A proposal of objective image quality assessment method based on processed images comparison

Léo PERNET-MUGNIER
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai
leo.pernet-mugnier@insa-lyon.fr

***Abstract*—As the images are more and more used in everyday life applications, the automatic image assessment demand has grown bigger and bigger, especially because the amount of data is too enormous to devolve this task to human beings. Nowadays, a lot of very efficient methods exist. However, all the branches have not been explored. This paper proposes a new objective Image Quality Assessment (IQA) method, designed to be very modular and adaptive depending the need of the client. Unfortunately, by lack of time, this solution is not finished and may need more work to be efficient.**

I. Introduction

Many IQA solutions exist today [1]. Some of them are very simple, like MSE and PSNR, which are basically matrix subtraction. A lot of other solutions try to imitate or simulate the Human Visual System (HVS), e.g. VIF [2], MAD [3] or FSIM [4]. They are actually the best existing solutions, according to [1]. Finally, some solutions are based on the observation of several elements of the image, like SSIM [5]. SSIM calculates difference scores in luminance, contrast and structure, and then gather them using some kind of balancing coefficients. Nevertheless, for almost all these solutions, parameters need to be manually specified, which could be hard for a non-specialized user.

In this paper, a process describes any kind of an automatic image modification, like filtering or transform. Our idea is to compare filtered images in addition to complete images. This way, we noticed that if we put an image through a process, a failure type could become more or less noticeable. To settle our system, on the other hand, we use some subjective image assessment data performed on the same set of images and we try to get close to it by weighting our different comparison values.

For this, we will use a two-level architecture: computing level and gathering level. This layering allows a great modularity and is easy to work with.

II. Background

A. Mean Square Error (MSE)

We will need to compare images two by two in the following. For this, will use MSE because it's the simplest way to proceed an objective assessment and its results are easy to understand, although it's also the less accurate method. It is computed as follows:

$$MSE = \frac{1}{WH}\sum_{j=1}^{H}\sum_{i=1}^{W}\left(I_{ref}(i,j) - I_{test}(i,j)\right)^2 \qquad (1)$$

Where W is the width of the image, H its height, I_{ref} is the reference image and I_{test} the image we want to test.

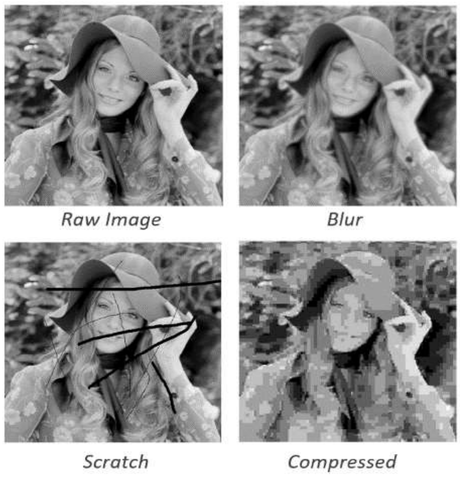

Fig. 1. the four images we use along this paper

When the 2 images have a lot of differences, the MSE is high. If we compare the same image, the MSE will be zero.

B. Comparing processed images

Processing image changes the way we perceive it. Our hypothesis is that it is also influences the image assessment. Unfortunately, we weren't able to find any paper about this, so we did a few tests to verify that processing image influences the image assessment. Here and along this paper, we use one image with three different common types of default: blur, scratch and compression. Those default are intentionally very pronounced to make more visible results. The compressed image results of a heavy jpeg compression. Figure 1 shows the original image and those three altered images.

We compare every altered image with the original image using MSE assessment, for its simplicity and also because we think its accuracy is sufficient here.

Concerning of the effects, we tried to pick them as much different as possible, so it could react differently depending on the type of image default. The three different processes we've chosen are:

- Discrete Fourier Transform (DFT) because it seems to bring up blur alteration.
- Derivative because it brings up scratch default.
- High pass filter because it brings up the jpeg compression alteration.

It's possible to consult how we implemented those processes in our MATLAB file, given in conclusion.

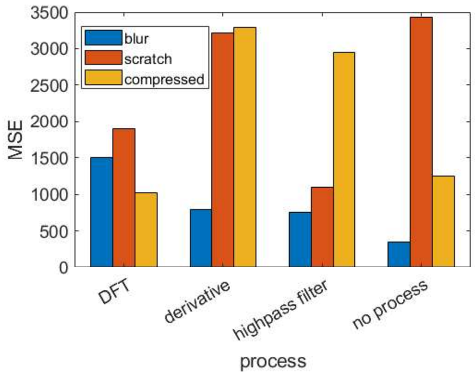

Fig. 2. MSE depending on three image default, treated with three processes and without process

Figure 2 shows the results of this test. First, we can see that, as expected, the image process influences the MSE a lot. The most important thing is that the different effects show different performances according to the type of default. For instance, without any processes, the MSE is very sensible to scratches, but with a high pass filter, it becomes much more sensible about jpeg compression, which confirms our intuition. N.B. in figure 2, the MSE values for *DFT* and *no process* sections have been multiplied, in order to have more visible results.

III. Subjective Assessment

In this part, we will perform a little subjective image assessment. The purpose of objective image assessment is to get as close as possible to the human subjective image comprehension, so we need to know how human beings perceive our image set. For this, as recommend in [6], we perform a Double-Stimulus Continuous Quality-Scale (DSCQS) assessment on the four images presented in figure1.

A. Setup

The experiment takes place in our dormitory, on evening. The test subjects are three males, students, aged between 22 and 26 years old. The advantage is that as they are in a familiar environment, we assume that the bias related to stress is less important than in a laboratory. On the other hand, the sample is very little, and with a very limited variety. However, we consider it’s good enough for this experiment.

We use a laptop with a 17.3” full HD anti-glare screen to display the images, with full brightness. The room is calm and bright, and every test subject does the experiment in the same conditions. They are three displayed images. Each is an assembly of the original image with a corrupted image. Every image is 1198*605 px.

The test subjects perform the experiment one by one. They are alone in the room with the experience manager. They have 10 seconds to see the double image, and then all the time they want to put the grade on a continuous scale, on a paper sheet, as described in [6]. As every scale measure 10cm, we just measure the place of their label to have a raw grade out of 100.

B. Results

Fig. 3 shows that, as expected, the original image got better grades than the altered ones. Also, compressed image has worse ratings than blurred and scratched images. However, because this experiment is very low accurate, we prefer not to draw conclusion concerning different perceptions of the different alterations.

By lack of time and also because it’s doesn’t seem important for this experiment, we don’t perform normalization or error correction on these results. We will assume that they give a rough idea if human perception.

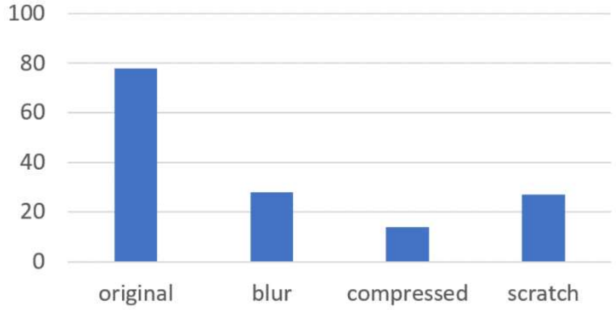

Fig. 3. Average subjective grades, out of 100, depending of default type.

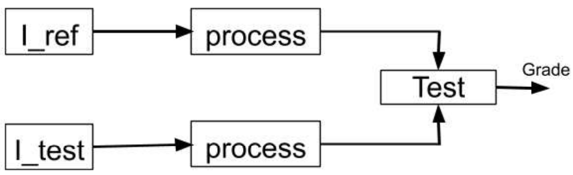

Fig. 4. The brick: the reference image and the test image go through a process and are then compared

IV. Our objective IQA method

We describe here a modular and flexible assessment method.

A. Computing Level: Bricks

As the basic idea of our method is to use several processes independently, we need an elementary brick, described in fig.4. Those bricks could use any process, as long as the process is the same for both test and reference images. Also, the test method is flexible and could be any. Still, here we are going to continue to use the same three processes as before: DFT, derivative and high pass filter. For testing, we will also continue to use MSE for its simplicity and rapidity.

B. Gathering Level

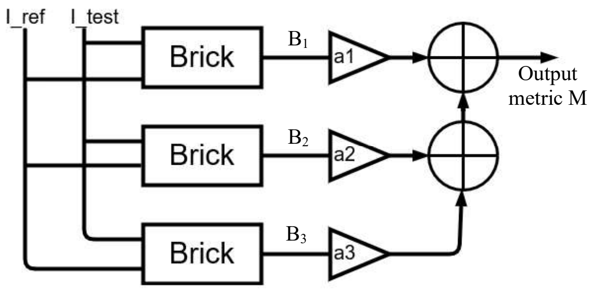

Fig. 5. Global view of our solution. Every brick is weighted by a coefficient. The number of bricks is flexible. However, in case of any changes, it’s important to re-compute the coefficients

figure 5 represents the gathering level of our method. This architecture shows a great flexibility: we can add or remove different processes and test types.

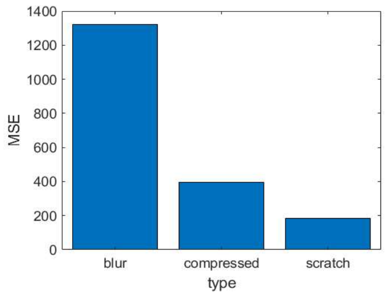

Fig. 6. MSE depending on corruption type. Here, the blur alteration is highlight.

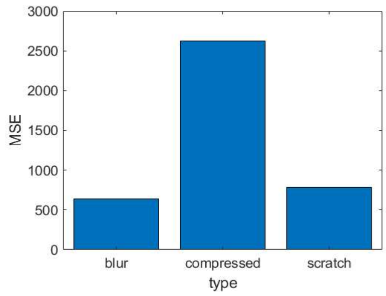

Fig. 7. MSE depending on corruption type. Here, jpeg compression highlight.

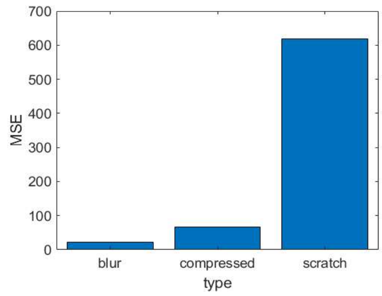

Fig. 8. MSE depending on corruption type. Here, the scratch alteration is highlight.

To compute the final metric, we use the expression:

$$M = \sum_{i=1}^{n} B_i * a_i \qquad (2)$$

With n the number of bricks, B_i the value computed by the brick, depending on $I_{ref,}$ the reference image and $I_{test,}$ the image we want to test. Finally, a_i is the corresponding weighting coefficient.

The most important problem is now how to find the good balancing parameters. To find them, we propose the following solution. First, we have to bind the experimental MSE values visible on figure 2 and the subjective grades on figure 3. For this, we need to normalize the subjective assessment results to obtain Difference Mean Opinion Scores (DMOS) [1]. Then, to link up objective and subjective results, we use a correlation, such as Spearman's rank, Kendall's rank, or other existing correlation metric [1]. Then only, the subjective and objective metrics are comparable, so we could try to adjust the coefficients to fit as much as possible the subjective results. Then, when the coefficients are computed, we could use our solution to guess the subjective grade of an image. For this last part, a decision tree machine learning algorithm could be used.

Unfortunately, again, we don't have enough time and skill to setup this parameter calculation system here.

C. Performing a few tests

As written above, we won't search the best coefficients in this paper. Nevertheless, we want to demonstrate that it's theoretically possible to play with the coefficients to adapt the result depending of the default type. For this, we tried to find the good coefficients to manipulate the MSE. As shown in figures 6,7,8, it works very well.

Conclusion

The MATLAB scripts used in this paper are available at this URL:

https://github.com/elpem/iqaAssignment

In this paper, starting from an intuition, we proved that it's possible to manipulate MSE to be more or less sensible to distinct image defaults, and we coded a basic tool to do it. Also, we proposed an idea to adapt this tool to objective image assessment, by using machine learning. This tool could be used for other uses, like recognizing the type of an image corruption.

Nonetheless, this paper went through many strong streamlining, e.g. we have tested our method only on three versions of the same image with very marked default.

Concerning layering in image assessment, we believe that, as HVS is very complex, layering and abstraction could be good way to simulate it. The solution discussed in this paper is very far to reach HVS complexity, but it might be integrated in more complex systems, as low layer levels.

Future work could relate to researching and implementing the best way to find the coefficients to perform efficient image assessment with our method. We also believe that with the same approach, we could recognize the default type on an image.

Acknowledgment

Thanks to my flat mates for giving their time for my experiment. Thanks also to all the canteens on Shanghai Minhang campus, for their fabulous contribution to science.

References

[1] P. Mohammadi, A. Ebrahimi-Moghadam and S. Shirani, "Subjective and Objective Quality Assessment of Image: a Survey", arXiv, 2014.

[2] H.R. Sheikh and A.C. Bovik, "Image information and visual quality", IEEE Trans. Image Processing, vol. 15, pp. 430-444, Feb. 2006.

[3] E.C. Larson and D.M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", J.Electron. Imag., vol. 19, pp. 011006:1-011006:21, Jan 2010.

[4] L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for imagequality assessment," *IEEE Trans. Image Processing*, vol. 20, pp. 2378-2386, Aug. 2011.

[5] G. Piella and H. Heijmans, "A new quality metric for image fusion," presented at IEEE Int. Conf. on Image Processing, Sept. 2003.

[6] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," ITU, Geneva, Switzerland, 2002.