Stochastic Sampling for Computing the Mutual Information of Two Images

M. Unser and P. Thévenaz

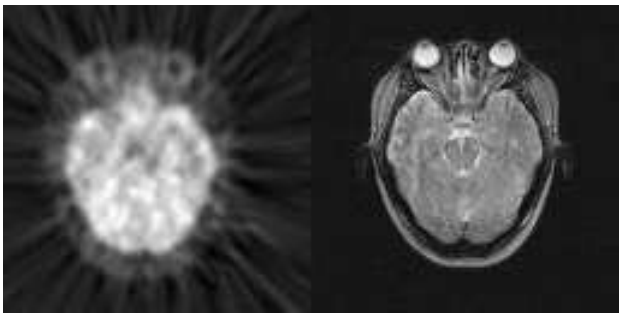

Fig. 1. Left: Positron Emission Tomography (PET) of a human brain, which depicts its functional activity. Right: Magnetic Resonance Image (MRI) of the same brain, which depicts its chemical composition.

Abstract— **Mutual information is an attractive registration criterion because it provides a meaningful comparison of images that represent different physical properties. In this paper, we review the shortcomings of three published methods for its computation. We identify the grid effect and the overlap problem as the most severe artifacts that these methods face, and propose a solution based on irregular sampling to solve for the grid effect. By implementing irregular sampling as stochastic sampling, we see that our solution covers the two problems at once, as the overlap problem ceases to be an issue, too.**

Keywords—**Entropy, spline, registration, multimodal alignment, Monte-Carlo sampling.**

I. Introduction

Since 1995, when it has been first proposed in [1] and [2], the registration criterion called "mutual information" (MI) has gained a wide acceptance in the medical community. Its purpose is to give a quantitative measure of the degree similarity between two images or volumes. By varying the relative position of the images (e.g., by rotating, translating, rescaling, and more generally warping one of the two), an optimal position is found such that MI is maximized. This position is said to give the best alignment.

The maximization of the grayscale correlation of two images plays a similar role but has a significant drawback: it results in the expected alignment only when the two images are close to being a copy of each other, up to noise and up to the deformation due to their misregistration; but correlation fails when the images are too different from a photometric point of view, for example when they are anticorrelated. By contrast, MI is impervious to such effects because it measures *how well*, on average, any given graylevel of one image can predict the graylevel of the other image at the same position. This measure of the *quality* of prediction is performed without ever attempting to give any *specific* prediction. For this reason, MI can cope with pairs of images for which there would exist no bijective relation between graylevels. This is typically the case when one attempts to register two images that were acquired by very different modalities, such as PET vs. MRI, a problem which is illustrated in Figure 1.

Corresponding author: Philippe Thévenaz, STI/BIO-E/LIB, Bldg. BM-Écublens 4.137, CH–1015 Lausanne VD, Switzerland, E-mail: philippe.thevenaz@epfl.ch, URL: http://bigwww.epfl.ch/

The discrete entropy[1] $H\{f\}$ of image f and the discrete mutual entropy $H\{f,g\}$ of images f and g are closely related to their discrete mutual information $I\{f,g\}$, because the latter is defined as

$$\begin{aligned} I\{f,g;\mathbb{F},\mathbb{G}\} &= H\{f\}+H\{g\}-H\{f,g\} &(1)\\ &= \sum_{\phi\in\mathbb{F}}\sum_{\gamma\in\mathbb{G}}\frac{1}{N}\,h(\phi,\gamma)\,\log_2\!\left(\frac{N\,h(\phi,\gamma)}{h_f(\phi)\,h_g(\gamma)}\right), &(2)\end{aligned}$$

where h is the joint histogram of both datasets, and where $h_f(\phi)=\sum_{\gamma\in\mathbb{G}}h(\phi,\gamma)$ and $h_g(\gamma)=\sum_{\phi\in\mathbb{F}}h(\phi,\gamma)$ are the marginal histograms of the joint histogram h. The normalizing factor $N=\sum_{\phi\in\mathbb{F}}\sum_{\gamma\in\mathbb{G}}h(\phi,\gamma)$ is the total of all histogram entries, and ϕ, γ, run over the discrete sets of graylevel intensities $\mathbb{F}$, $\mathbb{G}$, of images f, g, respectively. We note that $\mathbb{F}$ and $\mathbb{G}$ participate to the definition of I; it would therefore be meaningless to compare $I\{\cdot,\cdot;\mathbb{F},\mathbb{G}\}$ to $I\{\cdot,\cdot;\mathbb{F}',\mathbb{G}'\}$ if $\mathbb{F}\neq\mathbb{F}'$ or $\mathbb{G}\neq\mathbb{G}'$. This is one of the reasons for which Equation 2 should never be thought of as a quantized version of the continuous form of the mutual information $\bar{I}=\iint p(\phi,\gamma)\,\log_2\!\left(\frac{p(\phi,\gamma)}{p_f(\phi)\,p_g(\gamma)}\right)\mathrm{d}\phi\,\mathrm{d}\gamma$.

In this paper, we concentrate on the computation of the joint histogram h and leave aside many aspects related to the optimization of I, to the specification of the geometric transformation $\mathbf{g}$ used to perform registration, and to the interpolation required for applying $\mathbf{g}$. It would appear at first that computing a histogram is a trivial operation, but it happens that MI exhibits a surprising sensitivity to the specifics of histogram computation—the Devil is in the details. Arbitrary choices (e.g., interpolation model, choice of the samples used in estimating h) are a necessary ingredient of any practical implementation of a registration method based on MI. In terms of information, the uncontrolled introduction of arbitrariness may have severe effects. In particular, we show experimentally that computing h by sampling the images on a regular lattice leads to undesired artifacts. By resorting to stochastic sampling, we effectively remove some degree of arbitrariness (the fact that the sampling was regular) and are able to mitigate these effects.

II. Previous Work

Three major ways to compute h have been published in the context of MI. The authors of [3] adopt a two-tiered probabilistic view where the entropy is seen as the expectation—computed as an empiric average—of the logarithm of a probability density, and where this same probability density is estimated by a superposition of Gaussian densities. Their final expression for the joint entropy of image $f(\cdot)$ and of transformed image $g(\mathbf{g}(\cdot))$ is

$$\begin{aligned} H\{f,g(\mathbf{g})\} &= -\sum_{\phi\in\mathbb{F}}\sum_{\gamma\in\mathbb{G}}p(\phi,\gamma)\,\log_2 p(\phi,\gamma)\\ &\approx \mathbf{E}\{-\log_2 p(\cdot,\cdot)\}\\ &\approx -\frac{1}{N_A}\sum_{\mathbf{x}_i\in\mathbb{A}}\log_2(p(f(\mathbf{x}_i),g(\mathbf{g}(\mathbf{x}_i)))),\end{aligned}$$

with

$$\begin{aligned} p(\phi,\gamma) \approx \frac{1}{N_B}\sum_{\mathbf{x}_j\in\mathbb{B}}\frac{1}{2\pi\sqrt{|\det(\Psi)|}}\exp\!\Big(-\frac{1}{2}\begin{pmatrix} f(\mathbf{x}_j)-\phi\\ g(\mathbf{g}(\mathbf{x}_j))-\gamma\end{pmatrix}^{\top}\\ \cdot\,\Psi^{-1}\begin{pmatrix} f(\mathbf{x}_j)-\phi\\ g(\mathbf{g}(\mathbf{x}_j))-\gamma\end{pmatrix}\Big).\end{aligned}$$

[1]By "entropy", one must understand the entropy of the data graylevels.

We observe that two independent sets of random samples are used: $\mathbb{A}$ and $\mathbb{B}$, with $N_A = \text{card}(\mathbb{A})$ and $N_B = \text{card}(\mathbb{B})$. In addition, we observe that no discrete set of intensities $\mathbb{F}$ or $\mathbb{G}$ is explicit; this is a hint to the fact that this approach approximates $\bar{I}$ rather than I. We note also that, contrary to the claim made in [3], the use of Gaussian densities has the unexpected technical consequence that the marginal probability density of the reference image is made to depend on the geometric transformation. This is unfortunate as, obviously, the true probability density of an image that never changes should itself never change. Finally, MI is computed according to Equation 1. This approach has not been implemented often, perhaps in reason of practical difficulties: for example, the computation of $p(\phi, \gamma)$ is costly because Gaussian densities have an infinite extension, and N_B has to remain modest, which leads to a bad approximation of $\bar{I}$. In the present paper, we call this approach "Gaussian Parzen" (GP).

The authors of [4] propose a very different approach that they call "Partial Volume" (PV). In their approach, an explicit joint histogram is constructed as a sum of independent contributions, while no explicit transformed image needs be built. Let $\mathbf{g}$ be the geometric operation applied to image g; then, the pixels $f(\mathbf{k})$ of the reference image are aligned with $g(\mathbf{g}(\mathbf{k}))$. If one would produce $g(\mathbf{g}(\mathbf{k}))$ by linear interpolation, one would have to compute a weighted sum of intensities $g([\mathbf{g}(\mathbf{k})] + \boldsymbol{\Delta}\mathbf{1})$, where $[\cdot]$ indicates rounding to the nearest integer, and where $\boldsymbol{\Delta}\mathbf{1}$ is a symbolic representation of the coordinate offsets needed to reach some close-range neighborhood of $[\mathbf{g}(\mathbf{k})]$. The PV method puts emphasis on the weights of this sum. More precisely, the joint histogram can be expressed as [5]

$$h(\phi,\gamma) = \sum_{\mathbf{k}_1 \in \mathbb{D}} \delta(f(\mathbf{k}_1) - \phi) \sum_{\mathbf{k}_2 \in \mathbb{D}} \delta(g(\mathbf{k}_2) - \gamma)\, \beta^1(\mathbf{g}(\mathbf{k}_1) - \mathbf{k}_2),$$

where $\beta^n(\mathbf{x})$ is the tensor-product B-spline of degree n. In this expression, we observe that only terms of the form $g(\mathbf{k})$ are computed, but that no term of the form $g(\mathbf{g}(\mathbf{k}))$ appears. Consequently, for this approach to the computation of h to be well-behaved, it is necessary that $\mathbb{D} \subseteq \mathbb{Z}^q$, with $q = 2$ for images, and $q = 3$ for volumes, respectively. Meanwhile, the discrete set $\mathbb{F}$ must be identical to the range of the quantized image f, same for $\mathbb{G}$ and g. (Some freedom in the specification of $\mathbb{F}$ and $\mathbb{G}$ can be recovered by re-quantizing the images.) Finally, MI is computed according to Equation 2.

In [6], we have proposed another method that is based on Parzen windows. We bypass the expectation process of GP and replace the Gaussian weight functions by B-splines. This allows us to produce a simple expression of the gradient $\partial I / \partial \mathbf{g}$. Contrary to GP, our expression is exact. Moreover, the marginal histogram of the reference image f does not depend anymore on the transformation $\mathbf{g}$. Our joint histogram is computed as

$$h(\phi,\gamma) = \sum_{\mathbf{x} \in \mathbb{D}} \beta^3(f(\mathbf{x}) - \phi)\, \beta^3(g(\mathbf{g}(\mathbf{x})) - \gamma). \tag{3}$$

We observe that interpolation of g is required since terms of the form $g(\mathbf{g}(\mathbf{x}))$ do appear. To avoid having to interpolate f, the condition $\mathbb{D} \subseteq \mathbb{Z}^q$ is assumed in [6]. Finally, MI is computed according to Equation 2 with explicit $\mathbb{F}$ and $\mathbb{G}$; these sets can be arbitrarily chosen. In the present paper, we call this approach "Spline Parzen" (SP).

III. Grid Effect

We now perform a simple experiment to investigate the behavior of MI. We take the classical (512×512) Lena image in the role of both f and g, and let the transformation $\mathbf{g}(\cdot; 0, y)$ be a vertical translation of y pixels. We let $\mathbb{D}$ be such that no border effect occurs, by ignoring a 5 pixel-wide margin on each side of $f(\cdot)$ and $g(\mathbf{g}(\cdot; 0, y))$. We present the resulting plot of I vs. y in Figure 2. The top curve corresponds to the PV approach, while the bottom curve corresponds to SP. Both satisfy $\text{card}(\mathbb{F}) = \text{card}(\mathbb{G}) = 200$. We observe a well-defined MI maximum at $v = 0$, a translation for which $f = g(\mathbf{g})$. From this observation, many papers of the literature move on to other topics such as optimization or implementation issues, but we think there is more to say on the definition of MI and of its associated joint histogram.

We now change slightly the conditions of the experiment, by introducing a constant horizontal offset of 2 pixels. In other words, we consider $g(\mathbf{g}(\cdot; x, y))$ with $x = 2$ instead of $x = 0$. We present the resulting plot of I vs. y in Figure 3, where it is obvious that the slight

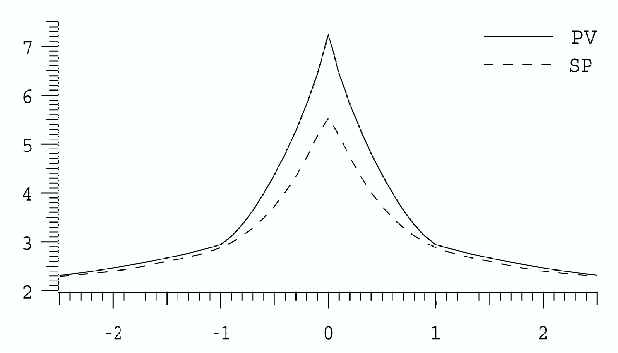

Fig. 2. Mutual information vs. translation of the Lena image for two different methods. Perfect registration is obtained for $y = 0$.

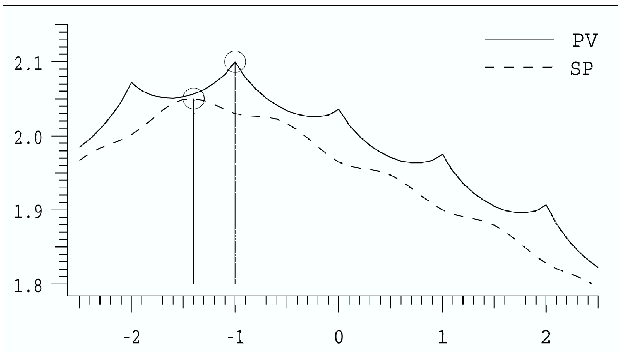

Fig. 3. Mutual information vs. translation of the Lena image for two different methods, with irreducible offset (see text).

change that we introduced produces dramatic effects. The top curve (PV) now exhibits many local extrema, which makes for a very difficult optimization problem. Moreover, the pattern of maxima clearly coincides with the grid of samples over which f is defined; as a matter of fact, the global maximum (for the range considered) is integer for the PV method, with $\hat{y}_{PV} = -1$. The SP method suffers less from these artifacts which are called "grid effect"; nevertheless, they are not totally suppressed. The global maximum $\hat{y}_{SP} \approx -1.4$ is likely to be more correct than $\hat{y}_{PV}$—we do not expect $\hat{y} = 0$ when $x = 2$—but is perhaps disturbed by the grid effect, too. The conditions of Figure 2 are exceptionally favorable, in the sense that the trajectory in the space of parameters—in the present case, the (x, y)-space—reaches the global optimum. In practice, we are never so lucky, and the conditions of Figure 3 are much more representative of a real registration task.

The specific interest of MI is to solve for registration problems where correlation methods fail; in other words, where f and g are very different. For example, translating the volume data of Figure 1 along an axis perpendicular to the displayed slice results in the even stronger grid effect of Figure 4. These curves have been obtained by setting again the arbitrary choice $\text{card}(\mathbb{F}) = \text{card}(\mathbb{G}) = 200$. The PV method has many local maxima which are strongly biased toward integer values, while the SP method is biased toward half integers. We see that the bias is lesser for SP than for PV, which reduces the risk of being trapped in a non-global optimum. It would be even better if there would be no bias at all; its removal is the topic of Section IV.

As large as it is, the grid effect is not the ultimate difficulty that classical methods encounter: difficulties could have been further reinforced, had we not applied a mask on $\mathbb{D}$ to keep constant the domain of overlap[2] between f and g. Allowing now $\text{card}(\mathbb{D})$ to vary proportionally to the area of overlap results in the curves of Figure 5,

[2]We were able to control the overlap in the present experiment because we knew in advance its extent, but this knowledge would not have been available in a real registration task.

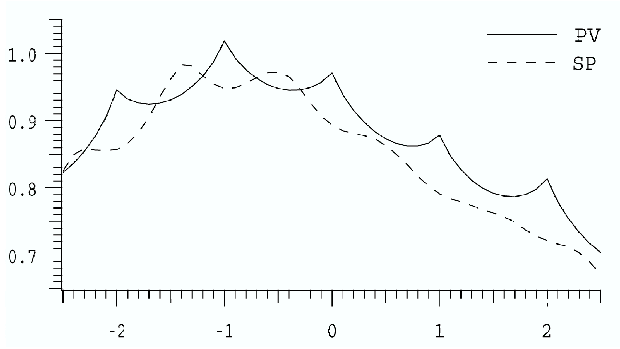

Fig. 4. Mutual information vs. translation of a pair of biomedical volumes for two different methods, with irreducible offset and dissimilar data (see text).

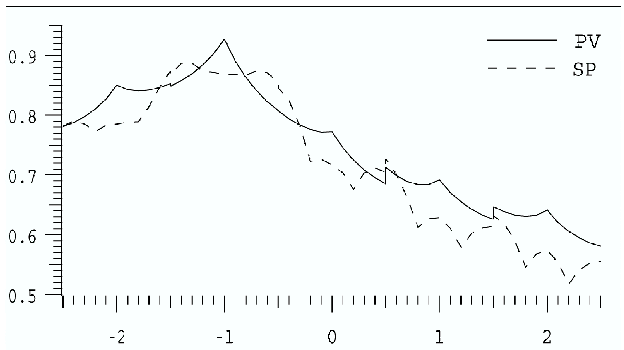

Fig. 5. Mutual information vs. translation of a pair of biomedical volumes for two different methods, with irreducible offset, dissimilar data, and varying overlap (see text).

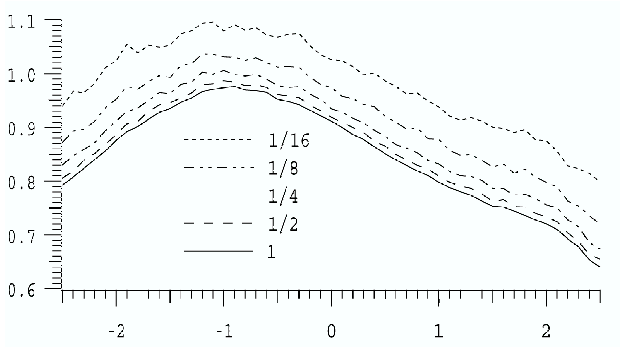

Fig. 6. Mutual information vs. translation of a pair of biomedical volumes for the proposed method. The number of samples card($\mathbb{D}$) decreases from bottom to top.

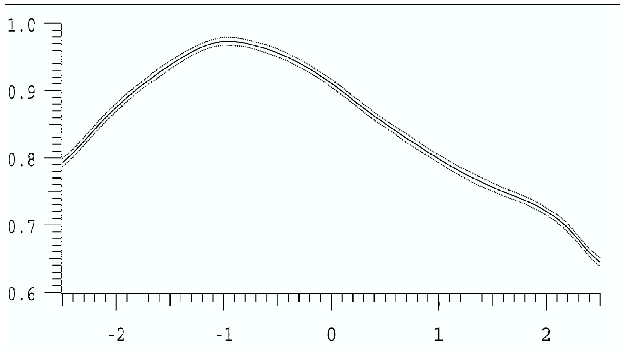

Fig. 7. Mutual information vs. translation of a pair of biomedical volumes when estimating the joint histogram by stochastic sampling. The dotted curve indicates ± 3 standard deviations (0.99998 percentile for a Gaussian distribution). The plain curve is the average of 100 realizations.

where we observe a severe loss of regularity—the overlap problem—in addition to the grid effect. The curves are not continuous anymore, which makes for a very difficult optimization problem. We propose in Section V a solution to the overlap problem that restores continuity.

IV. Irregular Sampling

We propose to remove the assumption $\mathbb{D} \in \mathbb{Z}^q$ in Equation 3. The claim of this paper is that irregular sampling does away with the grid effect. Thus, irregular sampling is extremely beneficial to the robustness and accuracy of registration; the cost is that interpolation is needed to compute $f(\mathbf{x})$, in addition to $g(\mathbf{g}(\mathbf{x}))$. We substantiate our claim by producing the bottom curve of Figure 6, where the conditions are the same as those of the SP curve of Figure 4, but for the fact that the samples $\{\mathbf{x}\} = \mathbb{D}$ are now realizations of a uniform distribution that covers at least the common support of f and $g(\mathbf{g})$. Those realizations of $\mathbf{x}$ that fall outside the common support are rejected. Obviously, the bottom curve of Figure 6 does suffer much less from non-global maxima than the curves of Figure 4 and is easy to optimize, particularly near the expected global optimum.

V. Overlap and Stochastic Sampling

Stochastic sampling offers the freedom to specify an arbitrary number of samples $N_D = \text{card}(\mathbb{D})$. On one hand, reducing the number of samples can accelerate the computations. On the other hand, since we have introduced a random process, our MI criterion is not deterministic anymore; a reduction in N_D leads to a greater variance VAR$\{I\}$, and to potential aliasing. Figure 6 shows what happens if we let the nominal N_D be reduced by a factor $\{1, 2, 4, 8, 16\}$, from bottom to top—The bottom curve has been built with the same nominal number of samples as in Figure 4, which corresponds to critical sampling (1 sample per voxel, on average). We see that the change in variance is not the only effect; the amount of mutual information (i.e., the measure of the quality of prediction) increases when reducing N_D. This has for consequence that we cannot allow for a prestored set $\mathbb{D}$ of coordinates, as comparing $I\{f, g(\mathbf{g}_1)\}$ to $I\{f, g(\mathbf{g}_2)\}$ would be meaningless because of a *systematic bias* if $\mathbf{g}_1$ and $\mathbf{g}_2$ are such that $N_{D_1} \neq N_{D_2}$—that is, if there is a change of overlap.

We have estimated the empirical standard deviation of MI for the bottom curve of Figure 6 by computing 100 realizations of this curve. We present the result in Figure 7, where we see that the departure from the average curve is very small and that its dependence on the actual value of I is weak. The largest standard deviation over $y \in [-2.5, 2.5]$ is only $\max(\sigma) = 0.0024$. Therefore, although non-deterministic, our proposed method succeeds in dramatically reducing the grid effect. In addition, it gracefully avoids the difficulties associated with the bias which other methods have to face when a change of overlap occurs. Finally, we note that it would be incorrect to infer from the partially jagged appearance of the curves of Figure 6 that their gradient $\partial I / \partial \mathbf{g}$ is erratic. In our case, an analytical form of this gradient is available; we use this form and we never estimate $\partial I / \partial \mathbf{g}$ by a finite-difference approach. This leads to a gradient that is the realization of a random variable (which exhibits a small variation), as opposed to the "gradient" of the realization of a random variable (which would be erratic and meaningless).

VI. Conclusion

Mutual information is a promising criterion for the registration of biomedical datasets. At its core, it relies on the computation of a joint histogram. Traditional implementations perform this computation by a regular sampling of the data, which is plagued by undesirable artifacts called "grid effect". Traditional implementations also suffer from another systematic bias, unrelated to the grid effect, which is referred to as the "overlap problem". We propose here to use irregular sampling to suppress the grid effect. By realizing this irregular sam-

pling as a random process, we are able to solve the overlap problem at the same time. We substantiate our claims by dramatic experimental evidence.

The proposed method is free from grid effect and from overlap effect. The gradient $\partial I/\partial \mathbf{g}$ can be computed easily and without approximation, because we use Parzen windows which are based on B-splines, are differentiable, and have a finite support. It can also easily be shown that the marginal histogram of the reference image f does not depend on the transformation. We expect that this combination of favorable properties will sensibly enhance the accuracy and robustness of the registration task.

References

[1] A. Collignon, D. Vandermeulen, P. Suetens, and G. Marchal, "3D Multi-Modality Medical Image Registration Using Feature Space Clustering," in *Proceedings of the Fourteenth International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine*, Nice, France, April 3-6 1995, pp. 195–204.

[2] P.A. Viola, *Alignment by Maximization of Mutual Information*, Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, March 1995.

[3] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-Modal Volume Registration by Maximization of Mutual Information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[4] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality Image Registration by Maximization of Mutual Information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, April 1997.

[5] R. Œuvray and M. Bierlaire, "On the Differentiability of the MI Estimator for Medical Image Registration," Tech. Rep. RO-020403, Swiss Federal Institute of Technology Lausanne, Lausanne VD, Switzerland, 2002, FSB/IMA/ROSO.

[6] P. Thévenaz and M. Unser, "Optimization of Mutual Information for Multiresolution Image Registration," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2083–2099, December 2000.