MULTI-MODAL MEDICAL IMAGE REGISTRATION: FROM INFORMATION THEORY TO OPTIMIZATION OBJECTIVE

Torsten Butz, Olivier Cuisenaire, and Jean-Philippe Thiran

Signal Processing Institute,
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
Web-page: http://ltswww.epfl.ch/~brain.
{torsten.butz,olivier.cuisenaire,jp.thiran}@epfl.ch

Abstract: A relatively large class of information theoretical measures, including e.g. mutual information or normalized entropy, has 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 enlight 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 is very general though and 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 increased 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 and explore an approach very similar to information theoretic 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 to use 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 theoretic derivation indicates clearly that we can very easily build multi-modal algorithms which automatically select and extract the optimal elements within a predefined class 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 how we can obtain normalized entropy [7], an overlap-invariant entropy measure for multi-modal medical image registration, from 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 modeled by a Markov Chain [4].

It's interesting to interpret classification as a Markov chain $C \to X \to F \to \widehat{C}$ as Fano's inequality [5] gives a lower bound of the error probability of miss-classification $P_e = Pr(C \neq \widehat{C} = \widehat{C}(F))$ [4]:

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

where C is a random variable (RV) modeling the learning sample of the classes. X is the RV of the obser-

0-7803-7503-3/02/$17.00 ©2002 IEEE

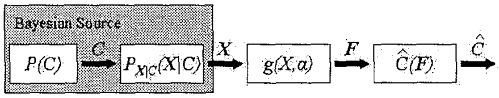

Fig. 1. Learning optimal features for classification with examples can be mathematically interpreted as a Markov chain [4]. C represents the random variable of the learning sample of the classes and X are the associated observations generated by its conditional probability density function $P_{X|C}(X|C)$. 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.

vations 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 modeling 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 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 modeled 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").

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 exactly is measured is the feature selection step. E.g. in an image, for each sample position generated from the RV S we can measure the intensity at that position, but also the gradient, Gabor response, etc. 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. 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].

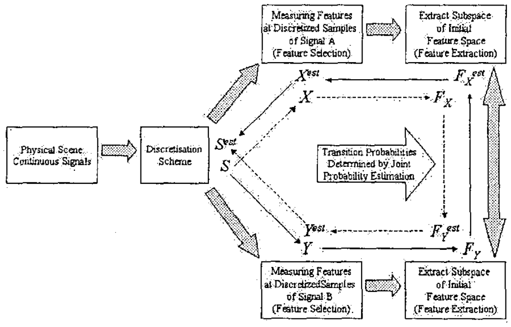
Physical Scene Continuous Signals
Discretisation Scheme
S^{est}
S
Measuring Features at Discretized Samples of Signal A (Feature Selection)
Extract Subspace of Initial Feature Space (Feature Extraction)
X^{est}
X
F_X^{est}
F_X
Transition Probabilities Determined by Joint Probability Estimation
Y^{est}
Y
F_Y^{est}
F_Y
Measuring Features at DiscretizedSamples of Signal B (Feature Selection).
Extract Subspace of Initial Feature Space (Feature Extraction)

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

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

$$S \rightarrow X \rightarrow F_X \rightarrow F_Y^{est} \rightarrow Y^{est} \rightarrow S^{est} \qquad (2)$$

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

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):

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

and

[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].

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

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 error probability when mapping one signal into 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 (the speaker's mouth motion) from a speech signal. These probabilistic mappings are modeled as the Markov chains of eq. 2 and 3.

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. In order to visualize this danger, let's re-write the lower bounds in a different way and 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 increasing 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 adding selectively 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 and (in the resulting intervals) minimize the bounds of eq. 4 and eq. 5.

For this aim we defined 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:

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

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:

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

The derivation was specific for 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 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 small 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. 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

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

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

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

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

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:

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

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 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.

3. ADDING MODELING ASSUMPTIONS

The previously developed theory is very general. No assumptions have been taken with respect to the employed features or the underlying probability distributions. Depending on the particular multi-modal signals, these generality can be abandoned in favor of a more specialized objective function in the sense of the error probabilities of eq. 4 and 5.

In everything that follows, we argue solely on the Markov chain of eq. 2. A completely analogue development is possible for eq. 3.

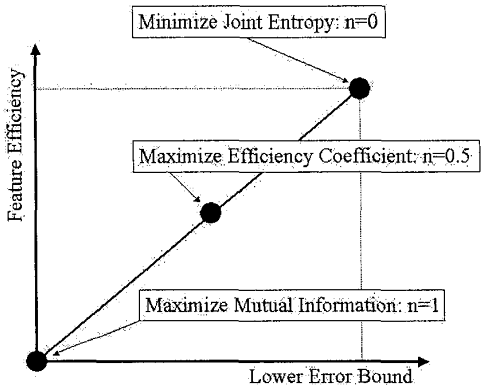
Minimize Joint Entropy: n=0
Feature Efficiency
Maximize Efficiency Coefficient: n=0.5
Maximize Mutual Information: n=1
Lower Error Bound

Fig. 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].

3.1. From Error Probability to Correlation ratio

Let's start with recalling Fano's inequality for the Markov chain of eq. 2 for multi-modal signals (eq. 4):

$$\begin{aligned} P_{e1} &= Pr(S^{est} \neq S) \\ &\geq 1 - \frac{I(F_X, F_Y) + 1}{\log|\Psi|} \\ &= 1 - \frac{H(F_Y) - H(F_Y|F_X) + 1}{\log|\Psi|} \qquad (13) \\ &\geq 1 - \frac{\log(\sqrt{2\pi e Var(F_Y)}) - H(F_Y|F_X) + \frac{13}{12}}{\log|\Psi|}, \qquad (14) \end{aligned}$$

where the last inequality approaches an equality when F_Y goes towards a discretized Gaussian [6]. It's important to note that in contrast to eq. 4, 5, 8 and 12, the last lower bound is not symmetric anymore with respect to the exchange of F_X and F_Y.

Instead of minimizing the lower bound of 13, we can minimize the weakened lower bound of eq. 14 by maximizing $\log(\sqrt{2\pi e Var(F_Y)}) - H(F_Y|F_X)$. Let's now assume that the probability density $P(f_Y|f_X)$ of the transition $F_X \to F_Y$ is characterized by

$$f_Y = E(F_Y|F_X) + \epsilon(E(F_Y|F_X)), \qquad (15)$$

where $\epsilon(.)$ is an additive Gaussian noise and $E(X|Y)$ is the conditional expectation of X knowing Y. The conditional probability $P(F_Y = f_Y|F_X = f_X)$ is given by

$$P(F_Y = f_Y|F_X = f_X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(f_Y - E(F_Y|F_X))^2}{2\sigma^2}}, \qquad (16)$$

with $\sigma^2 = E(Var(F_Y|F_X))$. Therefore we can easily calculate the conditional entropy $H(F_Y|F_X)$:

$$\begin{aligned} H(F_Y|F_X) &= -\sum_{f_X,f_Y} P(f_X,f_Y) \\ &\quad \cdot \log(\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(f_Y - E(F_Y|F_X))^2}{2\sigma^2}}) \\ &= \log(\sqrt{2\pi e E(Var(F_Y|F_X))}). \quad (17) \end{aligned}$$

This means that we can minimize the lower bound of eq. 14 by maximizing

$$\begin{aligned} &\log(\sqrt{2\pi e Var(F_Y)}) - H(F_Y|F_X) = \\ &\quad = \log(\sqrt{2\pi e Var(F_Y)}) \\ &\qquad - \log(\sqrt{2\pi e E(Var(F_Y|F_X))}), \end{aligned} \quad (18)$$

which is equivalent to maximizing its squared exponential

$$\eta_1(F_Y|F_X) = \frac{Var(F_Y)}{E(Var(F_Y|F_X))}, \quad (19)$$

or maximizing

$$\eta_2(F_Y|F_X) = 1 - \frac{E(Var(F_Y|F_X))}{Var(F_Y)}. \quad (20)$$

$\eta_2(F_Y|F_X)$ is just the correlation ratio as proposed in [16] for multi-modal medical image registration, when the employed features F_X and F_Y are the image intensities.

3.2. From Error Probability to Maximum Likelihood

In the previous section (chap. 3.1), only assumptions about the underlying transition probabilities were taken. On the other hand we didn't use any prior on the specific feature representations to be used. Let's now relax the prior on the probabilities, but assume that we can fix the feature representation F_Y. This means that also the entropy $H(F_Y)$ remains constant during the minimization of the lower error bound of eq. 13. Therefore we want to find the feature representation F_X so that the conditional entropy

$$H(F_Y|F_X) = -\sum_{f_X,f_Y} P(f_X,f_Y)\cdot\log P(f_Y|f_X) \quad (21)$$

is minimal.

Reversing the arguments of [17], it's easily shown that $H(F_Y|F_X)$ can be re-written as

$$H(F_Y|F_X) = -\frac{1}{|\Psi|}\prod_{s\in\Psi} P(f_Y(s)|f_X(s)). \quad (22)$$

This is up to the negative constant $-\frac{1}{|\Psi|}$ exactly the likelihood of getting a signal F_Y from a signal F_X for a given transition probability distribution $P(F_Y = f_Y|F_X = f_X)$ and a fixed feature space representation F_Y.

3.3. Feature Efficiency for Correlation Ratio

It's important to note that in chap. 3.1 and 3.2 we argued solely on the error probabilities and not on feature efficiency coefficients of the feature space representations. Using the definition of eq. 12 it's straight forward to construct specific feature efficiencies ("normalized entropies") for particular modeling assumptions.

In the case of maximum likelihood, maximizing feature efficiency is equivalent to maximizing the following expression:

$$e_n(F_X,F_Y) = \frac{(H(F_Y) - H(F_Y|F_X))^n}{H(F_X,F_Y)^{n-1}}, n\in[0,1], \quad (23)$$

where $H(F_Y)$ is constant though and allows therefore easy evaluation of $e_n(F_X,F_Y)$.

We can also find interesting expressions for the case of correlation ratio. Using the definition of eq. 12 and the modeling assumptions of chap. 3.1, we can write our generalized feature efficiency coefficient as follows:

$$e_n(F_X,F_Y) = \frac{(\log\frac{\sqrt{2\pi e E(Var(F_Y|F_X))}}{\sqrt{2\pi e Var(F_Y)}})^n}{H(F_X,F_Y)^{n-1}}, \quad (24)$$

with $n\in[0,1]$.

For $n = \frac{1}{2}$, the functionals of eq. 23, resp. 24, should correspond to an overlap invariant likelihood, resp. correlation ratio, just as normalized entropy is the overlap invariant expression for mutual information.

These expressions have a relatively complicated form as we used explicitly the definition of eq. 12 in order to emphasize the relation to normalized entropy. It's possible though to construct efficiency coefficients more tailored for likelihood or correlation ratio. For such a construction it would be important though to keep in mind that the feature efficiency coefficient has to find a tradeoff between minimizing the lower bound of the error probabilities and maximizing the efficiency of the employed features (summarized in fig. 3).

4. MULTI-MODAL MEDICAL IMAGE REGISTRATION

So far our developments have been completely unrelated to medical images or image registration. Everything has been derived in the very general context of a correspondent pair of multi-modal signals and their feature space representations. In the case of medical images we simply deal with image features, such as image intensity [1], [2], edgeness [18] etc. We have still to make the connection to the particular problem of medical image registration though. To do this we identify image registration as a special case of feature selection: Which transformation selects the best features (e.g. image intensities) for the voxels of the floating image so that the error bounds of the corresponding Markov chains are minimized?

Therefore for the very general case of completely unspecified Markov chains (eq. 2 and 3), the optimization objectives (mutual information resp. feature efficiency coefficient of order n) can be formalized as follows:

$$[\vec{t}^{opt}, F^{opt}_{\{X,Y\}}] = \\ = \arg \max_{\vec{t}\in\mathbf{R}^d, F_{\{X,Y\}}\subset\mathcal{F}_{\{X,Y\}}} I(T_{\vec{t}}(F_X), F_Y) \tag{25}$$

$$[\vec{t}^{opt}, F^{opt}_{\{X,Y\}}] = \\ = \arg \max_{\vec{t}\in\mathbf{R}^d, F_{\{X,Y\}}\subset\mathcal{F}_{\{X,Y\}}} e_n(T_{\vec{t}}(F_X), F_Y), \tag{26}$$

where $\vec{t} \in \mathbf{R}^d$ contains the registration parameters and d defines the specific transformation (e.g. $d = 6$ for rigid and $d = 12$ for affine registration). $\mathcal{F}_{\{X,Y\}}$ are the initial feature spaces from which the optimal image feature representations are selected/extracted. We made the arbitrary choice that F_X is the feature space representation of the floating and F_Y of the reference image. It's just for more clarity that we wrote $\vec{t}$ separately from F_X. In fact they can be implicitly included in the features F_X of the floating image.

Let's now add some assumptions to the general case of eq. 25. In particular let's assume that the modeling assumptions of eq. 16 are met. For this case we have shown in chap. 3.1 that minimizing the lower bound of the error probabilities is very closely related (up to the inequality of eq. 14) to maximizing the correlation ratio between the feature space representations. Thereafter we have also derived the general feature efficiency coefficient for this particular case. Therefore the optimization objective for the lower error bound is

$$[\vec{t}^{opt}, F^{opt}_{\{X,Y\}}] = \\ = \arg \max_{\vec{t}\in\mathbf{R}^d, F_{\{X,Y\}}\subset\mathcal{F}_{\{X,Y\}}} \eta_2(F_Y|T_{\vec{t}}(F_X)) \\ = \arg \max_{\vec{t}, F_{\{X,Y\}}} 1 - \frac{E(Var(F_Y|T_{\vec{t}}(F_X)))}{Var(F_Y)}, \tag{27}$$

which is the correlation ratio [16]. On the other hand the optimization objective for the feature efficiency coefficient using the definition of eq. 24 is

$$[\vec{t}^{opt}, F^{opt}_{\{X,Y\}}] = \\ = \arg \max_{\vec{t}, F_{\{X,Y\}}} \frac{(\log \frac{\sqrt{2\pi e E(Var(Y|T_{\vec{t}}(X)))}}{\sqrt{2\pi e Var(Y)}})^n}{(\log \sqrt{2\pi e Var(Y)})^{1-n}}. \tag{28}$$

Finally for the maximum likelihood expressions (chap. 3.2) we imposed that F_Y is a fixed feature space representation. Therefore we have to take F_X as features of the floating image. The resulting optimization objective is given by

$$\begin{aligned}[\vec{t}^{opt}, F^{opt}_X] &= \arg \min_{\vec{t}\in\mathbf{R}^d, F_X\subset\mathcal{F}_X} H(T_{\vec{t}}(F_X)|F_Y) \\ &= \arg \max_{\vec{t}\in\mathbf{R}^d, F_X\subset\mathcal{F}_X} P(T_{\vec{t}}(F_X)|F_Y).\end{aligned} \tag{29}$$

The first equality represents minimal conditional entropy and the second equality maximum likelihood optimization. The equivalence of these two objectives was shown in chap. 3.2. It's important to note that even though F_Y is fixed during this optimization it has not to represent image intensities. Furthermore the employed features F_X and F_Y of the initial images have not to be the same. For example to register an ultrasound (US) image onto a magnetic resonance (MR) data-set we might want to use image intensities for the MR image but rather a combination of image intensities and gradients to represent the corresponding information in the US data-set [19].

5. FEATURE EFFICIENCY FOR IMAGE QUANTIZATION

Let's consider a simple but illustrative example of feature extraction. It is closely related to simultaneous image registration and multi-channel segmentation of medical images as introduced in [20]. In this paper we want to show how the order n of eq. 12 influences the optimal number of uniform quantization levels of a pair of synthetic T1-T2 magnetic resonance images (MRI) [21], [22]. Uniform quantization can be seen as a very primitive way to reduce the corrupting noise in the MRIs and therefore potentially improve the performance of image registration algorithms. It's important though that the quantization will just reduce the noise, but as few as possible of the anatomical information in the images. This means that we have to find an order n for eq. 12, which on the one hand drops the unrelated information of both MR data-sets (noise), but keeps the related anatomical information when the number of uniform quantization levels are optimized with the corresponding efficiency coefficient of order n.

In fig. 4 a) and b), we show the initial T1 and T2 images resp. In images c) we have plotted the efficiency coefficients for different ns versus the logarithm of the number of bins. We see clearly that the maximum value lies at different numbers of quantization levels. In particular we can recognize the special case of $n = 0$ and $n = 1$. For $n = 0$ we minimize joint entropy and therefore the optimal number of bins is as expected 1 (section 2.4: just maximize feature efficiency). For $n = 1$ we actually maximize mutual information alone and therefore the optimal number of bins is the number of gray levels in the initial images (section 2.4: just minimize error bounds). For $n = 0.2$, 0.5 and 0.8, we lie somewhere in between these extreme cases, but $n = 0.8$ has also its global optimum where the number of uniform quantization levels equals the number of image intensities. $n = 0.2$ has it's optimum at 2 and $n = 0.5$ at 3 quantization levels.

In fig. 5, we have applied the optimal uniform quantization for $n = 0.2$ and 0.5 to the initial images. For the cases $n = 0$, 0.8 and 1 the results are trivially either completely black or the unchanged initial images. We can see that $n = 0.2$ adds some of the anatomical information to the images. Nevertheless lots of anatomy is lost in favor of a more efficient feature pair. On the other

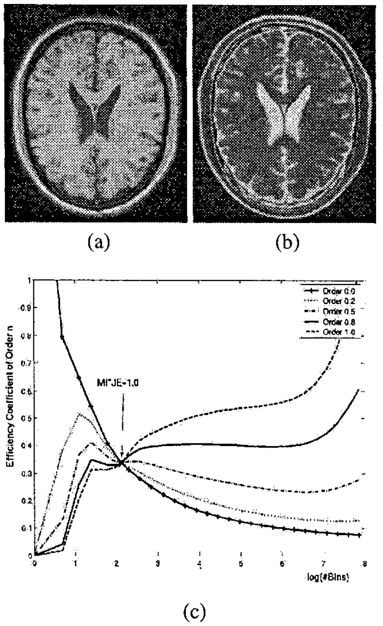

Fig. 4. In a) and b) we show a corresponding T1 and T2 data-set. In c) we show the efficiency coefficients for different orders n as a function of uniform quantization levels. We see that the maximum varies heavily for different ns and confirm the theoretical expectations of chap. 2.4

hand for $n = 0.5$ we have lots of anatomical information conserved, while the corrupting noise is rejected.

In [20] we have shown that the presented quantization task can easily be combined with image registration.

6. CONCLUSION

We have set-up a very general information theoretical framework for multi-modal signal processing. Mathematically it is based on Markov chains and the lower bounds of their error probabilities. Keeping the focus on medical image registration, we derived several widely used objective measures from the framework. In particular we showed how mutual information, normalized entropy ("Feature efficiency"), correlation ratio and likelihood are information theoretically related. Our derivations are very general and extend all these measures, which are classically applied as image intensity statistics, on general feature space representations of the initial datasets. We also extend the important concept of feature efficiency to a more general mathematical formulation.

Finally we give an illustrative example about the concept of feature efficiency, by applying a simple quantization step on medical images. We show that e.g. noise suppression can easily be integrated in multi-modal medical image registration. An example, where bias-correction is combined with registration is presented in [20].

It's important to note that the presented general framework opens the door towards a wide range of further developments about multi-modal signals and in particular medical images. For example 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.

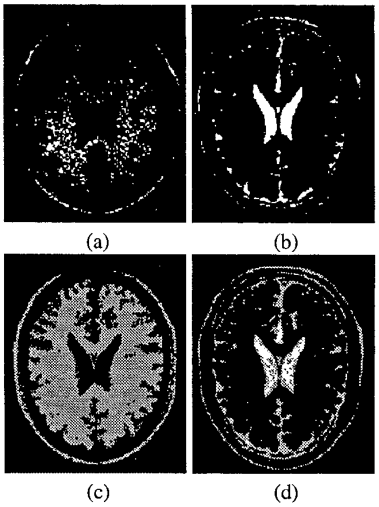

Fig. 5. In the images a) and b), respectively c) and d), we show the uniform quantization results at optimal feature efficiency (fig. 4 c)) for the feature efficiency coefficients of order 0.2 and 0.5 respectively.

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] A. Roche, G. Malandin, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for mulimodal image registration," in *Medical Image Computing and Computer-Assisted Intervention*. 1999, vol. 1496 of *Lecture Notes in Computer Science*, pp. 1115–1124, Springer-Verlag.

[17] A. Roche, G. Malandain, and N. Ayache, "Unifying maximum likelihood approaches in medical image registration," Tech. Rep., Inst. National de Recherche en Informatique et en Automatique, Sophia Antipolis, July 1999, Report No. 3741.

[18] 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.

[19] Alexis Roche, Xavier Pennec, Grǵoire Malandin, and Nicholas Ayache, "Rigid registration of 3-D ultrasound with MR images: A new approach combining intensity and gradient information," *IEEE Transactions on Medical Imaging*, vol. 20, no. 10, pp. 1038–1049, October 2001.

[20] Torsten Butz, Víctor Rodríguez Gil, and Jean-Philippe Thiran, "A general information theoretical framework for multi-modal medical image processing," Submitted to Medical Image Computing and Computer-Assisted Intervention (MICCAI 2002).

[21] 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.

[22] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, and A.C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, June 1998.