

# FROM ERROR PROBABILITY TO INFORMATION THEORETIC SIGNAL AND IMAGE PROCESSING

THÈSE N° 2798 (2003)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION D'ÉLECTRICITÉ

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

**Torsten BUTZ**

ingénieur physicien diplômé EPF  
de nationalité suisse et originaire de Teufen (AR)

acceptée sur proposition du jury:

Dr J.-P. Thiran, directeur de thèse

Dr S. Bengio, rapporteur

Dr M. Bierlaire, rapporteur

Dr O. Cuisenaire, rapporteur

Dr M. Kaus, rapporteur

Prof. M. Kunt, rapporteur

Prof. S. Warfield, rapporteur

Lausanne, EPFL  
2003

---

# Contents

---

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction and Preview</b>                           | <b>1</b>  |
| 1.1      | Preview of the Thesis . . . . .                           | 2         |
| 1.1.1    | Why Information Theory for Multi-modal Signals? . . . . . | 2         |
| 1.1.2    | Feature Efficiency Coefficient . . . . .                  | 6         |
| 1.1.3    | Multi-modal Medical Image Processing . . . . .            | 6         |
| 1.2      | Outline . . . . .   | 7         |
| <b>2</b> | <b>Probability Theory</b>                                 | <b>9</b>  |
| 2.1      | Random Variables . . . . .                                | 9         |
| 2.1.1    | Continuous Random Variables . . . . .                     | 10        |
| 2.1.2    | Discrete Random Variables . . . . .                       | 10        |
| 2.1.3    | Multivariate Random Variables . . . . .                   | 11        |
| 2.1.4    | Conditional Probabilities . . . . .                       | 11        |
| 2.2      | Expectation and Variance . . . . .                        | 12        |
| 2.3      | Probability Estimation . . . . .                          | 13        |
| 2.3.1    | Parametric Estimation . . . . .                           | 14        |
| 2.3.2    | Non-parametric Estimation . . . . .                       | 16        |
| 2.4      | Summary . . . . .   | 18        |
| <b>3</b> | <b>Information Theoretic Signal Processing</b>            | <b>19</b> |
| 3.1      | Signal Processing as a Stochastic Process . . . . .       | 19        |
| 3.2      | Some Information Theoretic Concepts . . . . .             | 21        |
| 3.2.1    | Shannon Entropy . . . . .                                 | 21        |
| 3.2.2    | Shannon Mutual Information . . . . .                      | 22        |
| 3.3      | Error Probability of Stochastic Processes . . . . .       | 23        |
| 3.4      | Error Probability: Bounds and Inequalities . . . . .      | 25        |
| 3.4.1    | Data Processing Inequality . . . . .                      | 25        |
| 3.4.2    | Fano's Inequality . . . . .                               | 26        |
| 3.5      | Summary . . . . .   | 26        |
| <b>4</b> | <b>From Error Probability to Optimization Objective</b>   | <b>29</b> |
| 4.1      | Signal Distortion . . . . .                               | 29        |
| 4.1.1    | Distortion as an Error Probability . . . . .              | 30        |
| 4.1.2    | Some Distortion Measures . . . . .                        | 31        |

|          |  |           |
|----------|--|-----------|
| 4.2      | From Error Probability to Feature Extraction . . . . .                         | 32        |
| 4.3      | Summary . . . . .  | 33        |
| <b>5</b> | <b>Target Applications</b>   | <b>35</b> |
| 5.1      | Medical Image Registration . . . . .   | 35        |
| 5.1.1    | Landmark Based Registration Methods . . . . .                                  | 36        |
| 5.1.2    | Segmentation Based Registration Methods . . . . .                              | 37        |
| 5.1.3    | Voxel Property Based Registration Methods . . . . .                            | 37        |
| 5.2      | Audio-video Joint Processing . . . . .   | 37        |
| 5.2.1    | Combination of Individual Features . . . . .                                   | 38        |
| 5.2.2    | Joint Processing . . . . .   | 38        |
| 5.3      | Statistical Brain Tissue Classification in Magnetic Resonance Images . . . . . | 38        |
| 5.3.1    | Semi-automatic Classification . . . . .  | 39        |
| 5.3.2    | Automatic Classification . . . . .   | 39        |
| 5.4      | Summary . . . . .  | 40        |
| <b>6</b> | <b>Multi-modal Feature Extraction</b>  | <b>41</b> |
| 6.1      | Multi-modal Stochastic Processes . . . . .                                     | 42        |
| 6.1.1    | From Error Probability to Multi-modal Signal Processing . . . . .              | 44        |
| 6.2      | Objective Functions for Multi-modal Signal Processing . . . . .                | 45        |
| 6.2.1    | Feature Efficiency . . . . .   | 45        |
| 6.2.2    | From Error Probability to Correlation Ratio . . . . .                          | 51        |
| 6.2.3    | From Error Probability to Maximum Likelihood . . . . .                         | 53        |
| 6.3      | Image Registration as Feature Selection . . . . .                              | 53        |
| 6.4      | Optimization . . . . .   | 54        |
| 6.4.1    | Parallel Genetic Optimization . . . . .  | 55        |
| 6.4.2    | Orthogonal Decomposition of Affine Transformations . . . . .                   | 55        |
| 6.4.3    | Parallel Steepest Gradient . . . . .   | 57        |
| 6.5      | Results . . . . .  | 57        |
| 6.5.1    | Multi-modal Medical Images . . . . .   | 57        |
| 6.5.2    | Speech-Video Sequences . . . . .   | 65        |
| 6.5.3    | Discussion . . . . .   | 68        |
| 6.6      | Summary . . . . .  | 68        |
| <b>7</b> | <b>Non-parametric, Non-supervised Classification</b>                           | <b>69</b> |
| 7.1      | Non-supervised Statistical Classification . . . . .                            | 69        |
| 7.1.1    | Finite Mixture Models . . . . .  | 70        |
| 7.1.2    | Markov Random Fields . . . . .   | 71        |
| 7.1.3    | Parametric Hidden Markov Models . . . . .                                      | 73        |
| 7.2      | Non-parametric Estimation of Classification Error . . . . .                    | 74        |
| 7.2.1    | An Analytical Expression of Classification Error . . . . .                     | 75        |
| 7.2.2    | Distortion Matrix for Classification . . . . .                                 | 78        |
| 7.2.3    | Distortion with Prior Information . . . . .                                    | 79        |
| 7.2.4    | Relationship to Information Potential . . . . .                                | 80        |
| 7.3      | Non-parametric Hidden Markov Models . . . . .                                  | 80        |
| 7.4      | Optimization . . . . .   | 82        |
| 7.4.1    | k-Nearest Neighbor Exchange . . . . .  | 82        |
| 7.4.2    | Feature Space Labeling . . . . .   | 84        |

---

|          |   |            |
|----------|---|------------|
| 7.4.3    | ICM for Non-parametric Hidden Markov Models . . . . . | 85         |
| 7.5      | Results . . . . .                                     | 85         |
| 7.5.1    | Distortion and Prior Information . . . . .            | 86         |
| 7.5.2    | Brain Segmentation . . . . .                          | 90         |
| 7.5.3    | Discussion . . . . .                                  | 92         |
| 7.6      | Summary . . . . .                                     | 92         |
| <b>8</b> | <b>Conclusion</b> . . . . .                           | <b>111</b> |
| 8.1      | Achievements . . . . .                                | 111        |
| 8.2      | Perspectives . . . . .                                | 112        |
| 8.2.1    | Other Interesting Error Bounds . . . . .              | 112        |
| 8.2.2    | Other Interesting Applications . . . . .              | 113        |
| <b>A</b> | <b>Appendix</b> . . . . .                             | <b>115</b> |
| A.1      | Translation of Chapter Quotes . . . . .               | 115        |
|          | <b>Bibliography</b> . . . . .                         | <b>119</b> |
|          | <b>Index</b> . . . . .                                | <b>129</b> |
|          | <b>Curriculum Vitæ</b> . . . . .                      | <b>131</b> |

---

# Abstract

---

The signal processing community is increasingly interested in using information theoretic concepts to build signal processing algorithms for a variety of applications. A general theory on how to apply the mathematical concepts of information theory to the field of signal processing would therefore be of great interest. This is one of the main goals of this thesis, namely to introduce a mathematical framework for information theoretic signal and image processing.

The framework is based on stochastic processes for information transmission and on the error probabilities associated to these transmissions. Within the developed model, the stochastic processes account for the signal processing tasks within probability space, and the error probabilities are the optimization functions that drive the algorithms towards the signal processing objectives. The resulting conceptual framework allows us to directly apply a large number of information theoretic concepts and formulae to signal processing, including lower error bounds for the error probabilities or concepts from rate-distortion theory. In order to illustrate the theoretic framework, we show that several existing information theoretic signal processing algorithms implicitly fit our general model. This allows us to study interesting relationships between several algorithms. More importantly, we also apply the theory to three important target applications, namely multi-modal medical image registration, audio-video joint processing, and non-parametric, non-supervised classification.

The first two applications are particular examples of the general concept of multi-modal feature extraction. Multi-modal feature extraction aims to determine those features in a pair of multi-modal signals that carry maximal mutual redundancy. This means that from the feature space representation of one signal we can predict the feature space representation of the second signal with low probability of error. After describing the mathematical basis, we illustrate the algorithm with examples of multi-modal medical image registration, where the algorithm adaptively extracts those features in the initial datasets which best perform the registration task. Again, this is done by determining those features which carry maximal mutual redundancy and therefore define optimally spatial registration. We also apply the model to audio-video signals to predict the localization of a speaker in a video scene from its corresponding speech signal. The resulting algorithms illustrate that the existence of features with large mutual redundancy in multi-modal signals can be used to improve multi-modal signal processing. Furthermore the general theory enables the construction of a wide range of completely new applications.

Another illustrative example of the general information theoretic signal processing framework consists of information theoretic classification. Even though the basic model for multi-modal feature extraction and classification is identical, the final mathematical expressions are different and complementary. This allows us to make very interesting analogies between these two distinct applications. In particular, it is interesting to see that in analogy to registration, also classification algorithms aim to minimize error probabilities. The entirely probabilistic nature of the classification framework allows us to add a hidden Markov random field to the algorithms, resulting in the promising concept

of *non-parametric* hidden Markov models. The classification algorithms are validated on synthetic and natural data. For instance, we apply the non-parametric hidden Markov model to the segmentation of medical images and obtain promising results in comparison to the state-of-the-art in this field.

In conclusion, the experimental results show that the introduced mathematical framework leads to interesting generalizations of existing signal processing tasks and to promising results for several newly derived signal processing algorithms.

---

# Kurzfassung

---

In der Signalverarbeitung tätige Forschungsgruppen machen in zunehmendem Maße Gebrauch von informationstheoretischen Ansätzen, um Signalverarbeitungsalgorithmen für eine Vielzahl von Anwendungen zu entwickeln. Eine allgemeingültige Theorie zur Anwendung informationstheoretischer Konzepte auf Problemstellungen der Signalverarbeitung wäre folglich von großem Interesse. Eines der wesentlichen Ziele dieser Doktorarbeit ist deshalb die Ausarbeitung und Anwendung eines mathematischen Rahmenwerkes für informationstheoretische Signal- und Bildverarbeitung.

Die Grundlagen bilden dabei stochastische Prozesse zur Informationsübertragung und die dazugehörigen Fehlerwahrscheinlichkeiten dieser Übertragungen. Die stochastischen Prozesse charakterisieren die Signalverarbeitungsalgorithmen innerhalb der Wahrscheinlichkeitstheorie, und die Fehlerwahrscheinlichkeiten stehen für die Optimierungsfunktionen, die diese Algorithmen zum Ziel der Signalverarbeitung führen. Das resultierende Modell läßt die Anwendung vieler informationstheoretischer Formeln, wie Grenzwerte für die Fehlerwahrscheinlichkeiten oder Konzepte der rate-distortion Theorie, im Bereich der Signalverarbeitung zu. Um den theoretischen Rahmen zu veranschaulichen, zeigen wir, daß mehrere existierende informationstheoretische Signalverarbeitungsalgorithmen anhand des entwickelten allgemeinen Modells interpretiert werden können. Dadurch lassen sich interessante wesentliche Zusammenhänge zwischen diesen Algorithmen studieren. Des weiteren wenden wir die in dieser Arbeit entwickelte Theorie zur Lösung von drei praktischen Problemen, der multi-modalen medizinischen Registrierung, der multimedialen Signalverarbeitung und der nicht-parametrischen, unüberwachten Datenklassifikation, an.

Die ersten zwei Anwendungen illustrieren exemplarisch ein allgemeines Konzept zur Bestimmung von Schlüsselmerkmalen in multi-modalen Signalen. Dieses Konzept zielt darauf ab, jene Merkmale in einem Paar multi-modaler Signale zu bestimmen, die maximale gegenseitige Abhängigkeit aufweisen. Das heißt, daß man von den Features eines Signals die Features des zweiten Signals mit minimaler Fehlerwahrscheinlichkeit voraussagen kann. Nach der Aufarbeitung der mathematischen Grundlagen veranschaulichen wir den Algorithmus mit Beispielen der multi-modalen medizinischen Bildregistrierung, in der dieser Algorithmus flexibel jene Merkmale in den Datensätzen bestimmt, die diese Registrierung am besten ausführen können. Dies sind jene Merkmale, die in beiden Datensätzen vorkommende Strukturen repräsentieren und deshalb die räumliche Registrierung optimal definieren. Außerdem wenden wir das Modell an multimedialen (audio-video) Signalen an, um so die sprechende Person in einer Videoszene anhand des dazugehörigen Audiosignals zu lokalisieren. Die resultierenden Algorithmen zeigen eine Verbesserung von vorhandenen multi-modalen Signalverarbeitungsalgorithmen, da sie die gegenseitige Abhängigkeit gewisser Merkmale ausnutzen und ausserdem die Merkmale mit maximaler Abhängigkeit adaptiv bestimmen. Darüber hinaus bietet die vorgestellte Theorie die Möglichkeit ein weites Spektrum von neuen multi-modalen Signalverarbeitungsalgorithmen zu entwickeln.

Ein weiteres illustratives Beispiel des allgemeinen informationstheoretischen Modells ist die in-

formationstheoretische Klassifikation von Daten. Obwohl die mathematischen Grundlagen hierzu identisch sind zur multi-modalen Bestimmung von Schlüsselmerkmalen, ergeben sich unterschiedliche mathematische Implementierungen, die sich gegenseitig komplementär ergänzen. Dies erlaubt es, sehr interessante Zusammenhänge zwischen diesen unterschiedlichen Anwendungen zu analysieren. Vor allem zeigen wir, daß sowohl Registrierungs- als auch Segmentierungsalgorithmen durch Fehlerwahrscheinlichkeiten, beziehungsweise deren Grenzwerte, charakterisiert werden können. Die Theorie zur nicht-parametrischen Klassifikation ist vollständig wahrscheinlichkeitstheoretisch, und so lassen sich die Algorithmen mit versteckten Markov Feldern kombinieren, was zu dem vielversprechenden Konzept der *nicht-parametrischen* versteckten Markov Modelle führt. Die Klassifikationsalgorithmen validieren wir an synthetischen und realen Daten. Außerdem wenden wir das nicht-parametrische Markov Modell auch zur Segmentierung medizinischer Datensätze an. Dabei erhalten wir vielversprechende Resultate im Vergleich zum existierenden State-of-the-Art auf diesem Gebiet.

Zusammenfassend zeigen die experimentellen Resultate, daß die vorgestellte Theorie eine vielversprechende Grundlage für eine Vielzahl von Signalverarbeitungsproblemen bildet.

---

# Version Abrégée

---

Les groupes de recherche en traitement des signaux empruntent de plus en plus les concepts de la théorie de l'information pour résoudre une grande variété de problèmes. Une théorie générale sur l'application des concepts mathématiques de la théorie de l'information dans le domaine du traitement des signaux serait donc de grand intérêt. Ceci est un des buts principaux de cette thèse, c.à.d. de présenter un cadre mathématique pour le traitement des signaux et images par la théorie de l'information.

Ce cadre est basé sur des processus stochastiques pour la transmission de l'information et sur les probabilités d'erreur associées à ces transmissions. Dans ce modèle, les processus stochastiques représentent les tâches algorithmiques du traitement des signaux dans l'espace des probabilités, et les probabilités d'erreur sont les fonctions d'optimisation qui conduisent les algorithmes du traitement des signaux vers ces objectifs. Le cadre conceptuel résultant permet d'appliquer directement un grand nombre de formules de la théorie de l'information au traitement des signaux, telles que des bornes inférieures de la probabilités d'erreur ou les concepts de la théorie de débit-distortion. Afin d'illustrer le cadre théorique, nous démontrons que plusieurs algorithmes du traitement des signaux par la théorie de l'information s'intègrent à notre modèle général. Ceci nous permet d'étudier des relations intéressantes entre plusieurs algorithmes. Mais plus important, nous appliquons notre théorie à trois applications pratiques: le recalage des image médicales multi-modales, le traitement joint des signaux audio-vidéo, et la classification non-paramétrique, non-supervisée.

Les deux premières applications sont des exemples particuliers du concept général de l'extraction multi-modale des caractéristiques. Cette extraction vise à déterminer les caractéristiques dans une paire de signaux multi-modaux qui portent une redondance mutuelle maximale. Ceci signifie qu'à partir des caractéristiques du premier signal on peut prévoir les caractéristiques du deuxième signal avec une probabilité d'erreur minimale. Après avoir décrit la base mathématique, nous illustrons l'algorithme avec des exemples du recalage médical multi-modal, où l'algorithme extrait de manière adaptative ces caractéristiques des données initiales qui effectuent le recalage le mieux, et nous appliquons le modèle aux signaux audio-vidéo pour déterminer la localisation de la personne parlante dans une scène vidéo à partir de l'audio correspondant. Les résultats des algorithmes montrent une amélioration des solutions existantes du traitement des signaux multi-modaux et une possibilité prometteuse de construire une variété de nouvelles applications.

La classification par la théorie de l'information est un autre exemple illustratif du modèle théorique et général. Bien que le modèle de base soit identique à celui de l'extraction multi-modale des caractéristiques, le formalisme mathématique résultant est différent et complémentaire. Ceci nous permet d'étudier des analogies très intéressantes entre ces applications différentes. De plus, le cadre entièrement probabiliste de la classification nous permet d'ajouter aux algorithmes un champ de Markov caché, ayant pour résultat le concept prometteur des modèles de Markov cachés *non-paramétriques*. Les algorithmes de classification sont illustrés sur des données synthétiques et réelles.

En particulier, nous appliquons le modèle de Markov caché non-paramétrique au problème de la segmentation des images médicales et obtenons des résultats prometteurs en comparaison à l'état de l'art dans ce domaine.

En conclusion, les résultats expérimentaux démontrent que le cadre mathématique qu'on a présenté mène à des généralisations intéressantes pour plusieurs applications du traitement des signaux existantes et à des résultats prometteurs pour les nouveaux algorithmes développés.