Feature-space Mutual Information for Multi-modal Signal Processing with Application to Medical Image Registration

Torsten Butz and Jean-Philippe Thiran

Signal Processing Institute,
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
http://ltswww.epfl.ch/~thiran
JP.Thiran@epfl.ch

ABSTRACT

A relatively large class of information theoretical measures, such as mutual information or normalized entropy, have been used in multi-modal medical image registration. Even though the mathematical foundations of the different measures were very similar, the final expressions turned out to be surprisingly different. Therefore one of the main aims of this paper is to enlighten the relationship of different objective functions by introducing a mathematical framework from which several known optimization objectives can be deduced. Furthermore we will extend existing measures in order to be applicable on image features different than image intensities and introduce "feature efficiency" as a very general concept to qualify such features.
The presented framework is very general and not at all restricted to medical images. Still we want to discuss the possible impact of our theoretical framework for the particular problem of medical image registration, where the feature space has traditionally been fixed to image intensities. Our theoretical approach though can be used for any kind of multi-modal signals, such as for the broad field of multi-media applications.

1 INTRODUCTION

The signal processing community has recently been paying an increasing attention to integrated approaches for dealing with multi-modal signals. In particular the use of information theoretic quantities, such as mutual information, has had a big success. For example the medical imaging community is very reliant upon mutual information to parametrically register multi-modal medical images [1, 2]. But also other applications, such as audio-video (multi-media) processing, started to benefit from integrating different signals which are physically of completely different nature and to explore their mutual (but unknown) relationship [3].

In this paper, we describe our recent developments in multi-modal signal processing. They are in fact closely related to the approach of information theoretical feature extraction and selection for classification [4]. Therefore we will start with a short review of this topic, before transposing it into the framework of multi-modal signals. Using Fano's inequality [5] and the data-processing inequality [6], we derive a probabilistic reason for using mutual information for multi-modal signal processing. This information theoretical framework shows that the restriction to a particular kind of signal features (such as gray levels for multi-modal medical images) can naturally be abandoned. In fact the presented information theoretical derivation indicates clearly that we can very easily build multi-modal algorithms which automatically select and extract the optimal elements within a predefined family of features.

In order to get a more intuitive feeling and interpretation about the developed approach, we will describe some of its possible implications for multi-modal medical image registration. For example we will show that normalized entropy [7], an overlap-invariant entropy measure for multi-modal medical image registration, can be seen as a particular case of our framework. This gives a more general explanation on when to use mutual information and when to use normalized entropy, also for applications outside the medical imaging community.

2 Why mutual information for multi-modal signals?

Our mathematical derivation is highly related to information theoretical feature extraction and selection for classification. Therefore we first want to recall the justification to use mutual information in this field. Afterwards we present our own derivation in the case of multi-modal signals which will lead to a probabilistic interpretation of mutual information in the context of multi-modal signals.

2.1 Fano's Inequality for Classification

As shown in fig. 1, the task of classifying a signal into a set of classes can be modelled by a Markov Chain [4].

It is interesting to interpret classification as a Markov chain $C \rightarrow X \rightarrow F \rightarrow \widehat{C}$ as Fano's inequality [5]

*This work is supported by the Swiss National Science Foundation under grants number 21-55580.98 and 2053-064947.01

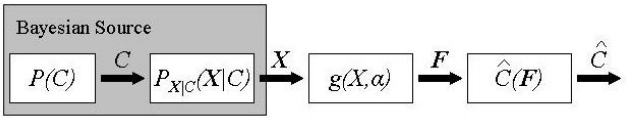

Figure 1: Learning optimal features for classification with examples can be mathematically interpreted as a Markov chain [4]. C represents the random variable of the learning sample of the classes and X are the associated observations generated by its conditional probability density function $P_{X|C}(X|C)$. The features F are extracted from X with the feature extractor $g(.,\alpha)$ and are used to estimate the output $\widehat{C}$ of the classifier.

gives a lower bound of the error probability of missclassification $P_e = Pr(C \neq \widehat{C} = \widehat{C}(F))$ [4]:

$$\begin{aligned} P_e &\geq \frac{H(C|F) - H(P_e)}{\log(|\Psi| - 1)} \\ &\geq \frac{H(C) - I(C,F) - 1}{\log|\Psi|}, \end{aligned} \tag{1}$$

where C is a random variable (RV) modelling the learning sample of the classes. X is the RV of the observations from the Bayesian source and is conditioned on the discrete RV C. F is a RV representing the features extracted from the initial RV X with a feature extractor $g(.,\alpha)$, characterized by α. Finally $\widehat{C}$ is the RV modelling the probability distribution of the output of our classifier. $H(.)$ is the Shannon entropy of a RV, $I(.,.)$ is the Shannon mutual information [8] of a pair of RVs and $|\Psi|$ is the number of elements in the range of C (e.g. for classification the number of classes).

No hypothesis about the specific classifier has been taken for eq. 1. So the inequality just quantifies how well we can classify at the best when using a specific feature space F. Unfortunately it is impossible to find an upper bound for the probability of error when we use Shannon's expression of entropy [9]. Hence the best we can do is minimizing this lower bound, so that a suitable classifier can do well. Observing that $H(C)$ as well as $\log|\Psi|$ is constant, we have to maximize the mutual information $I(C,F)$ in order to minimize this lower bound.

Therefore in the sense of error probability, we have to select/extract those features F that contain the largest information about the classes C.

2.2 Fano's Inequality for Multi-modal Signal Processing

We want to show that we can associate Markov chains with multi-modal signals as well. This allows us to build feature related quality measures for multi-modal signal processing algorithms in the same sense as the probability of error of eq. 1.

In fig. 2 we schematically show the realization of two signals of different modality from the same physical scene. Sampling the obtained continuous signal into a discrete representation can be modelled by a RV S which is uniformly distributed over the set of possible measurement "positions". Or more specifically, the RV S generates the possible sampling positions of the signals: in an image the pixel/voxel coordinates and in a video sequence the time coordinate of the frames. For instance a 3D image contains $n_x \times n_y \times n_z$ voxels, the probability that a certain measurement had been performed at coordinates (i,j,k) is $P(s=(i,j,k)) = \frac{1}{n_x \cdot n_y \cdot n_z}, \forall s \in \Im$ ("for all voxels in the image").

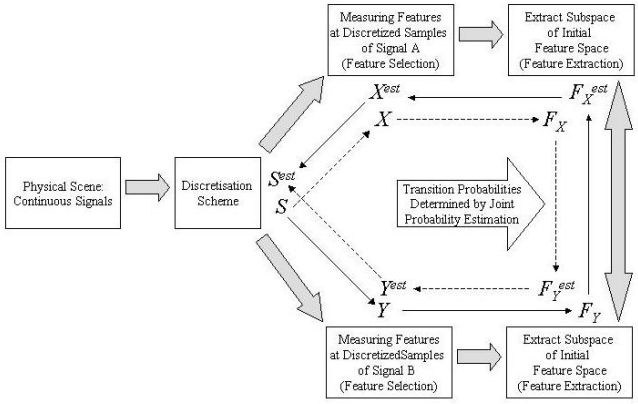

Figure 2: Markov chains can be built from a pair of multi-modal signals. They use the joint probability between the final features (F_X and F_Y) as the connecting block.

This initial random variable S can be seen as the starting block of two related Markov chains (Fig. 2): starting from S we can model the specific measurement X (resp. Y) of the initial signals as RVs conditioned on the outcome of S. What is exactly measured is the feature selection step. For instance in an image, for each sample position (i,j,k) generated from the RV S we can measure the intensity at that position, but also the gradient, Gabor response, or why not the image intensity at position $(i+i_0, j+j_0, k+k_0)$, i_0, j_0 and k_0 being constants, etc. We will come back to this in section 3. Furthermore S gives a physical correspondence between X and Y as we measure both signals at the same position in the sampling space of S[1]. Obviously X and Y can also model multi-dimensional feature spaces, which might ask for an additional feature extraction step. This means we project the measured features into lower dimensional sub-spaces of X and Y. Such sub-spaces are again RVs and we denoted them F_X and F_Y in fig.

[1] Sometimes S is not identical for both signals. For example two images of different modality might have different dimensions. For such cases we just want to make reference to interpolators which can build the bridge between the two respective sampling spaces [10], [11].

2. The physical correspondence of the measurements X and Y, resp. F_X and F_Y (both are conditioned on the same sampling RV S), makes it possible to link the two signals probabilistically through a joint probability distribution [12].

Interpreting the realization of multi-modal signals as a stochastic process as described above allows the construction of two related Markov chains:

$$S \rightarrow X \rightarrow F_X \rightarrow F_Y^{est} \rightarrow Y^{est} \rightarrow S^{est} \quad (2)$$
$$S \rightarrow Y \rightarrow F_Y \rightarrow F_X^{est} \rightarrow X^{est} \rightarrow S^{est}. \quad (3)$$

Just as for the case of classification, we can find lower bounds of the probabilities of error $P_{e1} = Pr(S^{est} \neq S)$ (Markov chain eq. 2) and $P_{e2} = Pr(S^{est} \neq S)$ (Markov chain eq. 3) that the final outcomes of the Markov chains S^{est} (the estimated value of S) are not the initial values S. We get respectively for eq. 2 and eq. 3:

$$\begin{aligned} P_{e1} &= Pr(S^{est} \neq S) \\ &\geq \frac{H(S|Y^{est}) - H(P_{e1})}{\log(|\Psi| - 1)} \\ &\geq \frac{H(S|Y^{est}) - 1}{\log|\Psi|} \\ &= \frac{H(S) - I(S, Y^{est}) - 1}{\log|\Psi|} \\ &= \frac{\log|\Psi| - I(S, Y^{est}) - 1}{\log|\Psi|} \\ &= 1 - \frac{I(S, Y^{est}) + 1}{\log|\Psi|} \\ &\geq 1 - \frac{I(F_X, F_Y^{est}) + 1}{\log|\Psi|} \end{aligned} \quad (4)$$

and

$$\begin{aligned} P_{e2} &= Pr(S^{est} \neq S) \\ &\geq \frac{H(S|X^{est}) - H(P_{e2})}{\log(|\Psi| - 1)} \\ &\geq \frac{H(S|X^{est}) - 1}{\log|\Psi|} \\ &= \frac{H(S) - I(S, X^{est}) - 1}{\log|\Psi|} \\ &= \frac{\log|\Psi| - I(S, X^{est}) - 1}{\log|\Psi|} \\ &= 1 - \frac{I(S, X^{est}) + 1}{\log|\Psi|} \\ &\geq 1 - \frac{I(F_Y, F_X^{est}) + 1}{\log|\Psi|}. \end{aligned} \quad (5)$$

For a detailed derivation the reader is referred to [13].

The mutual informations $I(F_Y, F_X^{est})$ and $I(F_X, F_Y^{est})$ are determined form the same joint probability distribution estimated by non-parametric probability estimation [12] (for example joint histogramming). From the symmetry of mutual information it follows that both lower bounds are equal, so that minimizing them simultaneously equals maximizing the mutual information between the feature representations of the multi-modal signals.

In fact eq. 4 and 5 give simply a lower bound of the probability of making an error when mapping the sampling space of one signal into the sampling space of the second signal of a corresponding multi-modal couple. For example it estimates the minimal error probability when generating a magnetic resonance image from a computer tomography image or when estimating a video sequence (e.g. the speaker's mouth motion) from a speech signal. These probabilistic mappings are modelled as the Markov chains of eq. 2 and 3.

It is therefore clear that to select, within a family of features, those features that best capture the relationship between the two RVs it is necessary to chose those with the highest mutual information. The optimal feature selection therefore appears as an optimization problem.

2.3 Feature Efficiency

There exists one danger though when simply maximizing the mutual information in order to minimize the lower bounds of eq. 4 and 5. To visualize this danger, let us use that for any pair of random variables X and Y we have $H(X,Y) \geq I(X,Y)$ and $\frac{H(X)+H(Y)}{2} \geq I(X,Y)$ to weaken them:

$$\begin{aligned} P_{\{e1,e2\}} &\geq 1 - \frac{I(F_X, F_Y) + 1}{\log|\Psi|} \\ &\geq 1 - \frac{H(F_X, F_Y) + 1}{\log|\Psi|} \end{aligned} \quad (6)$$

and

$$\begin{aligned} P_{\{e1,e2\}} &\geq 1 - \frac{I(F_X, F_Y) + 1}{\log|\Psi|} \\ &\geq 1 - \frac{H(F_X) + H(F_Y) + 2}{2 \cdot \log|\Psi|}. \end{aligned} \quad (7)$$

Eq. 6 and 7 both indicate that the error bounds can be decreased by selecting those features that increase the marginal entropies $H(F_X)$ and $H(F_Y)$ without considering their mutual relationship (this is equivalent to maximizing the joint entropy $H(F_X, F_Y)$, as we also have $H(F_X, F_Y) \geq H(F_X)$ and $H(F_X, F_Y) \geq H(F_Y)$). This would result in adding superfluous information to the feature space RVs F_X and F_Y. What we really want though is selecting the features that selectively add the

information that determines the mutual relationship between the signals while discarding superfluous information. Mathematically we want to maximize the bounds of eq. 6 and 7 but also minimize the bounds of eq. 4 and eq. 5.

For this aim we define a *feature efficiency coefficient* which measures if a specific pair of features is efficient in the sense of explaining the mutual relationship between the two multi-modal signals while not carrying much superfluous information. The problem of efficient features in multi-modal signals is closely related to determining efficient features for classification. Our proposed coefficient $e(X,Y)$ of a pair of RVs X and Y is defined as follows:

$$e(X,Y) = \frac{I(X,Y)}{H(X,Y)} \in [0,1]. \tag{8}$$

Maximizing $e(X,Y)$ still minimizes the lower bound of the error probabilities, but also minimizes the joint entropy $H(X,Y)$ which results in maximizing the weakened bounds of eq. 6 and 7. Looking for features that maximize the efficiency coefficient of eq. 8 will therefore look for features which are highly related (large mutual information) but haven't necessarily much information (marginal entropy)[2].

Interestingly there is a functional closely related to $e(X,Y)$ that has already been widely used in multi-modal medical image processing, even though it's derivation was completely different. It was called normalized entropy $NE(X,Y)$ [7] and was derived as an overlap invariant optimization objective for rigid registration:

$$NE(X,Y) = \frac{H(X)+H(Y)}{H(X,Y)} = e(X,Y)+1 \in [1,2]. \tag{9}$$

The derivation was specific for medical image registration and arose from the problem that mutual information might increase when images are moved away from optimal registration when the marginal entropies increase more than the joint entropy decreases. This is equivalent to our mathematically derived problem of feature efficiency above, but for the special case of image registration. Obviously maximizing $NE(X,Y)$ of eq. 9 is equivalent to maximizing the efficiency coefficient of eq. 8.

2.4 Generalizing Feature Efficiency

We want to introduce a short chapter that should enlarge the vision of feature efficiency for multi-modal signals.

It is very interesting to note that in the early years of information theoretical multi-modal signal processing, joint entropy $H(.,.)$ was also an optimization objective of choice. Interestingly this statistic had to be minimized in order to get for example good registration. Looking at the deduced error bounds of eq. 4, 5 and particularly 6, one realizes that minimizing joint entropy does *not* minimize these error bounds of eq. 4 and 5. On the contrary, it actually maximizes the weakened bound of eq. 6 and therefore contradicts error bound minimization. The result were very "efficient" features, but with relatively large error bounds (e.g. mapping a black on a white image). This results for example in disconnecting the images during the registration process. We employed the same property in the previous chapter but only in combination with error bound minimization to separate the superfluous information in the signals from the predictive information.

These arguments are very general. Nevertheless they could have resulted in other definitions for feature efficiency than eq. 8, such as

$$e(X,Y) = \frac{I(X,Y)}{H(X)+H(Y)}, \tag{10}$$

$$e(X,Y) = \frac{I(X,Y)^{\frac{2}{3}}}{H(X,Y)^{\frac{1}{3}}}. \tag{11}$$

While the first example is a variant equivalent to eq. 8, as it simply uses the weakened inequality of eq. 7 instead of eq. 6, the second is an extension of $e(X,Y)$, that can be generalized as follows:

$$e_n(X,Y) = \frac{I(X,Y)^n}{H(X,Y)^{1-n}}, n \in [0,1]. \tag{12}$$

We call an element of this class of functions the *feature efficiency coefficient of order n*. The three cases of $n = 0$, $n = 1$ and $n = \frac{1}{2}$ represent:

- $n = 0$: We emphasize entirely on the feature efficiency without caring about the resulting lower bound of the error probabilities (minimizing joint entropy). The algorithm will always converge towards image representations where all the voxels of an image has been assigned the same single feature value.
- $n = 1$: We emphasize on minimizing the lower error bound without caring about the efficiency of the features (maximizing mutual information). The algorithm would converge towards an image representation where each voxel has been assigned a different feature value.
- $n = \frac{1}{2}$: We put equal emphasize on minimizing the lower error bound and on feature efficiency (maximizing normalized entropy).

The two objectives of on the one hand minimizing the lower error bounds and on the other hand maximizing

[2] Because of the range $[0,1]$ of $e(X,Y)$, this functional is sometimes called "normalized measure of dependence" [14].

feature efficiency are therefore contradictory. The user has to choose an appropriate order n of eq. 12 for a given problem. For example order $\frac{1}{2}$ has shown to be very interesting for medical image registration [7, 15]. In fig. 3 we show a quantitative sketch of feature efficiency for different orders of n.

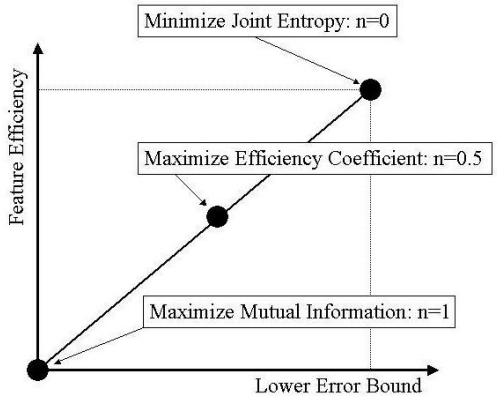

Figure 3: The sketch puts the efficiency coefficients for different orders n into a quantitative relationship. The contradictory optimization objectives of minimizing the lower error bound, but maximizing the feature efficiency have to be combined in a suitable way for a given problem. In the case of medical images, $n = \frac{1}{2}$ has shown to work fine, as it results into an optimization functional equivalent to normalized entropy [7].

3 Multi-Modal Medical Image Registration

We will now explicitly build the bridge to the particular problem of medical image registration. To do this we simply interpret image registration as a particular case of feature selection. At first sight this might look quite strange, but is actually quite intuitive: which geometric transformation selects the features (e.g. image intensities) for the voxel coordinates of the floating image so that the error bound of eq. 4 and 5 is minimized, respectively the efficiency coefficient of eq. 8 is maximized? This is exactly the feature selection example that we introduced in section 2.2.

The Markov chain model for multi-modal signals (eq. 2 and 3) revealed several interesting points about medical image registration, which are summarized as follows:

- Multi-modal image registration can be looked at as a parametric feature selection to minimize the lower error bounds of the Markov chains of eq. 2 and 3.
- Using maximum MI as the feature selection objective function is equivalent to minimizing the lower bounds of eq. 4 and 5.
- Using maximum normalized entropy as the feature selection objective function is equivalent to determining efficient features in the sense of eq. 8.
- The choice of image intensities for image registration might be appropriate in some cases, but the described framework allows the use of more complex features or feature vectors.

Mathematically we can formulate the general framework of feature efficiency (eq. 8) for multi-modal image registration as follows:

$$[\vec{t}^{\,opt}, \vec{F}_X^{opt}, \vec{F}_Y^{opt}] = \arg \max_{\vec{t} \in \mathbf{R}^m, \vec{F}_X \subset \mathcal{F}_X, \vec{F}_Y \subset \mathcal{F}_Y} e(\vec{F}_X, T_{\vec{t}}(\vec{F}_Y)), \tag{13}$$

where $\mathcal{F}_X$ and $\mathcal{F}_Y$ are the initial feature space representations of the reference and floating image, from which we want to extract and select the most efficient features (i.e. which have the largest efficiency coefficient). $\vec{t}$ is the vector of transformation parameters of the geometric transformation $T_{\vec{t}}$ of the floating image. m is determined by the particular transformation model, e.g. m is 12 for affine and 6 for rigid-body registration. It has just an instructive reason that we write the transformation $\vec{t}$ separately from $\vec{F}_Y$. In fact the transformation space $\mathbf{R}^m$ of $\vec{t}$ spans a sub-space of the much larger feature space $\mathcal{F}_Y$ of the floating image.

The possible applications of the general framework developed in chap. 2 might still appear in some obscurity. Therefore we would like to test the approach with three clinically important applications[3]. We used the *BrainWeb* magnetic resonance simulator [17] to get the utilized datasets and to control the results for their accuracy.

3.1 Affine registration of multi-modal medical images

Multi-modal affine registration is a very important task for medical image registration as it gives a good initialization for several non-rigid registration algorithms [18, 19]. In this study, we registered 3D MR-scans onto a 3D CT reference image of another patient (inter-patient) and compared the intensity based MI and the MI between other features extracted from the images, namely the norm of the gradient of the images [20]. The results are shown in figure 4. For rigid registration the quality of the results is comparable, while for affine transformations the global maximum of intensity based MI might not correspond to a good registration and the presented feature space defines a much better result. The edgeness defined by the gradient emphasizes contours in the medical images while the intensity based MI overemphasizes the volumetric information in the scans and therefore risks to neglect finer but important features in the images. An example is the skull and the brain: the brain covers lots of volume while the human skull is

[3] In what follows, we used a massively parallel, multi-scale genetic optimization [16] scheme to maximize mathematical expressions of the form of eq. 13. Unfortunately we don't have space to describe the specific implementation in more detail.

a relatively fine but anatomically important structure. Therefore the intensity based registration favors the statistical matching of the brain. On the other hand the gradient based MI reflects the statistical presence of surfaces. As a result, the skull and the brain have about the same importance and a compromise for their fitting is obtained. Figure 4, in particular the images e/f), shows a significant improvement with this approach. Please refer to [20] for more details on this.

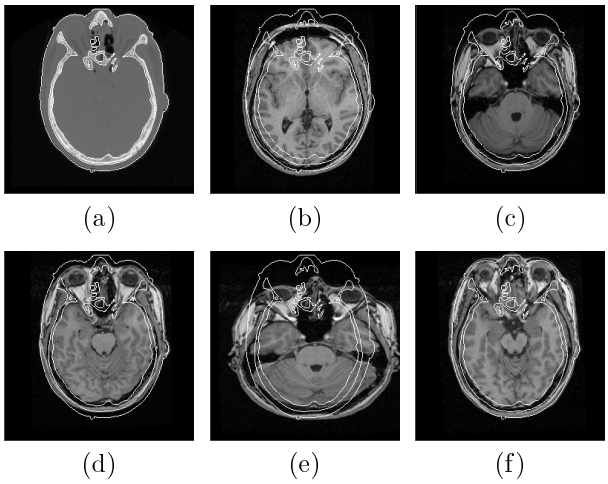

Figure 4: a) is one slice of the CT-target image. In b), the contours of the target image are superposed on the floating MR-scan. In c) and d) we see the results after a rigid optimization, when using resp. the intensity based MI and the edgeness MI. In e) and f) we show the corresponding results for affine registration. In e) and f) we recognize a significant improvement with the gradient based MI; resp. that the global maximum of intensity based MI doesn't correspond to good registration.

3.2 Registration and Quantification

Medical images are more or less noisy representations of the patient's anatomy. The noise has a negative impact on statistical image registration. Some approaches to minimize the influence of noise are based on initial filtering of the datasets (e.g. anisotropic filtering [21]), or even on anatomical segmentation to extract the information of the images that is really relevant for registration. In this example we therefore try a very naiv way to extract the representative anatomical information in the medical images while discarding the redundant noise. We use simple image intensity quantification which varies the number of bins for the joint probability estimation. But decreasing the number of bins obviously decreases the marginal entropies of the image representations, therefore simply maximizing the MI of eq. 4 and 5 is dangerous. We rather use the efficiency coefficient of the quantified images to find the optimal number of quantification intervals as well as the geometrical registration parameters. Mathematically we can write the

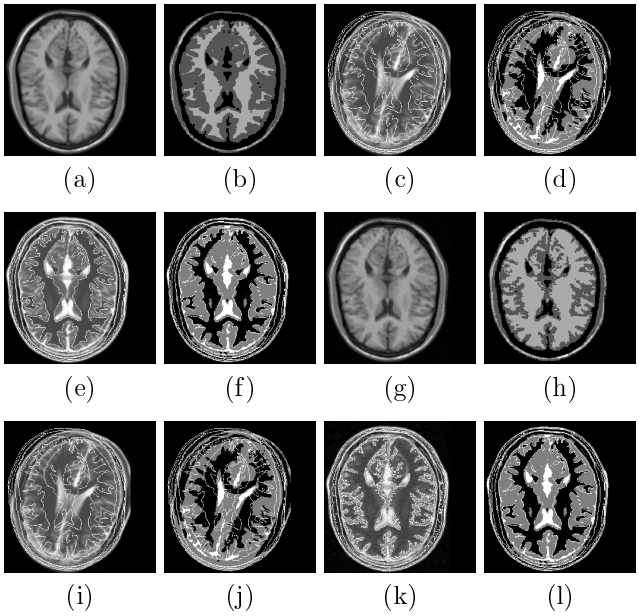

Figure 5: Image a) shows the reference image and c) the initial floating image. In e) we show the rigidly registered result. Images b), d) and f) show the quantized outputs of a), c) and e) with the optimal number of bins. Images g) to l) show an experiment equivalent to a) to f), but with noisier datasets. The contours of b), resp. h), are outlined in c) to f), resp. i) to l).

optimization objective as follows:

$$[\vec{t}^{\,opt}, n_X^{opt}, n_Y^{opt}] = \arg \max_{[\vec{t}, n_X, n_Y] \in [\mathbf{R}^m, \mathbf{Z}^+, \mathbf{Z}^+]} e(Q_{n_X}(X), T_{\vec{t}}(Q_{n_Y}(Y))), \quad (14)$$

where X and Y are the probability densities of the reference and floating image intensities respectively. n_X and n_Y are the number of bins used for the density estimation of X and Y and $\vec{t}$ is the vector of the parameters of the geometric transformation of the floating image. m is the dimension of $\vec{t}$ and is determined by the particular transformation model, e.g. for rigid-body we have 6 and for affine 12 parameters. Results for rigid registration are shown in fig. 5.

This shows nicely that the quantification task during the registration converges towards anatomical segmentation of the initial images. Therefore the feature efficiency coefficient of eq. 8 is not only capable of registering the images correctly, but of simultaneously extracting the anatomical information in the datasets. The labels associated to the different quantification intervals represent the anatomical information present in both initial images, while the noise is redundant information that can be discarded by the efficiency coefficient during the registration. This is equivalent to what several pre-processing algorithms aim to do (e.g. anisotropic filtering).

3.3 Image Registration with Bias-correction

Interventional imaging modalities suffer frequently from a large bias field. This makes image registration particularly difficult. Nevertheless it would be of particular interest to register pre-operatively acquired scans of different modalities onto the interventional datasets. In this section we want to show that the presented framework allows easily to register images with large bias fields. The approach simply combines minimum entropy bias-correction [22] with MI based image registration. From the developed theory, one can recognize immediately that MI is not appropriate for this task as minimizing entropy contradicts obviously the maximum MI principle of eq. 4 and 5. Therefore maximizing directly the mutual information would not correct the bias field even though the error bounds of eq. 4 and 5 would be minimized. Just as in chap. 3.2, this is a typical example of inefficient features. Rather than maximizing MI, we want to maximize the efficiency coefficient of the bias-corrected image intensities/features (eq. 8).

The resulting mathematical formalism can thereafter be written as follows: Let's $\vec{p} \in \mathbf{R}^{m_1}$ parameterize the polynomial bias-correction of [22], where m_1 is determined by the degree of the polynomials. Furthermore we have to determine the parameters $\vec{t} \in \mathbf{R}^{m_2}$ of the geometric transformation, where m_2 is the number of parameters that determines the transformation. In our specific application, we optimized over rigid-body transformations (m_2 equaled 6). Mathematically we have:

$$[\vec{t}^{opt}, \vec{p}^{opt}] = \arg \max_{[\vec{t},\vec{p}] \in \mathbf{R}^{m_1+m_2}} e(X, T_{\vec{t}}(P_{\vec{p}}(Y))), \quad (15)$$

where the parameters $\vec{p}^{opt}$ specify the optimal bias-correction and $\vec{t}^{opt}$ determines the optimal rigid transformation. Here X and Y are the probability densities of the image intensities. Fig. 6 presents the results.

We showed that registration of medical images with large bias-field fail when we use the initial images without any bias-correction. Thereafter we show that bias-correction can naturally be incorporated in the registration algorithm as a feature selection step. Feature efficiency converges thereafter nicely towards good registration and bias-correction.

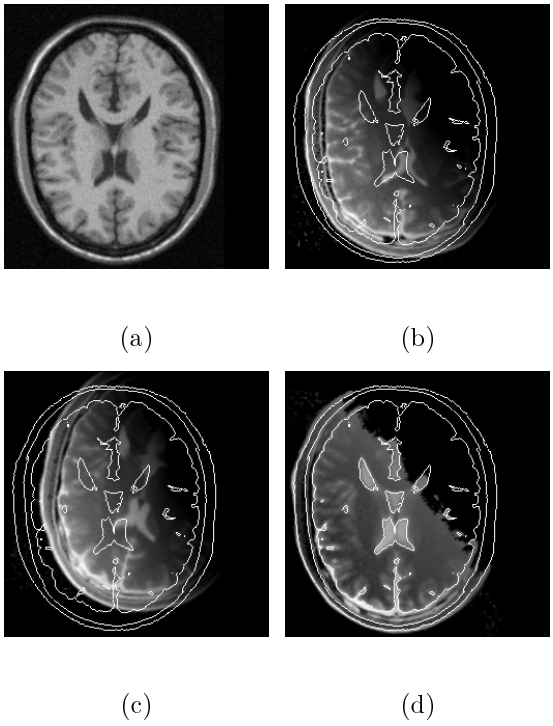

Figure 6: We rigidly registered the image of b) onto the reference image shown in a). c) shows the bad result without simultaneous bias-correction and d) shows good registration with simultaneous bias-correction.

4 Conclusion

Based on Markov chains, we introduced a very general model of multi-modal signals. Using information theoretical foundations, we derived a feature based lower error bound on mapping one signal into a corresponding second signal of different modality. Starting form this framework we show that normalized entropy is at the basis of a much more general concept than overlap invariant image registration, called feature efficiency. We discussed the implications of this framework on multi-modal medical image processing, keeping a close focus on image registration. Finally we applied the developed theory on three potentially valuable applications, both dealing with compensation of artifacts during the registration process. The first shows how the concept of feature efficiency can be used to suppress noise in the datasets to make the algorithms more reliable. The second example is showing how the normally large bias fields of interventional MR-scans can be corrected during the optimization to make registration of such heavily degraded images possible.

It's important to note that the presented general framework opens the door towards a wide range to further developments about multi-modal signal processing. Recent developments of this work in the field of multimedia (joint audio and video signal processing) can be found in [23]. Moreover it would be interesting to derive an upper bound of the error probabilities of eq. 4 and 5 or to incorporate spatial prior information into the proposed Markov chains.

References

[1] P. Viola and W.M. Wells III, "Alignment by maximization of mutual information," in *Fifth Int. Conf. on Computer Vision*, 1995, pp. 16–23.

[2] 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.

[3] Trevor Darrell, John W. Fisher III, Paul Viola, and William Freeman, "Audio-visual segmentation and the "cocktail party effect"," in *International Conference on Multimodal Interfaces*, 2001.

[4] J.W. Fisher III and J.C. Principe, "A methodology for information theoretic feature extraction," in *World Congress on Computational Intelligence*, March 1998.

[5] Robert M. Fano, *Transmission of Information: A Statistical Theory of Communication*, MIT Press and John Wiley & Sons, 1961.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[7] C. Studholme, D.J. Hawkes, and D.L.G. Hill, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognition*, vol. 32, pp. 71–86, 1999.

[8] C.E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[9] Deniz Erdogmus and José C. Principe, "Information transfer through classifiers and its relation to probability of error," in *Proceedings of the International Joint Conference on Neural Networks, Washington D.C., USA*, July 2001.

[10] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part I - Theory," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 821–833, February 1993.

[11] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part II - Efficient design and applications," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 834–848, February 1993.

[12] L. Devroye and L. Györfi, *Non-parametric Density Estimation*, John Wiley & Sons, 1985.

[13] Torsten Butz and Jean-Philippe Thiran, "Multimodal signal processing: An information theoretical framework," Tech. Rep. 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), 2002, http://ltswww.epfl.ch/~brain/.

[14] J.N. Kapur and H.K. Kesavan, *Entropy Optimization Principles with Applications*, Academic Press, Inc., 1992.

[15] Mark Holden, Derek L.G. Hill, Erika R.E. Denton, Jo M. Jarosz, Tim C.S. Cox, and David J. Hawkes, "Voxel similarity measures for 3D serial MR brain image registration," in *Information Processing in Medical Imaging (IPMI)*, 1999, vol. 1613, pp. 466–471.

[16] David E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989.

[17] R.K.-S. Kwan, A.C. Evans, and G.B. Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1085–1097, November 1999.

[18] A. Guimond, A. Roche, A. Ayache, and J. Meunier, "Multimodal brain warping using the demons algorithm and adaptative intensity corrections," Tech. Rep., Inst. National de Recherche en Informatique et en Automatique, Sophia Antipolis, 1999.

[19] Jean-Philippe Thiran and Torsten Butz, "Fast non-rigid registration and model-based segmentation of 3d images using mutual information," in *Proc. SPIE Medical Imaging, San Diego*, 2000, pp. 1504–1515.

[20] Torsten Butz and Jean-Philippe Thiran, "Affine registration with feature space mutual information," in *Medical Image Computing and Computer-Assisted Intervention*. October 2001, vol. 2208 of *Lecture Notes in Computer Science*, pp. 549–556, Springer-Verlag.

[21] Guido Gerig, Olaf Kübler, Ron Kikinis, and Ferenc A. Jolesz, "Nonlinear anisotropic filtering of mri data," *IEEE Transactions on Medical Imaging*, vol. 11, no. 2, pp. 221–232, June 1992.

[22] Bostjan Likar, Max A. Viergever, and Franjo Pernus, "Retrospective correction of mr intensity inhomogeneity by information minimization," in *Medical Image Computing and Computer-Assisted Intervention, Pittsburgh, USA*, October 2000, pp. 375–384.

[23] Torsten Butz and Jean-Philippe Thiran, "Feature space mutual information in speech-video sequences," in *Proc. IEEE Int. Conf. on Multimedia and Expo 2002, Lausanne, Switzerland*, August 2002.