Evaluation of the Quality of Ultrasound Image Compression for a Robotic Tele-Echographic System

C Delgorge, C Rosenberger, G Poisson, P Vieyres

Laboratoire Vision & Robotique, Universite d'Orleans, France

Abstract

The subject of this paper is to propose a method for the evaluation of ultrasound image compression. Compression techniques are used for image transmission in the frame of two tele-operated robotic chains dedicated to tele-echography : OTELO and TERESA. Our objective is to define a statistical criterion to measure the image quality with the same reliability than the one provided by a medical assessment. An initial psychovisual experiment is proposed to medical experts, and represents our reference value for the comparison of statistical evaluation criteria. We propose to fusion different some statistical criteria and to exploit the medical expert judgments during a training phase. Two methods are tested for this approach : linear combination of the criteria with a genetic algorithm and learning the experts judgment by a support vector machine. We show the benefit of this methodology through some experimental results.

1. Introduction

The OTELO European project is dedicated for real-time ultrasound image acquisition and medical diagnosis (see Figure 1) [1]. The robotic tele-operated chain TERESA has been designed for experts in charge of the study and follow up of the astronauts cardiovascular system in micro-gravity environment [2]. For both systems, a light weight robot holds and moves a real probe on a distant patient according to the expert hand's movements and permits an image acquisition using a standard ultrasound device. The choice of compression techniques for image transmission enables a compromise between flow and quality. Transmitted images are the only feedback information available to the medical expert to remotely control the distant robotized system and to propose a diagnosis. The diagnosis made by the specialist strongly depends on the quality of these images. An important task concerns also the evaluation of the quality of the compressed images.

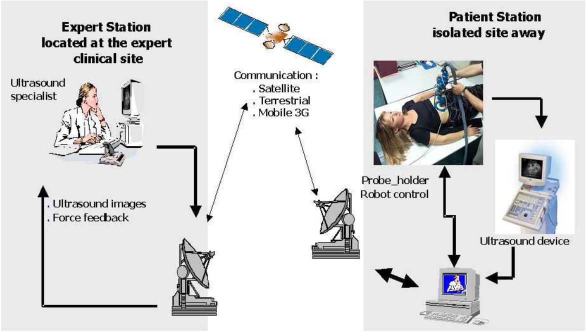


Figure 1. The tele-echography chain OTELO

There are several methods to evaluate the quality of a compressed image : statistical measures and psycho-visual studies. In the image processing literature, the most frequently used measures are the mean square error (MSE) and the signal to noise ratio (SNR)[3]. They are part of the pixel difference-based distortion measures set and are very popular due to their mathematical implementation facility. Others criteria can also be found such as statistical measures: Linfoot, based on the power spectral density [4] or the Moran-I statistics [5]. 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) [6] [7]. 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

0276–6547/05 $20.00 © 2005 IEEE

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 determine if a fusion of these criteria allows an improvement of the evaluation quality.

Section 2 presents the psychovisual test. Section 3 focuses on the evaluation criteria that we tested on compressed ultrasound images. Section 4 shows the fusion step with the genetic algorithm and with the support vector machine. Section 5 illustrates the efficiency of the proposed method thanks to experimental results. Conclusion is discussed in section 6.

2. The psychovisual evaluation

We performed a study to evaluate the quality of ultrasound image compression according to expert judgment. We defined a psychovisual test and proposed it to 10 experts, all specialized in ultrasonography. For this test, experts have to compare and sort from worst to best the compressed ultrasound images with respect to the original one. The goal is not to compare the performance of these compression methods, but to quantify the specialist's perception of the image quality. The protocol is detailed in [8]. Results obtained are considered as our absolute reference in term of image evaluation.

3. Statistical quality criteria

The advantage of a psychovisual method, such as the one developed in the previous section, is that 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 21 criteria among the ones studied in [9]:

- distance measures : The Minkowsky Mean absolute error D1; The Minkowsky Mean square error D2; The Minkowsky Modified infinity norm D3; The Neighborhood error (8 neighbours) D4; The Neighborhood error (24 neighbours) D5; The Multiresolution error D6.
- correlation measures : The Normalized cross correlation C1; The Image fidelity C2; The Czekonowski correlation C3.
- spectral measures : The Spectral phase error S1; The Spectral phase-magnitude error S2; The Block spectral magnitude error S3; The Block spectral phase error S4; The Block spectral phase-magnitude error S5; The Block spectral error S6.
- the Peak signal to noise ratio measure P1.
- the Contrast measure T1.
- human visual system based measures : The Absolute norm Human Visual System H1; The *L2* norm Human Visual System H2; The similarity H3; The DCTune error H4.

These criteria are real values and have different ranges.

3.1. Similarity function

As we have relative measures, we can compare the quality of different compression results. The criteria are sorted according to their own variation (e.g. the PSNR values are ranked from their highest to their lowest values, the Minkowski errors are ranked from their lowest to their highest values). For each screen of the psychovisual study, the 5 compression results are sorted according to the average score given by the medical experts. Given this sorting, we can extract 10 comparisons results for each pair of compression results given by the medical experts and by using an evaluation criterion. In order to define the similarity between each criterion and our reference given by the experts' scores, an absolute difference is measured between the criterion comparison and the expert's one. We define the cumulative similarity of correct comparison (SCC):

$$SCC = \sum_{k=1}^{15} \sum_{i=1}^{10} |A(i,k) - B(i,k)| \qquad (1)$$

where $A(i,k)$ and $B(i,k)$ are respectively the expert and the criterion results for the ith comparison of page k. A comparison result is a value in $\{-1, 1\}$. If a compression result is better than another one, the comparison value is set to 1 otherwise it equals -1. In order to more easily compare this error measure, we also define the similarity rate of correct comparison (SRCC), which represents the absolute similarity of comparison referenced to the maximal value :

$$SRCC = (1 - \frac{SCC}{SCC_{max}}) * 100 \qquad (2)$$

where SCC_{max} corresponds to the biggest difference of the 150 comparison results. In our case, $SCC_{max} = 150 * 2 = 300$.

Regarding the values obtained for this SCC measure, the four best criteria are D5, T1, S2 and S1. We can reach in this case a maximal value of 65.3%. That means that this criterion is able to reproduce the ability of a medical expert to compare two compression results in 65.3% cases. One can notice that the $PSNR$ criterion that is very often used for the comparison of compression results is only ranked at the 9th place.

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 proposed and compared two methods based on this approach. The first one consists in combining linearly different criteria so as to optimize the similarity measure of comparison. The second method uses a "support vector machine" (SVM) and realizes a training for the comparison of compression results.

4. Fusion of criteria

The objective is to fusion several statistical criteria to improve the similarity of their results with the expert judgment. We present in this section the principle of the two chosen methods.

4.1. Linear combination with GA

In the case of four combined criteria, the goal is then to determine the optimal values (a, b, c, d) of a linear combination giving the closest behavior to the medical assessment. We propose here to use the quadratic similarity of sorting computed on the 75 compression results. A genetic algorithm is used as an optimization method.

Genetic algorithms determine solutions of functions by simulating the evolution of a population until survival of best fitted individuals [10]. Survivors are individuals obtained by crossing-over, mutation and selection of individuals from the previous generation. A genetic algorithm is defined by considering five essential data :
1. *Genotype* : a set of characteristics of an individual such as its size. A vector of parameters (a, b, c, d) is considered as an individual.
2. *Initial population* : a set of individuals characterized by their genotypes. It is composed of a set of random values of parameters.
3. *Fitness function* : this function provides to quantify the fitness of an individual to the environment by considering its genotype. We take a quadratic similarity of sorting with the expert evaluation on the 75 compression results.
4. *Operators on genotypes* : they define alterations on genotypes in order to evaluate the population during generations. There exists three types of operators : Individual mutation, Selection of an individual, Crossing-over.
5. *Stopping criterion* : this criterion allows to stop the evolution of the population. We choose to consider the stability of the standard deviation of the evaluation criterion of the population.

4.2. Learning with SVM

Considering a set of pairs $\{x_i, y_i\}_{i=1,\cdot\ell}$ with $x_i \in \mathbb{R}$ 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. The objective is to learn from the knowledge of the training set $\{x_i, y_i\}_{i=1,\cdot\ell}$ a function f that predicts 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.

Furthermore, we are interested in knowing which statistical criteria are relevant for predicting the compression quality. While 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. For more details about this variable ranking procedure, refer to [11].

5. Experimental results

The data collected from the psychovisual evaluation campaign are used as a reference to identify the method that reproduce expert judgment : the goal is to determine a criterion or a combination of criteria able to compare two compressed ultrasound images. We first present the results obtained by the fusion with a genetic algorithm, then the fusion with a support vector machine is explained.

5.1. Genetic algorithm

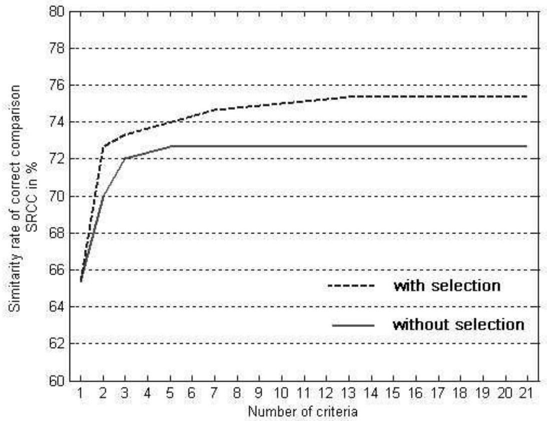
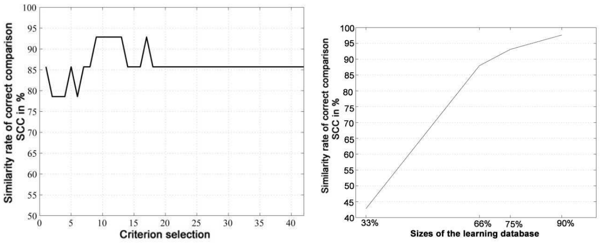


Figure 2. Fusion results with GA.

We used a population of size 20.000 and 1000 iterations. The mutation probability is set to 0.05 (that is to say that 5% of the 20.000 individuals will mute at each generation), the selection probability is set to 0.08 (8% of the individuals are selected to survive at the next generation) and the crossing-over is tested during 20 tries. Figure 2 presents the evolution of the similarity rate of comparison by merging different criteria. Given a number N of criteria to fusion, we take the N best criteria derived from the previous analysis. For example, by merging the three best criteria, we obtain a similarity rate of comparison equal to 72%. A second curve (in dotted line) shows the results of the fusion when using the criterion selection. In this case, instead of only determining the N linear coefficients, we also determine the best criteria to use. For N criteria to fusion, we have to determine $2.N$ values by using the previous genetic algorithm. In this case, we obtain a higher value of the similarity rate of correct comparison (75.3%).

5.2. SVM

The results of the criteria selection is shown on figure 3(a) : the learning database represents 90% of the data and the selection is done among the 21 criteria studied. The maximum rate of correct comparison is obtained for 5 criteria and is equal to $92,8\%$. Figure 3(b) presents the rate of good comparison regarding to the number of compression results used in the learning set. If 66% of the 150 examples is used for learning, we obtain a good comparison of 86.6%.



(a) (b)

Figure 3. Fusion results with SVM : (a) 95% of examples are used in the learning database.

6. Conclusion

In this paper we have presented a comparison of some evaluation criteria to quantify the quality of image compression. We implemented a psychovisual study involving 10 medical experts to identify the statistical criteria having the best behavior compared to the medical assessment. This study allows us to select three criteria among the 21 tested ones : neighborhood error (24 neighbors), contrast measure and spectral phase-magnitude error. The best similarity rate obtained with a single criterion is 65.3%. A genetic algorithm performs a linear combination of the criteria. The proposed criterion provides a higher value of the similarity rate of correct comparison (75.3%). A learning of the criteria evaluation is then done thanks to a support vector machine. The similarity obtained is of 92.8%. The optimal number of criteria is determined to 5. The performance of the new criterion (performed by the svm) provides an improvement of about 30% compared to the best criterion from our survey. This methodology improves significantly the possibility to evaluate the quality of ultrasound image compression results.

A prospect for this study is to use this criterion for the comparison of ultrasound image compression best fitted for a mobile robotized tele-echography system.

References

[1] Otelo C. Otelo : mobile tele-echography using an ultra-light robot. http://www.bourges.univ-orleans.fr/Otelo/, 2001-2004.

[2] Courreges F, Smith N, Poisson G, Vieyres P, Gourdon A, Szpieg M, Merigeaux O. Real-time exhibition of a simulated space tele-echography using an ultra-light robot. I SAIRAS juin 2001;.

[3] Turaga DS, Chen Y, Caviedes J. No reference psnr estimation for compressed pictures. Signal Processing Image Communication 2004;19:173–184.

[4] Fernandez-Maloigne C. Couleur numerique et psychometrie. Computer Art Journal 2004;1(1).

[5] Chen TJ, Chuang KS, Wu J, Chan SC, Hwang IM, Jan ML. A novel image quality index using moran i statistics. Physics in Medicine and Biology 2003;48:131–137.

[6] Lamminen H, Ruohonen K, Uusitalo H. Visual tests for measuring the picture quality of teleconsultations for medical purposes. Computer Methods and Programs in Biomedicine 2001;65:95–110.

[7] Kassai B, Leizorovicz A, M.Cucherat, S.Sonie, Boissel J, Gueyffier P, F., R. SN, S. G. A systematic review of the accuracy of ultrasound in the diagnosis of asymptomatic deep venous thrombosis : preliminary results. Journal of Thrombosis and Haemostasis juillet 2003;I(P1443):supplement 1.

[8] Delgorge C, Rosenberger C, Rakotomamonjy A, Poisson G, Vieyres P. Evaluation of the quality of ultrasound image compression by fusion of criteria with a support vector machine. EUSIPCO septembre 2005;.

[9] Avcibas I, Sankur B, Sayood K. Statistical evaluation of image quality measures. Journal of Electronic imaging avril 2002;11(2):206–223.

[10] Wall P. A Genetic Algorithm for Resource-Constrained Scheduling. Ph.D. thesis, MIT, 1996.

[11] Rakotomamonjy A. Variable selection using svm-based criteria. Journal of Machine Learning Research Special Issue on Variable Selection 2003;3:1357–1370.

Address for correspondence:

Cecile Delgorge
Laboratoire Vision & Robotique
IUT de Bourges / Batiment recherche
63 av de Lattre de Tassigny / 18020 Bourges cedex / France
tel.: ++33-2-48-23-84-70
fax: ++33-2-48-23-84-71
Cecile.Delgorge@bourges.univ-orleans.fr