No-Reference Video Quality Ranking

Ranking the quality of the videos in a you-tube search results list

Shannon McKenna
EE368, Spring 2013
Stanford University
Stanford, CA
scmckenna@gmail.com

Abstract—We evaluate three different no-reference video quality metrics to determine if they correspond well to perceived quality for a range of content and distortion types. Two of the metrics evaluate blur and one of the metrics evaluates blockiness, a common compression artifact. The results of the metrics are compared to the existing LIVE video database, which includes subjective quality scores generated through a study with 38 test subjects.

Keywords—video quality, no-reference, blur, blockiness,

I. INTRODUCTION

A video quality metric is a formula or algorithm that assigns a rating or score to a video that reflects the perceived quality of the video. The most common and simplest existing video quality metrics are mean squared error (MSE) and the peak signal-to-noise ratio (PSNR). These video quality metrics are full-reference metrics that require the original, uncorrupted, error-free source video in order to compute the metric [1]. In practice; however, it is not always possible to have the full-reference video. In addition, the simplest metrics such as MSE and PSNR do not correspond well with the subjective human perception of quality.

The motivation for this paper is to develop a video quality metric that is suitable for ranking the quality of you-tube videos that could help users identify the best quality video in their search results list. To determine the ranking of each video in the list, a no-reference video quality metric must be determined for each video. The metrics should correlate well with human perception of quality.

A no-reference video quality metric is a metric that does not require an original video stream as a reference when calculating the metric. A variety of no-reference video quality metrics have been studied and presented in the literature; however, a reliable no-reference video quality metric that corresponds well to subjective human rankings for a wide variety of video types has yet to be identified [2]. The difficulty in developing a no-reference video quality metric has to do with the complex nature of the human visual system (HVS) and the difference between how humans perceive an image and the distortions that are easy to mathematically detect.

A major source of distortion and poor visual quality is blurriness [1]. Bluriness in a video, whether from an out-of-focus camera, motion-blur, or another source, is perceptible by humans and contributes to the perception of poor video quality.

Two different metrics for blurriness are presented in this paper: a spatial domain based metric and a frequency domain based metric. The spatial domain blur metric implemented in this paper is based on the analysis of the spread of the edges in an image and is presented in [3]. The frequency domain blur metric is based on the slope of the power spectrum and is presented in [4].

A second major source of distortion and poor visual quality in videos are compression artifacts, which for many compression schemes is a blockiness appearance. A metric for blockiness evaluated in this paper that is based on the block-edge gradient, and is presented in [5].

The different metrics presented in this paper were tested on the LIVE video database. The videos in this database have been subjectively ranked by 38 test subjects [5, 6]. The results of the subjective ranking have been used to evaluate how well existing video quality metrics corresponded with human perception of quality.

II. LIVE VIDEO QUALITY DATABASE

The LIVE video quality database is composed of ten uncompressed source videos of natural scenes. For each source video, there are 15 corrupted versions: 4 videos simulating wireless distortions, 3 videos simulating IP distortions, 4 videos with H.264 compression, and 4 videos with MPEG-2 compression at different bit rates.

All of the source videos and their corrupted versions are publicly available in YUV 4:2:0 format with no audio. Seven of the videos have a frame rate of 25 frames per second, 3 of the videos have a frame rate of 50 frames per second, all of the videos were about ten seconds long. The 10 source videos are meant to be diverse in content and include “a wide range of objects, textures, motions, and camera movements” [6].

Each of the videos in the database has an associated Difference Mean Opinion Score (DMOS). The DMOS was derived through subjective testing of 38 test subjects. Each individual viewed all 150 of the videos in the database in random order and gave a score on a scale of 0 to 100 for each video. The DMOS is the difference between the score given to the uncorrupted source video and the score given to the corrupted video. For a low DMOS, there is less perceived difference between the corrupted video and the source video, which means better video quality. A higher score means a
greater perceived difference between the corrupted video and source video, indicating lower quality. The video database and subjective testing was originally put together to evaluate the performance of existing full-reference video quality metrics.

III. NO-REFERENCE BLURRINESS METRICS

An image can appear blurry due to an unfocused camera, motion blur, or compression. Blur in an image corresponds to a decrease in sharpness of an image. In the spatial domain, this corresponds to an increase in edge-width. In the frequency domain, this corresponds to a decrease in amplitude of high frequency content.

Two no-reference bluriness metrics were evaluated and compared for effectiveness. An edge-width metric was implemented to evaluate blur in the spatial domain and a power-spectrum slope metric was implemented to evaluate blur in the frequency domain.

A. Edge-width bluriness metric

The edge-width based blur metric implemented for this paper is presented in [3]. The metric does not require any information about the type of the content or the blurring content and is meant to be implemented for no-reference images. All computation is in the spatial domain which results in low computational complexity. The basic steps of the edge-width based blur metric are:

1. Apply a Sobel filter in the vertical direction to detect all vertical edges.
2. Find the start and end positions of each edge using local maximum and minimum Y (luminance) values of the image.
3. Find the average edge width for each frame of the video. Then, find the average edge width for the entire video by taking the average of the frame edge-widths.

Based on [3], it is sufficient to only consider blur along vertical edges. In the implementation for this paper, a pixel was considered part of an edge if the filtered image pixel value (using the Sobel filter) had a magnitude greater than 0.7.

B. Power-Spectrum Slope Bluriness Metric

The power-spectrum slope based blur metric implemented for this paper is presented in [4]. A blurred image is less sharp than an in-focus image. The reduced sharpness corresponds to less high-frequency content. The slope of the power spectrum can be used to compare the relative amount of high frequency content in images. In general, images with a steeper slope will have less high frequency content, and appear blurrier. Images with a flatter slope will have similar low and high frequency content, indicating that the image has sharper edges and greater level-of-detail. The basic steps of the power-spectrum slope based blur metric are:

1. Compute the power spectrum of a frame by taking the squared magnitude of the two-dimensional fast fourier transform (FFT) of the frame.
2. Calculate the distance from the origin for each pixel, i.e., \( d = \sqrt{u^2+v^2} \).
3. For each pixel, plot the distance from the origin vs the log of the power. Fit a straight line to the data for the frame.
4. Repeat steps 1-3 for all frames in a video. Take the absolute value of the average slope of the fitted line over all frames as the blur metric for the video.

The power spectra of most natural images have a slope of approximately -2. A blurred image generally has a slope that ranges from -2 to -10 [4].

C. Results of Bluriness Metrics

The results of the two bluriness metrics are presented for the H.264 and MPEG-2 compression videos for all ten source videos in Fig. 1 and Fig. 2; and in Table 1 for the “MC” video only. The “MC” video pans a wall with a calendar and a moving toy train.

The edge-blur metric proved unsatisfactory in determining blur between videos with different content. As shown in Fig. there is similar edge width for all versions of the same source video, but varying edge width between the source videos.

As shown in Table 1, the edge-width is slightly larger for videos with a higher DMOS (meaning worse perceived video quality). This indicates that increasing edge-width corresponds to blur and thus, a decrease in image quality. However, the change is the edge-width is very slight among videos with identical content but different distortions and the change in edge-width is very large between the different source videos even if there was similar perceived quality (as shown in Fig. 1).

The metric does not correspond strongly enough to blur and is too dependent on image content to be very useful for a no-reference application. In a no-reference application, the metric cannot vary so widely with videos of different content but similar quality.

The frequency based blur metric proved somewhat more effective. As shown in Fig. 2, there is not too much variation in the power spectrum slope for the different source videos. For most of the source videos, the original, uncorrupted version had the flattest slope (the slope closest to zero). The slope was increasingly steeper for versions with a higher DMOS, indicating that the slope of the power spectrum does correlate with perceived quality. However, the range of the slope was not as great as indicated in the literature. [4] indicated a natural image would have a slope around -2, which is what we observed pretty consistently for the uncorrupted videos, and a slope between -2 and -10 for blurry images. In all of the videos in the LIVE database, the steepest slope was -3.04. The slope of the power spectrum for some uncorrupted videos was as high as -2.4 and -2.5. With such a small range between an uncorrupted video and the steepest slope (indicating the least amount of high frequency energy), the metric is not very robust.

The frequency based blur metric shows promise for no-reference applications because it is not as dependent on the content of the video as the edge-width metric and corresponds to the reasonably well to the subjective ratings in the LIVE
### Table I. Summary of Video Quality Metric Results for "MC" Video

<table>
<thead>
<tr>
<th>“MC” Video</th>
<th>Subjective DMOS</th>
<th>Standard Deviation of DMOS</th>
<th>Edge-Width</th>
<th>Power-Spectrum Slope</th>
<th>Blockiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>NA</td>
<td>NA</td>
<td>4.38</td>
<td>-1.86</td>
<td>0.001</td>
</tr>
<tr>
<td>1 – Wireless Distortion</td>
<td>78.3</td>
<td>10.0</td>
<td>4.79</td>
<td>NA</td>
<td>0.012</td>
</tr>
<tr>
<td>2 – Wireless Distortion</td>
<td>69.2</td>
<td>8.1</td>
<td>4.69</td>
<td>-2.33</td>
<td>0.011</td>
</tr>
<tr>
<td>3 – Wireless Distortion</td>
<td>59.5</td>
<td>9.9</td>
<td>4.62</td>
<td>-2.30</td>
<td>0.014</td>
</tr>
<tr>
<td>4 – Wireless Distortion</td>
<td>57.8</td>
<td>10.3</td>
<td>4.71</td>
<td>-2.27</td>
<td>0.011</td>
</tr>
<tr>
<td>5 – IP Distortion</td>
<td>73.3</td>
<td>9.1</td>
<td>4.74</td>
<td>-2.23</td>
<td>0.011</td>
</tr>
<tr>
<td>6 – IP Distortion</td>
<td>58.5</td>
<td>11.3</td>
<td>4.64</td>
<td>-2.36</td>
<td>0.022</td>
</tr>
<tr>
<td>7 – IP Distortion</td>
<td>54.1</td>
<td>10.0</td>
<td>4.55</td>
<td>-2.26</td>
<td>0.021</td>
</tr>
<tr>
<td>8 – H.264 Compression</td>
<td>47.4</td>
<td>10.8</td>
<td>4.57</td>
<td>-2.45</td>
<td>0.022</td>
</tr>
<tr>
<td>9 – H.264 Compression</td>
<td>48.8</td>
<td>7.8</td>
<td>4.78</td>
<td>-2.48</td>
<td>0.023</td>
</tr>
<tr>
<td>10 – H.264 Compression</td>
<td>57.7</td>
<td>9.7</td>
<td>4.65</td>
<td>-2.52</td>
<td>0.023</td>
</tr>
<tr>
<td>11 – H.264 Compression</td>
<td>67.8</td>
<td>7.4</td>
<td>4.93</td>
<td>-2.57</td>
<td>0.023</td>
</tr>
<tr>
<td>12 – MPEG2 Compression</td>
<td>30.9</td>
<td>8.0</td>
<td>4.43</td>
<td>-1.96</td>
<td>0.029</td>
</tr>
<tr>
<td>13 – MPEG2 Compression</td>
<td>40.5</td>
<td>9.8</td>
<td>4.54</td>
<td>-2.05</td>
<td>0.044</td>
</tr>
<tr>
<td>14 – MPEG2 Compression</td>
<td>52.5</td>
<td>9.9</td>
<td>4.65</td>
<td>-2.15</td>
<td>0.069</td>
</tr>
<tr>
<td>15 – MPEG2 Compression</td>
<td>64.8</td>
<td>9.5</td>
<td>4.88</td>
<td>-2.24</td>
<td>0.101</td>
</tr>
</tbody>
</table>

**Fig. 1.** Edge-width Blur Metric Results for the Ten Source Videos

**Fig. 2.** Power-Spectrum Slope Blur Metric Results for the Ten Source Videos
database for videos from the same source. However, some additional work is required to refine the metric so that it can accurately discriminate quality because there isn’t a very large range in slopes for videos with DMOSs that varied pretty widely.

In an attempt to better understand at what level of blur the slope will start to more dramatically separate from that of the undistorted videos, I applied a horizontal and vertical lowpass filter to a frame of the “MC” video and calculated the slope. I applied a filter impulse response of varying lengths to simulate varying degrees of blur. The resulting slopes are shown in Table 2.

As shown in Table 2, more blur (corresponding to a longer impulse response length), does not necessarily lead to a steeper slope. The blur also eliminates enough of the lower frequency content to keep the power-spectrum slope from getting very steep. The slope begins to flatten out again for a very long filter impulse response, the opposite of what we expected.

<table>
<thead>
<tr>
<th>Impulse Response Length</th>
<th>Power-Spectrum Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vertical Blur</td>
</tr>
<tr>
<td>No Filter</td>
<td>-1.83</td>
</tr>
<tr>
<td>2</td>
<td>-2.60</td>
</tr>
<tr>
<td>3</td>
<td>-2.59</td>
</tr>
<tr>
<td>5</td>
<td>-2.48</td>
</tr>
<tr>
<td>8</td>
<td>-2.37</td>
</tr>
<tr>
<td>15</td>
<td>-2.22</td>
</tr>
</tbody>
</table>

IV. NO-REFERENCE BLOCKINESS METRIC

A. Edge-Gradient Blockiness Metric

The blockiness metric implemented in this paper is presented in [5]. A “blockiness” appearance is a common result of compression and can negatively affect the perceived video quality. Both MPEG2 and H.264 (included in the LIVE video database) are block-based compression schemes. The blockiness metric evaluated is based on the block-edge gradient. The basic steps of the blockiness algorithm are:

1. Compute the “activity” along each edge of an 8x8 pixel block. The activity is defined as the standard deviation, \( \sigma_{\text{edge}} \), of the pixels along the edge.

2. Compute the gradient corresponding to each edge. The gradient, \( \Delta_{\text{edge}} \), is:

   \[
   \Delta_{\text{edge}} = \text{mean} | I_{\text{edge}}(n) - E_{\text{edge}}(n) |; \ n = 0, \ldots, 7 \quad (1)
   \]

   where \( I_{\text{edge}}(n) \) is the magnitude of the edge pixels within the block and \( E_{\text{edge}}(n) \) is the magnitude of the pixels bordering the block-edge.

3. If at least one edge of the block satisfies:

   \( \sigma_{\text{edge}} < 0.1 \)

   \( \Delta_{\text{edge}} > 2.0, \)

   Increment a block counter, \( C_b \), by 1.

4. Repeat the above for all blocks within a frame.

5. Repeat steps 1 through 4 for all frames in the video.

   Average the overall blockiness, \( B_f \), from each frame.

   The constants in step 3 were the recommended constants in [5]. The threshold for \( \sigma_{\text{edge}} \) was chosen because blocks with high activity may mask the blockiness artifacts and should not be included in the blockiness count. The threshold for \( \Delta_{\text{edge}} \) was chosen to ensure the gradient is large enough to be perceptible.

B. Results of the Blockiness Metric

The results of the Blockiness Metric are presented in Fig for the four H.264 and the four MPEG2 compressed videos for all 10 source videos. As shown in Fig. 3, the blockiness metric does not heavily depend on the source content and, in general, there tends to be higher blockiness ratings for higher DMOSs. Table 1 shows that the metric tends to correlate well with DMOS for videos with the same compression type. Further work could be done with this metric to account for compression type if that piece of information were available for the no-reference video.

Table 1 shows the blockiness ratings for the “MC” source video for all 15 distortion types. The table shows that the blockiness ratings for MPEG2 compression are significantly greater than the ratings for H.264 compression. This indicates that the metric is more dependent on the compression type than the actual level of blockiness from the compression. For example, video 11, an H.264 compressed video, has a DMOS of 67.8, but a lower blockiness rating than video 12, an MPEG2 compressed video, which has a DMOS of only 30.9. The MPEG2 video has a much better DMOS, but a worse blockiness rating.

The edge-gradient based blockiness ratings do tend to correlate well with DMOS for videos with the same compression type. Further work could be done with this metric to account for compression type if that piece of information were available for the no-reference video.
V. CONCLUSIONS

In this paper, we presented three different no-reference video quality metrics and compared their performance to the subjective quality scores of the LIVE video database. One metric evaluated an edge-width metric was used to evaluate blur in the spatial domain, but proved to be too sensitive to video content to be satisfactory for no-reference application. The frequency based blur metric used the slope of the power spectrum to compare the amount of low frequency and high frequency content of an image. The frequency based blur metric was more successful than the edge-width metric. The metric corresponded well with the DMOS for the corrupted versions of the same source video and may be applicable between different source videos with further work. The blockiness metric calculated the number of blocks in a frame that exceeded a certain gradient threshold. The blockiness metric correlated well with the DMOS for similar compression types, but did not correlate well with DMOS between compression types. The blockiness metric may be suitable for no-reference use if the compression scheme is known beforehand and can be taken into account.

ACKNOWLEDGMENTS

I would like to acknowledge the Professor Girod and the teaching staff for EE368 Spring 2013. In addition, I would like to acknowledge my project mentor, Mina Makar.

REFERENCES