Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Signal Processing 85 (2005) 875–902
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

# From error probability to information theoretic (multi-modal) signal processing

Torsten Butz[a], Jean-Philippe Thiran[b,*]

[a]*ImaSys SA, PSE - Bâtiment C, 1015 Lausanne, Switzerland*
[b]*Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Institute, 1015 Lausanne, Switzerland*

Received 23 April 2004; received in revised form 10 November 2004

**Abstract**

We propose an information theoretic model that unifies a wide range of existing information theoretic signal processing algorithms in a compact mathematical framework. It is mainly based on stochastic processes, Markov chains and error probabilities. The proposed framework will allow us to discuss revealing analogies and differences between several well-known algorithms and to propose interesting extensions resulting directly from our formalism. We will then describe how the theory can be applied to the rapidly emerging field of multi-modal signal processing: we will show how our framework can be efficiently used for multi-modal medical image processing and for joint analysis of multi-media sequences (audio and video).


*Keywords:* Stochastic process; Information theory; Markov chain; Error probability; Multi-modal; Multi-media; Medical images

## 1. Introduction

The signal processing community has proved to be increasingly reliant upon information theoretic concepts to develop algorithms for a wide variety of important problems ranging e.g. from audio-visual signal processing to medical imaging [1–10]. When comparing the proposed algorithms, two facts are particularly surprising. First, the range of practical problems which are solved with the fundamentally very compact mathematical concepts of information theory seems to be very broad and unrelated. For example, mutual information has been very successful in multi-modal medical image registration, but has also been successfully used for information theoretic classification and feature extraction. The second striking fact is that the mathematical expressions governing the final algorithms seem not to be very related, even though the employed fundamental concepts were identical. For instance in [1], feature extraction by

*Corresponding author. Tel.: +41 21 693 4623; fax: +41 21 693 7600.
*E-mail addresses:* torsten.butz@ima-sys.ch (T. Butz), jp.thiran@epfl.ch (J.-P. Thiran).

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.11.027

maximization of mutual information has been derived from the general concept of error probabilities for Markov chains. On the other hand, information theoretic image registration algorithms with all their different optimization objectives [6–9] never made reference to error probabilities or Markov chains.

With this paper, we pursue two main goals. On the one hand, we propose a common mathematical framework for a large class of information theoretic signal processing algorithms. Concretely, we will show that stochastic processes and their error probabilities are very fundamental concepts underlying a large class of very distinct algorithms, such as physiological quantization or overlap-invariant multi modal medical image registration. Their mathematical exploitation allows deep insight into information theoretic signal processing and widely opens the door towards very interesting generalizations of existing algorithms.

The second main aim of this paper consists of using the developed framework to construct promising algorithms for the emerging field of multi-modal signal processing. It will lead to a multi-modal framework closely related to information theoretic feature extraction/selection. This important relationship will indicate how we can unify to a large extent multi-modal medical image processing, e.g. multi-channel segmentation and image registration, and extend information theoretic registration to features other than image intensities. The framework is not at all restricted to medical images though and we will illustrate this by applying it to multi-media sequences as well.

The paper is structured as follows: first, we recall some information theoretic concepts which build our mathematical framework (Section 2). In order to familiarize the reader with the mathematics and indicate the important role of error probability in information theoretic signal processing, we will hereafter re-derive known algorithms of quantization (Section 3.1) and classification (Section 3.2). In Section 4 we transpose our framework to multi-modal signal processing which allows to derive optimization objective functions used in multi-modal medical image processing and to study their mathematical relationships. A short section on genetic optimization (Section 5) will build the bridge to the final section where we show some interesting results in medical imaging (Section 6.1) and multi-media signal processing (Section 6.2). The discussion section (Section 7) will wrap up the results, before the conclusion in Section 8.

## 2. Some information theoretic concepts

We want to start by recalling some important information theoretic concepts which will be used extensively hereafter. All the presented notions are well known and widely used in several fields of information technology and computer science. We would also like to emphasize that the random variables throughout this work refer to discrete random variables, except when they are concretely specified to be continuous.

### 2.1. Stochastic process and error probability

A stochastic process is an indexed sequence of random variables (RV) with, in general, arbitrary mutual dependencies [11]. For the specific case of information theoretic signal processing, we construct the following stochastic process.

Let us define the discrete random variables $X$ and $X^{\text{est}}$ on the same set of possible outcomes $\Omega_X$. Let us also define $N$ discrete random variables $Y_i$ on $\Omega_{Y_i}$ for $i = \{1, \ldots, N\}$ resp. We consider the following stochastic process:

$$X \to Y_1 \to \cdots \to Y_N \to X^{\text{est}} \to E, \tag{1}$$

where, in general, the transition probabilities are conditioned on all the previous states. $E$ is a binary RV defined on $\Omega_E = \{0, 1\}$ and is 1 if the estimation $X^{\text{est}}$ of $X$ from $Y$ is considered as an error. This stochastic process will be the fundamental model of our following developments and examples.

An important characteristic of Eq. (1) is the probability of error $P_e = P(E = 1)$, which equals the expectation $\mu_E$ of $E$:

$$P_e = \mu_E = 1 \cdot P(E = 1) + 0 \cdot P(E = 0). \tag{2}$$

We can use the conditional probabilities defining the transitions of Eq. (1) to write

$$P_e = \sum_{x\in\Omega_X}\sum_{y_1\in\Omega_{Y_1}}\cdots\sum_{y_N\in\Omega_{Y_N}}\sum_{x^{\text{est}}\in\Omega_X} P(E=1|x^{\text{est}},y_N,\ldots,y_1,x)\cdot P(x^{\text{est}}|y_N,\ldots,y_1,x)\cdot\ldots\cdot P(y_1|x)\cdot P(x). \quad (3)$$

If any of the random variables is defined over a continuous interval and is therefore given by a continuous probability density function, the corresponding sum in Eq. (3) has to be replaced by an integral.

The error probability $P_e$ has a close connection to the well-known concept of signal distortion [11]. While signal distortion is defined in a deterministic way, error probability however, is, a probabilistic quantity. In Section 3.1 we will show with two examples that the probabilistic concept of Eq. (3) incorporates to a large extent the deterministic theory of distortion.

It is important to note that so far no hypothesis about the specific transition probabilities has been set. This is in particular the case for the error variable $E$: the fact that for example $x^{\text{est}}\neq x$ does not necessarily imply that $E=1$ with probability one. This generality might look quite artificial and impractical, but we want to show that specific hypotheses about the different steps in the stochastic process of Eq. (1), including the last one, will result in well-known mathematical formulas of quantization, classification and multi-modal signal processing. Nevertheless, a complete study of Eq. (1) would go beyond the scope of this paper. Therefore, we want to restrict ourselves to the case where all the transition probabilities of Eq. (1) besides the last one are Markovian [11]:

$$P(x^{\text{est}}|y_N,\ldots,y_1,x)=P(x^{\text{est}}|y_N),$$

$$P(y_N|y_{N-1},\ldots,y_1,x)=P(y_N|y_{N-1}),$$

$$\vdots$$

$$P(y_1|x)=P(y_1|x). \quad (4)$$

This implies that the stochastic process $X\to Y_1\to\cdots\to Y_N\to X^{\text{est}}$ forms a Markov chain. The Markovian condition is obviously not fulfilled for the last transition probability, as the error probability has at least to depend on the input to the chain $x$ and on its final output $x^{\text{est}}$. In what follows, we suppose that the Markovian conditions of Eq. (4) are fulfilled, so that the error probability $P_e$ becomes

$$P_e = \sum_{x\in\Omega_X}\sum_{y_1\in\Omega_{Y_1}}\cdots\sum_{y_N\in\Omega_{Y_N}}\sum_{x^{\text{est}}\in\Omega_X} P(E=1|x^{\text{est}},y_N,\ldots,y_1,x)\cdot P(x^{\text{est}}|y_N)\cdot P(x^{\text{est}}|y_{N-1})\cdot\ldots\cdot P(y_1|x)\cdot P(x). \quad (5)$$

Furthermore, we will consider the special case that $P(E=1|x^{\text{est}},y_N,\ldots,y_1,x)$ of Eq. (5) equals just $P(E=1|x^{\text{est}},x)$, i.e. the error probability of the chain only depends on the relationship between the input value and its estimated value after having gone through the chain.

### 2.2. Fano's inequality and the data processing inequality

The model proposed in the previous section and, in particular, the exact evaluation of the error probability (Eq. (5)) requires knowledge of all the transition probabilities. Sometimes though we might not have access to the data to such a deep extent. We still want to at least approximately estimate the error probability $P_e=P(E=1)$. When the process $X\to Y_1\to\cdots\to Y_N\to X^{\text{est}}$ fulfills the Markovian conditions of Eq. (4) and when the probability of error is given by $P_e=\Pr(X^{\text{est}}\neq X)$, we can use an expression known as Fano's inequality [12] to compute a lower bound of $P_e$ as a function of the input RV $X$ and the last transmission RV $Y_N$ only.

Let us state Fano's inequality, as will be used extensively later in this paper. Let $A\to B\to A^{\text{est}}$ be a Markov chain. $A$ (and therefore $A^{\text{est}}$) has to be finitely (or countably infinitely) valued. Then we have a lower bound on the error probability $P_e=\Pr(A^{\text{est}}\neq A)$ such that the output of the chain, $A^{\text{est}}$, is not the input $A$:

$$P_e \geqslant \frac{H(A|B)-H(P_e)}{\log|\Omega_A-1|} \geqslant \frac{H(A|B)-1}{\log|\Omega_A|} = \frac{H(A)-I(A,B)-1}{\log|\Omega_A|}, \quad (6)$$

where $|\Omega_A|$ is the number of elements in the range of $A$, and $H(.)$ stands for the Shannon entropy of one RV and $I(.,.)$ for the Shannon mutual information between a pair of RVs. Therefore, for the case of the Markov chain $X \to Y_1 \to \cdots \to Y_N \to X^{\text{est}}$ as introduced in Section 2.1 and under the assumption that $P_e = \Pr(X^{\text{est}} \neq X)$, this lower bound is written as

$$P_e \geqslant \frac{H(X|Y_N) - 1}{\log|\Omega_X|} = \frac{H(X) - I(X, Y_N) - 1}{\log|\Omega_X|}. \quad (7)$$

There is another very useful inequality which is applicable within the previously mentioned assumptions: the data-processing inequality [11] which states that if $A \to B \to C$ is a Markov chain, we have

$$I(A, B) \geqslant I(A, C),$$

$$I(B, C) \geqslant I(A, C). \quad (8)$$

The combination of these expressions allows to build a large number of resulting inequalities, such as

$$\begin{aligned} P_e &\geqslant \frac{H(X) - I(X, Y_N) - 1}{\log|\Omega_X|} \\ &\geqslant \frac{H(X) - I(X, Y_{N-1}) - 1}{\log|\Omega_X|} \\ &\geqslant \frac{H(X) - I(X, Y_{N-2}) - 1}{\log|\Omega_X|} \geqslant \cdots, \end{aligned}$$

$$\begin{aligned} P_e &\geqslant \frac{H(X) - I(X, Y_N) - 1}{\log|\Omega_X|} \\ &\geqslant \frac{H(X) - I(Y_1, Y_N) - 1}{\log|\Omega_X|} \\ &\geqslant \frac{H(X) - I(Y_2, Y_N) - 1}{\log|\Omega_X|} \geqslant \cdots, \\ &\vdots \end{aligned} \quad (9)$$

Under the specified assumptions, these expressions allow one to focus on a specific transition within the Markov chain $X \to Y_1 \to \cdots \to Y_N \to X^{\text{est}}$. This is of great interest if we have particular knowledge about one of the transitions in the chain or if some other transition is not sufficiently understood.

The concept of bounding the error probability $P_e$ is by far not exploited with the presented inequalities. There are other entropies, especially Renyi entropy [13], that can be more appropriate for specific applications [14]. Furthermore, the estimation of an upper bound could be very interesting [14], or specific assumptions on the RVs can result in very interesting specialized expressions. All this is almost a research domain on its own closely related to rate-distortion theory. In this paper we restrict ourselves to the theory described so far in this section.

## 3. From error probability to optimization objective

This section will clarify how the general concepts of Section 2 can be tailored for specific problems in the field of signal/information processing. Obviously, it is not possible in this paper to study all existing algorithms of information processing within the context of stochastic processes. Furthermore, we do not pretend that all existing algorithms in this field necessarily fit this model. Nevertheless, we want to show that quite a large class of existing solutions can be derived and interpreted within the mathematical concept of error probability. This allows a deep insight into several algorithms and outlines quite revealing analogies.

Another aim is to familiarize the reader with the presented formalism, so that the derived theory of Section 4 for multi-modal signals will be easily understood. Therefore, we first study several algorithms of distortion theory (Section 3.1) and classification (Section 3.2), before transposing the mathematics into the field of multi-modal signals (Section 4).

### 3.1. From error probability to distortion

We will first outline some important analogies between the error probability of Eq. (5) and the classical distortion [11,15] defined as follows:

Let us have two signal sequences $S$ and $S^{\text{est}}$ of equal length $n$. The distortion $D(S, S^{\text{est}})$ between these two sequences is defined by

$$D(S, S^{\text{est}}) = \frac{1}{n} \sum_{i=1}^{n} d(s_i, s_i^{\text{est}}), \quad (10)$$

where $d(.,.)$ is called the distortion measure between the samples $s_i$ and $s_i^{\text{est}}$.

Let us show that the error probability of Eq. (5) is in fact a probabilistic extension of the classical distortion of Eq. (10). To do this we show under which conditions the error probability of Eq. (5) becomes identical to the distortion defined above. Let us write down the most basic stochastic process of Eq. (1) and its resulting error probability $P_e$:

$$X \to X^{\text{est}} \to E,$$

$$P_e = \sum_{x\in\Omega_X} \sum_{x^{\text{est}}\in\Omega_X} P(E=1|x^{\text{est}},x)\cdot P(x,x^{\text{est}}), \quad (11)$$

where $X$, resp. $X^{\text{est}}$, is an RV on the set $\Omega_X$ modeling probabilistically the initial sequence $S$, resp. $S^{\text{est}}$, and can be estimated from the sequence by density estimation (e.g. histogramming [16,17]). Therefore the set $\Omega_X$ has to span the whole range of possible values of $S$ and $S^{\text{est}}$. Furthermore, $(X, X^{\text{est}})$ is a bi-variate RV on $\Omega_X^2$ that models the co-occurrences of the samples $s$ and $s^{\text{est}}$ in the bivariate sequence $(S, S^{\text{est}})$. Its probabilities can also be estimated by density estimation. In the case of joint histogramming, we have

$$P(x,x^{\text{est}}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{s_i,x}\cdot\delta_{s_i^{\text{est}},x^{\text{est}}}, \quad \forall (x,x^{\text{est}})\in\Omega_X^2, \quad (12)$$

where $\delta_{a,b}$ is the Kronecker delta defined by

$$\delta_{a,b} = \begin{cases} 1 & \text{if } a=b,\\ 0 & \text{otherwise.}\end{cases}$$

By analogy to Eq. (10), we can identify $P(E=1|x^{\text{est}},x)$ with a distortion measure $d(x,x^{\text{est}})$, so that $P(E=1|x^{\text{est}},x)=\alpha\cdot d(x,x^{\text{est}})$. $\alpha$ is a constant to ensure that $0\leqslant P(E=1|x^{\text{est}},x)\leqslant 1, \forall (x,x^{\text{est}})\in\Omega_X^2$. Thereafter, we can rewrite the error probability of Eq. (11) as follows:

$$\begin{aligned} P_e &= \sum_{x\in\Omega_X}\sum_{x^{\text{est}}\in\Omega_X} P(E=1|x^{\text{est}},x)\cdot P(x,x^{\text{est}})\\ &= \frac{1}{n}\sum_{x\in\Omega_X}\sum_{x^{\text{est}}\in\Omega_X}\sum_{i=1}^{n} P(E=1|x^{\text{est}},x)\cdot\delta_{s_i,x}\cdot\delta_{s_i^{\text{est}},x^{\text{est}}}\\ &= \frac{1}{n}\sum_{x\in\Omega_X}\sum_{x^{\text{est}}\in\Omega_X}\sum_{i=1}^{n} P(E=1|s_i^{\text{est}},s_i)\\ &= \frac{\alpha\cdot|\Omega_X|^2}{n}\sum_{i=1}^{n} d(s_i,s_i^{\text{est}}) \end{aligned} \quad (13)$$

which is up to the constant $\alpha\cdot|\Omega_X|^2$, identical to the classical definition of distortion (Eq. (10)).

### 3.1.1. Hamming distortion

Let us look at two well-known distortion measures, starting with the Hamming distortion. Let us assume that whenever $x^{\text{est}}\neq x$, we observe an error with probability 1. Then we can re-write the conditional error probability as

$$P(E=1|x^{\text{est}},x)=\alpha\cdot d(x,x^{\text{est}})=1-\delta_{x^{\text{est}},x}. \quad (14)$$

In this case the condition that $0\leqslant P(E=1|x^{\text{est}},x)\leqslant 1$ is naturally fulfilled for all $(x,x^{\text{est}})\in\Omega_X^2$, and $\alpha$ can arbitrarily be set to 1. In this specific case the error probability $P_e$ is the probability that $x^{\text{est}}\neq x$ ($P_e=\Pr(x^{\text{est}}\neq x)$).

### 3.1.2. Squared error distortion

Let us now assume that the probability that the estimate $x^{\text{est}}$ is considered different from $x$ increases quadratically with the distance $|x^{\text{est}}-x|$. Then we can write

$$P(E=1|x^{\text{est}},x)=\alpha\cdot d(x,x^{\text{est}})=\alpha\cdot|x-x^{\text{est}}|^2, \quad (15)$$

which leads to the squared error distortion [18]. In this case, $\alpha$ cannot just be set to 1 to fulfill the necessary probability conditions. But as we want to minimize Eq. (13), where $\alpha$ is just a constant, its specific value has no influence on the optimization results and can therefore be chosen arbitrarily positive.

There are mainly two very interesting and practical advantages of identifying classical distortion as a probability of error:

- First of all, perceptual studies normally do not come up with a deterministic model of apparent distortion. This means that a given distortion in a signal (e.g. image) might be perceived differently by two different observers. So quite naturally, subjective studies will result in

differences within the test population. The presented approach to distortion can incorporate such differences quite easily in a probabilistic way.

- The proposed approach can easily be extended to multi-channel distortion models. We have simply to identify $S$ and $S^{\text{est}}$ as vector sequences which contain not only the scalars $s$ and $s^{\text{est}}$ of the sequences but also other information such as e.g. local variation, contrast, etc. The proposed probabilistic interpretation of distortion can naturally combine different channels in a very general probabilistic manner, and can also consider channel interactions probabilistically (independent channel assumption is not required).

For a nice summary about perceptual distortion models, the reader is referred to [19].

### 3.2. From error probability to classification error

Information theoretic concepts are widely used in classification. In this section, we want to study two particularly interesting and complementary approaches of classification which can both be interpreted within the presented framework of stochastic processes and error probabilities. First, we will show that the information theoretic clustering algorithm as presented in [10] is easily interpreted as minimization of an error probability. And second, we will recall information theoretic feature extraction for classification [1] as it will be closely related to the proposed methodology for multi-modal signal processing.

#### 3.2.1. Information theoretic clustering

Let us consider the data set $S$ of $n$ data samples, each containing a feature vector or scalar $y_i \in \Omega_Y = \mathbb{R}^d$, where $d$ is the dimension of the feature vectors $y_i$. $Y$ is the corresponding continuous random variable over $\Omega_Y = \mathbb{R}^d$ and can be estimated by density estimation from the data $S$ [16]. A classification algorithm aims to classify the $n$ samples into $n_c$ classes, each being labeled by one of the symbols of $\Omega_c = \{1, \ldots, n_c, t\}$. The class label $t$ is associated to the feature subspace $\Omega_Y^{\bar{S}} \subset \mathbb{R}^d$, for which no feature vector of the samples $S$ exist: $\Omega_Y^{\bar{S}} = \{y \in \mathbb{R}^d : \nexists s_i \in S \text{ with feature vector } y\}$. Let us finally call $\mathbf{c} \in \mathscr{C}$ a class label map over the data, $S$.

The random variable over $\Omega_c$ associated to the resulting classes (clusters) is denoted $C$. Furthermore, we consider a random variable different from $C$, called $C^{\text{est}}$, also over $\Omega_c$, which models an estimation of the initial random variable $C$. Very naturally, we can build the following stochastic process:

$$C \to Y \to C^{\text{est}} \to E, \tag{16}$$

where $E$ is an error random variable being 1 whenever the estimated value $c^{\text{est}}$ is considered a wrong estimate of the initial class label $c$ and 0 otherwise. We now consider the special case where all the transition probabilities of Eq. (16), besides the last one, are first order Markov transitions. Furthermore, the final transition depends only on the initial input to the process $c$ and on its final output $c^{\text{est}}$. Therefore, the whole process is defined by the following probability densities:

$$P(C = c),$$

$$P(Y = y|C = c),$$

$$P(C^{\text{est}} = c|Y = y, C = c) = P(C^{\text{est}} = c|Y = y),$$

$$\begin{aligned} &P(E = 1|C = c, Y = y, C^{\text{est}} = c^{\text{est}}) \\ &\quad = P(E = 1|C = c, C^{\text{est}} = c^{\text{est}}), \end{aligned}$$

$$\begin{aligned} &P(E = 0|C = c, Y = y, C^{\text{est}} = c^{\text{est}}) \\ &\quad = P(E = 0|C = c, C^{\text{est}} = c^{\text{est}}). \end{aligned} \tag{17}$$

A key quantity of the stochastic process of Eq. (16) is the probability of error $P_e$ of the transmission from $C$ to $C^{\text{est}}$. If all the transitions of Eq. (17) are known, this quantity can be calculated explicitly as follows:

$$\begin{aligned} P_e = \sum_{c^{\text{est}} \in \Omega_c} \int_{\Omega_Y} \sum_{c \in \Omega_c} & P(E = 1|c^{\text{est}}, c) \\ & \cdot P(c^{\text{est}}|y) \cdot P(y|c) \cdot P(c)\,\mathrm{d}y. \end{aligned} \tag{18}$$

So far, we do not know the transition probabilities of Eq. (17). But it is possible to estimate them in a non-supervised manner and non-parametrically for any given classification label map $\mathbf{c}$ on the data $S$. Using Gaussian kernel density

estimation [20,21], we get

$$P(y|c) = \sum_{y_1 \in S_c} \frac{1}{|S_c|} G(y - y_1, \sigma_1^2) \tag{19}$$

and

$$P(c^{\text{est}}|y) = \begin{cases} k \cdot \sum\limits_{y_2 \in S_{c^{\text{est}}}} \frac{1}{|y_2|} G(y - y_2, \sigma_2^2) \\ \quad \text{if } c^{\text{est}} \neq t, \\ \int_{\Omega_Y^{\bar{S}}} G(y - y_2, \sigma_2^2)\, \mathrm{d}y_2 \quad \text{if } c^{\text{est}} = t, \end{cases} \tag{20}$$

where $G(x-a,b)$ denotes a Gaussian kernel with expectation $a$ and variance $b$. $|y|$ is the number of samples with feature vector $y$ (therefore $\sum_{y \in \Omega_Y} |y| = n$) and $k$ is a normalization constant. $S_c$ is the subset of $S$ that contains the samples classified to class $c$ and $|S_c|$ is the cardinality of $S_c$. We have therefore $\sum_{c \in \Omega_c} |S_c| = n$ and $|S_t| = 0$. Eq. (20) justifies the introduction of the class-label $t$ which simply ensures that we include the tails of the Gaussian kernels in the probability estimations, i.e. that $\sum_{c \in \Omega_c} P(c|y) = 1$. In addition to this, the probability $P(c)$ is given by the fraction $|S_c|/n$. Obviously, we could have used transition probabilities other than the one of Eq. (19). For example, we could have used kernels other than the Gaussian kernels for the estimation, or we could assume that the probability density of the random variable $C^{\text{est}}$ equals the density of $C$. In this particular case, the transition probabilities $P(y|c)$ and $P(c^{\text{est}}|y)$ would have to fulfill the Bayes theorem. In the general case of Markov chains, this is not necessary though, and as we will show, the transitions of Eqs. (19) and (20) result in particularly nice mathematical expressions.

Using the transitions of Eqs. (19) and (20), the equation for the error probability $P_e$ (Eq. (18)) can be re-written for any given class label map $\mathbf{c}$ on the data sequence $S$ as follows:

$$\begin{aligned} P_{e|\mathbf{c}} &= \sum_{c \in \Omega_c} \int_{\Omega_Y} \sum_{c^{\text{est}} \in \Omega_c} P(E = 1|c^{\text{est}}, c) \\ &\quad \cdot P(c^{\text{est}}|y) \cdot P(y|c) \cdot P(c)\, \mathrm{d}y \\ &= k \cdot \sum_{c \in \Omega_c \backslash \{t\}} \sum_{c^{\text{est}} \in \Omega_c \backslash \{t\}} \frac{|S_c|}{n} \cdot P(E = 1|c^{\text{est}}, c) \\ &\quad \cdot \int_{\Omega_Y} \left( \sum_{y_1 \in S_c} \frac{1}{|S_c|} G(y - y_1, \sigma_1^2) \right) \\ &\quad \cdot \left( \sum_{y_2 \in S_{c^{\text{est}}}} \frac{1}{|y_2|} G(y - y_2, \sigma_2^2) \right) \mathrm{d}y \\ &\quad + \sum_{c \in \Omega_c \backslash \{t\}} \frac{|S_c|}{n} \cdot P(E = 1|t, c) \\ &\quad \cdot \int_{\Omega_Y} \left( \sum_{y_1 \in S_c} \frac{1}{|S_c|} G(y - y_1, \sigma_1^2) \right) \\ &\quad \cdot \left( \int_{\Omega_Y^{\bar{S}}} G(y - y_2, \sigma_2^2)\, \mathrm{d}y_2 \right) \mathrm{d}y \\ &= k \cdot \sum_{c \in \Omega_c \backslash \{t\}} \sum_{c^{\text{est}} \in \Omega_c \backslash \{t\}} \frac{P(E = 1|c^{\text{est}}, c)}{n} \\ &\quad \cdot \sum_{y_1 \in S_c} \sum_{y_2 \in S_{c^{\text{est}}}} \frac{1}{|y_2|} \cdot G(y_1 - y_2, \sigma_1^2 + \sigma_2^2) \\ &\quad + \sum_{c \in \Omega_c \backslash \{t\}} \frac{P(E = 1|t, c)}{n} \\ &\quad \cdot \sum_{y_1 \in S_c} \int_{\Omega_Y^{\bar{S}}} G(y_1 - y_2, \sigma_1^2 + \sigma_2^2)\, \mathrm{d}y_2. \end{aligned} \tag{21}$$

For the general case, $\sigma_1^2$ and $\sigma_2^2$ can be chosen independently. In what follows, we have chosen to restrict ourselves to $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The generalization for different variances is straightforward though. Let us also assume that $P(E = 1|t, c) = 1, \forall c \in \Omega_c \backslash \{t\}$. Then, by identification of the different terms, we can re-write Eq. (21) as

$$\begin{aligned} P_{e|\mathbf{c}} = &\sum_{c \in \Omega_c \backslash \{t\}} \sum_{c^{\text{est}} \in \Omega_c \backslash \{t\}} P(E = 1|c^{\text{est}}, c) \cdot P(c^{\text{est}}, c) \\ &+ P(t), \end{aligned} \tag{22}$$

where

$$\begin{aligned} &P(c^{\text{est}}, c) \\ &= \frac{k}{n} \sum_{y_1 \in S_c} \sum_{y_2 \in S_{c^{\text{est}}}} \frac{1}{|y_2|} \cdot G(y_1 - y_2, 2\sigma^2) \end{aligned} \tag{23}$$

is the probability that a data sample of $S$ with initial class label $c$ is transmitted to an output label

$c^{\text{est}}$ and where

$$P(t) = \frac{1}{n} \sum_{y_1 \in S} \int_{\Omega_Y^{\bar{S}}} G(y_1 - y_2, 2\sigma^2)\, \mathrm{d}y_2 \tag{24}$$

is the probability that any sample gets transmitted into the out-layer $\Omega_Y^{\bar{S}}$. It is important to note that $P(t)$ depends on the data to be classified but is independent of the specific class label map $\mathbf{c}$ and therefore stays constant during any classification scheme that can be constructed with the derived formalism.

The probabilities of Eq. (23) can nicely be written in a matrix of size $n_c \times n_c$, noted $\Gamma$, and representing a theoretical, non-parametrically determined transmission matrix. Its trace, $\mathrm{Tr}(\Gamma)$, gives the probability that $c^{\text{est}}$, the output from the stochastic process of Eq. (16), equals its input $c$, and the sum of the off-diagonal elements plus $P(t)$ gives the probability that the output is a different class label than the input: $P(c \neq c^{\text{est}}) = \sum_{i \neq j} \Gamma_{ij} + P(t) = 1 - \mathrm{Tr}(\Gamma) + P(t)$. Furthermore, we have from the normalization condition of probability densities that $\sum_{i,j} \Gamma_{ij} + P(t) = 1$.

From an implementation point of view, we can simply compute the transmission matrix $\Gamma$ using Eq. (23) which can be done very efficiently. Hereafter, we can calculate the value for $P(t)$ by $P(t) = 1 - \sum_{i,j} \Gamma_{ij}$ in order to avoid the problem of integrating over the infinite tails of the Gaussian kernels. But it is important to note that for a given set of samples $S$, $P(t)$ is just a constant, and therefore its estimation is, in general, not even necessary.

There is one entity we have not specified yet. It is the probability of the final transition of Eq. (16), $P(E = 1|c^{\text{est}}, c)$. As we will show, it is a very interesting quantity whose choice can depend on the particular application and allows the incorporation of some prior information about the data $S$ to be classified.

A very important quantity of information theory is called distortion. The distortion accounts for the fact that not all errors in information transmission, image compression, etc., are of equal importance. For example in image compression, a large error in the pixel values is much more significant than small ones. This resulted in squared error and other distortion measures which are in comparison to Hamming distortion more sensitive to "large" than to "small" errors.

In our case, we can associate very similar properties to this term $P(E = 1|c^{\text{est}}, c)$ of Eq. (22), which can be interpreted as a discrete distortion measure for misclassification. In analogy to the transmission probabilities of Eq. (23), we can re-write $P(E = 1|c^{\text{est}}, c)$ as a matrix of size $n_c \times n_c$, noted $\Lambda$. Because of the close analogy to information theoretic distortion, we call $\Lambda$ the "distortion matrix" of classification. Furthermore, we can define a matrix $\widetilde{\Lambda}$ for the probabilities $P(E = 0|c^{\text{est}}, c)$, whose elements are determined by the condition that probabilities sum up to 1: $P(E = 1|c^{\text{est}}, c) + P(E = 0|c^{\text{est}}, c) = 1$, implying that $\widetilde{\Lambda}_{ij} = 1 - \Lambda_{ij}$.

In the most general form, $\Lambda$ has only to fulfill the condition $0 \leqslant \Lambda_{ij} \leqslant 1, \forall (i,j) \in (\Omega_c \backslash \{t\})^2$. In practice though, a suitable choice allows to penalize more significant classification errors and favor less significant ones. In particular, it is generally worse to misclassify elements of relatively small classes, while the misclassification of one single element of a large class is less significant. Furthermore, we normally consider that we do not commit an error when the output $c^{\text{est}}$ equals the input to the stochastic process $c$. Therefore, the diagonal elements of $\Lambda$ are 0: $\Lambda_{ii} = 0, \forall i \in \{1, \ldots, n_c\}$. These considerations resulted in the following definition for $\Lambda$:

$$\Lambda_{ij} = \gamma \cdot (1 - \delta_{ij}) \sum_{c \in \Omega_c / \{t\}} \frac{1}{|S_c|}, \tag{25}$$

where $\gamma$ is a normalization constant.

The matrix definitions of the transmission matrix $\Gamma$ and of the distortion matrix $\Lambda$ result in a very compact notation for the transmission error probability $P_{e|\mathbf{c}}$:

$$P_{e|\mathbf{c}} = \sum_{i,j} \Lambda_{ij} \cdot \Gamma_{ij} + P(t), \tag{26}$$

$$P_{\bar{e}|\mathbf{c}} = \sum_{i,j} \widetilde{\Lambda}_{ij} \cdot \Gamma_{ij}, \tag{27}$$

where the probability $P_{\bar{e}|\mathbf{c}}$ corresponds to the probability of correct transmission (or the "hit

probability") through the stochastic process of Eq. (16), implying that $P_{e|\mathbf{c}} + P_{\bar{e}|\mathbf{c}} = 1$.

The classification objective of finding the most representative label map $\mathbf{c}$ of the data $S$ with respect to the error probability $P_{e|\mathbf{c}}$ can therefore be written compactly by

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}\in\mathscr{C}} P_{e|\mathbf{c}}, \tag{28}$$

which we call the minimum error probability principle. Equivalently, we can also apply the maximum hit probability principle

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}\in\mathscr{C}} P_{\bar{e}|\mathbf{c}}. \tag{29}$$

It is interesting to look at a distortion matrix other than the one of Eq. (25). Let us consider the following distortion:

$$\Lambda_{ij} = \frac{1-\delta_{ij}}{|S_1|\cdot|S_2|\cdots|S_{n_c}|}. \tag{30}$$

This distortion results in the following expression for the error probability $P_{e|\mathbf{c}}$:

$$\begin{aligned} P_{e|\mathbf{c}} &= \sum_{c_1\in\Omega_c}\sum_{c_2\in\Omega_c}\frac{1-\delta_{c_1,c_2}}{|S_1|\cdot|S_2|\cdots|S_{n_c}|} \\ &\quad\cdot\frac{k}{n}\sum_{y_1\in S_{c_1}}\sum_{y_2\in S_{c_2}}\frac{1}{|y_2|}\cdot G(y_1-y_2,2\sigma^2)+P(t) \end{aligned} \tag{31}$$

$$\begin{aligned} &= \sum_{c_1\in\Omega_c}\sum_{c_2\in\Omega_c} P(E=1|c_1,c_2)\cdot P(c_1,c_2)+P(t) \\ &= \sum_{c_1\in\Omega_c}\sum_{c_2\in\Omega_c}\Lambda_{c_1c_2}\cdot\Gamma_{c_1c_2}+P(t) \\ &= \frac{k}{n}\cdot\frac{1}{|S_1|\cdot|S_2|\cdots|S_{n_c}|}\sum_{y_1\in S}\sum_{y_2\in S}M(y_1,y_2) \\ &\quad\cdot\frac{1}{|y_2|}\cdot G(y_1-y_2,2\sigma^2)+P(t), \end{aligned} \tag{32}$$

where $M(y_1,y_2)$ is defined to be 1 if $y_1$ and $y_2$ are coming from samples with different class labels and 0 if not. It is easily seen that when the number of measurements for a given feature value $y$, $|y|$, always equals one, then the expression of Eq. (32) equals, up to the constants $k/n$ and $P(t)$, exactly the clustering evaluation function of [10] which was shown to have a close relationship to information potential [5] and second-order Renyi entropy [22].

### 3.2.2. Information theoretic feature extraction

We now want to shortly recall the basic concept of information theoretic feature extraction as presented in [1]. Let us assume that we have a set $S$ of $n$ class prototypes, each labeled by one of the class symbols of $\Omega_C = \{1,\ldots,n_c\}$, where $n_c$ is again the number of classes. Furthermore, we have a multi-dimensional feature vector $y_i \in \Omega_Y$ associated to every sample $s_i$ within the set $S$. The RV modeling probabilistically the classes of the prototypes is called $C$ and is defined over the set $\Omega_C$. Its feature space representation, denoted $Y$, is defined over $\Omega_Y$. Feature extraction aims to extract the subspace $F_Y$ of the initial feature space $Y$ that is most significant for the specific classification task. Formally, we can represent this by the following stochastic process:

$$C \to Y \to F_Y \to C^{\text{est}} \to E. \tag{33}$$

The initial transition $C \to Y$ can be interpreted as a feature selection step. The transition $Y \to F_Y$ corresponds to feature extraction, where we select a sub-space $F_Y$ of $Y$. Hereafter, we estimate the class $C^{\text{est}}$ from the final feature representation $F_Y$ and evaluate whether the whole transmission process from $C$ to $C^{\text{est}}$ can be considered as erroneous.

One approach to select and extract the optimal features $Y$ and $F_Y$ for a particular classification task would be to minimize directly the error probability $P_e$ of the selected learning prototypes. This would have the disadvantage, however, that the optimization would get the most relevant features with respect to the chosen classification algorithm ($F_Y \to C^{\text{est}}$). We wish to determine those features for which any "suitable" classification algorithm can obtain "good" results. Therefore, we would need to neglect the classification step $F_Y \to C^{\text{est}}$ during the optimization. To do this, we can profit from Fano's inequality of Section 2.2. In the context of the stochastic process of Eq. (33), this inequality is re-written as

$$P_e \geqslant \frac{H(C|F_Y)-1}{\log n_c} = \frac{H(C)-I(C,F_Y)-1}{\log n_c}. \tag{34}$$

Note that $H(C)$, the entropy of the chosen prototypes as well as $n_c(=|\Omega_C|)$ stays constant, as the number of classes as well as the initial set of

samples stay fixed during the optimization. Therefore, we have to maximize the mutual information $I(C, F_Y)$ in order to minimize the lower bound on the error probability $P_e$, which ensures that a particular classification algorithm can perform well.

It is important to note that we estimated a lower bound of the error probability. Upper bounds might appear more suited for the described problem of classification and, in fact, can be estimated in several cases for a fixed classifier. For example, if it is known from the beginning that a maximum likelihood classifier will be employed, the error probability $P_e$ is upper-bounded by the conditional entropy $H(F_Y|C)$: $P_e \leqslant H(F_Y|C)$. But information theoretic feature extraction attempts to extract those features which are most suited for the given classification task independently of a particular classifier.

Next, we would like to introduce information theoretic multi-modal signal processing by multi-modal feature extraction. Its close relationship to feature extraction for classification will be easily recognized, as we will also use lower error bounds as optimization objectives. Furthermore, the same arguments as for feature extraction will justify the use of a lower bound instead of an upper bound.

## 4. From error probability to multi-modal signal processing

There are several possibilities to apply the presented framework to multi-modal signals. We want to explore one specific approach which we used extensively in several applications of multi-modal signal processing. It is based on one very basic but intuitive hypothesis: *a pair of multi-modal signals originates from the same physical source, even though the signals might have suffered from distortions, delays, noise and other artifacts which hide their common origin.* As we will show, the approach seems to be particularly suited to derive and compare a large number of existing optimization objectives particularly well known in the multi-modal medical imaging community.

### 4.1. Multi-modal stochastic processes

First we will outline how we can build Markov chains from multi-modal signals. The resulting chains should fulfill the conditions required to apply the theory of Section 2. In particular, the conditional error probability $P(E = 1|\ldots)$ has to be 1, whenever the output from the chain differs from the corresponding input (Hamming distortion). We use the fact that multi-modal signals originate from the same physical reality, even though the concrete representations of this reality may be quite different (Fig. 1). We can therefore expect that there exist features in a couple of multi-modal signals which reflect this physical correspondence statistically.

Let us consider the example of an audio–video sequence and assume we know a pair of features which show a statistical dependence. If now we take the feature value of one signal at a randomly chosen time in the sequence and the feature representation at an arbitrarily chosen second time in the other signal of different modality, we should be able to tell if the two measurements are likely to originate from the same physical reality i.e. were acquired at the same physical time or not.

In direct analogy, we also want to mention multi-modal medical images, where the spatial coordinates of the images represent the physical correspondence and take over the role of the time coordinate in audio–video sequences.

A couple of multi-modal signals are initially given by two signal sequences $S_X$ and $S_Y$, both of length $n$, and taking on values in the sets $\Omega_X$ and $\Omega_Y$, respectively.[1] For instance, a 3D image contains $n = n_x \times n_y \times n_z$ samples (or voxels) showing different intensities or colors. Let us define a uniform RV $O$ on the set $\Omega_O = \{1, \ldots, n\}$ labeling the samples in $S_X$ and $S_Y$, this RV is used to model the fact that we will consider a random selection of pairs of samples in the sequences $S_X$ and $S_Y$. Therefore, for 3D images

[1] Sometimes the sampling coordinates are not the same in both signals. For example, two images of different modality might have different dimensions and therefore different numbers of samples. For such cases, we just want to make reference to interpolators which can build the bridge between the two respective sequences [23,24].

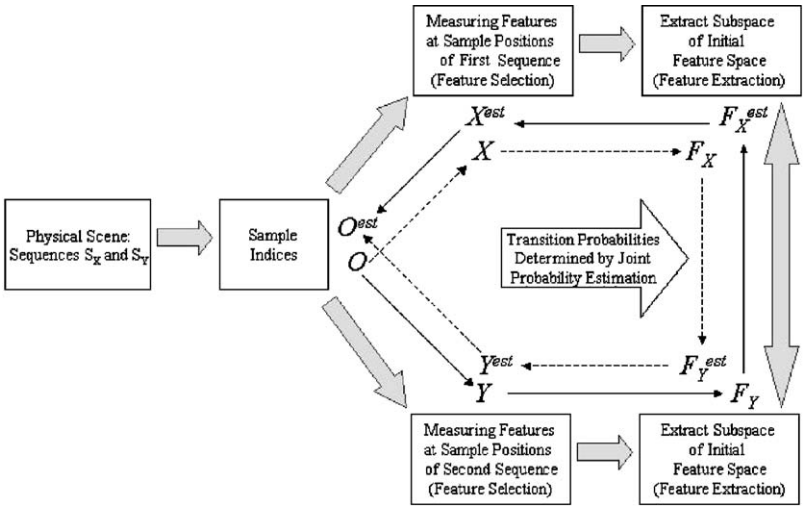


Fig. 1. Markov chains can be built from a pair of multi-modal signals. The connecting block between a couple of multi-modal signals (the transition probabilities for $F_X \to F_Y^{\text{est}}$ and $F_Y \to F_X^{\text{est}}$) is obtained by joint probability estimation.

we have $P(o=(i,j,k)) = 1/n = 1/(n_x \cdot n_y \cdot n_z)$, $\forall o \in \Omega_O$ ("for all voxels in the image"). Starting from the RV $O$, we can build the following two Markov chains (see Fig. 1):

$$O \to X \to F_X \to F_Y^{\text{est}} \to Y^{\text{est}} \to O^{\text{est}} \to E, \quad (35)$$

$$O \to Y \to F_Y \to F_X^{\text{est}} \to X^{\text{est}} \to O^{\text{est}} \to E, \quad (36)$$

where $X$ (resp. $Y$) models the specific feature values of the samples in the sequence $S_X$ (resp. $S_Y$) as an RV conditioned on the outcome of the sample position $o$. Which sequence features are exactly considered represents a feature selection step. For example, in an image, for each sample position generated from the RV $O$ we can consider the intensity at that position, but also the gradient, Gabor response, etc. 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 by $F_X$ and $F_Y$ in Fig. 1. As the considered features of both sequences $S_X$ and $S_Y$ originate from the same sampling label $o$, we can link the two sequences through a joint probability estimation [16,17]. When on the average the chosen feature values $f_X$ and $f_Y$ of $F_X$ and $F_Y$ reflect maximally the fact that they originate from the same sampling position $o$, then we minimize the error probability of the Markov chains. Therefore, we want to select and extract those features $F_X$ and $F_Y$ from the initial sequences $S_X$ and $S_Y$ that show as much as possible of this physical correspondence.

Until now we have constructed a couple of related Markov chains for general multi-modal signals. Let us now see what we can say about the corresponding error probabilities $P_{e1} = \Pr(O^{\text{est}} \neq O)$ (Markov chain Eq. (35)) and $P_{e2} = \Pr(O^{\text{est}} \neq O)$ (Markov chain Eq. (36)) when we use Fano's inequality (Eq. (6)) and the data processing inequality (Eq. (8)) [25]:

$$\begin{aligned} P_{e1} &= \Pr(O^{\text{est}} \neq O) \\ &\geqslant \frac{H(O|Y^{\text{est}}) - H(P_{e1})}{\log(n-1)} \\ &\geqslant \frac{H(O|Y^{\text{est}}) - 1}{\log n} \\ &= \frac{H(O) - I(O, Y^{\text{est}}) - 1}{\log n} \quad (37) \\ &= \frac{\log n - I(O, Y^{\text{est}}) - 1}{\log n} \quad (38) \\ &= 1 - \frac{I(O, Y^{\text{est}}) + 1}{\log n} \quad (39) \\ &\geqslant 1 - \frac{I(F_X, F_Y^{\text{est}}) + 1}{\log n} \quad (40) \end{aligned}$$

and

$$\begin{aligned} P_{e2} &= \Pr(O^{\text{est}} \neq O) \\ &\geqslant \frac{H(O|X^{\text{est}}) - H(P_{e2})}{\log(n-1)} \\ &\geqslant \frac{H(O|X^{\text{est}}) - 1}{\log n} \\ &= \frac{H(O) - I(O, X^{\text{est}}) - 1}{\log n} \qquad (41) \\ &= \frac{\log n - I(O, X^{\text{est}}) - 1}{\log n} \qquad (42) \\ &= 1 - \frac{I(O, X^{\text{est}}) + 1}{\log n} \qquad (43) \\ &\geqslant 1 - \frac{I(F_Y, F_X^{\text{est}}) + 1}{\log n}. \qquad (44) \end{aligned}$$

To get Eq. (38) from Eq. (37) (resp. Eq. (42) from Eq. (41)), we used the fact that $O$ is a uniform random variable over the set $\Omega_O = \{1, \ldots, n\}$ of $n$ possible sampling positions in the sequences $S_X$ and $S_Y$, and has therefore entropy of $\log n$. The last inequality follows directly from the data-processing inequality for Markov chains [11].

The probability densities of $F_X$ and $F_X^{\text{est}}$, resp. $F_Y$ and $F_Y^{\text{est}}$, are both estimated from the same data sequences $S_X$, resp. $S_Y$. Therefore, the estimations of the mutual informations $I(F_Y, F_X^{\text{est}})$ and $I(F_X, F_Y^{\text{est}})$ are equal, and we can write $I(F_Y, F_X^{\text{est}}) \approx I(F_X, F_Y^{\text{est}}) \approx I(F_X, F_Y)$. The value of this mutual information $I(F_X, F_Y)$ is determined from the joint probability density which is estimated by non-parametric probability estimation [16,17] from the sequences $S_X$ and $S_Y$ (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 $F_X$ and $F_Y$ of the multi-modal signals.

### 4.2. Objective functions for multi-modal signal processing

Using the example of multi-modal medical image registration, we will show that it is possible to derive a large class of objective functions of image registration and to theoretically determine their relationships and differences. In this respect, we will show how to derive normalized entropy, correlation ratio and likelihood directly from Eqs. (40) and (44). We will also generalize normalized entropy to the general concept of *feature efficiency* for multi-modal signal processing [26,27].

#### 4.2.1. Feature efficiency

Taking a closer look at Eqs. (40) and (44) reveals an important danger when simply maximizing the mutual information $I(F_X, F_Y)$ in order to minimize the lower error bounds of $P_{e_1}$ and $P_{e_2}$. In order to visualize this danger, let us re-write the lower bounds in a different way and use the fact that for any pair of discrete random variables $A$ and $B$ it can be shown [11] that $H(A, B) \geqslant I(A, B)$ and $(H(A) + H(B))/2 \geqslant I(A, B)$ to weaken them:

$$\begin{aligned} P_{\{e1,e2\}} &\geqslant 1 - \frac{I(F_X, F_Y) + 1}{\log n} \\ &\geqslant 1 - \frac{H(F_X, F_Y) + 1}{\log n} \qquad (45) \end{aligned}$$

and

$$\begin{aligned} P_{\{e1,e2\}} &\geqslant 1 - \frac{I(F_X, F_Y) + 1}{\log n} \\ &\geqslant 1 - \frac{H(F_X) + H(F_Y) + 2}{2 \cdot \log n}. \qquad (46) \end{aligned}$$

Eqs. (45) and (46) 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 $H(F_X, F_Y) \geqslant H(F_X)$ and $H(F_X, F_Y) \geqslant H(F_Y)$). 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 find a suitable trade-off between maximizing the bounds of Eqs. (45) and (46) and minimizing the bounds of Eqs. (40) and (44). Feature pairs which carry information that is present in both signals (large mutual information), but *only* information that is present in *both* signals (low joint entropy),

are the most adapted features for several multi-modal signal processing tasks such as for multi-modal image registration. The described feature efficiency coefficient is a functional that extracts these features from multi-modal signal pairs.

For this let us 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 $e(A,B)$ for any pair of RVs $A$ and $B$ (in particular also for the feature space RVs $F_X$ and $F_Y$) is defined as follows:

$$e(A,B)=\frac{I(A,B)}{H(A,B)}\in[0,1]. \tag{47}$$

Maximizing $e(A,B)$ signifies a trade-off between minimizing the lower bound of the error probability by maximizing the mutual information $I(A,B)$, and also minimizing the joint entropy $H(A,B)$ (resulting in maximizing the weakened bounds of Eqs. (45) and (46)). Looking for features that maximize the efficiency coefficient of Eq. (47) will therefore result in finding features which are highly related (large mutual information) but have not necessarily much information (marginal entropy).[2]

Interestingly, there is a functional closely related to $e(A,B)$ that has already been widely used in multi-modal medical image processing, even though its derivation was completely different. It was called normalized entropy $\mathrm{NE}(A,B)$ [29] and was derived as an overlap invariant optimization objective for rigid registration:

$$\begin{aligned}\mathrm{NE}(A,B)&=\frac{H(A)+H(B)}{H(A,B)}\\&=e(A,B)+1\in[1,2].\end{aligned} \tag{48}$$

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 $\mathrm{NE}(A,B)$ of Eq. (48) is equivalent to maximizing the efficiency coefficient of Eq. (47).

It is very interesting to note that in the early years of information theoretic multi-modal signal processing, joint entropy $H(A,B)$ 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 Eqs. (40), (44) and particularly (45), one realizes that minimizing joint entropy does *not* minimize these error bounds. On the contrary, it actually maximizes the weakened bound of Eq. (45) and therefore contradicts error bound minimization. The result showed 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 section 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. (47), such as

$$e(A,B)=\frac{I(A,B)}{H(A)+H(B)} \tag{49}$$

or

$$e(A,B)=\frac{I(A,B)^{2/3}}{H(A,B)^{1/3}}. \tag{50}$$

While the first example is a variant equivalent to Eq. (47), as it simply uses the weakened inequality of Eq. (46) instead of Eq. (45), the second is an extension of $e(A,B)$, that can be generalized as follows:

$$e_k(A,B)=\frac{I(A,B)^k}{H(A,B)^{1-k}},\quad k\in[0,1]. \tag{51}$$

[2]Because of the range $[0,1]$ of $e(A,B)$, this functional is sometimes called "normalized measure of dependence" [28].

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

- $k = 0$: We put emphasis entirely on the feature efficiency without caring about the resulting lower bound of the error probabilities (minimizing joint entropy). The algorithm will always converge toward signal sequence representations where the same single feature value is assigned to all the samples.
- $k = 1$: We put emphasis on minimizing the lower error bound without caring about the efficiency of the features (maximizing mutual information). The algorithm would converge toward signal representations where all samples get assigned a different feature value.
- $k = \frac{1}{2}$: We put equal emphasis on minimizing the lower error bound and on feature efficiency (maximizing normalized entropy).

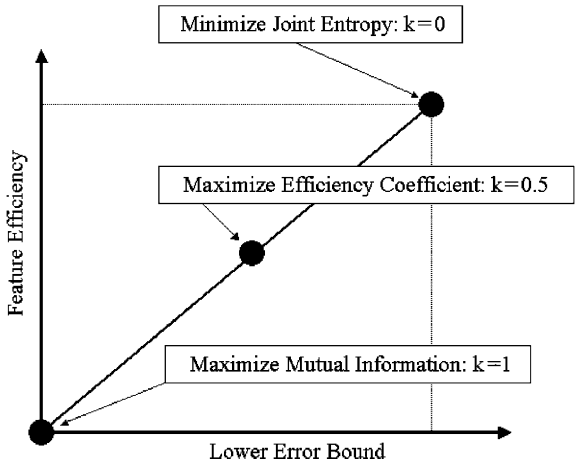


Fig. 2. The sketch puts the efficiency coefficients for different orders $k$ into a quantitative relationship. The contradictory optimization objectives of minimizing the lower error bound, but maximizing the feature efficiency, have to be combined in a suitable way for a given problem. In the case of medical images, $k = \frac{1}{2}$ has been shown to work well, as it results into an optimization functional equivalent to 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 $k$ of Eq. (51) for a given problem. For example, order $\frac{1}{2}$ has shown to be very interesting for medical image registration [29,30]. In Fig. 2 we show a quantitative sketch of feature efficiency for different orders of $k$. In fact, this trade-off between feature efficiency and error probability has an interesting analogy in rate-distortion theory, where on the one hand we want to transmit as little information as possible, but, on the other hand, we want to keep the transmission error as small as possible.

Let us very add a synthetic example that illustrates well how the feature efficiency of Eq. (51) varies with different orders $k$. For this, we take the initial magnetic resonance (MR) images of Fig. 3a and b and plot the feature efficiency coefficients of their image intensities for different orders $k$ versus the number of uniform image quantization levels (Fig. 3c). Image content-dependent optimal quantization can be looked at as image segmentation. Our approach of quantizing both images with the same number of uniform bins is very crude, but very illustrative in the context of feature efficiencies. In Section 6.1, we show results with a more practical quantization scheme (Fig. 7).

In Fig. 3c we see that the maximum feature efficiency varies significantly with its order $k$. For $k = 0$, the optimum suggests using one single quantization bin (very efficient image representation) while for $k = 1$ we would keep all the initial data, including the un-related image noise. For intermediate levels, in particular for $k = \frac{1}{2}$, we get an intermediate optimal number of bins (7 bins) which conserves the anatomical information of the initial scans, while discarding the un-related noise of the images (Fig. 4). This is exactly the behavior outlined in Fig. 2 in the special context of image quantization.

#### 4.2.2. Correlation ratio

We will show now that optimization objectives other than mutual information and normalized entropy can be derived from the proposed frame-

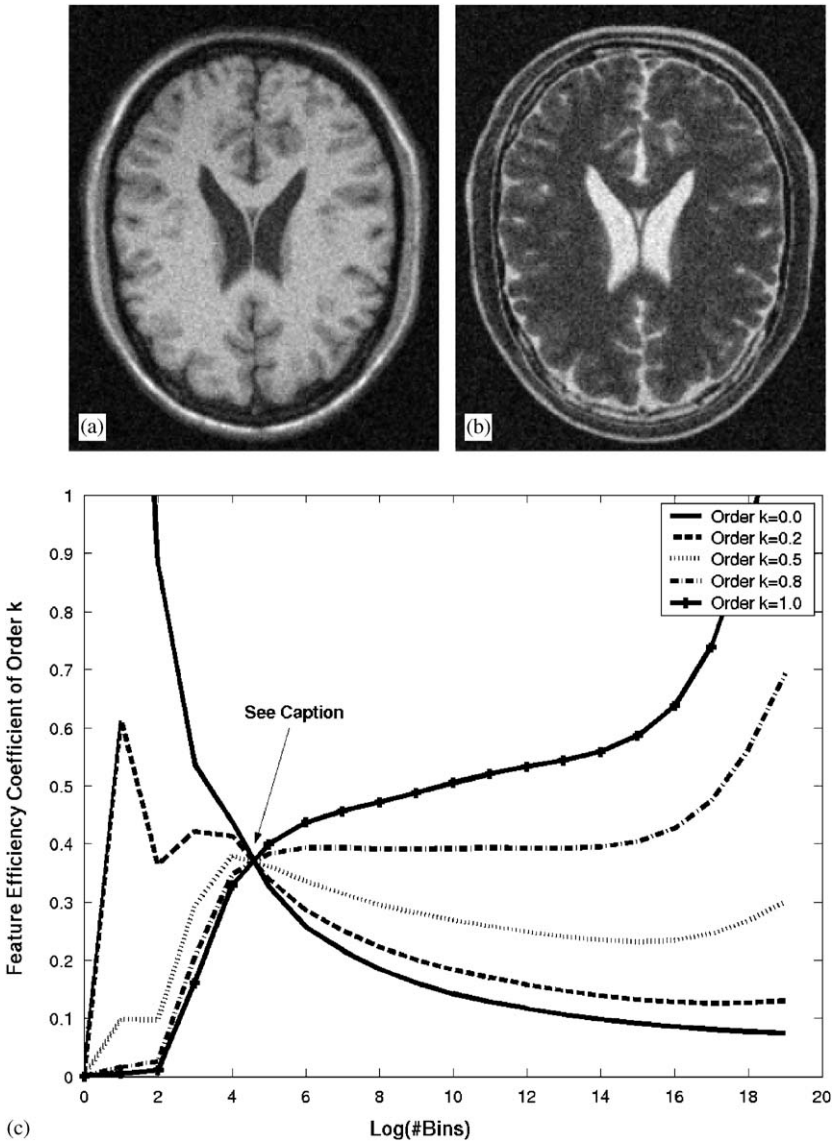


Fig. 3. In (a) and (b), we show a corresponding T1 and T2 data set. In (c), we show the feature efficiency coefficients of Eq. (51) for different orders $k$ as a function of uniform quantization levels. We see that the maximum varies heavily with the order $k$. For $k = 0.2$, the optimum lies at 2 quantization intervals (Fig. 4a and b), while for $k = 0.5$ we get 7 levels which conserve anatomical information while discarding unrelated noise (Fig. 4c and d). All the different graphs cross where mutual information times joint entropy equals one (see arrow).

work, starting with correlation ratio [9]. Let us start by recalling Fano's inequality for the Markov chain of Eq. (35) for multi-modal signals (Eq. (40)) and then lower the bound under the condition that $F_Y$ is characterized by a continuous Gaussian probability density:

$$\begin{aligned} P_{e1} &= P(O^{\text{est}} \neq O) \\ &\geqslant 1 - \frac{I(F_X, F_Y) + 1}{\log n} \\ &= 1 - \frac{H(F_Y) - H(F_Y|F_X) + 1}{\log n} \end{aligned} \tag{52}$$

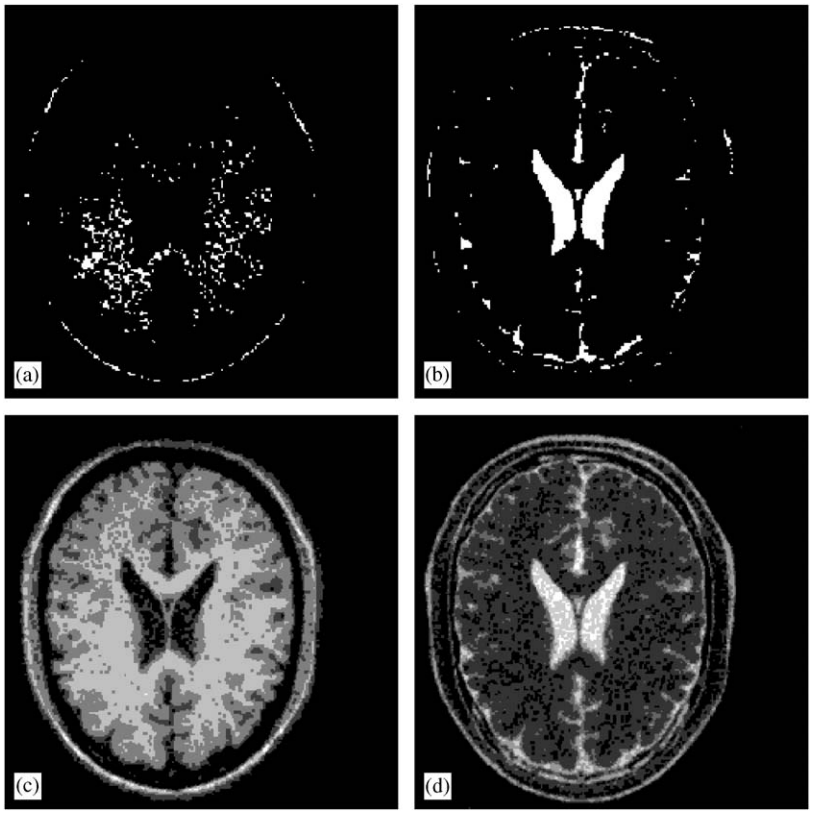


Fig. 4. In the images (a) and (b), respectively (c) and (d), we show the uniform quantization results at optimal feature efficiency (Fig. 3c) for the feature efficiency coefficients of order $k = 0.2$ and 0.5, respectively. The former conserves very limited but efficient information, while the latter keeps most of the anatomically relevant structures.

$$\geqslant 1 - \frac{\log(\sqrt{2\pi e \operatorname{Var}(F_Y)}) - H(F_Y|F_X) + 1}{\log n}, \tag{53}$$

where $F_X$ is a discrete random variable so that Fano's inequality still holds. It is important to note that in contrast to Eqs. (40), (44), (47) and (51), the last lower bound is not symmetric anymore with respect to $F_X$ and $F_Y$.

Instead of minimizing the lower bound of Eq. (52), we can minimize the weakened lower bound of Eq. (53) by maximizing $\log(\sqrt{2\pi e \operatorname{Var}(F_Y)}) - H(F_Y|F_X)$. Let us now assume that the probability density function $P(f_Y|f_X)$ of the transition $F_X \to F_Y$ is given by

$$F_Y = E(F_Y|F_X) + \mathrm{N}(0, E(\operatorname{Var}(F_Y|F_X))), \tag{54}$$

where $\mathrm{N}(\mu, \sigma^2)$ is an additive Gaussian noise of mean $\mu$ and variance $\sigma^2$ and $E(F_Y|F_X)$ is the conditional expectation of $F_Y$ knowing $F_X$. Then the conditional probability $P(F_Y = f_Y|F_X = f_X)$ is given by

$$P(F_Y = f_Y|F_X = f_X) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(f_Y - E(F_Y|F_X))^2/2\sigma^2}, \tag{55}$$

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

$$\begin{aligned} H(F_Y|F_X) &= -\sum_{f_X} \int_{f_Y} P(f_X, f_Y) \\ &\quad \cdot \log\left(\frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \mathrm{e}^{-(f_Y - E(F_Y|F_X))^2/2\sigma^2}\right) \mathrm{d}f_Y \\ &= \log(\sqrt{2\pi e E(\operatorname{Var}(F_Y|F_X))}). \end{aligned} \tag{56}$$

This means that we can minimize the lower bound of Eq. (53) by maximizing

$$\begin{aligned}&\log(\sqrt{2\pi e\,\mathrm{Var}(F_Y)}) - H(F_Y|F_X)\\ &\quad= \log(\sqrt{2\pi e\,\mathrm{Var}(F_Y)})\\ &\qquad- \log(\sqrt{2\pi e E(\mathrm{Var}(F_Y|F_X))}),\end{aligned} \tag{57}$$

which is equivalent to maximizing its squared exponential

$$\eta_1(F_Y|F_X) = \frac{\mathrm{Var}(F_Y)}{E(\mathrm{Var}(F_Y|F_X))}, \tag{58}$$

or maximizing

$$\eta_2(F_Y|F_X) = 1 - \frac{E(\mathrm{Var}(F_Y|F_X))}{\mathrm{Var}(F_Y)}. \tag{59}$$

It is important to note that $\eta_2(F_Y|F_X)$ is just the correlation ratio as proposed in [9] for multi-modal medical image registration, when the employed features $F_X$ and $F_Y$ are the image intensities.

#### 4.2.3. Maximum likelihood

In the previous paragraph on correlation ratio, only assumptions about the underlying transition probabilities were taken. On the other hand, we did not use any prior on the specific feature representations to be used. Let us now relax the prior on the transitions, but assume that we can fix the feature representation $F_X$. In direct analogy to Eq. (52), we have also

$$\begin{aligned}P_{e1} &= P(O^{\mathrm{est}} \neq O)\\ &\geqslant 1 - \frac{I(F_X,F_Y)+1}{\log n}\\ &= 1 - \frac{H(F_X) - H(F_X|F_Y)+1}{\log n},\end{aligned} \tag{60}$$

where $H(F_X)$ remains constant during the minimization as $F_X$ is fixed. Therefore, we want to find the feature representation $F_Y$ so that the conditional entropy

$$H(F_X|F_Y) = -\sum_{f_Y f_X} P(f_Y,f_X)\cdot\log P(f_X|f_Y) \tag{61}$$

is minimal. Let us now use histogramming to estimate the joint probabilities $P(f_Y,f_X)$:

$$P(f_Y,f_X) = \frac{|(f_X,f_Y)|}{n}, \tag{62}$$

where $|(f_X,f_Y)|$ is the number of samples that have feature value $f_X$ in one modality and feature value $f_Y$ in the other modality. Therefore, we can rewrite Eq. (61) as follows [31]:

$$\begin{aligned}&H(F_X|F_Y)\\ &\quad= -\sum_{f_Y f_X}\frac{|(f_X,f_Y)|}{n}\cdot\log P(f_X|f_Y)\\ &\quad= -\frac{1}{n}\sum_{o\in\Omega_O}\log P(f_X(o)|f_Y(o))\\ &\quad= -\frac{1}{n}\log\left(\prod_{o\in\Omega_O} P(f_X(o)|f_Y(o))\right),\end{aligned} \tag{63}$$

where $\Omega_O$ indexes the $n$ data samples (Section 4). Eq. (63) is, up to the negative constant $-1/n$, exactly the log-likelihood of obtaining a signal $F_X$ from a signal feature representation $F_Y$ for a given transition probability distribution $P(F_X = f_X|F_Y = f_Y)$ and a fixed feature space representation $F_X$.

#### 4.2.4. Image registration as feature selection

In the previous paragraphs, we derived several optimization objectives within the framework of error probability. This allows us to analyze their mutual relationship on a theoretical basis and facilitates the choice of the optimization objective for a particular problem. For example, we derived the differences between mutual information and normalized entropy, also for applications outside medical image registration. Even more generally, these developments show a very general concept of multi-modal signal processing based on feature selection and extraction which determines those feature space representations of the initial signal sequences that confirm most of the basic and natural hypothesis that multi-modal signals have the same physical origin. On the other hand, however, it might not seem very clear yet how the presented framework is actually applied to a particular problem, such as image registration. Let us therefore have a closer look at multi-modal medical image registration.

So far, we have just been discussing feature selection and extraction but not image transformations, such as rigid or affine deformations. The step from the proposed framework of multi-modal feature selection and extraction to image registration is straightforward though. In fact, it is possible to identify image registration as a special case of feature selection. We want to select the image transformation that best reflects the fact that the images are acquired from the same physical scene. This transformation will minimize the error probabilities $P_{\{e_1,e_2\}}$.

If $S_Y$ is the sequence of the floating image and $S_X$ of the reference image, the corresponding Markov chains can be re-written:

$$\begin{aligned} O \to X &= S_X(o) \to F_X = X \\ &= S_X(o) \to F_Y^{\text{est}} = S_Y(T(o)) \\ &\to Y^{\text{est}} = S_Y(o) \to O^{\text{est}} \to E, \end{aligned}$$

$$\begin{aligned} O \to Y &= S_Y(o) \to F_Y \\ &= S_Y(T(o)) \to F_X^{\text{est}} = S_X(o) \\ &\to X^{\text{est}} = F_X^{\text{est}} = S_X(o) \to O^{\text{est}} \to E. \end{aligned} \tag{64}$$

We see that most of the transitions are deterministic and that several of them are parameterized by the transformation parameters $T$ of the floating image. In fact, $T$ can be looked at as the optimization parameters of a feature selection step of the floating image (sequence) $S_Y$. The optimal $T$ should confirm the most of the basic multi-modal hypothesis, that the two multi-modal images (sequences) $S_X$ and $S_Y$ have the same physical origin. It is important to note that "the same physical origin" is a flexible hypothesis. This means that sometimes two brain images, even though of different patients, are considered to come from the same physical scene, which is not a particular patient's anatomy but the brain anatomy in general.

## 5. Optimization

The presented framework of feature selection and extraction for multi-modal signal processing leads quite frequently to very demanding optimization objectives (Section 6). The resulting objective functions can have very distinct shapes depending on the chosen feature space representations from which the optimal representation should be selected and extracted. As a result it is mostly not clear that local optimization schemes would be sufficient to lead to robust results. This is why we use a globally convergent genetic optimization algorithm [32]: to study the global behavior of the optimization functions and to ensure that we avoid local optima to get "good" results for a specific application, such as image registration. Hereafter, we refine the results locally using the steepest gradient algorithm.

The problem with this approach is that genetic optimization is very time consuming. Therefore we parallelized an existing genetic optimization library [33] for distributed memory architectures, using the MPICH implementation of message passing interface (MPI) [34,35].

## 6. Results

This paper is mainly dedicated to multi-modal signal processing. Therefore, we only want to show results in this field and discard examples of quantization or classification. In particular, we want to illustrate the field of multi-modal medical image registration and show how the feature-based framework enlarges the vision of image registration and results in a unifying framework for multi-modal medical image processing in general.

The second example is based on speech-video sequences, where we show the importance of adequate features to interpret both signals of a multi-modal sequences simultaneously.

### 6.1. Multi-modal medical images

We will show two examples for medical image registration [36] which show the importance of the feature-based framework for multi-modal signal processing. First of all, we want to show that the quality of registration results can depend heavily on the employed features. For example, to use rather edge instead of intensity information

for the registration can be very beneficial for some cases.

The second point will show that the proposed framework enlarges the view of image registration and leads to an integrated approach on multi-modal medical image processing in general. We will show examples where image registration, segmentation and artifact correction of medical images can be incorporated into one generalized algorithm. In the paragraphs on registration with quantization and registration with bias correction, we used synthetic MR-scans from the *BrainWeb* database [37,38].

#### 6.1.1. Feature space image registration

Conventionally, multi-modal medical image registration determined optimal transformation parameters by maximizing image intensity-based information theoretic quantities, particularly those re-derived in Section 4.2 [8,9,39]. In the sense of error probability, this seems to be the right thing to do. Nevertheless, it is important to note that maximizing mutual information minimizes the error probability on the average. This means that the statistical matching of large structures is much more emphasized than other small but anatomically important regions of the patients anatomy. Therefore, we propose to rather use the edgeness information in the images [40], so that the probability estimation does not use the volumetric information anymore, but rather information on the volume boundaries.

We considered affine registration to maximize mutual information between the two feature space representations of the initial images (their gradients) and compared the results to the conventional intensity-based mutual information registration. The results for CT-MR inter-subject registration are shown in Fig. 5. As we can see, while the rigid registration of the MR image with the CT image is achieved correctly when considering the voxel intensities (Fig. 5c and i) and the edgeness (Fig. 5d and j) as feature space, the maximization of the intensity-based mutual information yields aberrant results in case of non-rigid (affine) registration (Fig. 5e and k) while using edgeness as a feature space for this optimization leads to correct and stable results (Fig. 5f and l). The interpretation of those results is presented in Section 7.

Besides improving the robustness for MR-CT registration, we will show that edgeness information combined with globally convergent genetic optimization makes mutual information-based image registration applicable to the registration of blood vessel images in the human retina. In Fig. 6, we show how three partial views of the retinal vascular system can be combined to provide a virtually extended view. The different intensity distributions in the images are caused by an injected contrast agent which enables the study of the retinal blood flow for diabetic retinopathy.

#### 6.1.2. Image registration and quantization

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 data sets (e.g. anisotropic filtering [41]) 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 naive way of extracting the representative anatomical information in the medical images while discarding the dispensable noise. We use simple image intensity quantization which varies the number of bins for both axes of the joint probability distribution independently. But decreasing the number of bins obviously decreases the marginal entropies of the image representations, therefore simply maximizing the mutual information of Eqs. (40) and (44) is dangerous. We rather use the feature efficiency coefficient of order $\frac{1}{2}$ (Eq. (51)) of the quantized images to find the optimal number of quantization intervals as well as the geometrical registration parameters. Let us recall again that the feature efficiency of order $\frac{1}{2}$ is equivalent to the widely used normalized entropy. Mathematically, we can write the optimization objective as follows:

$$[\vec{t}^{\,\mathrm{opt}}, q_X^{\mathrm{opt}}, q_Y^{\mathrm{opt}}] = \arg \max_{\vec{t}\in\mathbf{R}^d, q_X\in\mathbf{Z}^+, q_Y\in\mathbf{Z}^+} e_{1/2}(Q_{q_X}(X), T_{\vec{t}}(Q_{q_Y}(Y))), \tag{65}$$

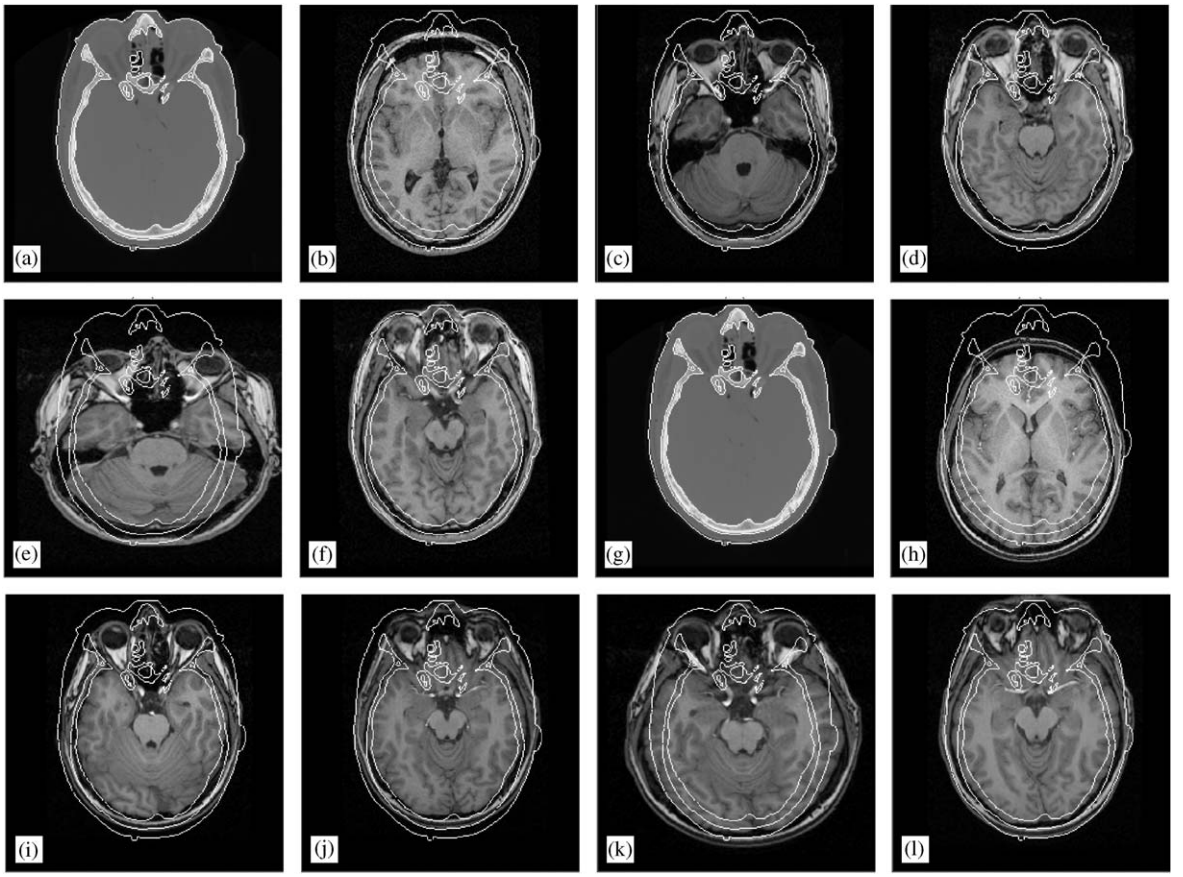


Fig. 5. (a) Is the CT-target image. In (b), the contours of the target image are superposed on the floating MR-scan. In (c) and (d), we see the results after a rigid optimization, when using resp. the intensity-based mutual information and the edgeness mutual information. In (e) and (f), we show the corresponding results for affine registration. Figures (g)–(l) show the results for a second MR-scan. In (e) and (f) (resp. (k) and (l)), we recognize a significant improvement with the edgeness-based mutual information (resp. that the global maximum of intensity-based mutual information does not correspond to good registration).

where $X$ and $Y$ are the RVs associated to the image intensities of the reference and floating image, respectively. $q_X$ and $q_Y$ are the number of bins used for the density estimation of $X$ and $Y$ and $\vec{t}$ are the parameters of the geometric transformation $T$ of the floating image. $d$ 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. 7.

In order to indicate more the benefits and importance of combining segmentation and registration into one single optimization scheme, let us sketch the mutual information and feature efficiency of order $\frac{1}{2}$ for the initial noisy images of Fig. 7g and i their quantization results with respect to translations away from their optimal registration. The plots are shown in Fig. 8. In particular, two important facts of the theoretical expectations outlined in Section 4.2 can be reconfirmed. First of all mutual information is as expected unable to segment the initial noisy images during the registration task: mutual information of the initial data is larger than mutual information of the optimally quantized data (Fig. 8a). And second, feature efficiency of order $\frac{1}{2}$ has a much more pronounced maximum at optimal registration for the quantized images than for the original data

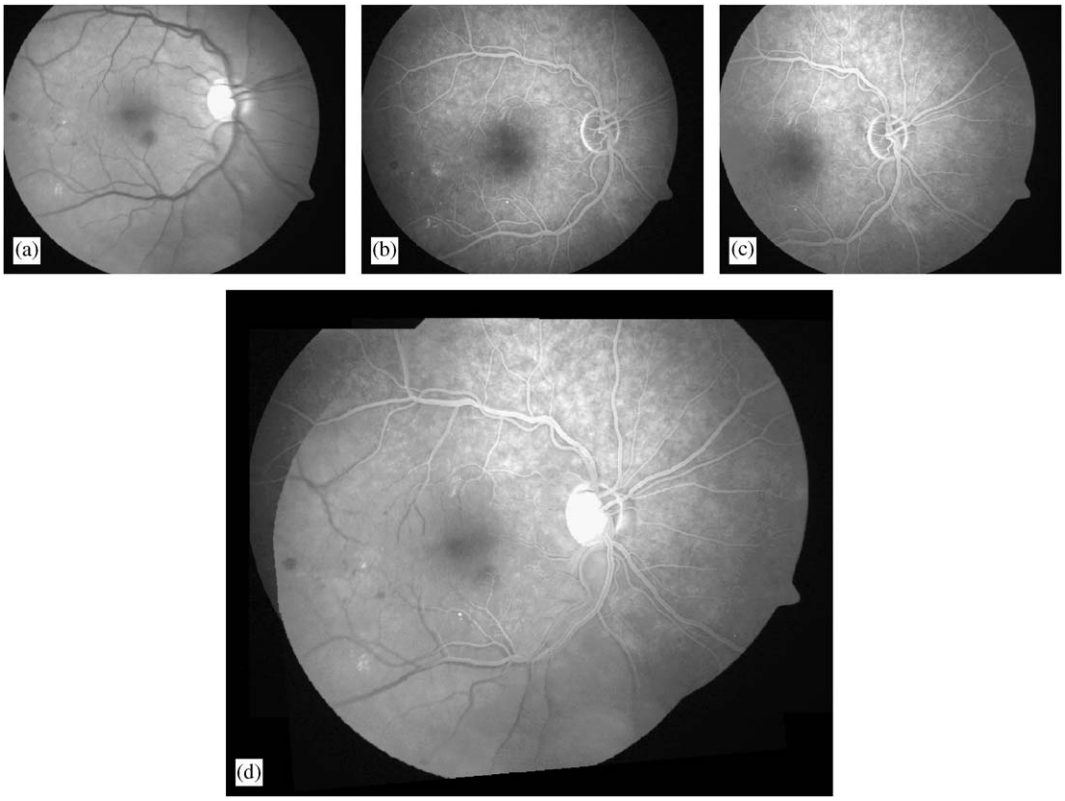


Fig. 6. The figures (a)–(c) had first to be registered in order to reconstruct the extended view shown in (d).

(Fig. 8b). The second point is in fact to a smaller degree also true for mutual information.

*6.1.3. Image registration with bias correction*

Interventional imaging modalities suffer frequently from a large bias field. Bias field is a standard term used in magnetic resonance imaging to define a smooth variation of the gray level values along the acquired image. This may be due to different causes linked to the MR scanner, such as poor RF coil uniformity, static field inhomogeneity, RF perturbation, etc. [42] 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 data sets. In this section, we want to show that the presented framework easily allows to register images with large bias fields. The approach simply combines minimum entropy bias-correction [42,43] with mutual information-based image registration. From the developed theory, one can recognize immediately that mutual information is not appropriate for this task as minimizing entropy contradicts obviously the maximum mutual information principle of Eqs. (40) and (44). Therefore, maximizing directly mutual information would not correct the bias field even though the error bounds of Eqs. (40) and (44) would be minimized. Just as in the previous paragraph, this is a typical example of inefficient features. Rather than maximizing mutual information, we want to maximize the efficiency coefficient of the bias-corrected image intensities/features (Eq. (47)).

The resulting mathematical formalism can thereafter be written as follows: let $\vec{p} \in \mathbf{R}^{d_1}$ parameterize the polynomial bias-correction of [42,43], where $d_1$ is determined by the degree of the polynomials. Furthermore, we have to determine the parameters $\vec{t} \in \mathbf{R}^{d_2}$ of the geometric

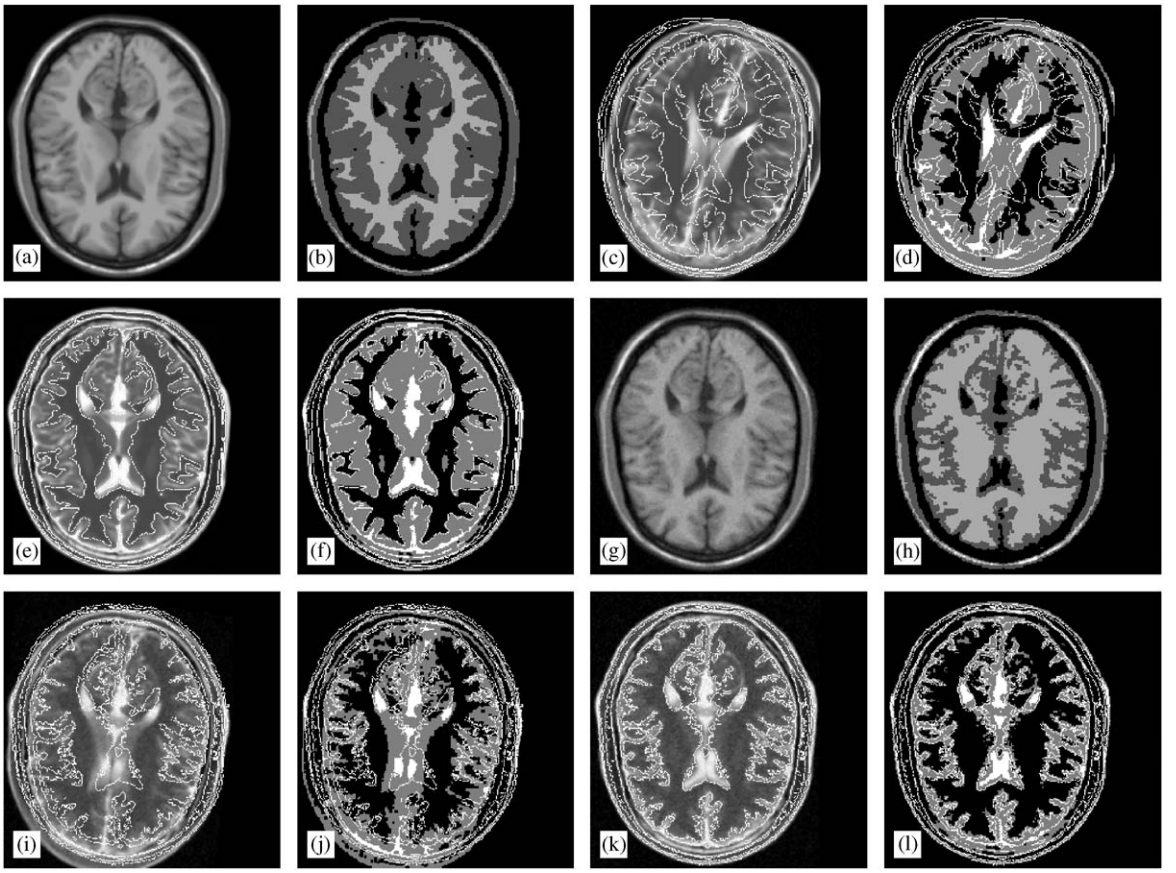


Fig. 7. Image (a) shows the reference image and (c) the initial floating image. In (e), we show the rigidly registered result. Images (b), (d) and (f) show the quantized outputs of (a), (c) and (e) with the optimal number of bins. Images (g)–(l) show an experiment equivalent to (a)–(f), but with noisier data sets. The contours of (b), resp. (h), are outlined in (c)–(f), resp. (i)–(l).

transformation $T$, where $d_2$ is the number of parameters that determines the transformation. In our specific application, we optimized over rigid-body transformations ($d_1$ equaled 6). Mathematically, we have

$$[\vec{t}^{\,\mathrm{opt}}, \vec{p}^{\,\mathrm{opt}}] = \arg \max_{[\vec{t},\vec{p}]\in\mathbf{R}^{d_1+d_2}} e_{1/2}(X, T_{\vec{t}}(P_{\vec{p}}(Y))), \tag{66}$$

where the parameters $\vec{p}^{\,\mathrm{opt}}$ specify the optimal bias-correction and $\vec{t}^{\,\mathrm{opt}}$ determines the optimal rigid transformation. Here $X$ and $Y$ refer to the RVs associated to the image intensities. Fig. 9 presents the results.

### 6.2. Speech-video sequences

In this application, we want to determine the region in a video scene that contains the speaker's mouth, i.e. where the motion seen in the image sequence corresponds to the audio signal [44].

With the framework presented before we will find the features in the audio and video signal that minimize the lower bounds on the error probabilities of Eqs. (40) and (44) in the region of the speaker's mouth [45]. The sampling RV $O$ of the Markov chains of Eqs. (35) and (36) now refers to the time index in the sequence and not to spatial coordinates as for medical image registration, as

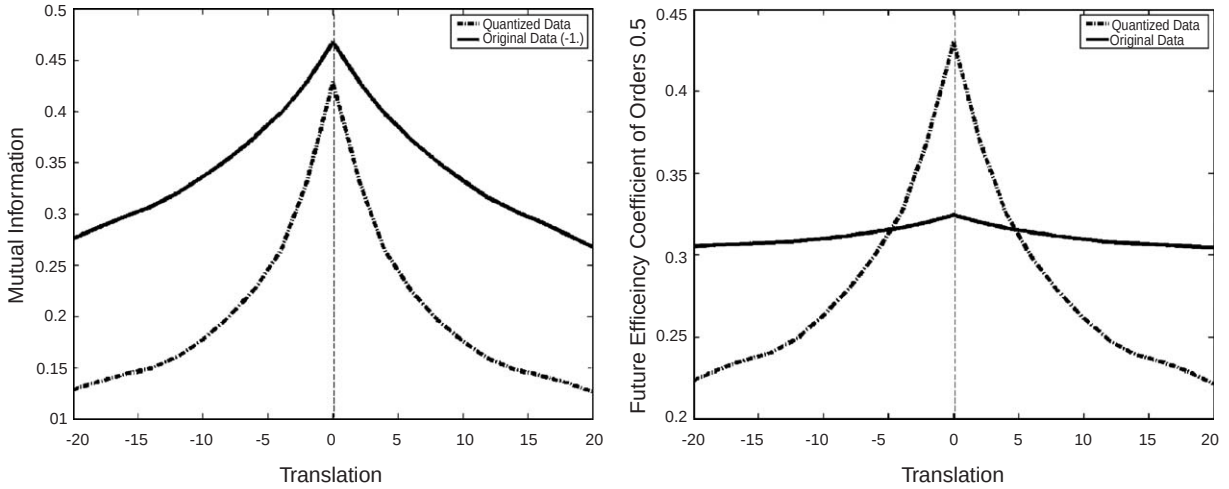


Fig. 8. Figure (a) compares mutual information of the original noisy images of Fig. 7g and k with the mutual information of their quantization results for translations away from optimal registration. For better comparison, the plot of the mutual information between the original data sets in (a) was moved down by one unity (−1). In (b) we see feature efficiency of order $\frac{1}{2}$ for the same images as for (a). Plot (a) shows that maximization of mutual information is unable to perform simultaneous registration and segmentation, contrary to the feature efficiency of (b). Also, we see that simultaneous segmentation improves the general behavior of the optimization objectives, in particular, for feature efficiency of order $\frac{1}{2}$. Let us recall that feature efficiency of order $\frac{1}{2}$ is equivalent to the well-known and widely used normalized entropy.

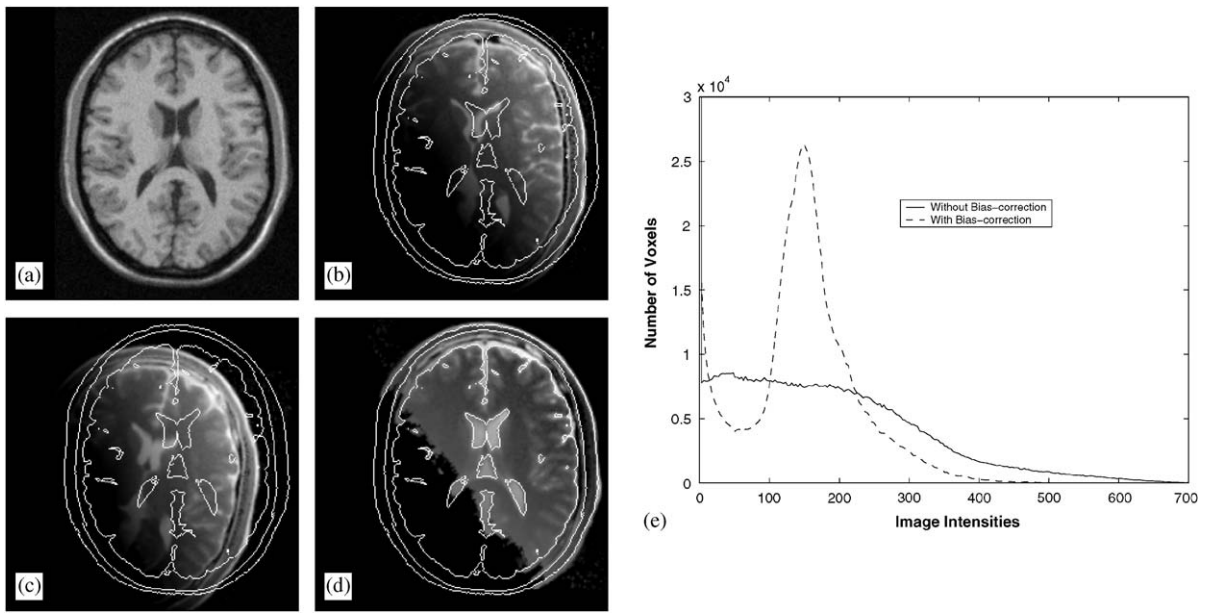


Fig. 9. We rigidly registered the image of (b) onto the reference image shown in (a). (c) Shows the bad result without simultaneous bias-correction and (d) shows good registration with simultaneous bias-correction (contours of the reference image (a) are shown in white). In (e) we show the histograms of (c) and (d), respectively, showing the effect of the bias correction on the grey level distribution.

the signal acquisition is now performed along a time interval.

From Eqs. (40) and (44) it follows that small lower error bounds in the region of the speaker's mouth are equivalent to a large feature space mutual information $I(F_X, F_Y)$ in this region. $F_X$ and $F_Y$ stand for the audio and video features, respectively. On the other hand, a large bound should result in the regions where the movements are not caused by the speaker's lips and are therefore unrelated to the speech signal. So that is where $I(F_X, F_Y)$ should be small.

To represent the information of the audio signal, we first converted it into a power spectrum (Fig. 10a). In order to deal with this multi-dimensional audio signal, we included a linear feature extraction step in the algorithm. As for any couple of RVs $A$ and $B$, we have $H(A) \geqslant I(A, B)$ and from Eqs. (40) and (44) we get a weakened lower bound for the error probabilities $P_{\{e_1,e_2\}}$:

$$P_{\{e1,e2\}} \geqslant 1 - \frac{I(F_X, F_Y) + 1}{\log n} \geqslant 1 - \frac{H(F_X) + 1}{\log n}. \quad (67)$$

Therefore, we looked for the linear combination of the power spectrum coefficients $W(f_i, t)$ (Fig. 10a) that carries most entropy. The finally obtained audio-feature is therefore defined by

$$F_X(t) = \sum_i \alpha_i^{\text{opt}} W(f_i, t), \quad (68)$$

with

$$\vec{\alpha}^{\text{opt}} = \arg \max_{\vec{\alpha}:|\vec{\alpha}|=1, \alpha_i \geqslant 0} H\left(\sum_i \alpha_i \cdot W(f_i, t)\right). \quad (69)$$

In Fig. 10b, we show for one sequence the weights $\alpha_i^{\text{opt}}$ that maximize the entropy of Eq. (68) and therefore define the audio-features $F_X$ of the audio signal.

We want to show two important points about the presented theory. First of all there exist features that relate the mouth movements of a speaker directly to the corresponding speech signal. On the other hand, we want to show that the choice of a particular feature representation is very crucial for the performance of the algorithm. There are features that contain lots of information (have lots of entropy), but are unrelated to the other signal. Other features represent this dependency much better and yield very good results.

The straightforward approach to quantify the dependency (in the sense of Eqs. (40) and (44)) between an audio and video signal of a speaker would consist of calculating the mutual information between the intensities of each pixel in the frame and the audio-feature of Eq. (68). In Fig. 11, we show the corresponding results.

We can see that this straightforward approach does not lead to the result we could have expected. It seems that the pixel intensities of the speaker's mouth do not carry much information about the audio signal. Instead, we propose a local feature

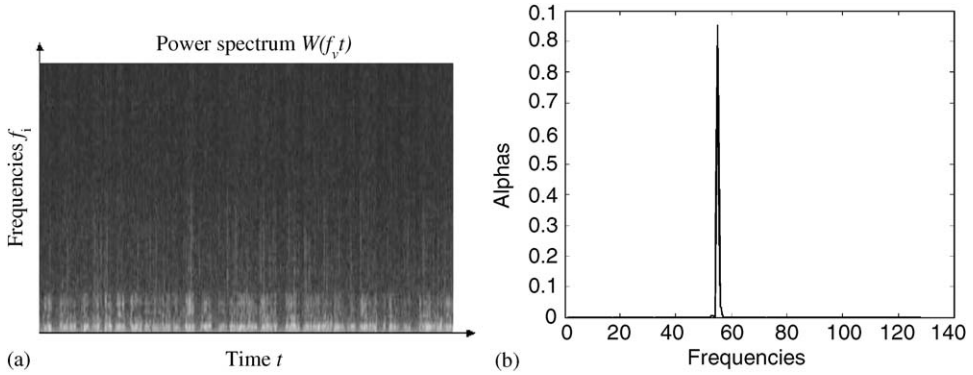


Fig. 10. (a) The power spectrum of the video sequence. At each time point we have the power coefficients of several frequencies. (b) The alphas for which the weighted sum of Eq. (68) has maximum entropy.

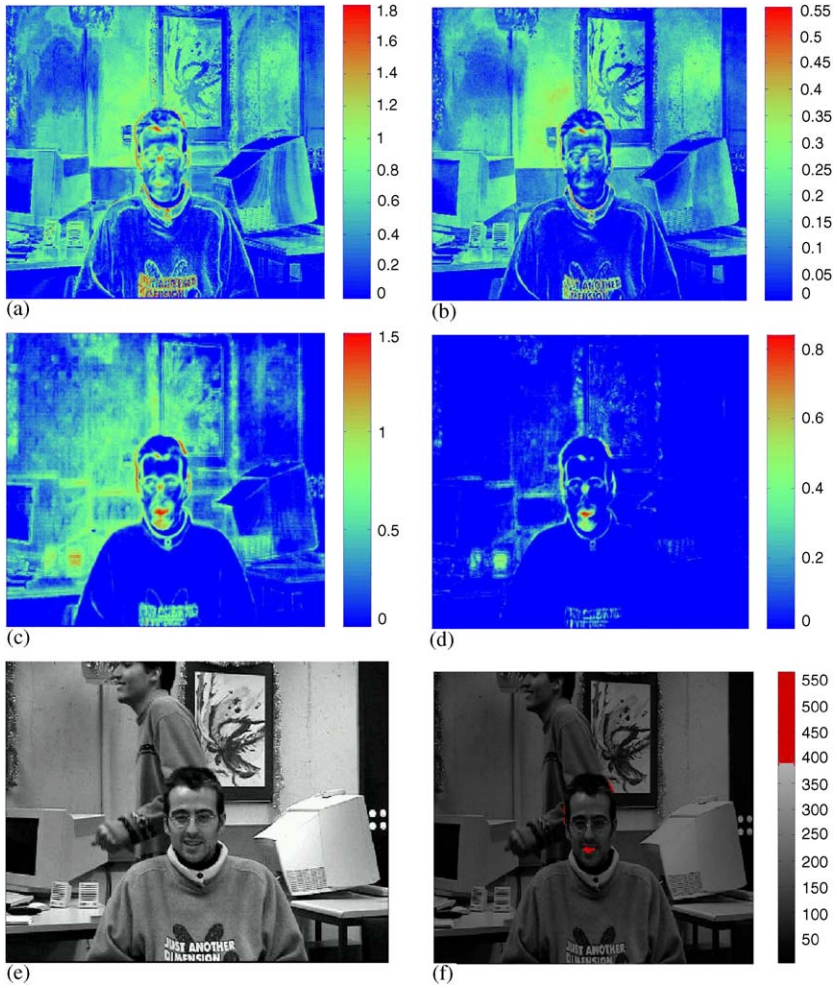


Fig. 11. (a) Intensity entropies for each pixel of the sequence. It shows that our sequence contained lots of motion in the background of the scene (people passing, waving arms, etc.). (b) Calculation of mutual information between the pixel intensities and the audio-feature. We see that there is not a particularly high mutual information in the region of the speaker's mouth. (c) Entropy of the video-features $F_Y$ for each pixel in the video scene. (d) Relation of this video information to the extracted audio-features $F_X$ of Eq. (68) by calculating the feature space mutual information $I(F_X, F_Y)$ for each pixel. (e) A typical frame of the sequence. (f) Thresholded image of (d) superposed on the frame of (e). It shows that the mutual information maxima lie clearly at the speaker's mouth.

that is more related to intensity changes (and therefore also to motion in the scene) than to the intensities themselves:

$$F_Y(i,j,t) = \sum_{l,m=-1}^{1} g_{t+1}(i+l,j+m) - g_{t-1}(i+l,j+m), \quad (70)$$

where $g_t(i,j)$ stands for the intensity of a pixel at coordinates $(i,j)$ in the frame at time $t$.

Thereafter we calculated for each pixel in the scene the mutual information between the resulting audio- and video-feature $I(F_X, F_Y)$. As shown in Fig. 11 a clear relationship between the speech and the speaker's mouth is obtained.

## 7. Discussion

We have shown that for medical image registration the choice of a good set of features to guide the registration is crucial. The results can vary significantly depending on what information the selected features carry. The comparison of Fig. 5 between image intensities and edgeness information shows that for some images the matching of boundaries can be more appropriate than the statistical matching of image intensities. Boundaries are small but very significant matching criteria, while image intensities might over-emphasize the large and therefore statistically important regions in the data sets. In terms of maximization of the mutual information between feature spaces, it appears that if we consider non-rigid registration using the voxel intensity as feature space, the maximum of MI is achieved not because of a good correspondence between the images but because of an increase in the marginal entropy of the MR image. This problem does not appear if we consider edgeness as feature (figure (f)), whose entropy is not increased by the geometric transformation.

Thereafter we used the developed framework to guide multi-modal signal processing in a unified framework. This means e.g. that image segmentation or bias correction can be incorporated into a generalized registration algorithm. We show that our framework is able to extract the information of the medical images, which is most important for the registration task, and to get rid of bias artifacts or background noise and information that can just corrupt the registration results.

The last experiment was performed on speech-video sequences, where we show two main points. First of all, we demonstrate that there exists a direct relationship between the speech signal and the video frames which can be explored e.g. multi-modal speaker localization. It is very important that our proposed approach does not make any hypothesis about their underlying relationship. The information theoretic framework rather quantifies non-parametrically their mutual dependency. As a second important point, we show that just as in the case of medical images, the choice of the right features is very crucial. Our framework gives a very flexible approach to construct feature selection algorithms such as proposed for medical image registration. Using this framework, we were able to re-confirm (Fig. 11) that motion estimation gives features much more related to the corresponding speech signal than pure image intensities (compare for example with [3]).

## 8. Conclusion

This paper presented two important points of information theoretic signal processing. The first one consisted of unifying a large class of algorithms in one single mathematical framework, using the information theoretical concepts of stochastic processes and their error probabilities. Combined with Fano's inequality and the data processing inequality, this mathematically compact framework allows the derivation and interpretation of optimization objectives which govern a wide range of information theoretic signal processing tasks.

The second main subject consisted of applying the introduced framework to the large field of multi-modal signal processing. We applied the theory successfully to several important and revealing problems of medical image and speech-video sequence processing.

## Acknowledgments

We would like to thank Conor Heneghan, Ph.D., from the Digital Signal Processing Laboratory, University College, Dublin, for providing the fundus images of Section 4 and Prof. Reto Meuli from the Radiology Department, University Hospital, Lausanne, for the MR-images of the same section. Also the support of the Swiss National Science Foundation projects 2153-055580.98, 20-64947.01 and the IM2 National Centre for Competence in Research is gratefully acknowledged.

## References

[1] J.W. Fisher III, J.C. Principe, A methodology for information theoretic feature extraction, in: World

Congress on Computational Intelligence, Anchorage, USA, March 1998.
[2] J.C. Principe, D. Xu, J.W. Fisher III, Learning from examples with information theoretic criteria, J. VLSI Signal Process. Syst. 26 (2000) 61–77.
[3] T. Darrell, J.W. Fisher III, P. Viola, W. Freeman, Audio-visual segmentation and the cocktail party effect, in: International Conference on Multimodal Interfaces, Beijing, China, 2000, pp. 32–40.
[4] A.O. Hero, B. Ma, O. Michel, Imaging applications of stochastic minimal graphs, in: IEEE International Conference on Image Processing, Thessaloniki, Greece, 2001, pp. 573–576.
[5] J.C. Principe, J.W. Fisher III, D. Xu, Information theoretic learning, Unsupervised Adaptive Filtering, Wiley, New York, 2000 (Chapter 7).
[6] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, R. Kikinis, Multi-modal volume registration by maximization of mutual information, Med. Image Analysis 1 (1) (1996) 35–51.
[7] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality image registration by maximization of mutual information, IEEE Trans. Med. Imaging 16 (2) (1997) 187–198.
[8] C. Studholme, D.J. Hawkes, D.L.G. Hill, A normalised entropy measure for multi-modality image alignment, in: SPIE Medical Imaging: Image Processing, San Diego, USA, 1998, pp. 132–142.
[9] A. Roche, G. Malandin, X. Pennec, N. Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Medical Image Computing and Computer-Assisted Intervention, Cambridge, USA, 1998, pp. 1115–1124.
[10] E. Gokcay, J.C. Principe, Information theoretic clustering, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 158–171.
[11] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[12] R.M. Fano, Transmission of Information: A Statistical Theory of Communication, MIT Press, Wiley, Cambridge, 1961.
[13] A. Renyi, Probability Theory, Elsevier Publishing Company, Amsterdam, 1970.
[14] D. Erdogmus, J.C. Principe, Information transfer through classifiers and its relation to probability of error, in: International Joint Conference on Neural Networks, Washington, DC, 2001.
[15] C.E. Shannon, A mathematical theory of communication, Bell Systems Technol. J. 27 (1948) 379–423.
[16] L. Devroye, L. Györfi, Non-parametric Density Estimation, Wiley, New York, 1985.
[17] L. Devroye, A Course in Density Estimation, Birkhäuser, Basel, 1987.
[18] S.P. Lloyd, Least squares quantization in PCM, Bell Laboratories Technical Note, 1957.
[19] S. Winkler, Vision models and quality metrics for image processing applications, Ph.D. Thesis, Swiss Federal Institute of Technology, 2000.
[20] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. 27 (3) (1956) 832–837.
[21] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076.
[22] A. Renyi, On measures of entropy and information, Fourth Berkeley Symposium on Probability Mathematics and Statistics 1 (1960) 547–561.
[23] M. Unser, A. Aldroubi, M. Eden, B-spline signal processing: Part I—theory, IEEE Trans. Signal Process. 41 (2) (1993) 821–833.
[24] M. Unser, A. Aldroubi, M. Eden, B-spline signal processing: Part II—Efficient design and applications, IEEE Trans. Signal Process. 41 (2) (1993) 834–848.
[25] T. Butz, J.-Ph. Thiran, Multi-modal signal processing: an information theoretical framework, Technical Report 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), 2002.
[26] T. Butz, O. Cuisenaire, J.-Ph. Thiran, Multi-modal medical image registration: from information theory to optimization objective, in: IEEE International Conference on Digital Signal Processing, Santorini, Greece, July 2002, vol. I, pp. 407–414.
[27] T. Butz, J.-Ph. Thiran, Feature-space mutual information for multi-modal signal processing, with application to medical image registration, in: XI. European Signal Processing Conference, Toulouse, France, September 2002, vol. I, pp. 3–10.
[28] J.N. Kapur, H.K. Kesavan, Entropy Optimization Principles with Applications, Academic Press, New York, 1992.
[29] C. Studholme, D.J. Hawkes, D.L.G. Hill, An overlap invariant entropy measure of 3D medical image alignment, Pattern Recogn. 32 (1999) 71–86.
[30] M. Holden, D.L.G. Hill, E.R.E. Denton, J.M. Jarosz, T.C.S. Cox, D.J. Hawkes, Voxel similarity measures for 3D serial MR brain image registration, in: Information Processing in Medical Imaging, Visegrd, Hungary, 1999, vol. 1613, pp. 466–471.
[31] A. Roche, G. Malandain, N. Ayache, Unifying maximum likelihood approaches in medical image registration, Technical Report 3741, Inst. National de Recherche en Informatique et en Automatique, Sophia Antipolis, July 1999.
[32] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, Reading, MA, 1989.
[33] M. Wall, GAlib 2.4.5: A C++ Library of Genetic Algorithm Components, Massachusetts Institute of Technology, 1999.
[34] W. Gropp, E. Lusk, A. Skjellum, Using MPI: Portable Parallel Programming with the Message-Passing Interface, second ed., MIT Press, Cambridge, MA, 1999.
[35] MPI Forum, A Message-Passing Interface Standard, University of Tennessee, 1999.
[36] J. West, J.M. Fitzpatrick, M.Y. Wang, B.M. Dawant, C.R. Maurer Jr., R.M. Kessler, R.J. Maciunas, Ch. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens,

D. Vandermeulen, P.A. van den Elsen, S. Napel, T.S. Sumanaweera, B. Harkness, P.F. Hemler, D.L.G. Hill, D.J. Hawkes, C. Studholme, J.B.A. Maintz, M.A. Viergever, G. Malandain, X. Pennec, M.E. Noz, G.Q. Maguire Jr., M. Pollack, Ch.A. Pelizzari, R.A. Robb, D. Hanson, R.P. Woods, Comparison and evaluation of retrospective intermodality brain image registration techniques, J. Comput. Assisted Tomogr. 21 (1997) 554–566.

[37] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, A.C. Evans, Design and construction of a realistic digital brain phantom, IEEE Trans. Med. Imaging 17 (3) (1998) 463–468.

[38] R.K.-S. Kwan, A.C. Evans, G.B. Pike, MRI simulation-based evaluation of image-processing and classification methods, IEEE Trans. Med. Imaging 18 (11) (1999) 1085–1097.

[39] P. Viola, W.M. Wells III, Alignment by maximization of mutual information, in: Fifth International Conference on Computer Vision, 1995, pp. 16–23.

[40] T. Butz, J.-Ph. Thiran, Affine registration with feature space mutual information, in: Medical Image Computing and Computer-Assisted Intervention, Utrecht, The Netherlands, 2001, pp. 549–556.

[41] G. Gerig, O. Kübler, R. Kikinis, F.A. Jolesz, Nonlinear anisotropic filtering of MRI data, IEEE Trans. Med. Imaging 11 (2) (1992) 221–232.

[42] E. Solanas, J.-Ph. Thiran, Exploiting voxel correlation for automated MRI bias field, in: Medical Image Computing and Computer-Assisted Intervention, Utrecht, The Netherlands, 2001, pp. 1220–1221.

[43] B. Likar, M.A. Viergever, F. Pernus, Retrospective correction of MR intensity inhomogeneity by information minimization, in: Medical Image Computing and Computer-Assisted Intervention, Pittsburgh, PA, 2000, pp. 375–384.

[44] J.W. Fisher III, T. Darrell, W.T. Freeman, P. Viola, Learning joint statistical models for audio-visual fusion and segregation, in: Advances in Neural Information Processing Systems, Denver, USA, November 2000.

[45] T. Butz, J.-Ph. Thiran, Feature space mutual information in video-speech sequences, in: IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 2002, vol. II, pp. 361–364.