A FRAMEWORK FOR EVALUATING THE IMPACT OF COMPRESSION ON REGISTRATION ALGORITHMS WITHOUT GOLD STANDARD

Tristan Glatard[1,2], *Johan Montagnat*[1], *Xavier Pennec*[2]

1. Université de Nice Sophia-Antipolis / CNRS / I3S
2. INRIA Sophia-Antipolis, Asclepios project

ABSTRACT

An evaluation of the impact of lossy compression on rigid registration algorithms for medical images is proposed. Due to the lack of gold standard for many clinical problems, the framework relies on a statistical procedure that estimates a reference from a large set of uncompressed images. The robustness, repeatability and accuracy of registration algorithms can then be derived for each compression ratio. Results are obtained thanks to a grid technology handling the computation cost of the method. Experiments reveal that the impact of compression is quite negligible below a significant compression ratio if the registration algorithm has a good multi-scale handling. Beyond this threshold, feature-based methods are highly penalized.

1. INTRODUCTION

With the generalization of digital image acquisition and manipulation devices, an increasing number of medical images are archived in digital warehouses. Manufacturers provide DICOM compliant devices interfaced to local storage facilities and PACS. The emergence of multi-sites PACS and technologies such as data grids ease the archiving of medical data at a large scale. Furthermore, recent regulations show a trend for long term archiving of patient data. Given the tremendous amount of radiology data acquired daily in clinical centers (tens of TBytes per year) and the need for long term archiving, optimizing storage space is increasingly needed.

Image compression can lead to drastic data size reduction. Compression algorithms, such as the well known JPEG, have been included in the DICOM standard. Compressed data size significantly depends on the data itself and the target Compression Ratio (CR) defined as the ratio between uncompressed and compressed data size. Lossless compression ensures a perfect reconstruction of the compressed data but leads to the lowest CR: typically 2 for any binary data; in the range of 3.3 to 3.9 for the brain Magnetic Resonance Images (MRI) with a large black background considered in this study. Compression with loss can achieve much better CR but at the cost of approximative reconstruction. In the medical area, the use of lossy compression should be considered with care given the sensitivity of image-based diagnosis and knowing that it will be impossible to recover the original data. Most often, in the current practice, only lossless JPEG is considered to compress DICOM data, if any compression is applied at all.

A trade-off has to be found between efficient image archiving and the quality of archived data. In the literature, a growing interest for multi-dimensional medical data compression recently appeared [1, 2, 3]. The authors often let to the user the choice of the compression factor and therefore the image quality. Some recent studies show that a reasonable level of lossy compression may remain acceptable in clinical routine though. For instance, Raffy *et al* [4] made a quantitative evaluation of the impact of an increasing compression factor on the ability to detect pulmonary nodules. The study shows that the detection performance of solid lung nodules did not suffer until a compression ratio of 48. This kind of study needs to be performed case by case for different image modalities and different detection tasks.

Another important question is the impact of lossy reconstruction on automated medical image analysis procedures as for instance image segmentation and Registration Algorithms (RAs). The goal of a RA is to estimate a transformation enabling the resampling of a floating image onto the geometry of a reference image, so that both images are best superimposed. In this paper, we are particularly interested in the impact of lossy compression on rigid RAs (RAs only considering the translation and the rotation).

This kind of study is difficult because in most real registration problems, there is no ground truth (gold standard) to evaluate the results. Phantom or simulated images may be used to provide a reference, but it is very difficult to produce realistic enough images. An alternative is the *Bronze-Standard* statistical method described in [5]. This method enables the use of a real image data set and can therefore be used for different imaging modalities and different body regions. It is very computationally intensive though.

In section 2, we describe an experimental framework to estimate the impact of compression on rigid RAs' *accuracy*, *repeatability*, and *robustness*. It relies on the Bronze-Standard method, sketched in section 2.1. It is then applied to the clini-

This work was funded by the AGIR French research program "ACI Masse de données"`http://www.aci-agir.org`.

cal problem of the follow-up of brain radiotherapy. The quantitative results obtained considering four different rigid RAs are studied in section 3. Grid technologies are exploited to handle the computational cost of this evaluation. The procedure not only provides quantitative information on these specific algorithms but also a framework for further algorithms and image databases testing.

2. EVALUATING THE IMPACT OF COMPRESSION

The founding hypothesis of this evaluation framework is to consider the transformations obtained from the uncompressed images as the reference for the evaluation. In absence of ground truth, this reference can be statistically built by exploiting a large number of longitudinal image sequences to register and different RAs. It is then called a *bronze standard*. Our goal is to estimate to what extent the compression makes the registration results deviate from their original locus. The *robustness* can be quantified by the size of the basin of attraction of the right solution or by the probability to find the right transformation. *Repeatability* (or precision) accounts for the errors due to parameters of the algorithm (mainly the initial transformation) and to the finite numerical accuracy of the optimization algorithm. It only measures the deviation from the average value, *i.e.* it does not take into account systematic biases, which are often hidden. *Accuracy* measures the error with respect to the truth (which may be unknown).

2.1. Building the Bronze-Standard reference

On uncompressed images, the reference is built using the statistical Bronze-Standard method. For each sequence of n images to be registered, let us denote $\bar{T}_{i,i+1}$ ($i \in [1, n-1]$) the $n-1$ unknown (exact) transformations relating image i to the following one. In our case-study, a sequence will correspond to images of the same patient. The unknown transformation $\bar{T}_{i,j}$ ($i, j \in [1, n]$) relating any pair of images is obtained by properly composing the free parameters . The registration of all the possible image pairs by m different methods yields a set of observations $\{T^k_{i,j}\}$.

The Bronze-Standard method considers the exact transformations as hidden variables of an overestimated system: $n-1$ transformations have to be estimated whereas $m \times n \times (n-1)$ observations are available. The exact transformations are estimated as the ones that minimize the prediction error of the observations:

$$\{\bar{T}_{i,i+1}\} = \arg\min_{\{\bar{T}_{i,j}\}} \sum_{i,j,k} \min\left(\mu^2\left({T^k_{i,j}}^{(-1)} \circ \bar{T}_{i,j}\right), \chi^2\right)^2$$

$$\text{with} \quad \mu^2(R(\theta,n),t) = \frac{\theta^2}{\sigma_\theta^2} + \frac{\|t\|^2}{\sigma_{\mathrm{t}}^2} \qquad (1)$$

θ is the angle of rotation R and n is the unitary vector defining its axis. t is the translation vector of the transformation. This criterion is based on a robust distance on rigid transformations. It includes a χ^2 threshold value to detect outliers. The Mahanalobis norm μ is normalized by the variances σ_θ^2 and σ_{t}^2 of the observations that have to be properly estimated. Computing those variances is not straightforward because they are used as input parameters of the norm μ defined in equation (1). Measured variances are thus re-injected in the minimization procedure which is iterated until they converge towards a stable estimation. The larger the number of registered images, the more accurate the estimated bronze standard. It is important to use various algorithms to prevent the results from being systematically biased by a specific registration technique. Results over several patients are averaged to obtain more significant estimations.

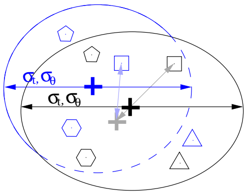

Fig. 1. Schema of transformations in the 2D plane. Each RA (identified by a given shape) produces transformations in the compressed (blue) and uncompressed (black) cases. Bronze standards are depicted by crosses. Ellipses represent covariances.

Fig. 1 diagrams the bronze standard notations. The reference for the evaluation is built exclusively from the uncompressed images (black items). Outliers are first removed thanks to the χ^2 test of equation 1. At this point, outliers may correspond to images for which the rigid assumption does not hold such as artifacted or strongly pathological ones. Those outliers are removed from the evaluation procedure because the transformations obtained at this step are going to be used to estimate the (bronze) reference of the whole study and they must be close enough to the truth. The bronze standard is then computed and the standard-deviations of the transformations (σ_θ and σ_t) are measured. They characterize the repeatability. The accuracy of each algorithm is finally obtained from the average distance between its measured transformations and the standard built from the remaining methods (gray cross and arrow on Fig. 1).

2.2. Evaluating the performance

The number of outlier transformations gives an estimation of the robustness of the algorithms w.r.t the compression. For each CR, outliers have to be identified by comparison to the transformations obtained from uncompressed images. Conversely, running the bronze standard procedure on the transformations obtained from the compressed images only could lead to a wrong detection of the outliers. Indeed, if compression leads to a similar bias for all the algorithms (*e.g.* by making a particular structure disappear from the images), the resulting transformation set could be considered as statistically consistent although it may be far from the truth. In the worst cases, the compression is likely to disturb the registration so much that algorithms converge towards the wrong local minimum. Those transformations cannot be included in the evaluation of the accuracy and repeatability because they would make it completely unstable and dependent only on a few number of outliers.

Outliers are detected with the χ^2 test included in the mean computation. Among the rejected transformations, a visual inspection has to be performed to determine whether they correspond to wrong local minima (when it is obvious that a manual registration can lead to a better result) or not. When a transformation is found to be in a wrong local minimum, the whole patient is removed. Otherwise, the absence of a specific algorithm for a given patient could bias the quantification of the accuracy of the remaining ones. To allow a fair comparison between the CRs, patients leading to a wrong local minimum in *any* of the CRs are excluded for the repeatability and accuracy studies. Indeed, it would be likely that high CRs would have been evaluated on less patients than lower ones, thus leading to potential artificial standard-deviation reduction. σ_θ and σ_t are also re-estimated from the uncompressed images after having removed those patients.

For each CR, the repeatability is given by the variances of the transformations obtained from the compressed images only. Repeatability is pictured by ellipses on Fig. 1. It is determined without performing any χ^2 test in the distance of equation (1): due to potential biases on compressed images, a transformation may be considered as an outlier for compressed images while it is an inlier for uncompressed images (and vice versa). This is the case of the blue triangle in Fig. 1.

The transformations obtained from the uncompressed images are considered as the reference for the evaluation. The accuracy of each algorithm is computed by measuring the mean distance of compressed transformations to the uncompressed reference. To avoid biases, the evaluated RA is excluded from the algorithms used to build the uncompressed reference. We should be aware that taking uncompressed images to build the reference does not imply that the accuracy is always lower for compressed images. It is for instance the case of the transformation of the algorithm depicted with a square on figure 1: compression has brought it closer to the bronze standard without compression. This could *e.g* be the consequence of a smoothing effect of the compression.

3. EXPERIMENTS

Experiments were made on a database of 65 images corresponding to 25 patients for which MRIs have been acquired at several time points to monitor the tumors growth. 126 registrations are required. This database has been compressed at CR=6,12,24,48 and 64, with the 3D-SPIHT algorithm [6].

Four different rigid RAs are used. Two of them, `Baladin` and `Yasmina` are intensity-based. `Baladin` uses a block-matching strategy and `Yasmina` uses a Powell optimization to maximize an intensity-based similarity measure. The two others (`CrestLines` and `PFRegister`) are feature-based (based on the crest lines extracted using the third derivatives of the images). The diversity of the RAs makes the procedure more robust against systematic biases of each algorithm.

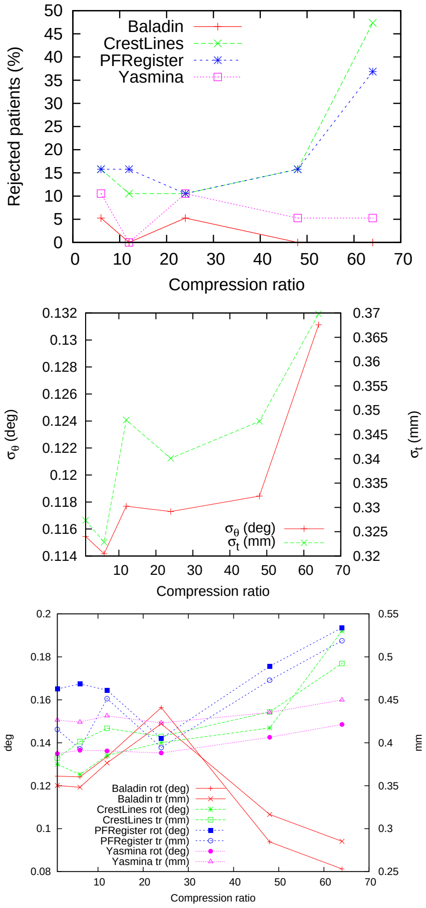

Fig. 2. Top: ratio of outlier patients w.r.t the CR ; Middle: mean variances of the transformations ; Bottom: accuracy of the algorithms.

A total of 3024 transformations (126 × 4 algorithms × 6 CRs) have to be computed. The total sequential execution time of this experiment is about 7.5 days. Thanks to a workflow-based grid implementation, the total duration of the experiment reduces to 18 hours on the shared EGEE production grid and to 4.2 hours on 60 dedicated CPUs spread over 3 clusters of a national grid. In practice, computation time reduction is very important as similar studies should be reproduced for every new image database or RA to evaluate.

3.1. Impact of compression on registration algorithms

Among the 25 patients of the database, 6 were removed by the χ^2 test on the uncompressed images (with χ^2=30). Those patients correspond to ones were the rigidity assumption is hardly valid, for instance because of high deformations in tumor areas. The reference was built from the uncompressed images of the remaining 19 patients.

The robustness of the algorithms was determined on those 19 patients (360 transformations per algorithm and per CR). The ratio of outlier patients is plotted on top of Fig. 2 for each algorithm. `Baladin` is the most robust method (at most 1 patient is rejected by the χ^2 test). `Yasmina` is also very robust, with 1 or 2 rejected patients. For those two algorithms, the behavior does not seem to be monotonic with respect to the CR: some patients are rejected for low CRs but are again accepted for higher ones and vice versa. The good robustness of those methods may be a consequence of their multi-scale strategy: they both use a pyramid of under-sampled images and initialize the input transformation of a given sampling level with the result of the upper one. The robustness of the crest-lines methods is lower, which can be explained by the extraction of the crest-lines at a single scale. The number of rejected patients is almost constant until a CR of 48, with 2 or 3 patients rejected. For a CR of 64, it highly increases up to almost 50% of rejected patients for `CrestLines`. At this CR, `PFRegister` performs a little bit better, with only 37% of rejected patients, which could be explained by a more robust matching of the crest-lines extracted by the previous one. The fact that feature-based methods are less robust to compression may come from the use of first to third order derivatives of the image, which are likely to be impacted by compression.

Among the patients rejected for at least one method, 4 were corresponding to wrong local minima for at least one CR. They were removed and the repeatability and the accuracy were evaluated on the remaining 296 transformations for each algorithm. The middle of Fig. 2 plots the evolution of the mean variances of the transformations. Despite a subtle improvement of 1% at CR=6, the main behavior is an impairment of 4 to 6% before a strong decline of 13% at CR 64.

The bottom of Fig.2 displays the accuracy of the algorithms w.r.t the CR. The accuracy of feature-based methods is highly reduced at CR=64. At this compression level, the mean error of `CrestLines` has increased by 48% for the rotation and 29% for the translation whereas the one of `PFRegister` has increased by 17% for the rotation and by 25% for the translation. `Yasmina` is quite insensitive to the compression: its mean error only increases by 10% for the rotation and by 5.5% for the translation. More surprisingly, after a brief rise until CR 24, the accuracy of `Baladin` is improving: at CR=64, it is 34% better than without compression for the rotation and 18.5% better for the translation. The fine behavior of `Baladin` and `Yasmina` can be explained by the fact that both algorithms include a multi-scale handling that may compensate the effects of potential noise introduced in the images. Moreover, in `Baladin`, only the most significant blocks (the ones with the largest standard deviations) are considered for the block-matching.

4. CONCLUSIONS

We presented an evaluation of the impact of the 3D-SPIHT compression algorithm on the rigid registrations of longitudinal images from a database of brain MRIs using 4 different registration methods. It is based on statistical method that is able to provide a *bronze* standard while no gold one is available on this particular clinical problem. Thanks to a grid implementation, we could perform the 3024 registrations required by the study in about 4 hours on dedicated resources whereas 7.5 days would have been needed on a single PC.

In our case, results show that the impact of 3D-SPIHT compression on robustness, repeatability and accuracy is quite negligible until a significant CR (64), in particular if the registration algorithm has a good multi-scale handling. Beyond this threshold, the methods based on crest-lines are highly penalized: half of the patients can be considered as outliers and their accuracy is lowered by 50%. Surprinsingly, compression improves the registration accuracy (up to 30% for `Baladin` on our setup) probably because the registration algorithm focuses on informative subsets of the image.

Thus, the bronze standard method is able to estimate the performances of rigid registration algorithms in the absence of gold standard and to evaluate the influence of parameters such as the compression ratio of the images. This kind of computation-intensive methods greatly benefit from grid technologies that speed up the experiments. Lossy compression does not seem to be problematic until a given compression ratio (48 in our study), which is coherent with the results found in [4] on another clinical problem. Evaluating the impact of other compression algorithms on different registration methods should be done to allow more general conclusions.

5. REFERENCES

[1] G. Menegaz and J.P. Thiran, “Lossy to lossless object-based coding of 3-D MRI data,” *IEEE Transactions on Image Processing*, vol. 11, no. 9, pp. 1053–1061, Sept. 2002.

[2] M. Unser, A. Aldroubi, and A. Laine, “Special Issue on Wavelets in Medical Imaging (editorial),” *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 285–288, Mar. 2003.

[3] A. Kassim, P. Yan, P. Yan, W. Lee, and K. Sengupta, “Motion Compensated Lossy-to-Lossless Compression of 4D Medical Images Using Integer Wavelet Transforms,” *IEEE Transactions on Information Technology In Biomedicine*, vol. 9, no. 1, pp. 132–138, Mar. 2005.

[4] P. Raffy, Y. Gaudeau, D. Miller, J.M. Moureaux, and R. Castellino, “Computer-aided Detection of Solid Lung Nodules in Lossy Compressed Multidetector Computed Tomography Chest Exams,” *Academic Radiology*, vol. 13, no. 10, pp. 1194–1203, Oct. 2006.

[5] T. Glatard, X. Pennec, and J. Montagnat, “Performance evaluation of grid-enabled registration algorithms using bronze-standards,” in *Medical Image Computing and Computer-Assisted Intervention*, Copenhagen, Oct. 2006, pp. 152–160.

[6] B.-J. Kim and W.A. Pearlman, “An Embedded Wavelet Video Coder Using Three-Dimensional Set Partitioning in Hierarchical Trees,” in *IEEE Data Compression Conference*, Snowbird, Utah, Mar. 1997, pp. 251–260.