Objective Video Quality Assessment for Tracking Moving Objects from Video Sequences

E. MENDI
Department of Applied Science,
University of Arkansas at Little Rock,
Little Rock, AR, 72204
USA
esmendi@ualr.edu

Y. ZHOU
Department of Information Science,
University of Arkansas at Little Rock,
Little Rock, AR, 72204
USA
yxzhou@ualr.edu

M. MILANOV
Department of Computer Science,
University of Arkansas at Little Rock,
Little Rock, AR, 72204
USA
mgmilanova@ualr.edu

J. TALBURT
Department of Information Science,
University of Arkansas at Little Rock,
Little Rock, AR, 72204
USA
jrtalburt@ualr.edu

Abstract: Video quality assessment has a great importance in several image processing applications. Recently, various objective video quality metrics have been proposed in order to predict the human visual perception and to achieve high correlation with the human perception of the image quality. In this paper, a novel objective quality metric is proposed for tracking moving objects in video sequences. The proposed metric particularly considers the moving objects in video sequences as visually important content. Foreground masks are produced by background subtraction based on an approximate median filter. Existing metrics are then modified by the weighting factors of the foreground masks. Our results show that our metrics have better performance than existing objective metrics.

Key-Words: Video Quality Assessment, Background Subtraction, Tracking Moving Objects from Video Sequences.

1 Introduction
Video quality assessment (VQA) is an important study for many applications. The industry’s need for accurate and consistent objective video metrics has become more critical with new digital video applications and services such as Internet video, surveillance, mobile broadcasting and Internet Protocol television (IPTV).

VQA methods fall into two categories: subjective assessment by humans and objective assessment by algorithms. Objective quality metrics are algorithms designed to characterize the quality of video and predict viewer opinion. Different types of objective metrics exist as illustrated in paper [1]. In the image processing community more than 50 years mean squared error (MSE) are being used as quasi – standard fidelity metrics. The MSE still continues to be widely used as a signal fidelity measure, but at the same time there are recent studies that have developed more advanced signal fidelity measures, especially in applications where perceptual criteria might be relevant. How well an algorithm performs is defined by how well it correlates with the human perception of quality. It is interesting to demonstrate how the video quality is measured for video records where the task is tracking moving objects. It is intuitively obvious that we need to use weighting factors for different regions and measure video quality. Only a small number of existing VQA algorithms detect motion and use motion information directly [2]. A heuristic weighting model is combined with the structural similarity (SSIM) based quality assessment method. The authors use the fact that the accuracy of visual perception is significantly reduced when the speed of motion is extremely large. In [3] a set of heuristic fuzzy rules are proposed that use both absolute and relative motion information to describe visual attention and motion suppression. In [2] the authors use the fact that the human visual system (HVS) is an optimal information extractor. In recent years, it has become clear that many problems in perception organization are difficult to solve without introducing the contextual information. We see and
hear the world in terms of meaningful causal interactions. Barlow’s hypothesis is that the purpose of early visual processing is to transform the highly redundant sensory input into more efficient factorial code. A perceptual system should be organized to transmit maximum information. This hypothesis and the other hypothesis that humans use consecutive approximation with increasing resolution for the selected regions of interest are implemented in [4].

Inspired from new cognitive image representation framework [4] we have developed improved VQA algorithm incorporating the model of motion as spatiotemporal weighting factors. In our video quality measure the weight increases with the information content and decreases with the perceptual uncertainty. The rest of the paper is organized as follows: Section 2 provides brief information about the tracking algorithm. An overview of existing and proposed quality metrics are presented in Section 3. We present the results of our approach in Section 4. Finally, in Section 5 the conclusions of this paper are summarized.

2 Tracking Algorithm

We have tracked moving cars from our traffic video data using background subtraction based on approximate median filter. Since the background is more likely to appear in our traffic data, approximate median, which is computationally efficient and fast, can be used. In this approach, background pixel is incremented by 1, if the input pixel is greater than the corresponding background. Similarly, if the input pixel is smaller than the background pixel, then corresponding background pixel is decremented by 1. In this way, background pixels converge to a value, where half of the input pixels are greater and the half of them is smaller than this value, which is the median. [5]

Background $B$ is estimated at a time $t$, for input frame $I$ as follows:

$$B(x, y, t) = \text{median}\{I(x, y, t - i)\} \quad (1)$$

where $i \in \{0, 1, ..., n - 1\}$ and $n$ denotes the previous frames.

Once background is estimated, foreground mask is obtained by applying a threshold $\tau$ to the absolute difference of estimated background and input frame:

$$|I(x, y, t) - \text{median}\{I(x, y, t - i)\}| > \tau \quad (2)$$

Estimated background and foreground mask of our traffic video data for $n = 20$ is given in Fig. 1.

3 Objective Video Quality Assessment

The most reliable way to measure of Video quality is perceptual quality based on subjective evaluation by orienting on human visual system (HVS). Subjective measures are determined by Mean Opinion Score (MOS) which relies on human perception. On the other hand, objective metrics are also very valuable to make meaningful quality evaluations. They are based on mathematical measurements which are practical to apply without need of human observers. Such methods are widely used in various image processing applications, including filter design, image compression, restoration, denoising, reconstruction, and classification [6]. Objective quality metrics can be classified into 3 metrics: Full Reference (FR), Reduced Reference (RR) and No Reference (NR). All these metrics are based on the availability of original non-distorted reference image which will be compared with the corresponding distorted image. In FR case, reference image information is available; in RR case, partial information of reference image is known and no information about the reference image is available in the NR case.

3.1 MSE

Consider two images $x = \{x_i | i = 1, 2, ..., N\}$ and $y = \{y_i | i = 1, 2, ..., N\}$ where $N$ is the number of pixels and $x_i$ and $y_i$ are the $i$th pixels of the images of $x$ and $y$, respectively; the MSE between these two images is:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$$
MSE is widely used as it is parameter free, computationally simple and mathematically convenient in the context of optimization. It also represents image energy measure that energy is preserved after any orthogonal linear transformation, such as the Fourier transform. However, MSE does not fit precisely with the perceived visual quality. Distorted images with the same MSE may have different visibility. [6] [7]

3.2 SSIM
To overcome limitations of MSE, a new objective quality metric SSIM [8] has been proposed. SSIM correlates well with human subjective perception [9]. Consider two images \( x = \{ x_i | i = 1, 2, ..., N \} \) and \( y = \{ y_i | i = 1, 2, ..., N \} \) where \( N \) is the number of pixels and \( x_i \) and \( y_i \) are the \( i \)th pixels of the images of \( x \) and \( y \), respectively. SSIM-SSIM(\(x,y\)) combines three comparison components, namely luminance-\(l(x,y)\), contrast-\(c(x,y)\) and structure-\(s(x,y)\): [8]

\[
SSIM(x,y) = f(l(x,y),c(x,y),s(x,y))
\]

Luminance, contrast and structure comparisons are defined as follows:

\[
l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad C_1 = (K_1L)^2
\]

\[
c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad C_2 = (K_2L)^2
\]

\[
s(x,y) = \frac{\sigma_{xy}^2 + C_3}{\sigma_x^2 \sigma_y^2 + C_3}, \quad C_3 = \frac{C_2}{2}
\]

where \( \mu_x \), \( \mu_y \), \( \sigma_x \), \( \sigma_y \) and \( \sigma_{xy} \) are means of \( x \) and \( y \), variances of \( x \) and \( y \) and correlation coefficient between \( x \) and \( y \). \( K_1 \) and \( K_2 \) are scalar constants that \( K_1,K_2<<1 \) and \( L \) is the dynamic range of the pixel values. Finally, SSIM index yields to:

\[
SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy}^2 + C_3)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}
\]

3.3 Weighted Objective Quality Metric
In human visual system, the importance of a visual event should increase with the information content, and decrease with the perceptual uncertainty [2], we incorporated foreground mask (2) as weighting function into the MSE and SSIM metrics to measure the motion feature of the moving car. At a time MSE is MSE(\(x,y,t\)) and SSIM is SSIM(\(x,y,t\)). The weighting function is:

\[
w(x,y,t) = ||f(x,y,t) - \text{median}[f(x,y,t-i)]|| > r
\]

We define weighted MSE as wMSE and weighted SSIM as wSSIM as follows:

\[
wMSE = \frac{\sum_i \sum_j \sum_k w(x,y,t)MSE(x,y,t)}{\sum_i \sum_j \sum_k w(x,y,t)}
\]

\[
wSSIM = \frac{\sum_i \sum_j \sum_k w(x,y,t)SSIM(x,y,t)}{\sum_i \sum_j \sum_k w(x,y,t)}
\]
4 Experimental Results

We demonstrated the weighted new objective quality metrics on an intuitive example. We used a traffic video data containing 23 frames from a ground sensor camera. We distorted the original reference video generated from 3 different processing: Blurring, Salt and Pepper noise and JPEG compression. Each process has also 3 distortion amount. Distortion types and amounts are summarized in Table 1.

A sample frame image from the video data and associated distortions are depicted in Fig. 2.

Fig. 3 shows the results of objective VQA. Fig.3.a, 3.c and 3.e show the MSE and wMSE scores of blurred, salt and pepper noise and JPEG compression distortions respectively. Corresponding SSIM and wSSIM scores are given in Fig.3.b, 3.d and 3.f. The x axis in the figures denotes the frame index (time), while the y axis denotes MSE & wMSE or SSIM & wSSIM. As shown in the figures, weighted metrics are more realistic and correlated with human perception. For instance, since there is no moving car in the first frame, MSE and SSIM give wrong scores, while weighted metrics give 0.0 and 1.0, respectively, as they give importance to only moving content. Similarly, in other frames, wMSE values are less than of MSE, and wSSIM values are greater than of SSIM. This is because visually important content such as the moving car is more considered by wMSE and wSSIM.

5 Conclusions

In this paper, we presented a novel objective quality assessment metric. In proposed metrics, moving objects from video sequences are particularly considered as visually important content. Background subtraction based on approximate median filter is used for tracking the moving objects. Then foreground masks are computed from the absolute difference of estimated background and input frame. Existing metrics MSE and SSIM are modified by the weighting factors of the foreground masks. We applied our approach to a traffic video data from a ground sensor. Our results show that our metrics are more realistic and correlated than existing metrics. In the future we will develop a subjective quality assessment to validate our metrics with human subjective perception.

Acknowledgement:

This paper was supported by the NSF grant # 0619069, the Development of an Interdisciplinary Arkansas Emulation Laboratory and by the System Research and Application (SRA) Contract # S280000131. The authors would like to thank the colleagues from QBASE in Dayton, Ohio for the reference video sequences provided.

References:

### Table 1: Distortion processing and amounts

<table>
<thead>
<tr>
<th>Distortion Type</th>
<th>Distortion 1</th>
<th>Distortion 2</th>
<th>Distortion 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blurring</strong></td>
<td>fil. size=6, std. dev = 6</td>
<td>fil. size=8, std. dev = 8</td>
<td>fil. size=10, std. dev = 10</td>
</tr>
<tr>
<td><strong>Salt and Pepper</strong></td>
<td>d (noise density) = 0.01</td>
<td>d (noise density) = 0.03</td>
<td>d (noise density) = 0.05</td>
</tr>
<tr>
<td><strong>JPEG Compression</strong></td>
<td>compression = 50%</td>
<td>compression = 70%</td>
<td>compression = 90%</td>
</tr>
</tbody>
</table>

Fig. 2: a) Sample reference frame, b) blurred of size 10 with standard deviation 10, c) salt and pepper noise with noise of 0.05, d) JPEG compression with 90%
Fig. 3: Objective VQA plots on a test video containing 23 frames, a) MSE and MSE with proposed weighting method for blurring distortion, b) SSIM and SSIM with proposed weighting method for blurring distortion, c) MSE and MSE with proposed weighting method for salt & pepper effect, d) SSIM and SSIM with proposed weighting method for salt & pepper effect, e) MSE and MSE with proposed weighting method for JPEG compression, f) SSIM and SSIM with proposed weighting method for JPEG compression