EVALUATION OF FUZZY IMAGE QUALITY MEASURES USING A MULTIDIMENSIONAL SCALING FRAMEWORK

Dietrich Vander Weken, Etienne Kerre
Ghent University
Fuzziness and Uncertainty Modeling Group
Krijgslaan 281-S9, Ghent, Belgium

Ewout Vansteenkiste, Wilfried Philips
Ghent University
Image Processing Group, TELIN Dept.
Sint-Pietersnieuwstraat 41, Ghent, Belgium

ABSTRACT
In this paper we present a comparison of fuzzy instrumental image quality measures versus experimental psycho-visual data. An existing as well as a new psycho-visual experiment we recently performed at our departments were used to collect data. The Multi-Dimensional Scaling (MDS) framework was applied in order to test which of our fuzzy image similarity measures correlates best to the human visual perception. Based on Spearman’s Rank Order Correlation coefficient we will show the $M_{6}^{p}$ and $M_{3}^{i}$ measures outperform their peers as well as the commonly used MSE and PSNR measures, in the case where image distortions are less trivial to distinguish with the bare eye.

1. INTRODUCTION
In this paper we focus on instrumental image quality measures that are based on similarity measures initially introduced to express the degree of equality between two fuzzy sets. Fuzzy similarity measures can be applied in different ways to digital images. First of all, we used the different similarity measures to construct neighborhood-based similarity measures which also incorporate homogeneity [5]. Secondly, we have illustrated how fuzzy similarity measures can be applied to image histograms [6]. In this case similarity measures turned out to be useful for the comparison of two different kinds of histograms.

In the first place, the similarity measures were applied directly to normalized histograms of the images considered. Secondly the similarity measures were applied to normalized ordered histograms. These combined similarity measures clearly outperform the classical image quality measures, like the Root Mean Square Error, in the sense of image quality evaluation.

In a next step we now want to confirm these results based on experimental psycho-visual data. In former research, a Multi-Dimensional Scaling (MDS) approach towards analyzing and modeling image quality variations has been shown successful [1]. This approach is based on the fact that image quality is often determined by several underlying attributes (such as noise and blur e.g.) and uses a multidimensional geometric model to describe the mutual relationships between different perceptual attributes, as well as the relationships between these attributes and the overall image quality. The dimensionality of this model is determined by the number of independently varying perceptual attributes.

The output or geometrical stimulus configuration can be used to create a perceptual distance matrix for the different stimuli, which can then be compared to the distance matrices created by our own fuzzy similarity measures. Herefore, Spearman’s Rank Order coefficient is used.

In the next sections we will first explain our psycho-visual experiment, Section 2. Then explain the fuzzy similarity measures in more detail, Section 3. We present our results in Section 4 and end with a concluding Section 5.

2. EXPERIMENTAL SETUP
A psycho-visual experiment for the image quality assessment has been described in detail in [2] for images artificially degraded by noise and blur. Three scenes (Wanda,
Terrace and Mondriaan), see Fig. 1, were used in an experiment where 16 combinations of 4 levels of blur (corresponding to no filtering and filtering with binomial filters of length 3, 5 and 9) and 4 levels of gaussian noise (corresponding standard deviations of 0, 7, 10 and 14) were shown on a calibrated display.

In a first experiment, dissimilarity scores for image pairs were collected on a scale from 0 to 10, using 5 different subjects. In a second experiment, 7 subjects were asked to rate blur, noisiness and overall quality for each image separately on the same scale. These data were then processed in the MDS framework. The stimulus configurations that resulted from that will be used as a reference for our own similarity measure comparisons.

Based on the former experiment, we constructed our own experiment that was slightly bigger and focused on artefacts that are more subtle because we believe the fuzzy measures are more suited for artefacts that are less trivial for the eye to see.

Again three scenes (Hill, Face and Barbara), Fig. 2, were used in the experiment but now the original image, a noisy one where gaussian noise was added with a standard deviation of $\sigma = 15$, together with 8 images denoised by eight state of the art filters, were shown. More details on the filters can be found in [9, 8, 7, 10, 11, 12, 13]. An example of the test images can be seen in Fig. 3 for Barbara. Image dissimilarity, on a scale from 0 to 5, and preference scores, on a scale from -3 to 3, were asked for image pairs. The attributes noisiness, bluriness and overall quality, on a scale from 0 to 5, were asked by displaying the images separately. 35 different subjects took part in this experiment of which 14 can be considered experts. The MDS framework was used in order to create a 2D-geometrical configuration and compute the distance matrix from that.

3. FUZZY SIMILARITY MEASURES

Many different similarity measures found in literature, originally introduced to express the degree of comparison between two fuzzy sets, were subject to an extensive theoretical study. We have investigated the similarity measures w.r.t. a list of properties which we considered to be relevant
in order to be useful in image processing applications [3, 4].

In the first place we required a similarity measure to be symmetrical and reflexive. Moreover, we investigated the behaviour of the different similarity measures w.r.t. an increasing amount of salt & pepper noise and an increasing amount of gaussian noise. Furthermore, we also investigated how the different similarity measures reacted to an increasing amount of enlightening or darkening. This investigation has resulted in a set of 14 similarity measures which satisfy the list of relevant properties [14].

The following similarity measure [15] is a well-known example which is also applicable in image processing:

\[
M_6(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

\[
= \frac{\sum_{(i,j) \in X} \min(A(i,j), B(i,j))}{\sum_{(i,j) \in X} \max(A(i,j), B(i,j))}
\]

with \( A \) and \( B \) two digital images over the universe \( X \) of pixels. Note that we use the minimum to model the intersection between two fuzzy sets, the maximum to model the union and the sigma count to model the cardinality of a fuzzy set.

Unfortunately, similarity measures which are applied directly to normalized digital images did not show a convincing perceptual behaviour. That’s why we need more advanced image comparison methods. First of all, we applied the pixel-based similarity measures to partitioned images in order to construct neighborhood-based similarity measures with a more robust behaviour. However, simply applying the similarity measures to corresponding image parts did not yield satisfactory results, so we looked for other techniques in order to improve the behaviour of the similarity measures.

We constructed neighborhood-based similarity measures which also incorporate characteristics of the human visual system [5]. The first property of the human visual system which was taken into account was the contrast sensitivity. Based on the pixel-based similarity measures which satisfy the list of relevant properties we built neighborhood-based similarity measures by which the homogeneity in the considered neighborhoods was taken into account.

Suppose the image is divided in \( N \) image parts of size \( 8 \times 8 \) pixels, and the similarity between the image part \( A_i \) of image \( A \) and the image part \( B_i \) of image \( B \) is denoted by \( M(A_i, B_i) \), then the similarity between the two images \( A \) and \( B \) is given by the weighted average of the similarities in the corresponding disjoint image parts. So, we have that

\[
M^h(A, B) = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot M(A_i, B_i),
\]

where the similarity \( M(A_i, B_i) \) is calculated using the pixel-based similarity measures restricted to the image parts \( A_i \) and \( B_i \) and the weight \( w_i \) is defined as the similarity between the homogeneity \( h_{A_i} \) of image part \( i \) in image \( A \) and the homogeneity \( h_{B_i} \) of image part \( i \) in image \( B \). We can prove these neighborhood-based similarity measures still satisfy the list of relevant properties.

Besides applying the similarity measures directly to pixels of the considered images or neighborhoods of pixels, we then investigated whether the similarity measures, having their origin in fuzzy set theory, could be applied in a meaningful way to the histograms of the considered images.

It is meaningful to compare two histograms in the framework of fuzzy set theory, because the histogram of an image can be transformed to a fuzzy set in the universe of gray levels by dividing the values of the histogram in every gray level by the maximum amount of pixels with the same gray value. In this way the most typical gray value gets membership degree 1 in the fuzzy set associated with the histogram and every other less typical gray value gets a smaller membership degree. Consequently, a normalized histogram is in accordance with the intuitive idea behind a fuzzy set: the most typical element in the universe gets membership degree 1 and all other less typical elements belong to the fuzzy set to a less extent which can be expressed by membership degrees smaller than 1.

A profound experimental study [6] of the applicability of similarity measures to normalized histograms resulted in 15 similarity measures which are appropriate for histogram comparison, i.e. they satisfy the list of relevant properties we impose to a similarity measure in order to be applicable in image processing.

Similarity measures can also be applied in a second way to associated histograms of digital images. The values of a histogram can be ordered in such a way the least occurring gray value is placed in the first position of the histogram and the remaining frequencies are ordered in increasing order. Again, the histogram is normalized analogously to the first case, and consequently the most typical gray value gets membership degree 1 in the fuzzy set associated with the histogram and all the other membership degrees are smaller than or equal to 1 and are ordered in increasing order. We can then apply the different similarity measures to these ordered and normalized histograms. If the similarity measures are applied to ordered histograms we obtain 22 similarity measures which satisfy the list of relevant properties.

In this experiment we selected the 7 most relevant measures of all those described above. Therefore, we considered the similarity measure \( M_6 \) applied in different ways to a pair of images:

- applied neighborhood-based \((M_6^a)\);
- applied neighborhood-based and incorporating homogeneity \((M_6^h)\);
- applied to normalized histograms \((H_6)\);
applied to ordered normalized histograms \((OH_b)\);

Furthermore, we considered the neighborhood-based similarity measures \(M^k_{4,6}\), \(M^k_{1,8c}\), and \(M^k_{1,8s}\), that also incorporate homogeneity. Again we refer to [3, 4, 5, 6, 14] for the exact details.

4. RESULTS

Fig. 4 shows the 2D (we consider blur and noise artefacts to be the two independent attributes in the experiment) geometrical output configuration as optimized by the MDS framework from the combined results of the 35 subjects in our psycho-visual experiment. Each point corresponds to one of the filters shown in Fig. 3. The standard deviations on the positions are also plotted as the little ellipses. All stimulus positions were found statistically significant and similar configurations were obtained for all three scenes. The black arrow shows the direction through which the overall image quality should be measured. The orthogonal projection of all points on this axes gives us a relative ranking of the images. In that way, one can easily see that the original image (3) comes out best, followed by the 3Dwdfit (10) en SADCT-AVGW (5) filter. The noisy image (2), GOA filter (6) and BiShrink 3 filter (7) are ranked worst.

What we now are interested in is to know which of our fuzzy similarity measures corresponds best to the obtained psycho-visual configuration. We do this as follows. From the stimulus positions, the euclidean distance matrix can be computed. Otherwise, using the similarity measures, we can also compute the different distance matrices directly from the input images. By comparing these matrices to the experimental euclidean distance matrix we obtain the similarity measure that fits the experiments best, or in other words that is closest related to human visual quality assessment.

The equivalence between two matrices is computed by Spearman’s Rank Order Correlation coefficient. Let \(D_p\) be the psycho-visual distance matrix and \(D_s\) be a \(N \times N\) similarity measure matrix, \(N\) being the number of input images used, let

\[
R_{ps} = 1 - 6 \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{(Rank[d_p(i, j)] - Rank[d_s(i, j)])^2}{N_D(N_D^2 - 1)}
\]

where \(Rank[d(i, j)]\) stand for the rank of matrix entry \(d(i, j)\) and is a number between 1 and \(N_D = N(N - 1)/2\) when we order all matrix elements ascendingly. The Spearman Rank Order Correlation coefficient \(D_{ps}\) is then given by

\[
D_{ps} = \sqrt{1 - R_{ps}^2}.
\]

One can easily see that the smaller this value, the bigger the equivalence between the matrices is. Table 1 shows the Spearman coefficients for the psycho-visual experiment described in [2]. Table 2 shows the Spearman coefficients for our new psychovisual experiment.

5. CONCLUSIONS

From the results of Spearman’s Rank Order Correlation coefficients for the psycho-visual experiment described in [2] we can see that the commonly used PSNR and MSE yield the lowest values. From the fuzzy similarity measures, the measure \(M^6_{1,8c}\) performs best. From the results of our own psycho-visual experiment one can observe that for the Barbara and Hill image the measure \(M^6_{1,8c}\) performs best (so out-performs the classical measures PSNR and MSE) and that the measure \(M^6_{1,8c}\) together with the measures \(M^6_{1,8s}\) and \(M^6_{1,8c}\), perform best for the Face image.

This can be explained by the nature of the different distortions. In the experiment described in [2] the distortions are rather large, and therefore easy detectable by the classical measures PSNR and MSE. In our own experiment the difference in visual quality between the different images is much more subtle and therefore more sophisticated measures are necessary to detect the differences in visual quality.

It is of course not surprising that the neighborhood-based similarity measures perform better than the histogram similarity measures, since the histogram similarity measures do not incorporate the spatial properties of the different gray values, but only consider the frequency of occurrence. In an overall conclusion \(M^6_{1,8c}\) is the best alternative to the classical
Table 1. Spearman’s Rank Order Correlation Coefficients for the psycho-visual experiment described in [2]. 7 fuzzy similarity measures were tested besides the known MSE and PSNR measures.

<table>
<thead>
<tr>
<th>Spearman’s Rank Order coefficient</th>
<th>Wanda</th>
<th>Terrace</th>
<th>Mondrian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_h^1$</td>
<td>0.651</td>
<td>0.682</td>
<td>0.702</td>
</tr>
<tr>
<td>$M_h^{18c}$</td>
<td>0.634</td>
<td>0.664</td>
<td>0.697</td>
</tr>
<tr>
<td>$M_h^6$</td>
<td>0.628</td>
<td>0.652</td>
<td>0.693</td>
</tr>
<tr>
<td>$H_6$</td>
<td>0.680</td>
<td>0.823</td>
<td>0.888</td>
</tr>
<tr>
<td>$O H_6$</td>
<td>0.561</td>
<td>0.504</td>
<td>0.685</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.511</td>
<td>0.461</td>
<td>0.551</td>
</tr>
<tr>
<td>MSE</td>
<td>0.511</td>
<td>0.461</td>
<td>0.551</td>
</tr>
</tbody>
</table>

Table 2. Spearman’s Rank Order Correlation coefficients for our own psycho-visual experiment. 7 fuzzy similarity measures were tested besides the known MSE and PSNR measures.

<table>
<thead>
<tr>
<th>Spearman’s Rank Order coefficient</th>
<th>Face</th>
<th>Barbara</th>
<th>Hill</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_h^1$</td>
<td>0.851</td>
<td>0.875</td>
<td>0.982</td>
</tr>
<tr>
<td>$M_h^{18c}$</td>
<td>0.850</td>
<td>0.875</td>
<td>0.981</td>
</tr>
<tr>
<td>$M_h^6$</td>
<td>0.850</td>
<td>0.873</td>
<td>0.973</td>
</tr>
<tr>
<td>$M_h^{13}$</td>
<td>0.845</td>
<td>0.867</td>
<td>0.970</td>
</tr>
<tr>
<td>$H_6$</td>
<td>0.896</td>
<td>0.947</td>
<td>0.998</td>
</tr>
<tr>
<td>$M_h^P$</td>
<td>0.880</td>
<td>0.838</td>
<td>0.963</td>
</tr>
<tr>
<td>$O H_6$</td>
<td>0.893</td>
<td>0.961</td>
<td>0.998</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.899</td>
<td>0.846</td>
<td>0.999</td>
</tr>
<tr>
<td>MSE</td>
<td>0.899</td>
<td>0.846</td>
<td>0.999</td>
</tr>
</tbody>
</table>

similarity measures when it comes to human visual perception. In future we want to continue testing this by comparing it to a full reference model known to have good performance such as NTIA-VQM or VQEG.

6. ACKNOWLEDGEMENTS

We would like to thank Prof. Dr. ir. Jean-Bernard Martens for the assistance on the MDS package and ir. Stefaan Lippens for the creation of the web-interface on which the psycho-visual experiment was performed.

7. REFERENCES


