Hanjalic, Alan, and Li-Qun Xu. "Affective video content representation and modeling. Multimedia," IEEE Transactions on 7.1, pp:143-154 (2005).
- [23] Kang, Hang-Bong. "Affective content detection using HMMs." Proceedings of the eleventh ACM international conference on Multimedia, (2003).
- [24] NAz, K. A. Y. A., and Helena Epps. "Relationship between color and emotion: A study of college students." College Student J 38.3, pp: 396 (2004).
- [25] P. Kortum, M. Sullivan. "The effect of content desirability on subjective video quality ratings." Human factors: the journal of the human factors and ergonomics society 52.1, pp. 105-118 (2010).
- [26] Palhais, Joana, Rui S. Cruz, and Mário S. Nunes. "Quality of Experience Assessment in Internet TV." Mobile Networks and Management. Springer Berlin Heidelberg, pp:261-274 (2012).
- [27] Witmer, Bob G., and Michael J. Singer. "Measuring presence in virtual environments: A presence questionnaire." Presence: Teleoperators and virtual environments 7.3, pp: 225-240 (1998).
- [28] O'Brien, Heather L., and Elaine G. Toms. "What is user engagement?A conceptual framework for defining user engagement with technology." Journal of the American Society for Information Science and Technology 59.6, pp: 938-955 (2008).
- [29] Wechsung, Schulz, Engelbrecht, Niemann, Möller. "All Users Are (Not) Equal-The Influence of User Characteristics on Perceived Quality, Modality Choice and Performance." In Proc. Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, Springer New York (2011).
- [30] Hilgard, "Hilgard's introduction to psychology." Philadelphia PA: Harcourt Brace College Publishers (1996).
- [31] S. D. Gosling, P. J. Rentfrow, and W. Swann. "A very brief measure of the big-five personality domains." Jrnl. of Research in Personality, 37, pp:504 –528 (2003).
- [32] Marcus, A., "Cross-cultural user-experience design." Diagrammatic Representation and Inference, Springer Berlin Heidelberg, pp: 16-24 (2006).
- [33] Sackl, A., Egger, S., Zwickl, P., & amp; Reichl, P. "The QoE alchemy: turning quality into money. Experiences with a refined methodology for the evaluation of willingness-to-pay for service quality." Fourth International Workshop on Quality of Multimedia Experience, (2012).
- [34] Piqueras, J. A., Kuhne, W., Vera-Villarroel, P., van Straten, A., & amp; Cuijpers, P. "Happiness and health behaviours in Chilean college students: a cross-sectional survey." BMC Public Health 11.1, pp: 443 (2011).
- [35] Kaiser, Peter K., Robert M. Boynton. "Human color vision" , (1996).
- [36] Wang, Yi-Shun, Hsin-Hui Lin, and Yi-Wen Liao. "Investigating the individual difference antecedents of perceived enjoyment in students' use of blogging." British Journal of Educational Technology 43.1, pp: 139-152 (2012).