Visual quality assessment for web videos

Tian Xia a,*, Tao Mei b, Gang Hua c, Yong-Dong Zhang a, Xian-Sheng Huab

a Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
b Microsoft Research Asia, Beijing 100190, China
c IBM Research, T.J. Watson Center, United States

Abstract

The advent of video-sharing sites such as YouTube has led to an unprecedented Internet delivery of community-contributed video content. However, most of these videos are not quality-controlled. This paper reports a first attempt towards assessing web videos in terms of visual quality with significant tests on 30k web videos. We regard the quality assessment as a two-class classification problem: features motivated from domain knowledge are extracted to be the visual representation while the overall quality is the two-class label. Observing that web videos are characterized by a much higher diversity of content, genres, capture devices, and skills than any other traditional video program, we propose to combine two types of domain knowledge to predict the perceived quality score. One of the domain knowledge types is the spatiotemporal factors affecting the overall perceived quality of web videos, including four spatial factors and two temporal factors. We study the effectiveness of various spatiotemporal factors and propose some novel spatial factors pertaining the characteristics of web videos. The other is the video editing style, including shot editing style, frame size, and black side ratio. Comprehensive experiments and evaluations over 30k web videos which add up to 1200 h in total demonstrated the effectiveness of the proposed approach. We show some preliminary results for application to filtering and re-ranking of retrieved web videos.

1. Introduction

The rapid development of Internet technologies has provided people with access to a gigantic amount of videos on the Web. Through video-sharing sites such as YouTube [1] and Youku [2], users can post videos on the Internet for free so that millions of people can watch and share. However, it is well known that the quality of web videos varies greatly, because they are uploaded without any quality control. As a result, a substantial portion of web videos are of low quality. These videos usually appear with blocking artifacts, blurring, unstable capturing, and so on, which largely lead to a negative user experience. Automatic quality assessment for web videos has a variety of applications. For example, a video search engine can re-rank or filter a text-based search result by considering the visual quality of each returned video. Moreover, the visual quality can also provide a metric for the generation of video summary or motion thumbnail. Therefore, it is desirable to explore an effective visual quality assessment algorithm for web video.

There are two types of objective quality assessment approaches according to the availability of the original signal, i.e., full-reference (FR) and non-reference (NR) [3]. FR assumes that a complete reference signal is known, while NR assumes the reference signal is not available. On one hand web video lacks a reference signal, NR methods are more suitable than FR. On the other hand, traditional quality assessment is designed according to the distortions during acquisition, compression, transmission, and reproduction. In other words, they are objective measures of the quality of the visual signal. This may not apply to web videos, because the low-quality content of web video is mainly due to non-professional skills or devices employed during the capture of the videos, which are more associated with users’ subjective perception. Therefore, a domain-specific NR quality assessment approach, which models the subjective perception of end users, is more desirable for web videos.

Domain-specific NR quality assessment has been addressed in several previous works. For example, several home video systems [4,5] apply specific knowledge to give an objective NR quality assessment.
assessments for home videos. These methods focus on selecting effective factors related to the overall perceived quality. However, they pay little attention to the video editing style (e.g., shot frequency, average shot length, frame size, etc.), which is also important to the visual quality. In addition, current methods are targeted on traditional video programs only, e.g., home videos, and TV programs. However, web video has its own characteristics. For example, its content covers many genre categories (e.g., movie, cartoon, news, MTV, etc.); its capture skill varies from professional to amateur operators; its capture devices cover professional cameras, web cameras, mobile phone cameras, etc. These specific characteristics make the quality assessment for web videos significantly different from the ones for traditional video programs.

In this paper, we focus on the domain-specific NR quality assessment for web videos. To the best of our knowledge, our work in this paper is the first to address the issue of evaluating visual quality for web video data. We propose a novel approach to visual quality assessment for web videos. The approach uses two types of domain knowledge to predict the perceived quality of the users. One is the spatiotemporal factors affecting the overall perceived quality. These factors include four spatial factors, i.e., blurring, blocking, dynamic range, and intensity contrast; and two temporal factors, i.e., jerkiness, and unstableness. The other type of domain knowledge is from the video editing style, including shot editing style, frame size, and black side ratio. These two kinds of domain knowledge are combined in a support vector machine (SVM) method. Meanwhile, several novel spatial factors are proposed in our approach, and the effectiveness of the various spatiotemporal factors is studied. There are about 30k web videos in our evaluation dataset. Results demonstrate the effectiveness of the proposed approach.

The main contribution of this paper is summarized below:

1. Based on the domain knowledge on web videos, we propose some effective features that reflect visual quality. Compared to the traditional features for photos or video programs, these novel features play a key role in domain-specific NR quality assessment for web videos.
2. This work reports a first attempt towards visual quality assessment for web videos; and the proposed method has been tested on 30k web videos.
3. Based on the proposed visual quality assessment method, we have discussed a wide variety of online video applications.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents the pipeline of the proposed approach to visual quality assessment for web videos. Section 4 shows a study on the domain knowledge for web video quality assessment from two aspects. Section 5 reports the experimental results. Section 6 describes two applications of the proposed approach in video search. Finally, Section 7 summarizes the main results and discusses the possible extensions and future work.

2. Related work

In this section, we will give a review on the literature of visual quality assessment (VQA) for digital images and videos. The problem of VQA can be formulated as a mapping from the original visual data to perceived quality, denoted as

\[ D^{\text{VQA}}(Q) \]

where \( D \) denotes the video or image data whose visual quality is to be assessed, and \( Q \) denotes the perceptual quality. \( Q \) can have multiple dimensions, however in this paper we only use a single value to indicate the overall quality of the video or image.

Strictly speaking, VQA consists of two parts, i.e., subjective VQA and objective VQA [6]. The former is entirely based on a human’s viewing test, while the latter aims to automatically predict visual quality that is consistent with a human’s subjective assessment. Since the focus of this paper is objective VQA, we will concentrate solely on it. As discussed in Section 1, the objective VQA method can be classified into two categories, namely FR and NR methods, according to the availability of the original signal. We will review the related work on objective VQA from these two aspects.

As for FR quality assessment, mean square error (MSE) and peak signal-to-noise ratio (PSNR) [3] are two popular measurement criteria for evaluating the difference between the reference signal and the candidate signal. However, they are not always in line with human perceived quality measurement. Therefore, some FR methods are proposed to define the visual quality measurement from the perspective of human visual system (HVS) [7,8], as well as perceived structural distortions [9]. In addition, the video quality experts group (VQEG) has completed the full reference television (FR-TV) project [6], but achieved limited success in FR quality assessment.

As for NR quality assessment, there are generally two kinds of approaches. The first one is aimed at some specific distortion type, e.g., some NR methods for image data [10,11] focus on the distortion derived from block DCT-based compression; a novel NR metric for video data [12] measures the video quality by pooling the spatial distortions of all frames; an NR video quality estimation method [13] evaluates the quality on the basis of the number of macro blocks containing errors. The second one aims to obtain the quality metric by means of machine learning, e.g., a binary classifier for both high-quality and low-quality classes is built to model the quality metric of image data [14].

It is observed that traditional FR and NR quality assessment methods such as the aforementioned ones are generally targeted on the degradation during acquisition, compression, transmission, and reproduction. While for web videos, low quality content is mainly due to non-professional capture skills or capture devices. Therefore, traditional FR and NR quality assessment methods do not fit the problem of visual quality assessment for web videos.

During recent years, a number of authors have considered the problem of domain-specific NR quality assessment. Several home video systems [4,5] have attempted to give an objective NR quality assessment taking the domain knowledge of home video into account. In [5], home video is partitioned into four kinds of segments, including blurred, shaky, inconsistent and stable, based on camera motion. Our previous work in [4] has proposed a more comprehensive method for visual quality assessment with spatiotemporal factors for home video. A subject region based NR quality assessment method is proposed from the perspective of photography esthetic [15]. It is primarily for photo quality evaluation; in addition, it also reports some preliminary efforts on video quality assessment.

However, current domain-specific NR quality assessment methods pay little attention to the video editing style (e.g., shot frequency, average shot length, frame size, etc.), which is also relevant to the visual quality, and very easy to extract from the structured video data. More importantly, current methods are targeted on photos or traditional video programs, e.g., home videos, and TV programs. They are not taking the specific characteristics of web videos into account. In this paper, we propose a novel domain-specific NR quality assessment approach for web video regarding web video’s characteristics.

3. Visual quality assessment approach

Since a video clip typically has a hierarchical structure, we first perform video structuring to get the sampled shots and its key...
frames. Then a set of spatiotemporal factors that intrinsically affect perceived quality are analyzed based on the characteristics of web videos, and some features reflecting the video editing style are extracted. Finally the spatiotemporal factors and the video editing style features are integrated into an overall quality metric using an SVM based fusion method. In the SVM based fusing process, the spatiotemporal factors and the video editing style are combined as the input feature to the SVM model, and a binary SVM classifier is built on the training set which contains training samples for both high quality and low quality classes. The overall quality metric of a web video under assessment is denoted by the extent to which it belongs to these two classes. The flow chart of the proposed approach is illustrated in Fig. 1.

We explore six spatiotemporal factors in the proposed approach, i.e., unstableness and jerkiness as temporal factors; blurring, blocking, dynamic range, and intensity contrast as spatial factors, in which blocking, dynamic range and intensity contrast are first proposed in this paper regarding the characteristics of web videos. We also propose to utilize the information regarding video editing style in our approach, and propose a 20-dimensional feature which is relevant to the perceived quality of web videos from the perspective of video editing style. It consists of a 13-dimensional shot editing style feature, black side ratio, and frame size. Following this, we will introduce some practical details in the proposed approach, including three modules: preprocessing, black side removal, and global motion estimation. These modules are very important to the efficiency and effectiveness of the proposed approach.

3.1. Preprocessing

A video can typically be split into a hierarchical structure which has three levels of granularity comprising frame level, shot level and scene level. Frame level has no temporal information and is very time consuming for assessing quality. Scene level is a high semantic level: it is still very difficult to partition a video clip into different scenes, especially for web video data which usually has rich content. Thus, we select the shot level to perform the quality assessment. Given a web video clip, we first detect shot boundary using a double threshold shot detector [16], then select some shots for further analysis according to a selecting rule which will be detailed as follows. For the selected shots, we extract key frames that are located in the middle of each shot for the analysis of spatial factors.

3.1.1. Shot selection

Fig. 2 shows the shot frequency distribution in our dataset of about 30k web videos, in which the shot boundary is detected using the algorithm proposed in [16]. We can see that the shot frequency distribution of web videos has two peaks at the two ends separately. It is not necessary to consider all the shots in quality assessment for the consideration of efficiency, and different shots within a video clip generally have consistent visual quality. Consequently, we only sample some shots for further assessment in the proposed approach. Since the shot with a single color or a background which often occurs at the beginning or ending of a video clip contains no information for quality assessment, we select shots trying to avoid these traits. The rule for shot selection in the proposed approach is as follows:

If \( \#\text{shot} < T_s \), select the middle shot; if \( \#\text{shot} \geq T_s \), select three shots by averagely sampling after excluding the first five and last five shots. In which, \( \#\text{shot} \) means the number of shots for the given video clip, \( T_s \) is a threshold on shot number and we set \( T_s = 30 \) in the proposed approach.

3.2. Black sides removal

For the selected shots, we extract key frames that are located in the middle of each shot for analyzing the spatial factors. From observation we find that some web videos, such as movie clips, have black sides which are horizontal or vertical or both. Fig. 3(a) illustrates a key frame with two horizontal black sides. The black sides are caused by the fact that web videos are framed with a black border, which is used to frame the content of the video. To remove these black sides, we propose a black side removal algorithm that uses a combination of morphological operations and thresholding.

\[ T_s = 30 \]

\( \#\text{shot} \)

\( \#\text{shot} \)
sides will distort the quality assessment a lot, since the region of black sides contains no visual content, but it can affect the calculation of spatial factors, especially for dynamic range and intensity contrast. Therefore, we first extract an effective region on a key frame through removing the black sides, as shown in Fig. 3(b).

Then the spatial factors will be analyzed purely on the extracted effective region.

3.3. Global motion estimation

In contrast to a single image, a video consists of a series of frame images, so temporal factors are designed from this perspective. They focus on the erratic camera motion. Temporal factors are usually calculated based on the parameters of an affine model for global motion estimation (GME). We use the method proposed in [17] for estimating an affine model due to its robustness, and improve the efficiency of it through a spatiotemporal down-sampling technique.

The affine model for GME is as follows:

\[
\begin{align*}
\nu_x &= a_1 + a_2 x + a_3 y, \\
\nu_y &= a_4 + a_5 x + a_6 y,
\end{align*}
\]

where \(a_i (i = 1, \ldots, 6)\) denotes the motion parameters for estimation and \((x, y)\) denotes the motion vector at location \((x, y)\). The spatiotemporal down-sampling technique will be discussed in Section 5.3.

4. Domain knowledge analysis for web video quality assessment

In this section, two types of domain knowledge for web video quality assessment are analyzed based on the characteristics of web videos, including spatiotemporal factors and video editing style.

4.1. Characteristics of web video

In order to assess the quality for web videos, certain characteristics should be identified. Different from home videos and TV programs, web videos have two distinctive characteristics as follows:

- Hybrid forms: since the majority of web video data comes from data uploaded by web users, its content covers many genre categorizes, e.g., movie, cartoon, photo2video (i.e., photo sequences with music), MTV, news, sports, etc.; its capture skill covers professional and amateur operations (however, home videos are almost all amateur videos, TV programs are all professional videos); its capture devices cover professional camera, web camera, mobile phone camera, etc.

- Some conversions: since web video is for online viewing, in order to guarantee smooth playback, video websites usually convert videos into one particular format (e.g., YouTube converts videos into the Flash Video format after uploading), and change the configuration of video to meet its standard (e.g., resolution, bit rate, etc.). Some users may also convert the video into lower resolution and lower bit rate in order to make its uploading faster. Both the conversions from websites and users can affect the visual quality of the web videos.

From the observation on web video data, we find that the low quality content usually results from four aspects as follows:

- Unskilled camera manipulation, such as shaky and jerky camera motion, defocusing.
- Poor photographic environment, such as dark lighting, varying illumination.
- Performance limit of the camera device, such as some web cameras and mobile phone cameras.
- Quality degrading conversions from video website or owner, such as reformatting, resolution reduction, bit rate reduction, etc.

---

Please cite this article in press as: T. Xia et al., Visual quality assessment for web videos, J. Vis. Commun. (2010), doi:10.1016/j.jvcir.2010.06.005
4.2. Spatiotemporal factors analysis

Based on the above observations and regarding the characteristics of web video data, we adopt six factors as shown in Fig. 4 which can be regarded as the intrinsic factors affecting the overall perceived quality. Among these factors, two are temporal factors, including unstableness and jerkiness; four are spatial factors comprising blocking, blurring, dynamic range and intensity contrast. Unstableness, jerkiness, and blurring are borrowed from our previous work on home video [4]; blocking, dynamic range, and intensity contrast are first proposed in this paper regarding the characteristics of web videos. For each factor, we calculate an unacceptable score denoted as $S_i(\theta)$ for shot $\theta$ on factor $i$. The higher $S_i(\theta)$ is, the more unacceptable the overall visual quality is.

4.2.1. Temporal factors

Compared with a single image, video has temporal characteristics, e.g., camera motion. Temporal factors are designed from this perspective; they focus on erratic camera motion. To be concise, we only give definitions to the unstableness and jerkiness here; for details refer to [4]. Both the two factors focus on three camera motions $P/T/Z$ (i.e., pan/tilt/zoom), which can be measured by the parameters in the affine model in (1)

\[ P = a_1, \quad T = a_4, \quad Z = \frac{1}{2} (d_2 + a_6). \]

(2) Calculate the MSDS for each block.
(3) Exclude the largest 10% and smallest 10% of MSDS, then average MSDS for all the remaining blocks to get $\text{MSDS}_{\text{avg}}$. Filtering out that 20% of blocks may avoid noise such as sharp edges.
(4) Calculate $S_i(\theta)$ following (5).

\[ S_i(\theta) = \min\{\text{MSDS}_{\text{avg}}, \text{MSDS}_{\text{thr}}\}/\text{MSDS}_{\text{thr}}, \quad (5) \]

where $\text{MSDS}_{\text{thr}} = 7$, which is set manually after testing on our web video data.

4.2.2. Blocking. The blocking factor represents the blur extent of the frame; the gradient histogram based approach proposed in [19] is adopted to define the blurring factor $S_i(\theta)$. This method first calculates the gradients in both horizontal and vertical directions on the gray image, and based on the histogram of gradients, a support vector machine classifier [20] is used to estimate the probability of being blurry, and we define this probability as $S_i(\theta)$. We used libsvm [21] to implement the algorithm, and during the training of SVM, we adopted a radial basis kernel, and sampled 216 blur frames and 326 clear frames, both from web video data, as the training data.

4.2.2.3. Dynamic range and intensity contrast. Dynamic range and intensity contrast are designed to focus on the poor lighting in a frame regarding the properties of rich content and genres of web video data; the poorer lighting is, the smaller the dynamic range and the lower the intensity contrast is.

Specifically, the calculation of dynamic range $S_i(\theta)$ and intensity contrast $S_6(\theta)$ consists of the following steps:

(1) Calculate the 256-bin intensity histogram of the frame, and normalize it to get a distribution denoted as $P(k), k \in [0, 255]$.
(2) Search the 0.05 cutting off points $p_l$ and $p_h$, bisecting point $p_m$ of cumulative distribution of $P(k)$ in (6).
(3) Calculate dynamic range $S_i(\theta)$ in (7).
(4) Calculate intensity contrast $S_6(\theta)$ in (8)

\[ S_i(\theta) = 1 - \frac{(p_h - p_l)}{255}, \quad (6) \]

\[ S_6(\theta) = 1 - \left( \frac{\sum_{k=p_l}^{k=p_m} P(k)}{0.45} - \frac{\sum_{k=p_l}^{k=p_m} P(k)}{0.45} \right)/255. \]

4.3. Video editing style analysis

Similar to film editing, which is an art of storytelling practiced by connecting shots to form an entire movie, video editing is the practice of assembling shots into a coherent whole [22]. Video editing style can be represented from many aspects of video cap-

---

**Fig. 4.** Samples of spatiotemporal factors: (a) unstableness, (b) jerkiness, (c) blocking, (d) blurring and (e) low dynamic range and intensity contrast.

Please cite this article in press as: T. Xia et al., Visual quality assessment for web videos, J. Vis. Commun. (2010), doi:10.1016/j.jvcir.2010.06.005
turing and editing, such as shot frequency, shot length, frame size, etc. It varies with different video genre categorizes and different capture skills.

As aforementioned in Section 1, very little attention has been paid to video editing style in visual quality assessment. However, video editing style has some implications with the perceived quality of the video. Based on the analysis in Section 4.1, the low quality content of web videos is partially due to non-professional capture skills, which will produce a different video editing style when compared with professional capture skills, which generally produce high quality videos. For example, movie clip has a very consistent video editing style, e.g., some movie clips have black sides, and movie clip generally has a high visual quality; amateurish video generally has less shot numbers than a professional video, and professional video generally has higher visual quality than amateurish video. We will further demonstrate this in our experiments. In this section, we propose a 20-dimensional feature which is relevant to the perceived quality of web videos from the perspective of video editing style.

The proposed 20-dimensional video editing style consists of three types of features. The first one is frame size (3D), which consists of width, height, and width-height ratio. The second one is the black side ratio (4D), including up-ratio, bottom-ratio, left-ratio, and right-ratio. As shown in Fig. 5, the calculation of the black side ratio is: up-ratio = 1/height, bottom-ratio = 1/height, left-ratio = 1/width, right-ratio = 1/width.

The third one is shot editing style (13D), which is calculated based on the result of shot boundary detection. Given a web video clip with N shots, for each shot Si, let |Si| denote its length in seconds, then the 13-dimensional shot editing style can be represented as below:

1. Average shot length: \( u = \frac{1}{N} \sum_{i=1}^{N} |S_i| \).
2. Standard deviation of shot length: \( \text{std} = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} (|S_i| - u)^2} \).
3. Maximal shot length: \( m = \max(|S_i|) \).
4. 10-Dimensional local shot length histogram. For each shot \( S_i \), calculate its local length \( |LS_i| = |S_i|/m \), \( |LS_i| \in (0, 1] \). Based on \( |LS_i|, i = 1, \ldots, N \), a 10-dimensional equal bin histogram can be calculated, the bin length is 0.1.

From above, we can see that the extraction of the 20-dimensional feature is mainly derived from the statistics of the shot boundary information. Therefore, the extraction of video editing style features is very computationally efficient. We will further demonstrate its effectiveness in the experimental section.

5. Experiments

Since it is difficult to give an objective no-reference quality validation method, we regard quality assessment as a two class (high quality as positive, low quality as negative) classification problem. We collect 30k videos from the web, invite people to label their visual quality, and split them into a training set and a testing set. We adopt a support vector machine classifier (SVM) [20], and use the implementation of libsvm [21]. The six spatiotemporal factors and the 20-dimensional video editing style feature are combined together as the input feature to SVM, and the quality score \( S(v) \) of video clip \( v \) is the probability of being the high quality category estimated by the trained SVM model. According to the quality score \( S(v) \), we rank the testing samples in descending order, thus a precision-recall curve can be drawn based on the ground truth, which is used to evaluate the proposed approach.

In this section, we perform an extensive study of the proposed approach. Firstly, we evaluate the effectiveness of various spatio-temporal factors on our web video dataset. Besides the six factors adopted in the proposed approach, we also study the performance of some factors proposed from the home video assessment system in [4] including color entropy and brightness. Secondly, we study the effectiveness of the video editing style in quality assessment for web videos. Finally, we test the proposed approach and compare with the quality assessment method for home video [4]. To the best of our knowledge, our work in this paper is the first to address the problem of evaluating visual quality for web video data.

5.1. Shot factor fusion

As aforementioned in Section 3, for a given video clip \( v \) from which \( n \) shots are selected, we can calculate the six spatiotemporal factors \( S(\theta_i), i = 1, \ldots, 6 \), for each shot \( \theta_i, i = 1, \ldots, n \). In order to get the final six spatiotemporal factors \( S_i(v), i = 1, \ldots, 6 \), for the video clip \( v \), we use average fusion to combine \( S(\theta_i), i = 1, \ldots, n \), as shown in (9)

\[
S_i(v) = \frac{1}{n} \sum_{i=1}^{n} S(\theta_i). \tag{9}
\]

5.2. Web video dataset

We collect 29,850 videos from the web over two separate periods, in which 11,518 videos are collected during one period, and denoted as dataset-A; 18,332 videos are collected during the other period, and denoted as dataset-B. We invite 10 average web users to label all these videos into one of the three categories, i.e., high,
fair, and low visual quality. Each video clip is labeled by three different subjects. All the subjects are graduate students who have different majors and often view online videos.\(^3\)

### 5.3. Spatial down-sampling for GME

Traditionally, GME is performed on two adjacent frames on full frame scale. Thus, for a shot containing \(n\) frames, GME needs to be run \((n - 1)\) times to obtain a complete description of global motion of this shot, and each GME is performed over full frame scale. This process is very time consuming. In fact, in our application we only need to come out with temporal factors to indicate how erratic the camera motion is, so we do not need to make the calculation so time consuming by performing GME in the traditional way.

Consequently, a spatiotemporal down-sampling technique is adopted to improve the efficiency of GME in our approach. Given a shot, we down-sample frames temporally with a step of three, then down-sample them spatially into 1/4 of the original size before performing GME. Temporal down-sampling is feasible and effective since the difference between two adjacent frames is usually very small. As for spatial down-sampling, although the accuracy of GME on a small scale may be somehow lower than on the original scale, we will further demonstrate that the difference of performance resulting from 1/4 spatial down-sampling is acceptable in the following experiments. Meanwhile, the adoption of spatial down-sampling in GME has a distinct computational advantage. It can largely reduce the processing time for GME by performing GME in the traditional way.

In the following, we will demonstrate that the difference of performance resulting from 1/4 spatial down-sampling is acceptable in our application through a simple test on six testing videos.

Table 1 summarizes the properties of the six testing videos. Table 2 presents the performance difference between full scale and 1/4 scale GME, in which Column \(i\) means the sign opposite ratio for parameter \(a_i\), in (1) compared between the two scales. Considering a video containing \(K\) frames, there are \(K - 1\) pairs for two adjacent frames. For each pair, six parameters can be estimated from GME. Therefore, \(K - 1\) sets of affine model parameters can be obtained and are denoted as \([a_0', a_1', \ldots, a_6']\), \(j = 1, \ldots, K - 1\). We adopt \([b_0', b_1', \ldots, b_6']\), \(j = 1, \ldots, K - 1\) to denote the parameters estimated on the other scale. The sign opposite ratio for parameter \(a_i\) is calculated as in (10)

\[
R_i = \frac{1}{K - 1} \sum_{j=1}^{K-1} \text{sign}(a'_i - b'_i).
\]  

where \(\text{sign}(x) = \left\{ \begin{array}{ll} 1, & x \geq 0 \\ 0, & x < 0 \end{array} \right.\)

In Column \(\text{diff/pixel}\), the quantity before the colon means the average \(\text{diff/pixel}\) on the original scale, the one behind the colon refers to the 1/4 scale. The \(\text{diff/pixel}\) between two frames is defined by

\[
\text{diff/pixel} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} |\text{cur}(x', y') - \text{pre}(x, y)|.
\]  

where \(M\) and \(N\) denote the width and height of the frame respectively, \(\text{cur}(x, y)\) denotes the grey value of the pixel located at \((x, y)\) on the current frame, \(\text{pre}(x, y)\) denotes the one on the previous frame.

Based on the above results, we can see that the adopted GME algorithm has relatively similar performance on the two scales separately. Therefore, the 1/4 spatial down-sampling for GME is applicable in our application.

### 5.4. Spatiotemporal factor evaluation

Since selecting appropriate factors plays a key role in domain-specific NR quality assessment for web videos, we study the effectiveness of various spatiotemporal factors on web videos. Besides the six factors adopted in the proposed approach, we also study the performance of some factors proposed from the home video assessment system in [4] including color entropy and brightness.

We test the eight spatiotemporal factors based on a subset of dataset-B. From dataset-B we choose the videos labeled by at least two people as high or low quality to construct our experimental dataset. This dataset only focuses on the two-graded high/low quality videos, thus it is a relatively clean dataset with little labeling noise, and we denote it as dataset-B-C. Dataset-B-C consists of 6629 high quality videos and 3976 low quality videos. We regard high quality video as a positive sample, and rank the testing samples according to \(S(f)\), then we can draw the precision-recall curves for different spatiotemporal factors as shown in Fig. 6. Generally speaking, a reasonable precision-recall curve should decrease monotonously.

Fig. 6 shows the performance of different spatiotemporal factors, from which we can draw some conclusions that:

- Color entropy is not an effective factor for web videos. Color entropy represents colorlessness of image content, it is effective on home video since home video is generally all color video, and the average quality is relatively high. While for web video, which covers color video and also gray video, the average quality is relatively low, and in such cases, color entropy is not a reasonable choice, so we will ignore this factor in the proposed approach.
- Brightness is not an effective factor for web videos. Brightness represents the degree of poor lighting artifact, it is not as effective on web video as it does on home video, since web video generally has rich content and genres which makes the assump-

---

\(^3\) People with different ages may have a different perception on visual quality; our subjects are mainly young people since they are more interested in web videos than old people.
tion that a dark image is likely to be a low quality image biased (e.g., some high quality movie clip is shot under dark environment). So we propose two new factors including dynamic range and intensity contrast which pay attention to the relative change of intensity instead of the average intensity as brightness does. We can see that these two factors outperform brightness on web videos.

- Blurring is both effective for home videos and web videos, so we adopt it as one of the spatial factors in our system, and it gets the best performance compared with other spatial factors.
- Blocking is effective for web videos, so it is adopted in our system. Blocking artifacts generally result from the coarse quantization of coefficients of block-based transformation during the video compression. It is specific for web videos since some web video data has to adopt coarse quantization to achieve a low bit rate for online viewing; meanwhile some web videos captured by low performance cameras (e.g., web camera, mobile phone camera) also have coarse quantization due to the limitations of the device.
- Though blocking, dynamic range, and intensity contrast are effective for web videos, the current performance is still not very satisfying. This is due to the rich content of web videos, current factors are not robust enough to handle it. For example, the MSDS method for measuring blocking artifacts prefers to recognize cartoon video as blocking since cartoon is rich of edge information.
- Temporal factors achieve better performance than spatial factors, so temporal factors contribute more to quality assessment than spatial factors for web videos.

As for the six spatiotemporal factors adopted in the proposed approach, we also explore four methods to fuse them into an overall quality metric $S(v)$, i.e., MIN fusion, MAX fusion, average fusion, and linear weighted fusion; they are all shown in Table 3. MIN fusion implies that the best quality among the six factors represents the final quality of the video, MAX fusion implies the opposite. Average fusion can also be regarded as a particular linear weighted fusion with $w_i = \frac{1}{6}, i \in [1, 6]$. In linear weighted fusion, we simply set $w_i = 0.2$ for the two temporal factors; and $w_i = 0.15$ for the four spatial factors.

**Fig. 7** shows the performance of different fusion methods, from it we can find that:

- MIN fusion performs very badly, while MAX fusion performs much better. This implies that a video having high quality on one factor is not likely to be a high quality video, while a video having low quality on one factor is likely to be a low quality video. Therefore, a high quality video is required to have high quality on all factors.
- Both linear weighted fusion and average fusion achieve the best performance. This implies that various spatiotemporal factors have rich complementary features among them. Therefore, we adopt an SVM based fusion method to explore the complementary nature of the various spatiotemporal factors in the proposed approach.

### 5.5 Video editing style evaluation

In this experiment, we study the effectiveness of the 20-dimensional video editing style in visual quality assessment for web video. As discussed above, we regard the quality assessment as a two class classification problem, and adopt a precision-recall curve for evaluation. The training set in this experiment is a subset of dataset-A. From dataset-A, we choose the videos labeled by at least two people as high or low quality to construct the training set. This training set only focuses on the two-agreed high/low quality videos, so it is a relatively clean dataset with little labeling noise, and we denote it as dataset-A-C. Dataset-A-C consists of 6710 high quality videos and 2831 low quality videos. The testing set is dataset-B-C as mentioned in Section 5.4. The 20-dimensional video editing style is the input feature vector to the SVM model. The performance is illustrated in **Fig. 8**, from which we can find that video editing style provides useful information in quality assessment for web videos.

![Precision-recall curves for different spatiotemporal factors.](image)

**Fig. 6.** Precision-recall curves for different spatiotemporal factors.

---

### Table 3: Fusion methods.

<table>
<thead>
<tr>
<th>Fusion methods</th>
<th>Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN fusion</td>
<td>$S(v) = \min {S_i(v) \mid i \in [1, 6]}$</td>
</tr>
<tr>
<td>MAX fusion</td>
<td>$S(v) = \max {S_i(v) \mid i \in [1, 6]}$</td>
</tr>
<tr>
<td>Average fusion</td>
<td>$S(v) = \frac{1}{6} \sum_{i=1}^{6} S_i(v)$</td>
</tr>
<tr>
<td>Linear weighted fusion</td>
<td>$S(v) = \sum_{i=1}^{6} w_i S_i(v), \sum_{i=1}^{6} w_i = 1$</td>
</tr>
</tbody>
</table>

---

Please cite this article in press as: T. Xia et al., Visual quality assessment for web videos, J. Vis. Commun. (2010), doi:10.1016/j.jvcir.2010.06.005
since there is little previous effort on the visual quality assessment for web videos. We denote the method in [4] as PQA-HV, which stands for Visual Quality Assessment for Home Video. PQA-HV focuses only on spatiotemporal factors for home video, including color entropy, brightness, blurring, jerkiness, and unstability, in which the last three factors are also adopted in our method. Therefore, PQA-HV takes the 5-dimensional factors as the input feature vector to the SVM model. Fig. 8 illustrates the performance comparison between the proposed approach and PQA-HV. We can find that the proposed approach, which combines both spatiotemporal factors and video editing style, achieves the best performance.

6. Applications based on video quality

In this section, we apply the proposed approach in the environment of a commercial video search engine. We focus on two applications, i.e., low quality video filtering, and video re-ranking through visual quality. We select 28 queries from the log of a commercial video search engine, most of which are from top queries, the remainder being from average queries. The query list is as below: 14 juillet, beautiful people, cute animals, german tv, girls animals, girls anime, hack my space, I hope you dance, intoxicated, iron maiden music videos, kubrick, latin dancing, linkin park faint, movie video, nrl.com, off road racing videos, omo, piscina, rakim, scary and funny, shining, superhero movies, taking my life away, the closer, the fast and the furious, trish tratus, up, britney spears.

We send the above 28 queries to a commercial video search engine. For each of them, we download the videos in the first four results pages; there are 20 videos on each page. Therefore, we have the top 80 videos for each query from this commercial video search engine.

Before we move further to the two applications, we first study the performance of the proposed approach applied as a low quality video detector on the resulting videos of the 28 queries. Firstly, we invite average web users to label these videos in the same protocol as presented in Section 5.2. We regard the videos labeled by at least two people as low quality to be real low quality videos, and all the other videos to be high quality videos, including videos labeled by at least two people as high or fair quality, thus this dataset has some labeling noise with high quality videos. Secondly, we use our trained model in Section 5.6 to detect the low quality videos in this dataset. The proposed approach achieves a precision of 74.7% and recall of 70.0% for low quality video. Table 4 shows the performance of the proposed approach applied as a low quality video detector in real video search environment.

We also illustrate some detecting results of three queries in Fig. 9. For each query, we only display the results on the first page, which lists the top 20 videos for the given query. In Fig. 9, each im-

---

Table 4

<table>
<thead>
<tr>
<th>Predict low</th>
<th>Actual low</th>
<th>Actual high</th>
</tr>
</thead>
<tbody>
<tr>
<td>284</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>122</td>
<td>1738</td>
<td></td>
</tr>
</tbody>
</table>

Since low quality video gives viewers negative browsing experiences, in this application, among the top 80 videos for each query, we filter out the low quality videos detected by the proposed approach. The performance of low quality video detection is studied above.

Here, we conduct a simple user study experiment to evaluate the filtering results in terms of relevance, quality, and overall satisfaction.
isfaction. Seven average web users are invited to do this user study. For each query, two video lists are provided, one is the original top 80 video list (denoted as A), and the other is the video list after low quality video filtering (denoted as B). Each user is required to compare the two video lists and decide if they are similar or one is better on three aspects separately, including relevance to query, visual quality, and overall satisfaction. On each aspect for a query, the decisions of the seven users are combined through majority rule.

Finally, we find that: (a) on the aspect of relevance, users agree that the original list is better for all the 28 queries; (b) on the aspect of quality, users agree that the filtered list is better for all the 28 queries; (c) on the aspect of overall satisfaction, users agree that the filtered list is better for 18 queries, the two lists are similar for 8 queries, and the original list is better for 2 queries. The results are shown in Table 5. This user study shows that filtering out the low quality videos from the original list will degrade the relevance, since low quality video does not mean there is zero relevance. It also shows that low quality video filtering is relatively helpful in increasing the satisfaction with search results, and this means that users pay more attention to the visual quality than the relevance in video search.

6.2. Video quality assessment for re-ranking

In order to balance the relevance and quality, in this application, we get the re-ranking score for each video by linear fusion of the original ranking score and the quality score. Since the original ranking score is not available and only the ranking order is known to us, we adopt a normalized rank in [23] to convert the ranking order into ranking score as in (12)

\[ S_i(v) = 1 - i/N. \]  

(12)

where \( S_i(v) \) is the original ranking score of video clip \( v \), \( i \) is the ranking order of \( v \) in the original ranking list, \( N \) is the number of the returned videos, 80 in our application. The quality score \( S(v) \) is the probability of being the high quality category estimated by the proposed approach. We get the re-ranking score by averaging \( S_i(v) \) and \( S(v) \), and re-rank the videos for the 28 queries separately.

In order to evaluate the re-ranking results, we conduct a simple user study experiment similar to the one in Section 6.1. In this user study, the seven average web users in the above user study are required to compare the original video list (denoted as A) and the re-ranking video list (denoted as B).

We find that: (a) on the aspect of relevance, users agree that the two lists are similar for 22 queries, the original list is better for 6 queries; (b) on the aspect of quality, users agree that the re-ranking list is better for 19 queries, the two lists are similar for 9 queries; (c) on the aspect of overall satisfaction, users agree that the re-ranking list is better for 25 queries, the two lists are similar for 3 queries. The results are shown in Table 6. This user study shows that re-ranking can balance the relevance and quality well, and increase the satisfaction with the search results.

| Table 6: Results of user study for re-ranking. |
|---|---|---|
| A is better | Relevance | Quality | Satisfaction |
| B is better | 6 | 0 | 0 |
| A and B are similar | 0 | 19 | 3 |

7. Discussion and conclusions

In this paper, we propose a novel domain-specific NR quality assessment method for web videos. The method combines two types of domain knowledge to predict the perceived quality score. One is the spatiotemporal factors, which are related to the overall perceived quality of web videos, and the other is the video editing style. We also study the effectiveness of various spatiotemporal factors. Based on the specific characteristics of web videos, several novel spatial factors are proposed. The proposed factors have been tested on a dataset consisting of 30k web videos. Results demonstrate the effectiveness of the proposed method.

From the experimental results, we find that some of the factors are not robust enough to measure the perceived quality due to the rich content of web videos. To address this issue, some classification algorithms could be developed to categorize web videos into several categories (e.g., cartoon, photo2video, amateur video, etc.) [24]. Then more specific factors are extracted within each category. In such cases, the performance of each factor will be improved, since the video content is more similar within a specific category. Moreover, compared to the huge amount of web videos on the Internet, the scale of our dataset is still a little small. Therefore, we plan to distribute a beta version of our quality assessment tool, and invite more people to use it, and collect their feedback for the evaluation of the proposed approach.

References


