EVALUATION OF THE QUALITY OF ULTRASOUND IMAGE COMPRESSION BY FUSION OF CRITERIA WITH A SUPPORT VECTOR MACHINE

C. Delgorge[1], C. Rosenberger[1], A. Rakotomamonjy[2], G. Poisson[1] and P. Vieyres[1]

[1] Laboratoire Vision et Robotique - UPRES EA 2078
IUT de Bourges - Bâtiment Recherche
63 av. de Lattre de Tassigny, 18020 Bourges Cedex, France
phone: +33 248238470, fax: +33 248238471
email: Cecile.Delgorge@bourges.univ-orleans.fr

[2] Laboratoire Perception, Systèmes et Informations, FRE CNRS 2645
INSA de Rouen, Avenue de l'Université, 76801 Saint Etienne du Rouvray
phone: +33 232959703 , fax: +33 232959708
email:alain.rakoto@insa-rouen.fr

ABSTRACT

In the framework of a robotized tele-echography, ultrasound images are compressed and sent from a patient station to an expert one. An important task concerns the evaluation of the quality of the compressed images. Indeed, transmitted images are the only feedback information available to the medical expert to remotely control the distant robotized system and to propose a diagnosis. Our objective is to measure the image quality with a statistical criterion and with the same reliability as the medical assessment. We propose in this work a new method for the comparison of compression results. The proposed approach combines different statistical criteria and uses the medical assessment in a training phase with a support vector machine. We show the benefit of this methodology through some experimental results.

1. INTRODUCTION

The goal of the teleoperated chain developed in the frame of the European project OTELO (mObile Tele-Echography using an ultra-Light rObot) is to allow an ultrasound expert to perform an echography examination on a remotely located patient with a teleoperated probe-holder robot. For such an emergency telemedicine application, a low bandwidth and real time examination are the main technical constraints. Due to a reduced bandwidth of the available communication links, an image compression is needed to deliver, from the patient station to the expert station, ultrasound images of 'acceptable' quality and in real time. In the framework of a robotized tele-echography, ultrasound images are compressed at the patient station and sent to the specialist. These received images are the only feedback information available to the medical expert to remotely control the distant robotized system [1]. The diagnosis made by the specialist strongly depends on the quality of these images. This work has been realized within the framework of the European project OTELO where we had to choose an image compression technique and a performance evaluation method.

There are many methods to evaluate an image quality. In the image processing literature, the most frequently used measures are the mean square error (MSE) and the signal to noise ratio (SNR)[2]. They are part of the pixel difference-based distortion measures set and they are very popular due to their mathematical facility. Others criteria can also be found such as statistical measures: Linfoot, based on the power spectral density [3] or the Moran-I statistics [4]. The important drawback of these criteria is the fact that they do not always correspond to the human visual system (HVS), which corresponds to an observer's visual perception.

Image quality, especially in medical specialty, is traditionally evaluated with a visual test where experts examine a large set of images and score each one on its quality (contrast, details) and its distortion. The most common psychovisual study is the Receiver Operating Characteristics Curves method (ROC method) [5] [6]. Such tests are time and human consuming ; they need a large database of images to test. Also, these qualitative and subjective evaluations may depend on the medical specialty. Psychovisual tests require a strict protocol which is very difficult to implement.

If mathematical criteria can easily offer a tool to evaluate the quality of a compressed image with respect to the original ultrasound image, the evaluation of a medical image echography diagnosis remains dependant on the specialist's ability to detect eventual pathologies in one given image. This subjective element in the clinical diagnosis has led us to define a psychovisual test whose results are set as our absolute reference. The goal of this work is to study the behavior of several statistical criteria compared to a clinical evaluation. Then, we propose to fusion the best criteria by taking into account the medical assessment. We then realize a training phase with a support vector machine to improve the evaluation quality.

Section 2 presents the evaluation criteria that we tested on compressed ultrasound images : first the psychovisual test is detailed, then 16 criteria are analyzed. Section 3 shows the learning step with the support vector machine. Section 4 illustrates the efficiency of the proposed method. Conclusion is discussed in section 5.

2. DEVELOPED METHOD

The goal of this study is to find out a statistical criterion close to a medical assessment for the evaluation of a compressed ultrasound image quality. First, we create a psychovisual test. This test allows us to collect a significant number of experts' scores, which we define as our *reference evaluations*. We performed a comparative study of the statistical evaluation criteria. Second, a fusion of the best statistical criteria was done using a support vector machine approach. The idea is to predict the index quality of a compressed im-

age as close as the experts' quality score.

2.1 The psychovisual evaluation : the expert reference

We performed a study to evaluate the quality of ultrasound image compression according to psychovisual measures.

The survey was performed on 15 ultrasound images, each one is compressed with 5 different compression techniques given an exhaustive database of 75 compression results. The goal of this work is not to compare the performance of these compression methods, but to quantify the specialist's perception of the image quality.

The test was held following a rigorous protocol regarding the lighting conditions around the examinee :
• the intensity of light falling on the video monitor and on the examinee's face is measured using an incident type exposure meter and set to $8.5 + / - 0.5$ and $10 + / - 0.5$, respectively.
• we use a single monitor for all the examinees, its contrast is fixed, its resolution is set to 1024x768 at 32 bits/pixel.

The whole test is composed of a sequence of 15 different screens. Each screen presents, for one particular image, the original image and 5 compression results. An illustration of such a screen is presented in figure 1. Experts have to compare and sort from worst to best the compressed ultrasound images with respect to the original one. A score ranging from 1 to 5 is given from worst to best quality, respectively. The test campaign was held in October 2004 and involved 12 medical experts, all specialized in ultrasonography. For each compression result, we measure the score average value given by the experts. We analyse for the whole data 15 sorting results, which is a permutation of $\{1, 2, 3, 4, 5\}$. The average standard deviation measured on these results is equal to 0.87. We can also conclude that answers are homogeneous and results consistent.

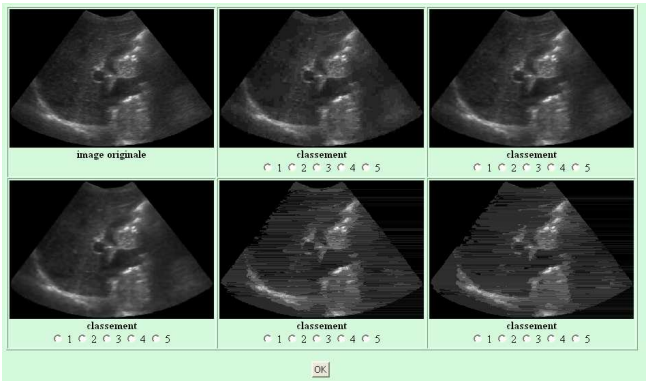


Figure 1: Screen example with 5 compressed images and a reference one presented to the expert

2.2 Statistical quality criteria

The advantage of a psychovisual method, such as the one developed in the previous section, is that the results are closely related to the medical expertise. However, this is a very time and manpower consuming approach. We study some statistical criteria, and compare them regarding the results of the previous psychovisual test. We selected 16 criteria among the ones studied in [7](see table 1).

D1	Minkowsky - Mean absolute error
D2	Minkowsky - Mean square error
D3	Minkowsky - Modified infinity norm
D4	Neighborhood error - 8 neighbours
D5	Neighborhood error - 24 neighbours
D6	Multiresolution error
C1	Normalized cross correlation
C2	Image fidelity
C3	Czekonowski correlation
S1	Spectral phase error
S2	Spectral phase-magnitude error
S3	Block spectral magnitude error
S4	Block spectral phase error
S5	Block spectral phase-magnitude error
P1	Peak signal to noise ratio
T1	Contrast measure

Table 1: Statistical criteria

2.3 Similarity function

As we have relative measures, we compare a sorting and not the score given to each compression result. Criteria are sorted according to their own variation (e.g. the PSNR values are ranked from their highest to lowest values, the Minkowski errors are ranked from their lowest to highest values). For each screen of the psychovisual study, we obtain a sorting of the five compression results, which is a permutation of $\{1, 2, 3, 4, 5\}$.

We can now express the comparison between each couple of 2 images among the 5 compression results in one screen. With this method, we obtain 10 comparison results per screen, where the value 1 is given to the image with best quality, and value -1 to the other.
For example, if we have the sorting values $\{3, 1, 5, 4, 2\}$ for the screen $\{image1, image2, image3, image4, image5\}$. We compare $image1$ and $image2$: $image1$ has a better quality than $image2$ as the experts give the rank 3 to $image1$ and the rank 1 to $image2$. We obtain the following comparison result $\{1 \ \ -1\}$.

We then have a set S of 150 comparisons of compression results for the whole test (e.g. 10 comparison results for each of the 15 screens).

The medical assessment is expressed by a vector Se of dimension 150 corresponding to the comparison result of each compression result for all the different screens. The average score of medical experts is used to determine this vector. A vector Sc can be also obtained for each statistical criterion by comparing each compression result given the value of the criterion. As for example, the comparison result of two compression results, will have the value 1 if the first result has a higher PSNR value than the secund one.

In order to define the similarity between each criterion and the reference given by the experts' scores, we define the good comparison rate (GCR) :

$$GCR = \frac{1}{150} \sum_{i=1}^{150} 1_{\{Sc_i = Se_i\}}$$

where Se_i and Sc_i are the expert values and the criterion values for the comparison i. This GCR measure represents the criterion fidelity to reproduce the expert judgment (a value of 1 or 100% means a perfect method).

In order to have a more reliable evaluation, we propose a methodology to fusion different evaluation criteria by taking into account the medical assessment. We use a Support Vector Machine (SVM) to achieve this goal.

3. LEARNING COMPRESSION QUALITY WITH SUPPORT VECTOR MACHINES

Suppose we have a set of pairs $\{x_i, y_i\}_{i=1,\cdot\ell}$ with $x_i \in \mathrm{R}^d$ being a vector of d statistical criteria describing the quality of a compression of a given image and y_i an index quality of a compression scheme. Our objective is to learn from the knowledge of the training set $\{x_i, y_i\}_{i=1,\cdot\ell}$ a function f that will be able to predict accurately the index quality of compression of a new image x. Thus, our idea is to use a supervised learning framework for achieving this goal but also to use this context for fusing different criteria and selecting the most useful ones.

For solving this learning problem, we have used a 2-norm Support Vector Machines. [8]. Hence, we are looking for a hyperplane in a space $\mathcal{H}$ defined as : $f(x) = \sum_{i=1}^{\ell} \alpha_i^\star y_i K(x_i, x) + b$ that maximizes the margin between the hyperplane and the projected data point x_i in $\mathcal{H}$. Hence $\alpha_i^\star$ are the solution of the following optimization problem :

$$\begin{array}{l} \max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j (K(x_i,x_j) + \frac{1}{C}\delta_{i,j}) \\ \text{with} \sum_i \alpha_i y_i = 0 \quad 0 \leq \alpha_i \end{array} \qquad (1)$$

where K is the kernel associated to $\mathcal{H}$, $\delta_{i,j}$ is the kronecker symbol and C a trade-off parameter between the margin width and the number of training examples located beyond the margin.

Furthermore, we are interested in knowing which statistical criteria are relevant for predicting the compression quality. Learning the decision function f, a criterion selection has been performed. The variable selection algorithm is a backward features ranking algorithm based on the influence of a given criterion on the margin [9] . Hence, each criterion has been weighted by a scaling factor σ and the sensitivity of the margin with regards to a criterion u is related to $|\sum_i \sum_j \alpha_i\alpha_j y_i y_j \frac{\partial K(x_i,x_j)}{\partial \sigma_u}|$. For more details about this variable ranking procedure, the reader is referred to [9]

4. EXPERIMENTAL RESULTS

We present experimental results of comparison with the sorting realized by a medical assessment and the 16 statistical criteria. The GCR between all the 16 criteria score and the expert score is measured, and represents a similarity of comparison we would like to maximize (see table 2).

D1	0.3933	D2	0.3800	D3	0.4267	D4	0.5200
D5	0.5200	D6	0.3800	C1	0.3800	C2	0.4867
C3	0.5000	S1	0.4067	S2	0.3333	S3	0.3800
S4	0.4133	S5	0.4133	P1	0.3800	T1	0.3467

Table 2: Good comparison Rate between each criterion and the experts'scores.

The best rate is obtained by D4 and D5 with a value of 0.52 (that means a 52% similarity to the medical assessment). Based on these results, we can select the 4 criteria presenting the highest value in the comparison namely : D4, D5, C3, and C2.

We can note that two criteria are pixel difference-based measures and two are correlation-based measures.

The Neighborhood Errors D4 and D5 are given by

$$\sqrt{\frac{1}{2(N-w)^2} \sum_{i,j=\frac{w+1}{2}}^{N-\frac{w-1}{2}} d(C,\tilde{C})^2 + d(\tilde{C},C)^2}$$

where $w = 3$ for D4 and $w = 5$ for D5 and represent the mean square error extended to a $w * w$ neighborhood. $d(.,.)$is a distance metric measured between the original image C and the compressed one $\tilde{C}$ of size N^2 pixels.

The Image Fidelity C2 is defined by

$$C2 = \frac{\sum_{i,j=0}^{N-1} C(i,j)\tilde{C}(i,j)}{\sum_{i,j=0}^{N-1} C(i,j)^2}$$

and represents the normalized cross-correlation measure.

The Czekonowski correlation C3 is given by

$$C3 = \frac{1}{N^2} \sum_{i,j=0}^{N-1} (1 - \frac{2 * min(C(i,j), \tilde{C}(i,j))}{(C(i,j) + \tilde{C}(i,j))})$$

We can notice that the P1 measure, also known as the PSNR (one of the most popular criterion), obtains a bad score: with a GCR rate equal to 38 % the PSNR ranks 10^{th} among the 16 studied criteria.

In this first experiment, we have used the same data as in section 2.3 namely, 150 compression results. Hence, we have run a SVM with a variable ranking at each run. We have analyzed the performance of our algorithm with respect to the ratio of examples in the learning set. Hence, for a given ratio, the learning and testing set have been built by splitting randomly all examples. Then, due to the randomness of this procedure, 10 trials have been performed with different random draws of the learning and testing set.

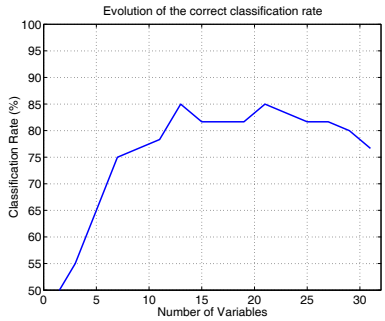
For each trial and run, the SVMs and variable ranking algorithm have been performed on a large range of hyperparameters values C and d the gaussian kernel bandwidth.

Figure 2(a) shows the results of the criterion selection, which concludes that the optimal number of criteria is 6. Figure 2(b) shows the GCR with respect to the number of compression results used in the training set. The learning is done with the 16 criteria measures. Table 3 resumes the GCR obtained by the 4 selected criteria and shows the best GCR obtained by the SVM.

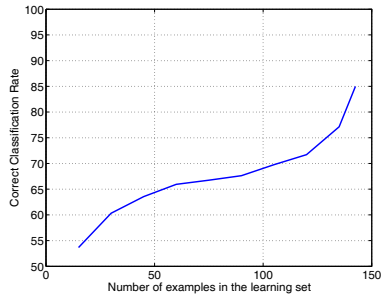
D4	D5	C3	C2	Fusion
0.52	0.52	0.50	0.48	0.85

Table 3: Good comparison Rate : the 4 selected criteria and the fusion criterion.

When only 25 compression results are used in the learning phase by the system, the good recognition rate is equal to 60% ; the best comparison rate obtained by one criterion among the 16 was 52% (Fig 2(b)).
When 95% of the whole set (150 compression results) is used in the learning database, the system obtains a successful score of 85% in the recognition of these same 150 results.



(a) Criterion selection



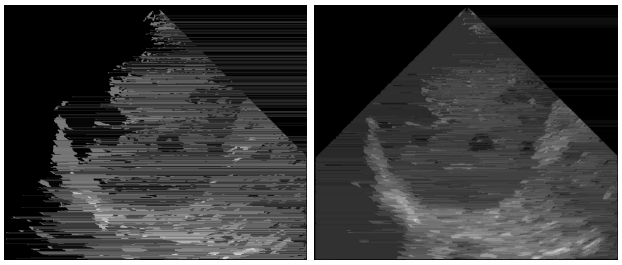
(b) Comparison results

Figure 2: Evolution of the correct classification rate

The fusion of all criteria allows a good improvement of the results.

We illustrate the efficiency of the proposed method for the comparison of two compression results.

The example (see figure 3) concerns an image compressed with the standard Jpeg-Ls related to two compression rates (a high compression at 0.98% and a very high compression at 1.77%). Visually and with no expertise in ultrasound images analysis, one can determine that the image B has a better quality. The learning phase was done with the 150 previous compression results. The fusion defines correctly image B as the best (see Table 4).



(a) image A (b) image B

Figure 3: Two compression results (image A and image B) of the same original ultrasound image

5. CONCLUSIONS AND PERSPECTIVES

We expose in this paper a comparison of some evaluation criteria to quantify the image compression quality. We use a psychovisual study with 12 medical experts to identify the statistical criteria having the best behavior compared to the medical assessment. This study allows us to select four criteria among the 16 tested ones : image fidelity, neighborhood errors and Czekonowski correlation. A support vector machine performs the fusion with the selected criteria and offers a significant improvement of the evaluation efficiency. The performance of the proposed criterion provides an improvement of about 30% compared to the best criterion from our survey for the quality evaluation of compression results. A perspective of this study is to use this criterion for the comparison of ultrasound image compression best fitted for a mobile robotized tele-echography system.

	Method A	Method B
D4	-1	**1**
D5	-1	**1**
C3	**1**	-1
C2	-1	**1**
Fusion	-1	**1**

Table 4: Comparison of two compressed ultrasound images by the four evaluation criteria and the fusion criterion (value for the best image is presented in bold face for each criterion).

Acknowledgement : this work was funded by the European Commission under OTELO project (IST 2001-32516).

REFERENCES

[1] Smith-Guerin N. and AlBassit L. and Courreges F. and Poisson G. and Delgorge C. and Arbeille Ph. and Vieyres P., "Clinical validation of a mobile patient-expert tele-echography system using ISDN lines", Conference on Information Technology Applications in Biomedicine, ITAB'03, Birmingham, 2003.

[2] Deepak S. Turaga and Yingwei Chen and Jorge Caviedes, "No reference PSNR estimation for compressed pictures", Signal Processing : Image Communication, 19, pp. 173-184, 2004.

[3] Christine Fernandez-Maloigne, "Couleur numerique et psychometrie", Computer Art Journal, 1(1), 2004.

[4] Tzong-Jer Chen et al., "A novel image quality index using Moran I statistics", Physics in Medicine and Biology, 48, pp. 131-137, 2003.

[5] H. Lamminen and K. Ruohonen and H. Uusitalo, "Visual tests for measuring the picture quality of teleconsultations for medical purposes", Computer Methods and Programs in Biomedicine, 65, pp. 95-110, 2001.

[6] B. Kassai et al., "A systematic review of the accuracy of ultrasound in the diagnosis of asymptomatic deep venous thrombosis : preliminary results", Journal of Thrombosis and Haemostasis, I-supplement 1, n P1443, 2003.

[7] Ismail Avcibas and Bulent Sankur and Khalid Sayood, "Statistical evaluation of image quality measures", Journal of Electronic imaging, 11(2), pp. 206-223, 2002.

[8] N. Cristianini and J. Shawe-Taylor, "Introduction to Support Vector Machine", Cambridge University Press, 2000.

[9] A. Rakotomamonjy, "Variable selection using SVM-based criteria", Journal of Machine Learning Research, 3, pp. 1357-1370, 2003.