République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Département d'Electronique

Laboratoire des Dispositifs de Communications et de Conversion Photovoltaïque

Thèse de Doctorat en Sciences

Présentée par

M^me^ Atidel LAHOULOU ép. BOURAOUI

Intitulée

MACHINE LEARNING TECHNIQUES FOR IMAGE QUALITY EVALUATION

Soutenue publiquement le **03 Décembre 2012**

Devant le jury :

	Nom	Prénom	Etablissement
Président	HAMAMI	Latifa	Professeur, E. N. P. (Algérie)
Rapporteurs	HADDADI	Mourad	Professeur, E. N. P. (Algérie)
	VIENNET	Emmanuel	Professeur, Univ. Paris13 (France)
Examinateurs	SMARA	Youcef	Professeur, U.S.T.H.B. (Algérie)
	KOUDIL	Mouloud	Professeur, E.S.I. (Algérie)
	GUESSOUM	Abderrezak	Professeur, Univ. de Blida (Algérie)
Membre invité	BOURIDANE	Ahmed	Professeur, Northumbria univ. (RU)

ملخـص : يعتبر قياس جودة الصور عنصرا بالغ الأهمية في كل من الخدمات المرئية و كذا أنظمة معالجة الصور التي تستهدف مراقبين بشرا. الهدف الأول لهذه الأطروحة هو تقديم تقييم كمي كامل و شامل للأداء التنبؤي لمجموعة متنوعة من المعايير الموضوعية لجودة الصور كاملة المرجع تـم اختـيارها على أسـاس كونـها الأكثر استـعمالا، و قد طبق منهـج التقـييم عـلى ستة قواعد للبيانات مخصـصة لتقييم جودة الصور، وهي مصنـفة حسب تقديـر ذاتي غير موضـوعي و يمكـن تحميلهـا على الانترنت. الهدف الثاني هو تحديد مميزات الصور الأكثر ملائمة لتقييم جودتها، و قد استخدمت طريقتان لاختيار المميزات بما في ذلك طريقة التقليل من المخاطر الهيكلية و كذا المناهج القائمة على الشبكات العصبية. الهدف الثالث لهذا البحث هو استغلال تقنيات التعلم الآلي تحت الإشراف و لاسيما نموذج البرسبترون متعدد الطبقات للتقدير الآلي لجودة الصور. النظام المبتكر يتعلم من خلال درجات الجودة الذاتية و يمكنه بناء نموذج قادر على إنتاج قياسات موضوعية تتوافق مع تلك الذاتية لأي صورة كانت يتم تقديمها له. الهدف الرئيسي من هذا التطوير هو تحسين الأداء التنبؤي وفقا للترابط، الرتابة و الدقة، و لتحقيقه تم استخدام دالة التكلفة الافتراضية القائمة على أساس الخطأ لتصميم المعيار الأول (الذي سميناه ECF)، ثم قمنا بتخصيص هذه الدالة على أساس الترابط لتطوير قياس جديد ثان (سميناه CCF). بعد مقارنة هذين المقياسين بالنسبة لثمانية عشر خوارزميا آخر خلال ثلاثة قواعد بيانية تبين أن خوارزميي ECF وCCF يأخذان بعين الاعتبار الطبيعة اللاخطية للنظام البصري البشري.ECF هو أكثر دقة من معظم القياسات قيد الدراسة في حين أن CCF يتفوق على جميعها من حيث الترابط و بالتالي الرتابة.

Résumé : L'évaluation de la qualité d'image présente un intérêt substantiel pour les services ainsi que pour les systèmes de traitement d'images où le dernier maillon de la chaîne est l'observateur humain. Le premier objectif de cette thèse est de fournir une évaluation statistique complète et approfondie des performances prédictives d'une large variété de mesures objectives de qualité avec référence complète sur un certain nombre de bases de données étiquetées avec des scores indiquant la qualité subjective des images. Le second objectif consiste à définir les attributs de l'image les plus pertinents pour l'évaluation de sa qualité. Deux méthodes de sélection de caractéristiques ont été utilisées, à savoir la minimisation du risque structurel et l'approche basée sur le modèle connexionniste. Le troisième objectif de ce travail de recherche est d'exploiter les techniques d'apprentissage supervisé, en particulier le modèle du perceptron multicouche, pour l'estimation automatique de la qualité de l'image. Le système apprend à partir des étiquettes de la qualité subjective et construit un modèle capable de continuer à fournir une mesure objective toujours corresponde à l'avis de l'homme à toute image qui lui est présentée. Le but principal était d'optimiser la performance prédictive des mesures développées en fonction de la corrélation, la monotonicité et la précision. La fonction de coût par défaut basée sur l'erreur a été employée pour la première mesure développée (que nous avons appelé ECF) et une fonction de coût personnalisée basée sur la corrélation a été proposée pour concevoir la deuxième mesure (que nous avons appelé le CCF). L'étude comparative de ces deux nouvelles métriques à dix-huit autres algorithmes de qualité d'image avec référence complète sur trois bases de données de qualité d'image montre que les algorithmes d'ECF et CCF prennent en considération les non-linéarités du système visuel humain. L'ECF est plus précise que la majorité des mesures étudiées, tandis que la CCF améliore largement les résultats de toutes les métriques concurrentes en termes de corrélation et de monotonicité.

Abstract: Image quality assessment presents a substantial interest for image services that target human observers. The first objective of this thesis is to provide a complete and thorough statistical predictive performance assessment of a variety of full-reference objective quality measures over number of subjectively rated image quality databases. The second is to define the image attributes that are the most relevant to its quality evaluation. Two feature selection methods have been used including the structural risk minimization and the neural network based approaches. The third objective of this research work is to exploit the supervised machine learning techniques, especially the multilayer perceptron based model, for automatic image quality appreciation. The system learns from the subjective quality scores and builds a model capable to further provide an objective measure that continues to match with the human opinion to any other image. The main target was to optimize the predictive performance of the developed measures according to correlation, monotonicity and accuracy. The default cost function based on error was employed for the first developed measure (that we called ECF) and a customized cost function based on correlation was proposed to design the second metric (that we called CCF). The comparative investigation to eighteen other full-reference image quality algorithms over three image quality databases shows that both ECF and CCF take into consideration the nonlinearities of the human visual system. The ECF is more accurate than the majority of the metrics under study, while the CCF outperforms all its counterparts in terms of correlation and hence monotonicity.

ملخـص

يعتبر قياس جودة الصور عنصرا بالغ الأهمية في كل من الخدمات المرئية و كذا أنظمة معالجة الصور التي تستهدف مراقبين بشرا. في الواقع، يمكن قياس جودة الصور بطريقتين مختلفتين: الأولى تدعى "التقييم الذاتي" و هو النهج المباشر نظرا للطبيعة الشخصية لتقييم جودة البيانات البصرية، أما الطريقة الثانية فتدعى "التقييم الموضوعي" التي تنتج آليا قيما كمية لجودة الصورة. الهدف الأول لهذه الأطروحة هو تقديم تقييم كمي كامل و شامل للأداء التنبؤي لمجموعة متنوعة من المعايير الموضوعية لجودة الصور كاملة المرجع تم اختيارها على أساس كونها الأكثر استعمالا، و قد طبق منهج التقييم على ستة قواعد للبيانات مخصصة لتقييم جودة الصور، وهي مصنفة حسب تقدير ذاتي غير موضوعي و يمكن تحميلها على الانترنت. كل هذه العمليات قد نفذت بالنسبة لأربعة أنواع من مشوهات الصور المتمثلة في ضغط JPEG ، ضغط JPEG2000 ، التمويه الضبابي والضوضاء التمويهي. الهدف الثاني هو تحديد مميزات الصور الأكثر ملائمة لتقييم جودتها، و قد استخدمت طريقتان لاختيار المميزات بما في ذلك طريقة التقليل من المخاطر الهيكلية و كذا المناهج القائمة على الشبكات العصبية. سمحت لنا هذه الخطوة بتطوير معيارين موضوعيين جديدين جزئيي المرجع لتقييم جودة الصورة، حيث يتطلب ذلك استخدام مميزات محددة لكل من صورة المرجع و صورة الاختبار. الهدف الثالث لهذا البحث هو استغلال تقنيات التعلم الآلي تحت الإشراف و لاسيما نموذج البرسبترون متعدد الطبقات للتقدير الآلي لجودة الصور. النظام المبتكر يتعلم من خلال درجات الجودة الذاتية و يمكنه بناء نموذج قادر على التعميم بعد التدريب، بعبارة أخرى يجب على النظام القدرة على إنتاج قياسات موضوعية تتوافق مع تلك الذاتية لأي صورة كانت يتم تقديمها له. الهدف الرئيسي من هذا التطوير هو تحسين الأداء التنبؤي وفقا للترابط، الرتابة و الدقة، و لتحقيقه تم استخدام دالة التكلفة الافتراضية القائمة على أساس الخطأ لتصميم المعيار الأول (الذي سميناه ECF)، ثم قمنا بتخصيص هذه الدالة على أساس الترابط لتطوير قياس جديد ثان (سميناه CCF). بعد مقارنة هذين المقياسين بالنسبة لثمانية عشر خوارزميا آخر خلال ثلاثة قواعد بيانية تبين أن خوارزميي ECF وCCF يأخذان بعين الاعتبار الطبيعة اللاخطية للنظام البصري البشري. ECF هو أكثر دقة من معظم القياسات قيد الدراسة في حين أن CCF يتفوق على جميعها من حيث الترابط و بالتالي الرتابة.

الكلمات الرئيسية: جودة الصورة، موضوعية، ذاتية، الأداء التنبؤي، اختيار المميزات، برسبترون متعدد الطبقات، دالة التكلفة.

Résumé

L’évaluation de la qualité d'image présente un intérêt substantiel pour les services ainsi que pour les systèmes de traitement d'images où le dernier maillon de la chaîne est l’observateur humain. En effet, la qualité d'image peut être mesurée de deux manières différentes. La première, appelée «évaluation subjective de la qualité", est l'approche évidente étant donnée la nature subjective de la qualité visuelle des médias. La seconde est appelée «évaluation objective de la qualité» qui permet de produire automatiquement des valeurs mesurant la qualité de l'image de manière quantitative. Le premier objectif de cette thèse est de fournir une évaluation statistique complète et approfondie des performances prédictives d'une large variété de mesures objectives de qualité avec référence complète sur un certain nombre de bases de données étiquetées avec des scores indiquant la qualité des images qui sont évaluées de manière subjective selon des protocoles strictes. Le second objectif consiste à définir les attributs de l'image qui sont les plus pertinents pour l’évaluation de sa qualité. Deux méthodes de sélection de caractéristiques ont été utilisées, à savoir la minimisation du risque structurel et l’approche basée sur le modèle connexionniste. Cela nous a permis de développer deux nouvelles métriques objectives de qualité d'image avec référence réduite où l’estimation de la qualité de l'image nécessite l'utilisation de seulement quelques uns des descripteurs de l’image de référence et celle de test. Le troisième objectif de ce travail de recherche est d'exploiter les techniques d'apprentissage supervisé, en particulier le modèle du perceptron multicouche, pour l’estimation automatique de la qualité de l'image. Le système apprend à partir des étiquettes de la qualité subjective issues des bases d’images utilisées et construit un modèle capable de généraliser après un certain temps d’entraînement. En d’autres termes, le modèle doit continuer à fournir une mesure objective toujours corresponde à l'avis de l'homme à toute image qui lui est présentée. L'objectif principal était d'optimiser la performance prédictive des mesures développées en fonction de la corrélation, la monotonicité et la précision. La fonction de coût par défaut basée sur l'erreur a été employée pour la première mesure développée (que nous avons appelé ECF) et une fonction de coût personnalisée basée sur la corrélation a été proposée pour concevoir la deuxième mesure (que nous avons appelé le CCF). L’étude comparative de ces deux nouvelles métriques à dix-huit autres algorithmes de qualité d'image avec référence complète sur trois bases de données de qualité d'image montre que les algorithmes d’ECF et CCF prennent en considération les non-linéarités du système visuel humain. L'ECF est plus précise que la majorité des mesures étudiées, tandis que la CCF améliore largement les résultats de toutes les métriques concurrentes en termes de corrélation et de monotonicité.

Mots clés: qualité d’image, objective, subjective, performance prédictive, sélection de variables, perceptron multicouche, fonction de coût.

Abstract

Image quality assessment presents a substantial interest for image services that target human observers. Indeed, Image quality can be measured in two different ways. The first, called “subjective quality assessment”, is the obvious approach given the subjective nature of the visual data quality. The second one is called “objective quality assessment” that automatically allow to produce values that score image quality. In fact, the first objective of this thesis is to provide a complete and thorough statistical predictive performance assessment of a variety of full-reference objective quality measures over number of subjectively rated image quality databases. The second is to define the image attributes that are the most relevant to its quality evaluation. Two feature selection methods have been used including the structural risk minimization and the neural network based approaches. This allowed us to develop two new objective reduced-reference image quality metrics where the image quality assessment requires the use of only a few features of the reference and the test images. The third objective of this research work is to exploit the supervised machine learning techniques, especially the multilayer perceptron based model, for automatic image quality appreciation. The system learns from the subjective quality scores and builds a model capable to further provide an objective measure that continues to match with the human opinion to any other image. The main target was to optimize the predictive performance of the developed measures according to correlation, monotonicity and accuracy. The default cost function based on error was employed for the first developed measure (that we called ECF) and a customized cost function based on correlation was proposed to design the second metric (that we called CCF). The comparative investigation to eighteen other full-reference image quality algorithms over three image quality databases shows that both ECF and CCF take into consideration the nonlinearities of the human visual system. The ECF is more accurate than the majority of the metrics under study, while the CCF outperforms all its counterparts in terms of correlation and hence monotonicity.

Keywords: image quality, objective, subjective, predictive performance, feature selection, multilayer perceptron, cost function.

Acknowledgements

*I would like to first thank **Emmanuel Viennet** for his patience, valuable advices and ceaseless efforts to keep my research and this thesis as simple and concise as possible. Next, I would like to thank **Mourad Haddadi** for the help he has provided along the way. A special note of thanks to **Ahmed Bouridane** for his suggestions, guidance and insurance.*

*I also want to acknowledge my thesis jury consisting of the jury president **Pr. Latifa Hamami**, my supervisor **Pr. Mourad Haddadi** both from Ecole Nationale Polytechnique (ENP) and the jury members **Pr. Youcef Smara** from Université des Sciences et de la Technologie Houari Boumediène (USTHB), **Pr. Ahmed Bouridane** from Northumbria University of Newcastle (UK), **Pr. Mouloud Koudil** from the Ecole Supérieure d'Informatique (ESI), **Pr. Abderrezak Guessoum** from Université Saad Dahlab de Blida (USDB). Thank you for accepting to assess my work.*

*I offer my special gratitude to my coach, my husband **Fayçal Bouraoui** for his tireless sacrifice he has made on my behalf. I would especially like to appreciate your belief in me and your support always with the same energy.*

*I am forever in debt to my friends in the L2TI laboratory (Université Paris 13) with whom I have spent 18 months. That period was an excellent social, cultural and scientific adventure. My particular greetings go to **Emmanuel**, **Azeddine**, **Anissa**, **Marie**, **Ghiles**, **Arnaud**, **Abdelhak**, **Chen**, **Bao**, **Bang**, **Aladine**, **Zehira** and **Vianey** for their great and valuable interaction.*

I have not had the honour to work within a team at ENP but I salute all the staff and the research teams. I consider myself one of them.

*I am very grateful to my family for their patience and support. A very special thanks to my **Mother** for all her support and advice my whole life. I would not be here without her, my uncles, aunts, brother, sister and their respective families. They always gave me the possibility to make my dreams come reality. Also, a pious thought to my grandparents.*

*To **RANIA**, my daughter, my big love.*

Thanks to you all – I could not have done this without a single one of you.

Table of Contents

List of Figures

List of Tables

CHAPTER

1

Chapter 1

General Introduction

1.1 Motivation

The primary source of information we constantly need to acquire about our world is visual. Indeed, the ease with which huge amounts of digital visual data is being transmitted and/or exchanged over the Internet, every minute, increases the use of visual information that have pervaded our everyday lives. Without being aware, we spontaneously make judgements about the perceptual quality of the visual content (images, videos, movies, 3D drawings, etc) we regularly see on our screens. We may open wide the eyes or may make a grimace showing disappointment when we see a visual content that we deem of very good or of poor quality, respectively. In an attempt to satisfy the increasing quality requirements of the target human observers, particular attention has been drawn on the quality measurement of the visual data. Moreover, assessing visual data quality enables to adjust the parameters of data processing techniques in order to maximize their quality or to reach a given satisfaction. Indeed, a great deal of research has been devoted to image and video content quality assessment.

Image and video quality assessment is an important component in most modern visual data processing systems. This research field is of increasing interest and therefore considerable effort has been committed to the development of new image/video quality evaluation tools in the last couple of decades. This interest goes hand-in-hand with the emergence of many new applications that require automatic real-time media quality assessment. Quality

monitoring of massive data transmission over networks and automated media quality measurement for printing systems are typical examples of such latest applications.

The challenge here relates to stringent requirements to consider the perceptual quality of image and video in the data storage and transmission devices. In addition, advanced visual data processing benchmarking tools rely on subjective criteria related to human vision. It is widely agreed that the Human Visual System (HVS) is able to quickly appreciate the quality of an image or a video sequence even if its original version is absent, which suggests that it is probably based on a high level interpretation of the visual data, using a lot of knowledge about the scene at hand.

Since visual data quality is subjective in nature, its evaluation based on experiments with subjective ratings is a broadly accepted solution. The first subjective quality rating tests have been conducted on images which we examine along this research work. Hence, different methodological options have appeared to construct subjective image quality databases. Subjective quality tests are, in general, divided into several sessions of limited duration and consist of a number of subjects invited to judge the quality of a set of images under particular conditions. Typically, images are divided into two groups: *reference* images considered as pristine versions and processed ones that were subject to some degradation and were used as *test* images. Whatever the experimental protocol adopted, observers are asked to bring their appreciation on the perceptual quality of the images they are shown. The obtained quality ratings are then processed and the average score over all observers is computed for each image of the database. The average score is commonly referred to as the Mean Opinion Score (MOS). In some cases where the reference images are also evaluated, the DMOS (Difference Mean Opinion Score) is derived instead of the MOS. It is the difference of mean opinion scores obtained on the reference and on the test images, respectively.

Although subjective assessment is the obvious and ultimate gauge of image quality, it is time-consuming, and cannot be implemented in systems where a real-time quality score for images is needed. To overcome these drawbacks, the objective approach, which consists in

developing objective image quality assessment models, has been well approved. Such models should be capable to automatically make quality estimations in agreement with the subjective human opinions. To check this condition, the quality measures are usually validated by comparing them to the human appreciation of the image quality, in particular the Mean Opinion Score (MOS) or the Difference Mean Opinion Score (DMOS). This has lead to the creation of a number of subjectively rated image quality databases. On the other hand, many image quality metrics have been developed in the last couple of decades.

Machine evaluation of visual data quality takes a potential advantage in a wide range of application environments where the human visual consumption is exponentially growing. Furthermore, the goal of objective image quality assessment models is to provide *computational* models that can *automatically* estimate the perceptual quality of images that an average human observer will report. In other words, solving the problem of image quality prediction requires matching image quality to human perception appreciation.

As a matter of fact, our knowledge about human perception mechanisms is still very limited. The existing models of the HVS aim at simulating some functionalities of our perception, in particular the lower level ones such as the contrast sensitivity functions, the perceptual decomposition into channels, visual masking and visual attention. Experiments leading to establish HVS models are generally performed under very restrictive conditions.

Given that the visual data quality appreciation is a nonlinear natural cognitive task that evolves by time and personal experience, and since mechanisms leading to the image quality evaluation are still ill-understood, developing machine learning techniques for image quality assessment can be considered as a legitimate choice that might help us determine how do humans make judgement about the quality of visual content and which factors mainly affect this process. Indeed, learning image quality assessment in machines allows us to exploit the subjective quality scores in order to construct a model capable to further give us an objective measure that continues to match with the human opinion to any other image.

There exists variety of image types for a variety of uses: satellite, medical, multispectral and synthesized images to name just only few ones. In this thesis, we will deal with the natural images for which there are many image quality databases intended for research purposes.

1.2 Research Questions

One important problem that slows down developments in the image quality field is the lack of good and complete benchmarking of the proposed metrics. Until now, contributions in the visual data quality field have only been checked at a reduced scale, i.e. use of singular image quality datasets, comparison to a little number of yet existing metrics, use of only one or a reduced number of measures of the algorithm's performance evaluation. Two questions are raised in the first part of this thesis. Firstly, *was there any significant improvement of the capability to objectively predict the perceived visual image quality having its original version at hand*? Secondly, *is there any significant difference between the existing subjective rating protocols for image quality databases elaboration*? We have attempted to answer these questions by providing a complete quantitative predictive performance evaluation of eighteen objective image quality metrics of different approaches over six available subjective image databases.

Another aspect of image quality assessment is that an image supplies a lot of information that is not all used for the evaluation of the quality of that image. Thereby, instead of using the whole image data, the image features are extracted which can consist in a large array of uneven importance for the quality appreciation task. Selecting only a reduced amount of image features makes computations easier and speeder. The crucial matter here is: *What are the image features that intervene into the image quality judgement*? And *which ones most affect this judgement*? The feature selection approach has been used in this research work to attempt to discern the image attributes that are relevant to the image quality evaluation from those which are not.

The image feature selection is a key step for developing new objective reduced-reference image quality measures. A variety of approaches is presented in chapter 2 that shows the numerous possible ways we have to design such measures. The question is *what is the best*

method to integrate the feature selection into the process of definition of a quality metric? The artificial neural networks with supervised learning have been the favourite candidate. They have allowed us to find a solution to the problem of *how to design image quality metrics with optimal predictive performance abilities based on correlation, monotonicity and accuracy measures*? This is done by customizing the neural network cost function in such a way that optimizes the models' predictive performances.

1.3 Contributions of the Thesis

1) A complete and thorough statistical evaluation of objective methods over subjective quality databases using the nonparametric Friedman's statistical analysis has been conducted.
2) Based on results of the quantitative evaluation, a set of image features is derived, and feature selection with neural networks is performed to reduce the size of the initial image attributes and get the vector that conveys more information about the image quality evaluation.
3) Feature selection is then made using the KXEN polynomial regression statistical modelling software.
4) Based on results of contributions 3 and 4, the set of selected features is used for the development of a reduced reference neural network based model image quality metric shown to outperform other full-reference image quality metrics in terms of accuracy.
5) Customization of the neural network's cost function to get more correlated and more monotonous image quality measure than its counterparts.

1.4 Layout of Dissertation

The dissertation is organized in seven chapters:

Chapter 2 entitled **"Objective versus Subjective Quality"** provides a review of the state-of-the-art objective image quality measurement approaches in the first part of the chapter. In the second part, the subjective rating tests as well as the methodological options available to

construct subjective image quality databases are outlined. A comprehensive description is then provided for a range of image quality databases available for the research community including the Toyama database (2000), LIVE database (2005), IVC database (2006), A57 database (2007), TID database (2008), and CSIQ database (2009). The subjective quality databases are intended for the predictive performance benchmarking of the objective quality models carried out in the next chapter.

Chapter 3 deals with the **"Quality Metrics Performance Evaluation and Comparison"**. It supplies a complete quantitative evaluation of the objective full-reference image quality models performance elucidated in chapter 2. Their predictive performance is then computed and compared using the non parametric Friedman test over the public and subjectively rated image quality databases described in chapter 3. To do this, three performance measures are computed: the Pearson's Correlation Coefficient (PCC) as an indication on the *correlation*, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) to quantify the prediction *accuracy*, the Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order Correlation Coefficient (KROCC) to indicate the *monotonicity* measure. All the calculations are first performed on the whole databases and then on the sets concerned with four degradation types namely the JPEG compression, JPEG2000 compression, noise and Gaussian blur.

Chapter 4 is an **"Overview on Machine Learning and Artificial Neural Networks"**. Besides a summary of the machine learning theory and the artificial neural networks principles, the key concepts pertaining to building models based on feed-forward multilayer perceptrons are presented. The basic issues, mainly the generalization capabilities, linked to the use of neural networks based methods in chapters 6 and 7 are then laid out.

Chapter 5 is dedicated to **"Feature Selection for Image Quality Assessment"**. The general process and the different approaches of feature selection are outlined at the beginning of this chapter. The issue related to combining and selecting low-level features for image quality assessment is then discussed and experimented. The aim is to identify the set of image attributes that are relevant to the evaluation of the quality of an image. Experiments

are carried out and validated using two different approaches: the artificial neural networks based approach and the structural risk minimization statistical based approach.

Chapter 6 presents two new **"Reduced Reference Multilayer Perceptron based Metrics"**. The proposed measures are based on the variance of the reference and the test images as well as their covariance. The standard cost function is first employed to develop the ECF (Error based Cost Function) metric which is shown to be more accurate than the eighteen full-reference measures compared in chapter 4. The cost function is then modified leading to the CCF (Correlation based Cost Function) metric that outperforms the aforementioned metrics to which it has been compared in terms of correlation and monotonicity. The comparative study has been conducted over three image quality databases including the LIVE (second release), TID and CSIQ databases.

The summary of the overall work achieved in the present thesis, the open issues and the perspectives are given as a general conclusion in chapter 7.

1.5 List of Publications

- **Journal Papers**

1. Atidel Lahouhou, Emmanuel Viennet, Azeddine Beghdadi, "Selecting Low-level Features for Image Quality Assessment by Statistical Methods", *Journal of Computing and Information Technology*, Vol. 18, No. 2, pp. 183-189, 2010.
2. Atidel Lahoulou, Emmanuel Viennet, Ahmed Bouridane, Mourad Haddadi, "Customizing Cost Function for Optimizing Image Quality Measures Performances", *International Journal of Electronics*, Vol. 99, No. 11, pp. 1533-1546, Nov. 2012.
3. Atidel Lahoulou, Ahmed Bouridane, Emmanuel Viennet, Mourad Haddadi, "Full Reference Image Quality Metrics Performance Evaluation over Image Quality Databases", *Accepted for publication in The Arabian Journal for Science and Engineering,* DOI: 10.1007/s13369-012-0509-6.

❖ **International Conferences with proceedings**

1. Atidel Lahouhou, Emmanuel Viennet, Azeddine Beghdadi, "Combining and Selecting Indicators for Image Quality Assessment", In *31st International Conference on Information Technology Interfaces*, Croatia, pp. 261-268, 22-25 June 2009.
2. Atidel Lahoulou, Emmanuel Viennet, Mourad Haddadi, "Variable Selection for Image Quality Assessment using a Neural Network based Approach", *IEEE European Workshop on Visual Information Processing, (EUVIP 2010),* Paris, France, pp. 45-49, 5-7 July 2010.
3. Atidel Lahoulou, Emmanuel Viennet, Ahmed Bouridane, Mourad Haddadi, "A Complete Statistical Evaluation of State-of-the-art Image Quality Measures", *The 7th International Workshop on Systems, Signal Processing and their Applications* (*Wosspa 2011*), ENP, Algiers, Algeria, pp. 219-222, 9-11 May 2011.

❖ **Technical Reports**

1. Atidel Lahoulou, Emmanuel Viennet, Ahmed Bouridane, Mourad Haddadi, "Technical Report: Full Numerical Results for image quality metrics performance benchmarking", April 2011. Available on: http://www-l2ti.univ-paris13.fr/~lahoulou/tech_report.html

CHAPTER

2

Chapter 2

Objective versus Subjective Quality

2.1 Introduction

Given the phenomenal rate at which image and video content is being generated and distributed, evaluation of the perceptual quality of the content becomes a critical task. Indeed, there is a wealth of research on both subjective and objective image quality measures to reliably predict either perceived quality across different scenes and distortion types or to predict algorithmic performance computer vision tasks.

The goal of objective image quality assessment models is to provide *computational* models that can *automatically* estimate the perceptual quality of images that an average human observer will report. In other words, solving the problem of image quality prediction requires matching image quality to human perception appreciation. The Video Quality Experts Group (http://www.vqeg.org) is a forum that validates and establishes subjective and objective approaches to visual data quality measurement.

In the next section, we give an overview of historical and most contemporary objective image quality assessment methods together with the key concepts involved for each approach. Particular focus will be given to the full reference image quality measures since they are the most widely used within the image processing research community. Section 2.3 is dedicated for the subjective quality rating tests; the methodological options available to construct subjective image quality databases. Section 2.4 provides a comprehensive

description for each of the subjective quality databases intended for predictive performance benchmarking of the objective quality models presented in the next chapter.

It is worth noting that although subjective assessment is time-consuming, and cannot be implemented in systems where a real-time quality score for an image or video sequence is needed, it is the obvious and ultimate gauge of image quality. The subjective approach is needed to establish the performance of the objective visual data quality assessment algorithms that should predict subjective image quality accurately and rapidly.

2.2 Objective Image Quality Measurement Methods

Machine evaluation of image and video quality takes a potential advantage in a wide range of application environments where the human visual consumption is exponentially growing. First, they can be employed for image quality *monitoring* during image acquisition, transmission and reproduction. Second, they can be deployed for *benchmarking* image processing algorithms designated for restoration and enhancement. Third, they can be embedded in compression and communication systems for parameters *optimization* [1]. Also, knowledge about the possible distortion processes is eventually important information that can be supplied about the environment for the development of image quality measures.

2.2.1 Families of image quality metrics

Depending on the amount of information available on the reference image during the quality assessment process of its distorted version(s), we distinguish three broad families of models in the literature as schematized in figure 2.1:

- Full reference models: where the reference image is available when evaluating its test version(s). The task reduces to a comparison of two images and image quality evaluation can be regarded as an *image fidelity* problem. The calculation should be fast and should correlate with human subjective appreciation.
- Reduced reference models: In practical situations, it is not often possible to have the reference image while evaluating its test version(s). In fact, some applications require massive image or video transmission via telecommunication networks. Thus, the

reference image is not entirely provided but only a feature vector giving relevant information to control the quality of the transmitted visual data. Methods based on these features are fast, but their relatively poor performances restrict their use to some specific applications.

- No reference models: also called “blind models”. They attempt to evaluate the quality of an image without accessing its reference. They are complicated to elaborate but are the ideal form and the most interesting ones for many applications.

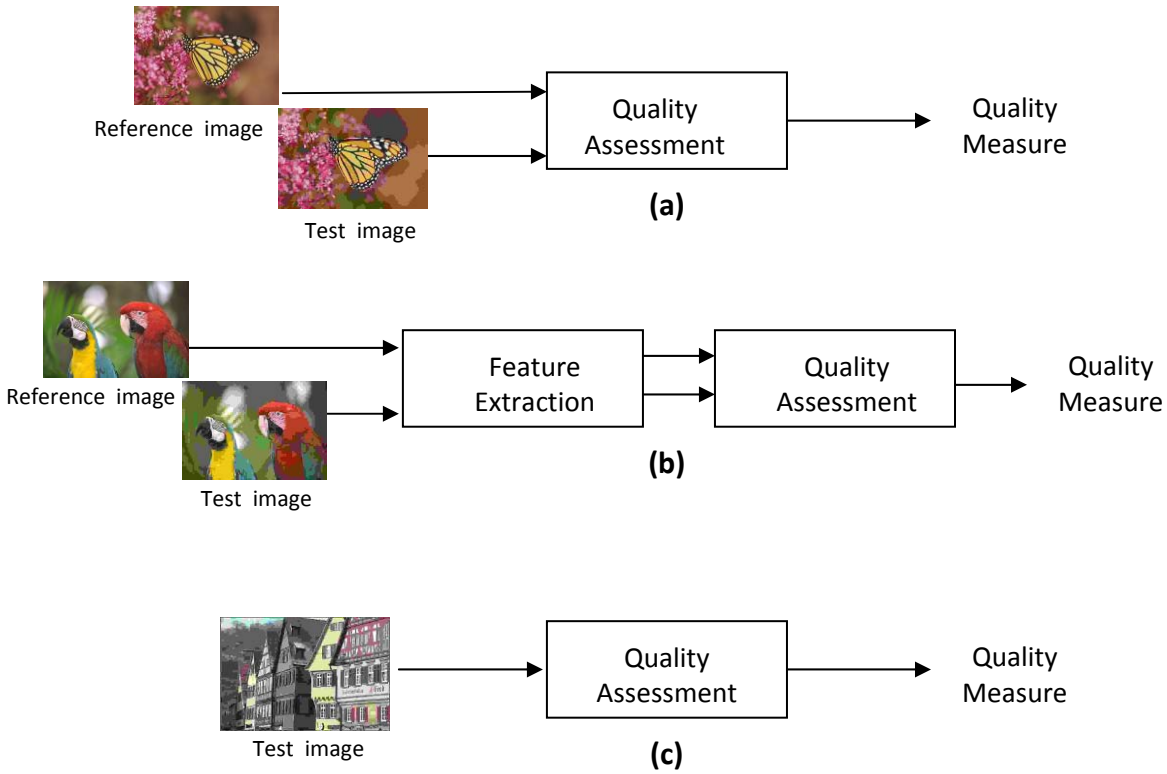

Figure 2.1: overview of full-reference (a), reduced-reference (b) and no-reference (c) image quality models [2].

The majority of the proposed state-of-the-art image quality methods fall in the first category. We can classify the different existing full reference image quality assessment metrics into two major classes: raw error-based measures and Human Visual System (HVS) inspired metrics.

2.2.2 Traditional raw-error based image quality measures

Initial investigations on objective image quality assessment focused, for decades, on raw mathematical metrics based on error quantification between two images. SNR, PSNR and

MSE were precursors, while the PSNR and MSE were and still are the most widely used methods to quantify the quality of an image with regard to its reference version. Their popularity is due to their simplicity and their very low computational cost.

SNR and Peak SNR were derived by considering hypothetically that an image distortion is only produced by additive noise which is independent from the signal. Thus, SNR is defined as the ratio of average signal power to noise signal power (eq. 2-1) while PSNR is defined as the ratio of peak signal power to noise signal power (eq. 2-2).
Let assume that $X = \{x_i \mid i=1, 2, \ldots, N\}$ and $Y = \{y_i \mid i=1, 2, \ldots, N\}$ are the reference and test images of size N respectively.

$$SNR = 10\log_{10}\left(\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N}(x_i - y_i)^2}\right)[dB] \qquad (2\text{-}1)$$

$$PSNR = 10\log_{10}\left(\frac{N \cdot Max(X)^2}{\sum_{i=1}^{N}(x_i - y_i)^2}\right)[dB] \qquad (2\text{-}2)$$

Knowing that $Max(X) = 2^l - 1$ represents the maximum value of pixel intensities for image X (and image Y as well) where l denotes the number of necessary bits to encode the image pixels.

For example: Max(X) = 255 = 2^8-1 for 8-bit image.

In the literature of image processing, PSNR is also expressed in terms of the mean squared error as follows:

$$PSNR = 10\log_{10}\left(\frac{Max(X)^2}{MSE}\right) = 20\log_{10}\left(\frac{Max(X)}{\sqrt{MSE}}\right)[dB] \qquad (2\text{-}3)$$

The higher the SNR (and PSNR) value, the better the similarity between test and reference images. However, a higher value of the MSE denotes a larger amount of error between the test and the reference images.

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are also raw error metrics but not usually employed amongst the image processing research community for image quality assessment but rather for objective quality estimator's performance assessment.

Despite of their simplicity and calculation convenience, the objective quality metrics described above (SNR, PSNR, MSE) are criticised for not correlating with the subjective quality appreciation. Their limitations have been exposed in [3] and reported later in almost all the papers on image and video quality. Furthermore, disadvantages of the raw error methods have become the ultimate argument to the considerable effort that has gone into developing new objective quality metrics by considering human visual system (HVS) characteristics.

2.2.3 HVS inspired image quality metrics

A new era for the image quality research community has been started by a Weighted version of the SNR that has been derived by T. Mitsa *et al.* [4] using the Contrast Sensitivity Function (CSF). WSNR (in dB) is therefore defined as the ratio of the averaged weighted signal power to the average weighted noise power. The development of Picture Quality Scale (PQS) for achromatic image coding [5] and Noise Quality Metric (NQM) for image restoration purposes [6] was also a big step in the design and the development of new image quality algorithms.

One of the findings on the HVS is that it is highly adapted to extract structural information from images. By following the assumption that an image quality degradation is not due to independent noise as assumed previously, but rather to the loss of structured information in images, then conceiving a quality metric that measures structural distortions should have good correlation with the perceived image dissimilarity [7].

In [8], Z. Wang *et al.* suggested that the HVS can be decomposed into three independent channels: Luminance, Contrast and Structure. The universal quality index (UQI) has then been constructed upon comparisons of the three image components pairs leading to the following three equations:

$$l(x,y)=\frac{2\bar{x}\cdot\bar{y}}{\bar{x}^2+\bar{y}^2} \quad \text{for luminance comparison} \qquad (2\text{-}4)$$

$$c(x,y)=\frac{2\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2} \quad \text{for contrast comparison} \qquad (2\text{-}5)$$

$$s(x,y)=\frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y} \quad \text{for structure comparison} \qquad (2\text{-}6)$$

Hence, the overall value of the UQI measure is obtained using the product of the three comparison equations as:

$$UQI=s(x,y)\cdot l(x,y)\cdot c(x,y) \qquad (2\text{-}7)$$

$$UQI=\frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y}\cdot\frac{2\bar{x}\cdot\bar{y}}{\bar{x}^2+\bar{y}^2}\cdot\frac{2\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad (2\text{-}8)$$

$$UQI_{\in[-1,1]}=\frac{4\sigma_{xy}\cdot\bar{x}\cdot\bar{y}}{\left(\sigma_x^2+\sigma_y^2\right)\cdot\left(\bar{x}^2+\bar{y}^2\right)} \qquad (2\text{-}9)$$

where $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$ and $\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$ are the mean pixel intensities values of images *X* and *Y*, respectively.

$\sigma_x=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2$ $\sigma_y=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2$ are the standard deviations of the pixel intensities values of images *X* and *Y*, respectively.

$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)$ is the covariance between pixel intensities values of images *X* and *Y*.

An instability problem of the UQI measure arises when the denominator of Equation (2-9) tends towards zero. The solution to this problem has been suggested in [7] by making some changes to the luminance, contrast and structure comparison definitions of Equations (2-4), (2-5) and (2-6). The changes consist of introducing two scalars in order to tune the system and ensure its stability. Equations (2-4), (2-5) and (2-6) are then rewritten as follows:

$$l(x,y)=\frac{2\bar{x}\cdot\bar{y}+C_1}{\bar{x}^2+\bar{y}^2+C_1} \quad \text{for luminance comparison} \qquad (2\text{-}10)$$

$$c(x,y)=\frac{2\sigma_x\cdot\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \quad \text{for contrast comparison} \qquad (2\text{-}11)$$

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\cdot\sigma_y+C_3} \quad \text{for structure comparison} \qquad (2\text{-}12)$$

where C_1 and C_2 are small positive constants tuned by the SSIM's authors.

The structural similarity index (SSIM) is then expressed by the following equation:

$$SSIM=\left[l(x,y)\right]^{\alpha}\cdot\left[c(x,y)\right]^{\beta}\cdot\left[s(x,y)\right]^{\gamma} \qquad (2\text{-}13)$$

where α,β,γ are positive non zero parameters used to define the importance of each of the three components.

If $\alpha=\beta=\gamma=1$ and $C_3=C_2/2$, SSIM is then given by:

$$SSIM_{\in[-1,1]}=\frac{\left(2\bar{x}\cdot\bar{y}+C_1\right)\cdot\left(2\sigma_{xy}+C_2\right)}{\left(\bar{x}^2+\bar{y}^2+C_1\right)\cdot\left(\sigma_x^2+\sigma_y^2+C_2\right)} \qquad (2\text{-}14)$$

The best SSIM score (SSIM=1) is achieved when the test image structure is the same as its original version and consequently is of the best quality.

Three variants of the SSIM have been suggested later including the multi-scale SSIM (MS-SSIM) [9], the SSIM with automatic down-sampling (MSSIM) [10], and the Complex Wavelets-SSIM based on the principle that structural information is more contained in the phase than in the magnitude of the signal [11].

Since 2004, a new approach to image quality assessment problems has emerged and new quality metrics have been developed using an information and communication theoretic framework. The Information Fidelity Criterion (IFC) [12] and its extensions VIF (Visual Information Fidelity index) and pixel-based VIF [13] metrics proposed by H. R. Sheikh *et al.* belong to a different class of image quality assessment methods which are built upon Natural Scene Statistics (NSS) models. The employed premise in this case is that visual fidelity can be accurately quantified if it is known how much Shannon information the test image brings about from its reference version. Following Shannon's communication scheme, the transmitter, the channel and the receiver correspond to the reference source image, the distortion model applied to it and the test generated distorted image, respectively. Furthermore, the distortion and HVS models have also been incorporated into the visual information fidelity index design. VIF has been well appreciated and commonly employed in the image quality community regarding the good correlation it presents with subjective quality judgment.

Some researchers chose to rely on advanced signal processing transforms rather than HVS-behavioural models when developing better quality predictors. Their argument is that the human vision is too complex to be understood and hence to be modelled or simulated. To overcome this drawback some well known signal processing techniques having similar features as the human perception can be used. For example, A. Shnayderman *et al.* [14] suggested a Multidimensional full reference quality metric based on the Singular Value Decomposition (M-SVD). Based on the same theorem, the R-SVD quality predictor has been developed later by A. Mansouri *et al.* [15] where the right singular vector matrix of the original image is used. Similarly, VSNR is a wavelet-based Visual Signal to Noise Ratio proposed by D. M. Chandler *et al.* [16] and most recently in 2010, the Riesz-transform based feature similarity metric (RFSIM) [17].

One can also notice a revival of the raw mathematical metrics including SNR and PSNR combined to some basic human visual features in the state-of-the-art of image quality evaluation. This has resulted in the VSNR metric evoked above as well as the HVS based PSNR (PSNRHVS) [18] and the modified PSNRHVS (PSNRHVSM) [19] developed by N. Ponomarenko *et al.* In the first case, only the Contrast Sensitivity Function (CSF) has been

taken into account as a visual feature. It was combined with the PSNR described at the beginning of this section. An improvement to the PSNRHVS metric has been brought thereafter by introducing the model of visual correlation between-coefficient contrast masking of DCT (Discrete Cosine Transform) functions based on the HVS.

2.3 Subjective Image Quality Rating Tests

Visual data quality assessment measurement can be either qualitative or quantitative. This section is about subjective quality evaluation where the aim is to provide accurate, consistent and reliable predictions of the perceptual image/video appreciation.

Subjectively evaluating the quality of content is an extremely difficult task due to the time and cost involved. Indeed, for the subjective test to be reliable a large number of human test observers should be invited to participate to the images and/or videos quality evaluation, under controlled psychometric experimental conditions. In addition, images are divided into two groups: *reference* images considered as pristine versions and processed ones that were subject to some degradation and were used as *test* images.

Depending on the experimental protocol which is adopted, observers are shown the images (either the pair reference/test, only the test version, both the reference and all its test versions, or all the test versions of the same reference) and instructed to rate their perceived quality, to score the degradation intensity or to order the images according to their similarity to a given reference.

It is known that perceptual quality may vary from one individual to another. This depends on observers' general experience (if he/she is expert in image processing or not), on their personal appreciation and may vary according to their mood. To alleviate this problem, the average score of the given individual ratings is computed over all observers and called the Mean Opinion Score (MOS). In some cases where the reference images are also evaluated, the DMOS (Difference Mean Opinion Score) is derived instead of the MOS. It is the difference of mean opinion scores obtained on the reference and on the test images,

respectively. Unfortunately, subjective image quality assessment methods present the two major following disadvantages:

- They are very expensive and obviously cannot be integrated in real time systems.
- The knowledge obtained in the form of quality scores (MOS or DMOS) cannot be generalized and thus the evaluation process cannot be modelled.

Despite their drawbacks, subjective image quality assessment measurements are essential to establish the performance of the automatic objective models introduced in section 2.2. This section deals with the methodology recommended by the International Telecommunications Union – Radio-communication Sector (http://www.itu.int/ITU-R/) to successfully perform such elaborate tests. Section 2.4 gives thorough description of the different image quality databases available for the research community including the Toyama database (2000), LIVE database (2005), IVC database (2006), A57 database (2007), TID database (2008), and CSIQ database (2009).

2.3.1 Methodology for subjective quality rating tests

In this subsection, we briefly cover the most important ITU-R BT.500-11 Recommendations [20] where the methodology for subjective quality rating tests of visual content is described.

a) Test configuration

The following factors related to the test configuration may alter and/or influence the human observers' judgements:

- General viewing conditions: the establishment of standardized viewing environments reduces the influence of outside world on the observation and evaluation of visual data quality. Factors commonly measured are the luminance of screen, the display brightness and contrast, chromaticity of image background, size of the screen and room illumination. These parameters are generally selected to define an environment slightly more critical than the typical home viewing situations.

- Observation distance: the distance between the observer and the display device should be set to four or six times the height of the screen.
- Display system: the display devices used in the tests must be tuned in such a way to reduce the observers' eyestrain. For this, the ITU-R BT.814 [21] and ITU-R BT.815 [22] recommendations specify how to measure the luminance and contrast of the screen. They are also reported in [20].

b) Observers

- Choice of human observers: at least 15 observers should participate to the experiments. They should be non-expert and should be checked for normal visual acuity and normal colour vision. Further research needs to be undertaken to assess some factors that may influence the perceptual quality appreciation; such as observers' occupation, intellectual level, gender and age. This is why it is desirable that these data be provided by the experimenters to facilitate further investigation of such factors.
- Instructions for the assessment: observers should be introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scales, the sequence and timing. Training sequences demonstrating the range and the type of the impairments to be assessed should be used with illustrating pictures other than those used in the test, but of comparable sensitivity.
- Grading scales: observers should give the scale very clearly. They should have numbered boxes or some other means to record their appreciations about quality according to the ITU-R scales given in table 2.1.
- Test session: a session should last up to half an hour. At the beginning of the first session, about five training presentations should be introduced to stabilize the observers' opinion. The data issued from these presentations must not be taken into account in the final results. If several sessions are necessary, about three training presentations are only necessary at the beginning of the following session. The diagram 2.2 shows how a typical test session is organized.

Class	Quality	Impairment
5	Excellent	Imperceptible
4	Good	perceptible but not annoying
3	Fair	slightly annoying
2	Poor	annoying
1	Bad	Very annoying

Table 2.1: ITU-R Quality and impairment "five-grade" scales.

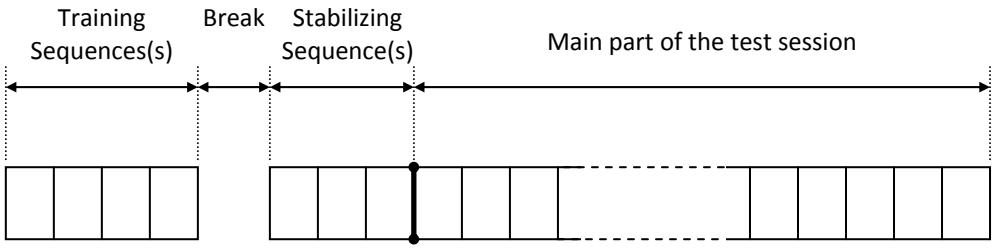

Figure 2.2: Structure of a typical test session [20].

c) Experimental protocols

International recommendations for subjective video quality assessment [20] include specifications for how to perform many different types of subjective tests. The most commonly used experimental methods can be single stimulus or double stimulus as illustrated in figure 2.3. Some examples are briefly provided in this sub-section and more details can be found in [20] and [23].

- Single stimulus: observers rate the quality of just the distorted image (or just the distorted video stream). An example of the single stimulus method is *single stimulus continuous quality evaluation* (SSCQE). Observers are shown images sequentially one by one with a latency time between two presentations allowing them to record their ratings.
- Double stimulus: observers rate the quality or change in quality between the reference and the test images (or video streams). The *double stimulus continuous*

quality scale (DSCQS) and *double stimulus comparison scale* (DSCS) are well-known methods in the category of double stimulus protocols. The two stimuli presentation time should be the same and they are alternate by a gray screen.

Other methods are employed for elaboration of most recent subjective image quality databases such as the Continuous rating system, pairwise sorting and linear displacement of the images. They are defined in section 2.4.

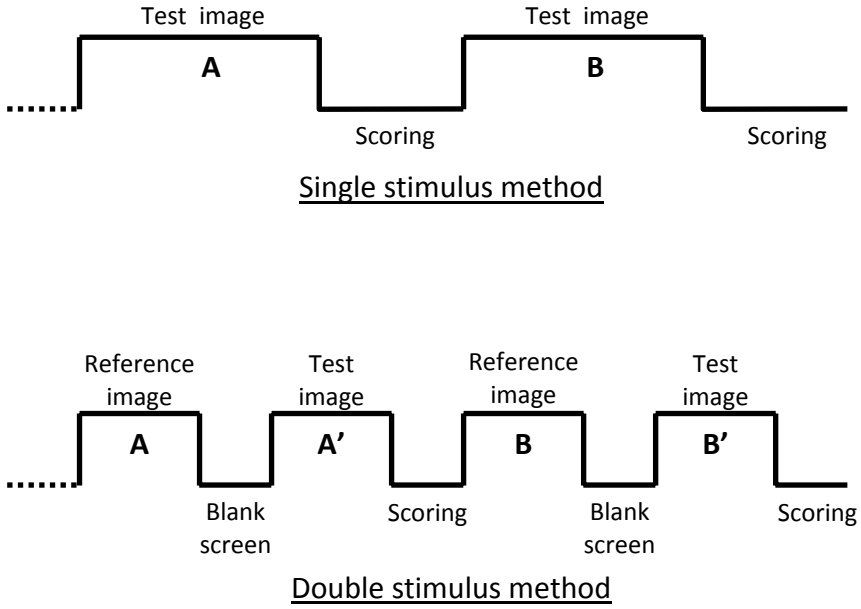

Figure 2.3: Presentation of the single stimulus and the double stimulus experimental methods.

d) Processing of the subjective ratings

Data should be collected from all test sessions. A realignment experiment is conducted upon completion of the series of psycho-visual testing. The aim of the realignment is to eliminate aberrant values reported by the observers in the 95% confidence interval. For each image i the mean opinion score $MOS(i)$ is further computed using the following formula:

$$MOS(i) = \frac{1}{N}\sum_{j=1}^{N} score_i\,(j) \qquad (2\text{-}15)$$

where N is the number of scores collected for image i and $score_i(j)$ is the score given by observer j to image i.

For each image i, all the mean opinion scores should have an associated confidence interval which is derived from the mean μ_i and the standard deviation σ_i. It is suggested to use the 95% confidence interval given by equations (2-16) to (2-18).

$$[MOS(i)-\mu_i, MOS(i)+\mu_i] \quad (2\text{-}16)$$

$$\mu_i = 1{,}96 \cdot \frac{\sigma_i}{\sqrt{N_i}} \quad (2\text{-}17)$$

$$\sigma_i = \sqrt{\sum_{i=1}^{N} \frac{(MOS(i)-score_i)^2}{(N-1)}} \quad (2\text{-}18)$$

2.4 Overview of the Subjective Image Quality Databases

All the image quality databases publicly available to the research community were built upon some extensive psycho-visual experiments performed under specific but different test conditions. In this section, we investigate six different image quality databases which share the following points [24]:

- ✓ For each image database, a set of *reference* images is considered. These images are assumed to be pristine originals so of perfect quality on which some image distortion algorithms have been applied to construct a set of *test* images.
- ✓ The reference images were altered with a single type of distortion.
- ✓ For each induced stimuli process, levels have been selected in order to achieve a large range of visual quality: from excellent quality where artefacts are not visible, to bad quality where distortions are annoying.
- ✓ All the tested images represent natural scenes except one synthesized image added to the TID image database.

2.4.1 Toyama image database (2000)

Toyama image database was published in 2000 by the Multimedia Information and Communication Technology (MICT) Laboratory at University of Toyama, Japan [25].

The database was built upon 14 high resolution color reference images. These images were distorted with the JPEG and JPEG2000 coders at different bitrates: 15, 20, 27, 37, 55 and 79 for JPEG and 12, 24, 32, 48, 72 and 96 for JPEG2000. This resulted in 196 test images for which the quality was evaluated.

According to the adjectival categorical judgment method, during the psychometric experiments, each of the 16 subjects was shown the images randomly one at a time (single stimulus) and was asked to assign each image with an adjective that indicates his / her perception of the quality of these images. The quality adjectives correspond to discrete numerical values from 1 to 5 as shown in table 2.1. Adjectives were then converted to the corresponding numerical values and the mean opinion score (MOS) was calculated as the average of the 16 scores for each image with subject reliability of 95% confidence interval. Higher value of MOS corresponds to higher visual quality of the image.

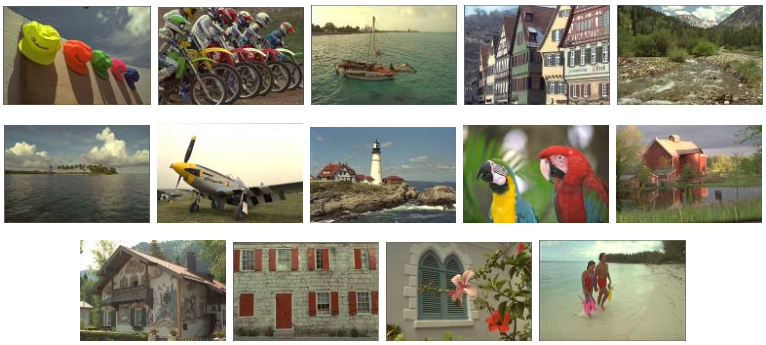

Figure 2.4: Reference image samples of the Toyama database.

<table>
<tr><td rowspan="12">TOYAMA2000</td><td colspan="4">Images</td></tr>
<tr><td colspan="2">Image formats</td><td colspan="2">24-bits/pixel RGB color (768x512)</td></tr>
<tr><td colspan="2">N. reference images</td><td colspan="2">14</td></tr>
<tr><td colspan="2">Distortion types</td><td>JPEG</td><td>JPEG2000</td></tr>
<tr><td colspan="2">N. distorted images</td><td>98</td><td>98</td></tr>
<tr><td colspan="2"></td><td colspan="2">Total : 196</td></tr>
<tr><td colspan="4">Test Methodology</td></tr>
<tr><td rowspan="4">Test Configuration</td><td>Display Devices</td><td colspan="2">CRT 17-inch (1024x768)</td></tr>
<tr><td rowspan="3">Test Conditions</td><td colspan="2">Standard :
ITU-R BT. 500-11 [20]</td></tr>
<tr><td>Viewing distance</td><td>4H (H: Picture height)</td></tr>
<tr><td>Room illumination</td><td>Low</td></tr>
<tr><td colspan="2">Observers</td><td colspan="2">16 (Non expert, college students)</td></tr>
<tr><td colspan="2">Method</td><td colspan="2">Single Stimulus (Adjectival categorical judgment)</td></tr>
<tr><td rowspan="3">Subjective ratings of images</td><td>Raw data</td><td colspan="2">Adjective scales corresponding to 5 quality levels:
Bad = 1, Poor = 2, Fair = 3, Good = 4, Excellent = 5</td></tr>
<tr><td>Final scores</td><td colspan="2">MOS</td></tr>
<tr><td>Scores' Scale</td><td colspan="2">1 .. 5</td></tr>
</table>

Table 2.2: Summary of Toyama image database description.

2.4.2 LIVE image database (2005)

LIVE image database was developed at the Laboratory for Image and Video Engineering in collaboration with the Center for Perceptual Systems at the University of Texas at Austin, USA. The first release was made available online in 2003 while release 2 on which we conducted the present study was published in 2005 [26].

In both the two releases, the database was created from 29 high resolution color reference images. In release 2, the perceptual quality of a total of 982 test images was subjectively estimated. The images were generated using five distortion types: JPEG, JPEG2000, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model.

Subjective testing was performed in seven sessions where observers were instructed to rate images and to provide their perception of quality. The method is similar to the one used for Toyama database. As the database contains a number of reference images that have been assessed as well, raw scores (from 1 to 5) were converted for each subject to difference mean opinion scores (DMOS). This latter represents the difference between the scores obtained for the reference image and its test version. A low DMOS means little degradation whereas an important value corresponds to severe distortions in the image.

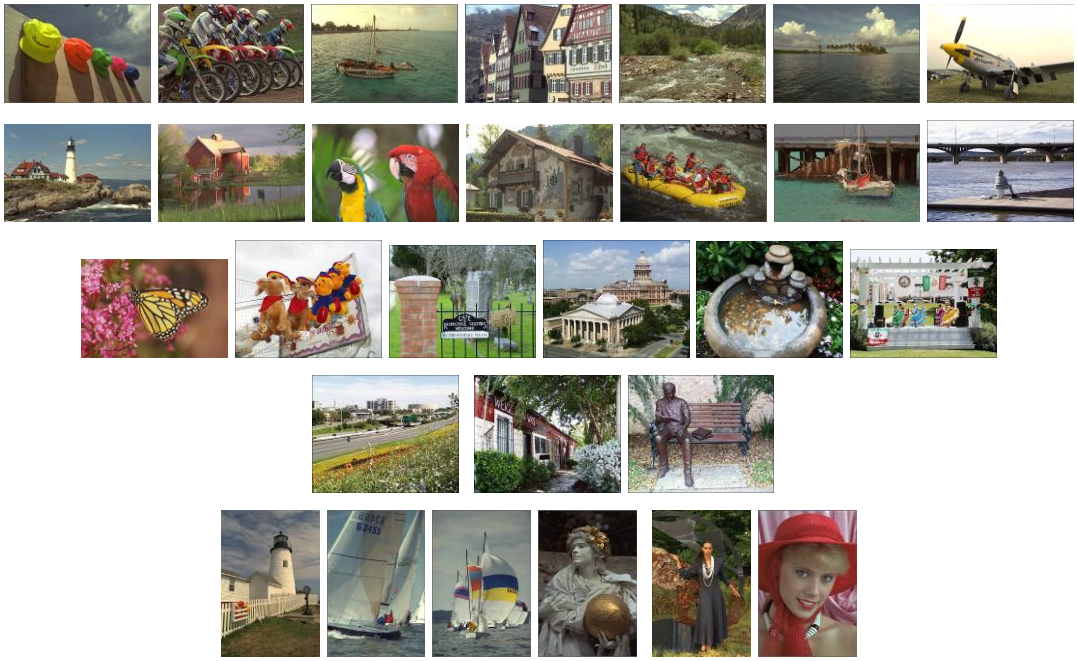

Figure 2.5: Reference image samples of the LIVE database.

<table>
<tr><td rowspan="11">L
I
V
E

2
0
0
5</td><td colspan="6">Images</td></tr>
<tr><td colspan="2">Image formats</td><td colspan="4">24-bits/pixel RGB color (typically 768x512)</td></tr>
<tr><td colspan="2">N. reference images</td><td colspan="4">29</td></tr>
<tr><td colspan="2">Distortions</td><td>JPEG2000</td><td>JPEG</td><td>White noise</td><td>Gaussian blur</td><td>Fast-fading</td></tr>
<tr><td colspan="2" rowspan="2">N. distorted images</td><td>227</td><td>233</td><td>174</td><td>174</td><td>174</td></tr>
<tr><td colspan="5">Total : 982</td></tr>
<tr><td colspan="6">Test Methodology</td></tr>
<tr><td rowspan="4">Test Configuration</td><td rowspan="2">Display Devices</td><td colspan="4" rowspan="2">CRT 21-inch (1024x768)</td></tr>
<tr><td rowspan="3">Test Conditions</td></tr>
<tr><td colspan="4">Standard :
ITU-R BT. 500-11 [20]</td></tr>
<tr><td>Viewing distance</td><td colspan="3">2-2.5 H (H: Screen height)</td></tr>
<tr><td colspan="2">Observers</td><td colspan="4">20-29 human observers</td></tr>
<tr><td colspan="2">Method</td><td colspan="4">Single Stimulus (Adjectival categorical judgment)</td></tr>
<tr><td rowspan="3">Subjective ratings of images</td><td>Raw data</td><td colspan="4">Adjective scales corresponding to 5 quality levels:
Bad = 1, Poor = 2, Fair = 3, Good = 4, Excellent = 5</td></tr>
<tr><td>Final scores</td><td colspan="4">DMOS</td></tr>
<tr><td>Scores’ Scale</td><td colspan="4">0 : for undistorted images
1 .. 100 : for distorted images</td></tr>
</table>

Table 2.3: Summary of LIVE image database description.

Subjective quality scores were then stretched to the [1..100] range with a subject reliability of 95% confidence interval. An update of the DMOS values has been made available online later using a realignment method. Details on the experiments and raw data processing can be found in [27].

It is worth noting that we use the second release of the LIVE database and the realigned DMOS values in the present study.

2.4.3 IVC image database (2006)

The IVC database has been released by the Image, Video and Communication Laboratory at the University of Nantes, France [28]. This database was derived from 10 square high resolution color reference images that were subjected to JPEG, JPEG2000, blurring and Local Adaptive Resolution based coding (LAR). Thus, 160 test images have been generated. The database also includes 25 monochromatic images for which we do not have any information about their generation process.

Subjective evaluations of images were carried out by 15 observers according to the double stimulus impairment scale method. Unlike the single stimulus method employed for Toyama and LIVE image databases which display the test images randomly, the double stimulus strategy is based on providing both the reference and the test images sequentially. Each observer is then asked to assess the artefact annoyance he/she felt on the distorted image with respect to the reference one. The impairment scales correspond to five classes marked with adjectives and numbers as shown in table 3.1. Ratings are reported in the form of the MOS.

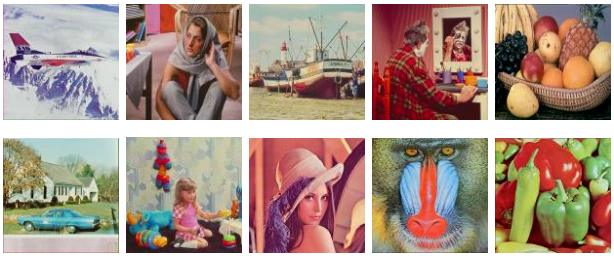

Figure 2.6: Reference image samples of the IVC database.

<table>
<tr><td rowspan="12">I
V
C

2
0
0
6</td><td colspan="5">Images</td></tr>
<tr><td colspan="2">Image formats</td><td colspan="4">24-bits/pixel RGB color (512x512)</td></tr>
<tr><td colspan="2">N. reference images</td><td colspan="4">10</td></tr>
<tr><td colspan="2">Distortions</td><td>JPEG2000</td><td>JPEG</td><td>Blur</td><td>LAR</td></tr>
<tr><td colspan="2">N. distorted images</td><td>50</td><td>50</td><td>20</td><td>40</td></tr>
<tr><td colspan="2"></td><td colspan="4">Total : 160</td></tr>
<tr><td colspan="6">Test Methodology</td></tr>
<tr><td rowspan="4">Test Configuration</td><td>Display devices</td><td colspan="4">One CRT standard definition TV monitor.</td></tr>
<tr><td rowspan="3">Test Conditions</td><td colspan="4">Standard :
ITU-R BT.500-11 [20]</td></tr>
<tr><td colspan="2">Viewing distance</td><td colspan="2">6H (H: Screen height)</td></tr>
<tr><td colspan="2">Room illumination</td><td colspan="2">Background luminance of 10.5 cd/m²</td></tr>
<tr><td colspan="2">Observers</td><td colspan="4">15 observers</td></tr>
<tr><td colspan="2">Method</td><td colspan="4">Double Stimulus Impairment Scale (DSIS)</td></tr>
<tr><td rowspan="3">Subjective ratings of images</td><td>Raw data</td><td colspan="4">The impairment scale from 1 to 5:
1 = very annoying, 2 = annoying, 3 = slightly annoying,
4 = perceptible but not annoying, 5 = not perceptible.</td></tr>
<tr><td>Final scores</td><td colspan="4">MOS</td></tr>
<tr><td>Scores' Scale</td><td colspan="4">1 .. 5</td></tr>
</table>

Table 2.4: Summary of IVC image database description.

2.4.4 A57 image database (2007)

The A57 database freely available on [29] was built upon a psychophysical scaling experiment accomplished on a set of 54 test images to measure the perceived distortions. The 3 reference grayscale images from which the database was derived were processed using six types of distortions: JPEG compression, JPEG2000 compression, Additive White Noise, Gaussian blurring, JPEG-2000 compression with Dynamic Contrast-Based Quantization (DCQ), quantization of the LH sub-bands of a 5-level 9/7 filters-DWT of the image (Contrast) [16].

Seven adult imaging expert observers participated to the image quality database assessments using a continuous rating system. This method was used to Measure the fidelity between two impaired images. This is done by presenting both the reference image and the set of its test versions to the observers who are asked to position the test images such that the ones which will be placed furthest from the reference were judged to be of lower visual fidelity.

The advantage of this method is that it gives the observers the opportunity to simultaneously compare multiple test versions of an image. This allowed them to see if an image is of better or lower quality relative to both the other distorted versions and the reference one and to make adjustments to previous judgements if this is necessary.

Figure 2.7: Reference image samples of the A57 database.

The MOSs over the observers and test images were derived from the obtained quality scores using the z-scores method. The values of the MOS span the range [0, 1] so that a score near to zero corresponded to the image containing an imperceptible artefact while a MOS value near to one denotes that the corresponding image was affected so that the distortion is very annoying.

<table>
<tr><td rowspan="11">A57
2007</td><td colspan="8">Images</td></tr>
<tr><td colspan="2">Image formats</td><td colspan="6">8-bits/pixel grayscale images (512×512)</td></tr>
<tr><td colspan="2">N. reference images</td><td colspan="6">3</td></tr>
<tr><td colspan="2">Distortions</td><td>JPEG</td><td>JPEG2000</td><td>Noise</td><td>Blur</td><td>DCQ</td><td>Contrast</td></tr>
<tr><td colspan="2" rowspan="2">N. distorted images</td><td>9</td><td>9</td><td>9</td><td>9</td><td>9</td><td>9</td></tr>
<tr><td colspan="6">Total : 54</td></tr>
<tr><td colspan="8">Test Methodology</td></tr>
<tr><td rowspan="4">Test Configuration</td><td>Display devices</td><td colspan="6">No display devices: the presented images were high-quality, physical printed versions of digital images of size 11 × 11 cm. they were placed on a large, solid gray table. [30]</td></tr>
<tr><td rowspan="3">Test Conditions</td><td colspan="6">Standard :
ITU-R BT.500-11 [20]</td></tr>
<tr><td colspan="3">Viewing distance</td><td colspan="3">45 cm</td></tr>
<tr><td colspan="3">Room illumination</td><td colspan="3">D65 lighting [1]</td></tr>
<tr><td colspan="2">Observers</td><td colspan="6">07 adult imaging-experts</td></tr>
<tr><td colspan="2">Method</td><td colspan="6">Double stimulus continuous rating system</td></tr>
<tr><td rowspan="3">Subjective ratings of images</td><td>Raw data</td><td colspan="6">Nature of raw data is not specified by the database authors.</td></tr>
<tr><td>Final scores</td><td colspan="6">MOS</td></tr>
<tr><td>Scores' Scale</td><td colspan="6">0 .. 1</td></tr>
</table>

Table 2.5: Summary of A57 image database description.

[1] International-standard Artificial Daylight defined by the International Commission on Illumination (CIE).

The A57 database is considered to be of limited statistical reliability according to its authors. This is due to the use of hard copies of images instead of digital ones and to the limited number of both reference images (3) and human observers (7) [16].

2.4.5 TID image database (2008)

TID2008 (Tampere Image Database) version 1.0 was published in 2008 [31]. By making TID database available online, the authors aimed to provide a tool for evaluation of full-reference image visual quality assessment metrics.

TID database includes 25 high resolution color images, 24 out of them are natural images while the last one is an artificial image synthesized by the database authors. The images were processed by 17 different distortion types at different levels including JPEG, JPEG2000 compression, Additive Gaussian noise and Gaussian blur. This has resulted in 1700 test versions of the reference images.

The MOS was obtained from the subjective scores collected from 838 observers from three countries including 251 in Finland, 150 in Italy and 437 in Ukraine. Part of the experiments has been carried out via Internet.

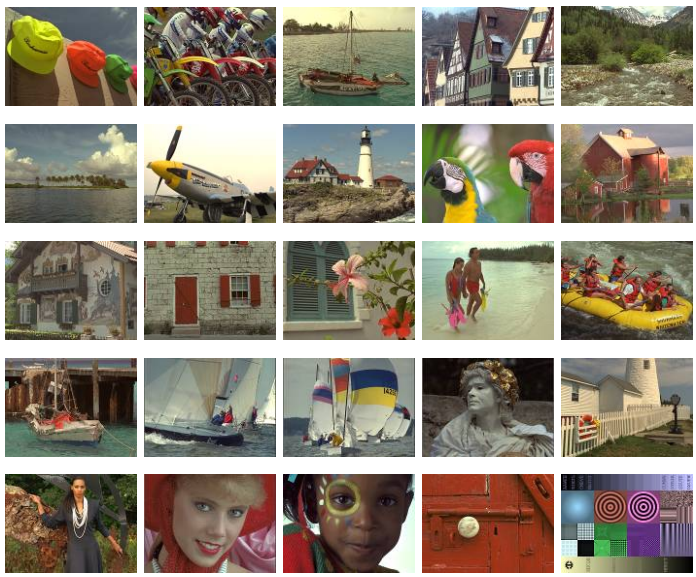

Figure 2.8: Reference image samples of the TID database.

T I D 2 0 0 8	**Images**						
	Image formats		24-bits/pixel RGB color (512x384)				
	N. reference images		25				
	Distortions		JPEG	JPEG2000	Additive Gaussian noise	Gaussian blur	17 distortion types
	N. distorted images		100	100	100	100	100
			Total : 1700				
	Test Methodology						
	Test Configuration	Display devices	LCD and CRT monitors, mainly 19-inch (1152x864 pixels)				
		Test Conditions	Standard : /				
			Viewing distance	/			
			Room illumination	/			
	Observers		838 observers 25% internet based experiments 75% conventional experiments (in room)				
	Method		Pairwise sorting (choosing the best image that visually differs less from the original between two considered).				
	Subjective ratings of images	Raw data	Nature of raw data is not specified by the database authors.				
		Final scores	MOS				
		Scores' Scale	0..9				

Table 2.6: Summary of TID image database description.

2.4.6 CSIQ image database (2009)

The Categorical Subjective Image Quality (CSIQ) database was developed at the Image Coding Analysis Laboratory at Oklahoma State University, USA [32]. It is the most recent and the latest image quality database up to the writing of this paper. It consists of 30 color high resolution square reference images that were distorted using six different image processing algorithms including JPEG and JPEG2000 compression, Gaussian blurring, Additive Gaussian white noise, Global contrast decrements, Additive Gaussian pink noise. This has resulted in a total of 900 distorted images out of which the subjective ratings of only 866 test images are provided. In this paper, only the labelled images were considered.

Thirty-five human observers rated each image. Their visual dissimilarity measurements were performed based on a linear displacement strategy. This consists of presenting simultaneously all the test versions of an image across a monitor array. The images are digital in order to overcome the inconvenience due to image printing used in the case of the A57 database. Observers are then asked to place these images so that the horizontal distance between two test images reflects the perceived dissimilarity between them. Subjective quality scores have been published in terms of difference scores in quality ranging from 0 to 9.

Figure 2.9: Reference image samples of the CSIQ database.

<table>
<tr><td rowspan="11">CSIQ2009</td><td colspan="7">Images</td></tr>
<tr><td colspan="2">Image formats</td><td colspan="5">24-bits/pixel RGB color (512x512)</td></tr>
<tr><td colspan="2">N. reference images</td><td colspan="5">30</td></tr>
<tr><td colspan="2" rowspan="2">Distortions</td><td>JPEG</td><td>JPEG2000</td><td>G. blurring</td><td>G. noise</td><td>Pink noise</td><td>Global contrast</td></tr>
<tr><td colspan="2" rowspan="2">N. distorted images</td><td>150</td><td>150</td><td>150</td><td>150</td><td>150</td><td>116</td></tr>
<tr><td colspan="6">Total : 866</td></tr>
<tr><td colspan="7">Test Methodology</td></tr>
<tr><td rowspan="4">Test Configuration</td><td rowspan="3">Display devices</td><td colspan="5" rowspan="3">4 calibrated LCD monitors (1920 x 1200) placed side by side with equal viewing distance to the observer</td></tr>
<tr></tr>
<tr></tr>
<tr><td rowspan="3">Test Conditions</td><td colspan="5">Standard : /</td></tr>
<tr><td colspan="3">Viewing distance</td><td colspan="2">80 cm (approximately)</td></tr>
<tr><td colspan="3">Room illumination</td><td colspan="2">/</td></tr>
<tr><td colspan="2">Observers</td><td colspan="5">35 different male and female observers (ages range from 21 to 35)</td></tr>
<tr><td colspan="2">Method</td><td colspan="5">linear displacement of the images</td></tr>
<tr><td rowspan="3">Subjective ratings of images</td><td>Raw data</td><td colspan="5">Nature of raw data is not specified by the database authors.</td></tr>
<tr><td>Final scores</td><td colspan="5">DMOS</td></tr>
<tr><td>Scores’ Scale</td><td colspan="5">0..9</td></tr>
</table>

Table 2.7: Summary of CSIQ image database description.

2.5 Conclusion

Given a rich literature for visual data quality assessment, a systematic summarization and comparison studies are necessary to facilitate the research and application of image quality techniques. A classification scheme of state-of-the-art objective image quality assessment methods would be of big importance to serve this purpose.

Despite their evident utility, research should not rely exclusively on objective quality models development since their predictive performance is usually evaluated in terms of their ability to predict visual image quality in a manner that agrees with subjective ratings. Similarly, subjective quality measures have their limitations but cannot be completely substituted since they are essential to establish the predictive capabilities of their objective counterparts. In this regard, it is understood that it may not be possible to fully characterize visual content quality evaluation systems' performance by objective means. Consequently, it is necessary to supplement objective measurements with subjective rating tests.

In the next chapter, we present a statistical study where we evaluate the predictive abilities of eighteen objective quality metrics over six subjective image quality databases and compare their performances.

CHAPTER

3

Chapter 3

Quality Metrics Performance Evaluation and Comparison

3.1 Introduction

This chapter deals with a performance evaluation and comparison of number of objective image quality state-of-the art full-reference models over six public subjectively rated image quality databases.

To provide a complete quantitative evaluation investigation on the objective full-reference image quality models performance (elucidated in section 2.2 of chapter 2), we have attempted to quantify how much the models are able to predict the raw subjective scores. Thus, we have measured the Pearson's Correlation Coefficient (PCC) as an indication on the correlation, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) to quantify the prediction accuracy, the Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order Correlation Coefficient (KROCC) to indicate the monotonicity measure.

All the calculations are first performed on the whole databases (described in section 2.4 of chapter 2) and then on the sets concerned with the four degradation types namely the JPEG compression, JPEG2000 compression, noise and Gaussian blur.

3.2 Predictive Performance Criteria

The Video Quality Experts Group (VQEG) Phase I FR-TV [33] suggests a definition for the performance of the image quality assessment algorithms. It consists in defining the performance in terms of several attributes namely *correlation*, *consistency* and *monotonicity* for comparison to be reliable.

3.2.1 Correlation

Let assume that the subjective image quality ratings represented by the mean opinion score (MOS) or by difference mean opinion score (DMOS) be called the true values X and the computed objective image quality measures be called the estimated values Y. Let also assume that X and Y are vectors of size N given by: $X = \{x_i \mid i=1, 2, \ldots, N\}$ and $Y = \{y_i \mid i=1, 2, \ldots, N\}$.

Correlation between X and Y vectors is represented by the Pearson's Correlation Coefficient (PCC) defined as the ratio of the covariance between X and Y to the product of their respective standard deviations. It assumes that the relationship between variables X and Y is linear and measures its strength.

$$PCC = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{3-1}$$

$$PCC = \frac{\sum_{i=1}^{N}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3-2}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the mean values of *X* and *Y* vectors, respectively.

3.2.2 Accuracy

Accuracy establishes the faithfulness of the estimated values *Y* to match the true values *X*. It is measured by the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) defined respectively as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \quad (3\text{-}3)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i| \quad (3\text{-}4)$$

Both of RMSE and MAE denote the average magnitude of error and consequently the nearest they are to zero, the more accurate is the measure.

3.2.3 Monotonicity

In the context of image quality evaluation, monotonicity can be interpreted as the ability of the measure to assess how well the relationship between qualitative and quantitative image quality scores can be described using a monotonic function. Spearman's and Kendall's rank order correlation coefficients, denoted by RHO and TAU respectively, are typical non-parametric estimators of monotonicity. Unlike the Pearson's correlation coefficient, they do not require any assumption on the linearity between variables. They are defined as:

$$RHO = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)} \quad (3\text{-}5)$$

$$TAU = \frac{\sum C - \sum D}{[N(N-1)]/2} \quad (3\text{-}6)$$

where $d_i = rank(x_i) - rank(y_i)$,

$\sum C$ is the number of concordances, $\sum D$ is the number of discordances and *[N(N-1)]/2* is the total number of pairs.

Pearson's, Spearman's and Kendall's correlation coefficient values lie between -1 and +1. Perfect correlations of -1 or +1 occur when each of the variables is a perfect linear (for PCC) or monotone (for SROCC and KROCC) function of the other.

3.3 Proposed Comparative Study

In this section, we have investigated eighteen image quality assessment algorithms described in chapter 2 and summarized in Table 3.1. For the study to be reliable, we have reused the source code provided in [34] or by the algorithms' authors. We have also made evaluation and comparison tests on six image quality databases subjectively rated and publicly available also described in chapter 2.

To provide a complete quantitative evaluation investigation on the aforementioned objective full-reference image quality models performance, we have attempted to quantify how much the models are able to predict the raw subjective scores. Thus, according to the VQEG recommendations, we have measured the Pearson's Correlation Coefficient (PCC) as an indication on the **correlation**, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) to quantify the **prediction accuracy**, the Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order Correlation Coefficient (KROCC) to indicate the **monotonicity** measure.

All the calculations are first performed on the whole databases and then on the sets concerned with the four degradation types namely the JPEG compression, JPEG2000 compression, noise and Gaussian blur [24].

3.3.1 On the use of a logistic function

The VQEG Phase I FR-TV [33] suggests the use of a logistic function for nonlinear mapping between the subjectively rated scores and the objectively predicted values before calculating the performance measures related to correlation and accuracy. This recommendation is valuable for image and video quality assessment but particularly to the High-definition television (HDTV) since a video quality metric generates as many objective

scores as the number of video sequences under examination, and it is so important to realign the behaviour of these objective scores to the subjective ones in order to fairly compare them [35]. The mapping is performed by applying a nonlinear transformation using a monotonic function whose parameters should be optimized for both the metric's generated values and the ratings given by a panel of human observers. This is reiterated for each quality metric over each considered data sample into a 95% confidence interval.

Symbol / year	**Metric's description**	**Reference**
MSE	Mean Squared Error	
PSNR	Peak Signal to Noise Ratio	
SNR	Signal-to-Noise Ratio	
WSNR(1993)	Weighted Signal-to-Noise Ratio	[4]
NQM (2000)	Noise Quality Measure	[6]
UQI(2001)	Universal image Quality Index	[8]
SSIM (2003)	Structural SImilarity Index	[7]
MS-SSIM (2003)	Multi-Scale SSIM index	[9]
VIF (2004)	Visual Information Fidelity	[13]
VIFP (2004)	Pixel-based VIF	[13]
IFC (2005)	Information Fidelity Criterion	[12]
M-SVD(2006)	Multidimensional IQM based on the SVD[2]	[14]
PSNRHVS(2006)	HVS based PSNR	[18]
PSNRHVSM(2007)	Modified PSNRHVS	[19]
VSNR(2007)	Visual Signal-to-Noise Ratio	[16]
MSSIM(2009)	Modified SSIM with Automatic down-sampling	[9]
R-SVD (2009)	Right singular vectors of the SVD based metric	[15]
RFSIM(2010)	Riesz-transform based Feature SImilarity Metric	[17]

Table 3.1: Summary of the full-reference IQMs being evaluated and compared.

[2] SVD denotes the Singular Value Decomposition theorem

There exist different logistic functions that have been used in the literature amongst especially the video quality assessment community. There are fitting functions that have been defined empirically like the ones suggested in [35-37] and psychometric functions inspired from the behaviour of the human visual system.

However, the fitting function described in [38] by H. R. Sheikh *et al.* is largely employed within the image quality assessment community and in the present work. It is a five parameters monotonic logistic function given by Equations (3-7) and (3-8) below:

$$Quality(x) = \beta_1 \log istic(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \qquad (3\text{-}7)$$

$$logistic(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \qquad (3\text{-}8)$$

where x and $Quality(x)$ denote the objective quality scores before and after nonlinear mapping, respectively.

In this work, the parameters β_1 to β_5 have been optimized for each database and each quality metric individually by means of the unconstrained multivariable optimization and the unconstrained nonlinear minimization techniques. For the first technique, Nelder-Mead Simplex method [39] has been employed (the Matlab's *fminsearch* function). For the second one, the Quasi-Newton Method with a cubic line search procedure has been applied (the Matlab's *fminunc* function). It uses the FBGS formula [40-43] for updating the approximation of the Hessian matrix using first order derivatives. FBGS (for Broyden, 1969, Fletcher, 1970, Goldfarb, 1970, and Shanno, 1970) is generally regarded as the best performing method. The performance measures concerned with the logistic fitting are the Pearson's correlation coefficient (PCC), the root mean squared and the mean absolute errors (RMSE and MAE).

Nevertheless, it is important to point out that there exist some work on image data [44, 45] as well as on video data [35, 46] where it has been suggested that such nonlinear mappings yield to higher correlation coefficients as well as they can provide different results under the influence of parameters tuning.

3.3.2 The significance of difference

After applying the nonlinear mapping on the PCC, RMSE and MAE but not on the Spearman's and Kendall's rank order correlation coefficients (SROCC and KROCC), we have studied the significance of the difference of the image quality predictive performance across the six databases and then between the eighteen objective image quality predictors. This is to answer the following two main questions: 1) Are there any significant differences between the existing subjective image quality databases? 2) Is there any significant improvement of the capability to objectively predict the perceived visual image quality having its original version at hand?

To achieve this aim, we have used a non parametric Two-Way Analysis Of Variance (ANOVA) called the Friedman test which is a measure of variability based on the median statistic. The choice of this statistical method is motivated by the following properties. The first one is that no assumption is made about the distribution of data being compared. Secondly, the data should not be independent; this means that it is possible to compare repeated records over the same samples. The third advantage of the Friedman test is that it is as robust as a parametric test when the sample size is greater than five which is the case in the present comparative study.

The first property of the Friedman test is known as the free-distribution property. It is very useful when the distribution of the data to be compared is not Gaussian; which is the case of the quality measures values computed on sets of images. Free-distribution tests use the ranks of the data rather than their raw values to calculate the statistic. However, this property may become a weakness when the sample size is less than five. This problem is not of concern in the present work since the sample size is 6 and 18 when assessing the variability of the databases and the metrics, respectively.

Thus, the Friedman test returns a probability value (p-value) for the null hypothesis "H_0: there are no significant differences between the variables". If the p-value is near zero, then H_0 is rejected and it can be concluded that at least two of the variables are significantly different from each other. In our work, we needed additional information about which pairs

of variables are significantly different, and which are not in the case where the null hypothesis is rejected. Therefore, we have performed a multiple comparison procedure.

The decision made depends on the confidence interval considered for the comparative study. If it is 95%, then the p-value should be lower than 0.05 to reject the null hypothesis. If the confidence interval is 99%, then the p-value should be lower than 0.01 to conclude that there is significant variability.

3.4 Experimental Results and Comments

As explained in section 3.3, the correlation, accuracy and monotonicity based performance investigation has been conducted on eighteen objective full-reference image quality estimators over six publicly available subjective image quality databases. The procedure is repeated to the whole image datasets, and then to the four image sets subject to the four degradation types considered in the present work (JPEG, JPEG2000, blur and noise). The Friedman statistical method has been employed to compute the probability (p-value) that there are no significant differences between the variables. In the first stage, the 99% confidence interval has been selected, so the p-value is compared to 0.01. If it is lower then a pairwise comparison is performed to know which pairs of variables are significantly different, and which are not.

It turned out that the Friedman analysis over the eighteen image quality metrics in the 99% CI is very tight when comparing the objective quality metrics. This leads us to extend the statistical test to the 95% confidence interval. The p-value is subsequently compared to 0.05. This allows us to make the tests at a larger scale and thus to get additional information on the predictive quality models that fall into the same performance range as the ones found in the 99% confidence interval analysis. The full comparative results have been published in [47] before logistic fitting and in [24] after applying a logistic function.

The Friedman's test results are summarized in tables 3.1 to 3.6 and F.1 to F.4 where the "x" symbol means that a significant difference has been detected between the two variables. The tables are symmetric since the multiple comparisons have been reiterated pairwise. In

tables 3.3, 3.5, 3.7 and tables in appendix F, it was more convenient to represent only the variables pairs that are significantly different than to make 18 x 18 matrices. In order to know if a measure X is "significantly better" or "significantly worst" than Y, we refer to the appendices A to E where the full numerical values are provided and also published in [48].

3.4.1 Friedman analysis of the correlation performance

The correlation performance is given by the Pearson's correlation coefficient (PCC) for which the absolute values after nonlinear fitting are depicted in appendix A.

a) Comparison of the image quality databases

Results related to the Friedman analysis of the PCC variability in the 99% CI over the six image quality databases are presented in table 3.2 below. They show that:

	All data sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		x				
LIVE	x				x	
IVC					x	
A57						
TID		x	x			x
CSIQ					x	

	JPEG coded images sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		x		x		x
LIVE	x					
IVC						
A57	x					
TID						x
CSIQ	x				x	

	JPEG2000 coded images sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		x				x
LIVE	x				x	
IVC						
A57						
TID		x				x
CSIQ	x				x	

	Gaussian blurred images sets				
	LIVE	IVC	A57	TID	CSIQ
LIVE					
IVC				x	
A57					
TID		x			x
CSIQ				x	

	Noised images sets			
	LIVE	A57	TID	CSIQ
LIVE			x	
A57				
TID	x			x
CSIQ			x	

Table 3.2: Pearson's Correlation Coefficient (PCC) variability over the 6 databases according to the Friedman test in the 99% CI.

- The predictive correlation performance of the overall studied metrics obtained over the Toyama database is always significantly worst than the one obtained on the LIVE and CSIQ databases for the JPEG and JPEG 2000 coded images sets knowing that the Toyama database supports only the two aforementioned degradations.

- The lowest correlations are obtained on the overall TID database. In addition, they are significantly worst than the ones obtained on the LIVE, IVC and the CSIQ databases over the entire datasets. This is related to the considerable number of aberrant values of PCC highlighted in frames in appendix A.

- The studied image quality metrics are always significantly less correlated with the subjective scores on the TID than on the CSIQ databases whatever the image set considered: all data, JPEG compressed images, JPEG 2000 coded images, noised images or Gaussian blurred images.

b) Comparison of the image quality metrics

- As can be noticed from the underlined values in appendix A, the best linear correlation coefficients are mostly achieved with the VIF metric amongst the 18 quality models under study. This is true over four databases namely Toyama, LIVE, IVC and CSIQ for all the image sets except the noised images one.

- Moreover, according to the Friedman test results, the predictive performance of the VIF model in terms of correlation is significantly higher than that of the R-SVD, MSE, PSNR and SNR depending on the case study as illustrated in table 3.3.

- The R-SVD metric records significant lower correlations than VIF and MS-SSIM on the entire data sets. This is due to the aberrant (framed) values of the R-SVD on Toyama, A57 and TID databases in appendix A.

	All data sets		
	MS-SSIM	VIF	R-SVD
MS-SSIM			x
VIF			x
R-SVD	x	x	

	JPEG coded images sets				
	MSE	PSNR	SNR	MS-SSIM	VIF
MSE					x
PSNR					x
SNR				x	x
MS-SSIM			x		
VIF	x	x	x		

	JPEG2000 coded images sets		
	SNR	VIF	R-SVD
SNR		x	
VIF	x		x
R-SVD		x	

	Gaussian blurred images sets	
	SNR	VIF
SNR		x
VIF	x	

Table 3.3: Pearson's Correlation Coefficient (PCC) variability over the 18 quality metrics according to the Friedman test in the 99% CI.

c) Preliminary conclusions

- For the noised images, no significant variability in correlation between the metrics in the 99% Confidence Interval (CI) has been detected.

- Unlike the Toyama and LIVE databases, lower linear correlation results are obtained on the TID database.

3.4.2 Friedman analysis of the accuracy performance

The image quality community usually assesses the accuracy performance of the quality models by means of the RMSE and the MAE on which the logistic function defined by Equations (3-7) and (3-8) has been applied. The full numerical error values of the 18 quality metrics over the 06 databases are given in appendices B and C, respectively.

a) Comparison of the image quality databases

- It is worth noting that the Friedman test gives similar results for the Root Mean Squared and the Mean Absolute Errors over the six databases in the 99% Confidence Interval.

- The RMSE and MAE variability depicted in table 3.4 indicate that the predictive accuracy performance of the quality models obtained over the LIVE database is always significantly inferior than the one obtained on the IVC, A57 and CSIQ databases whatever the image set considered. An exception is that the noise degraded images set is not contained in the IVC database. This statistical analysis result can be justified by the high error values recorded by the PSNR, SNR, WSNR, NQM, PSNR-HVS and PSNR-HVSM on the LIVE database for the five case studies as highlighted in bold in appendices B and C.

- The lowest error values and consequently the best accuracy performance of the image quality metrics is achieved over the A57 and the CSIQ databases. However, this superiority is found to be significant only to Toyama, LIVE and TID databases but not to IVC. This is true for the following case studies: all data, JPEG coded images, JPEG 2000 coded images and Gaussian blurred images sets.

	All data sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY				x		x
LIVE			x	x		x
IVC		x				
A57	x	x			x	
TID				x		x
CSIQ	x	x			x	

	JPEG coded images sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY				x		x
LIVE			x	x		x
IVC		x				
A57	x	x			x	
TID				x		x
CSIQ	x	x			x	

	JPEG2000 coded images sets					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY				x		x
LIVE			x	x		x
IVC		x				
A57	x	x			x	
TID				x		x
CSIQ	x	x			x	

	Gaussian blurred images sets				
	LIVE	IVC	A57	TID	CSIQ
LIVE		x	x		x
IVC	x				
A57	x			x	
TID			x		x
CSIQ	x			x	

	Noised images sets			
	LIVE	A57	TID	CSIQ
LIVE		x		x
A57	x		x	
TID		x		
CSIQ	x			

Table 3.4: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) variability over the 6 databases according to the Friedman test in the 99% CI.

b) Comparison of the image quality metrics

- Table 3.5 below illustrates that the MSE, PSNR and SNR models are significantly less accurate compared to the VIF quality measure according to the case study (JPEG compressed or Gaussian blurred images).

	JPEG coded images sets (MAE)	
	SNR	VIF
SNR		x
VIF	x	

	JPEG coded images sets (RMSE)		
	PSNR	SNR	VIF
PSNR			x
SNR			x
VIF	x	x	

	Gaussian blurred images sets (MAE)		
	MSE	SNR	VIF
MSE			x
SNR			x
VIF	x	x	

Table 3.5: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) variability over the 18 quality metrics according to the Friedman test in the 99% CI.

c) Preliminary conclusions

- Unlike the Toyama and LIVE databases, higher root mean squared and mean absolute errors are recorded on the TID database.

- There is no image quality metric over the 18 investigated which is more accurate than others for the JPEG2000 compressed and the noised images according to the Friedman test in the 99% confidence interval.

3.4.3 Friedman analysis of the monotonicity performance

The Spearman's and Kendall's Rank Order Correlation Coefficients (SROCC & KROCC) are the measures of the monotonicity performance. The appendices D and F contain the values of SROCC and KROCC, respectively.

a) Comparison of the image quality databases

The following comments about the monotonicity performance over the six image quality databases can be made from table 3.6.

	All data sets (SROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY					**x**	
LIVE			**x**	**x**	**x**	**x**
IVC		**x**			**x**	
A57		**x**				
TID	**x**	**x**	**x**			**x**
CSIQ		**x**			**x**	

	All data sets (KROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		x			**x**	
LIVE	x		**x**	**x**	**x**	**x**
IVC		**x**			**x**	
A57		**x**				
TID	**x**	**x**	**x**			**x**
CSIQ		**x**			**x**	

	JPEG coded images sets (SROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		**x**				**x**
LIVE	**x**		**x**	x	**x**	
IVC		**x**				
A57		x				
TID		**x**				
CSIQ	**x**					

	JPEG coded images sets (KROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		**x**				**x**
LIVE	**x**		**x**		**x**	
IVC		**x**				
A57						
TID		**x**				
CSIQ	**x**					

	JPEG 2000 coded images sets (SROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		**x**			x	
LIVE	**x**		**x**	**x**	**x**	
IVC		**x**				
A57		**x**				**x**
TID	x	**x**				**x**
CSIQ				**x**	**x**	

	JPEG 2000 coded images sets (KROCC)					
	TOY	LIVE	IVC	A57	TID	CSIQ
TOY		**x**				
LIVE	**x**		**x**	**x**	**x**	
IVC		**x**				
A57		**x**				**x**
TID		**x**				**x**
CSIQ				**x**	**x**	

	Noised images sets (SROCC)			
	LIVE	A57	TID	CSIQ
LIVE		x	**x**	**x**
A57	x			
TID	**x**			
CSIQ	**x**			

	Noised images sets (KROCC)			
	LIVE	A57	TID	CSIQ
LIVE			**x**	**x**
A57			x	
TID	**x**	x		
CSIQ	**x**			

	Gaussian blurred images sets (SROCC)				
	LIVE	IVC	A57	TID	CSIQ
LIVE			**x**	**x**	
IVC					
A57	**x**				**x**
TID	**x**				**x**
CSIQ			**x**	**x**	

	Gaussian blurred images sets (KROCC)				
	LIVE	IVC	A57	TID	CSIQ
LIVE			**x**	**x**	
IVC			x		
A57	**x**	x			**x**
TID	**x**				**x**
CSIQ			**x**	**x**	

Table 3.6: Spearman's and Kendall's Rank Order Correlation Coefficients (SROCC & KROCC) variability over the 6 databases according to the Friedman test in the 99% CI.

- The best results are obtained on the LIVE database. However, they are significantly superior to those gotten on the Toyama, IVC and TID databases whatever the cases study considered.

- The worst monotonicity results are recorded on the A57 database. They are significantly lower than those recorded on the JPEG 2000 compressed and the Gaussian blurred images of the LIVE and CSIQ data sets.

b) Comparison of the image quality metrics

	All data sets					
	SNR	MS-SSIM	VIF	MSSIM	R-SVD	RFSIM
SNR				x		
MS-SSIM					x	
VIF					x	
MSSIM	x				x	
R-SVD		x	x	x		x
RFSIM					x	

	JPEG coded images sets				
	SNR	MS-SSIM	VIF	PSNR-HVSM	MSSIM
SNR		x	x		x
MS-SSIM	x				
VIF	x				
PSNR-HVSM					
MSSIM	x				

	JPEG 2000 coded images sets			
	SNR	MS-SSIM	PSNR-HVSM	R-SVD
SNR		x		
MS-SSIM	x			x
PSNR-HVSM				x
R-SVD		x	x	

	Gaussian blurred images sets		
	SNR	VIF	M-SVD
SNR		x	
VIF	x		x
M-SVD		x	

Table 3.7: Spearman's and Kendall's Rank Order Correlation Coefficients (SROCC & KROCC) variability over the 18 quality metrics according to the Friedman test in the 99% CI.

- The VIF, MS-SSIM, SSIM, PSNR-HVSM and RFSIM image quality models are significantly more monotonous than SNR, M-SVD and R-SVD measures depending on the case study as

illustrated in table 3.7. We can also notice the number of the aberrant values of SROCC and KROCC obtained by the R-SVD algorithm. They are highlighted in frames in appendices D & E.

- For the noised images, no significant variability in monotonicity (for both SROCC and KROCC values) in the 99% confidence interval has been detected.

c) Preliminary conclusions

- Unlike the Toyama and LIVE databases, lower rank order correlation results are given on the TID database.

- There is no image quality metric over the 18 investigated which is more monotonous with the subjective quality scores (MOS or DMOS) than others for the noised images.

3.4.4 Friedman test results on the 95% confidence interval

The Friedman analysis of the eighteen image quality measures in the 99% CI has shown that the VIF model outperforms its counterparts. Looking for other metrics that have similar performance capabilities as the VIF, we have extended the test to the 95% CI where the probability value (p-value) that there are no significant differences between the quality metrics is compared to 0.05 instead of 0.01. If it is lower, then a multiple pairwise comparison is performed to know which pairs of objective metrics are significantly different.

- The extended Friedman test reveals that the MS-SSIM, MSSIM and VIFP quality measures have almost the same performance as the VIF model in terms of correlation, accuracy and monotonicity. This conclusion can be drawn from tables F.1 to F.4 in appendix F where the variability of the PCC, RMSE, MAE, SROCC and KROCC over the 18 full reference quality algorithms according to the Friedman analysis in the 95% confidence interval is summarized.

- For the noised images, no significant variability in correlation, accuracy and monotonicity in the 95% confidence interval has been detected. This is probably due to the fact that the nature of noise induced to images is different from one database to another. Another

possible interpretation is that the Friedman test might not be robust when the number of variables is less than 5 (noise is present in only 4 databases).

3.5 Interpretation of the Results

In this chapter we have carried out a complete quantitative predictive performance evaluation of eighteen state-of-the-art full reference image quality measures over six public subjectively rated image quality databases. We were particularly interested in four types of degradation including JPEG and JPEG2000 compression, noise and Gaussian blur.

Performance evaluation focused on three measures: correlation, accuracy and monotonicity. Thus we used the Pearson's Correlation Coefficient, the Root Mean Squared Error, the Mean Absolute Error, the Spearman's Rank Order Correlation Coefficient and the Kendall's Rank Order Correlation Coefficient. Statistical tests have been performed via the Friedman analysis to check if there are significant differences between the existing quality databases on the one hand and between the predictive capabilities of the investigated objective models on the other hand.

According to the Friedman's tests over the noised images in the 99% and 95% confidence intervals, the Noise degradation is ill represented in the existing subjectively rated quality databases.

Furthermore, the preliminary conclusions in the previous section reveal that unlike the TID database, the Toyama and LIVE ones record higher values in terms of correlation, accuracy and monotonicity. This is particularity due to the fact that the LIVE database (release 2) contains 20.5% of images (202 out of 984) of perfect quality and having a DMOS value equal to 0, the Toyama database includes 14.3% of non distorted images (28 out of 196). While in the TID database, 4% of images (68 out of 1700) are very severely degraded.

In the special case of images of perfect quality, the MSE, M-SVD and R-SVD tend towards zero while the values of UQI, SSIM, MS-SSIM, VIF, VIFP, MSSIM and RFSIM are equal to 1. However their PSNR, SNR, WSNR, NQM, PSNR-HVS and PSNR-HVSM values tend to infinity.

The reason for the root mean squared and the mean absolute errors being very high and their linear correlation coefficient (PCC) is the same as can be seen in bold in appendices A to C. This is inversed for the images of very poor quality which explains the difference between the three databases [24].

Coefficients of the logistic function defined by equations 3.7 and 3.8 are given in appendices G through L for each quality measure, for each dataset and for each image quality database.

3.6 Conclusion

According to the results of the investigation presented in this chapter, it can be concluded that the performance of the error-based image quality estimators depends on the structure of the image database on which it is applied. We mean by structure, the proportions of images at different levels of degradation, in particular the proportions of images of very good or very poor quality.

In order to assert the conclusion that Toyama, LIVE and TID databases are different, we have run the Friedman analysis only on IVC, A57 and CSIQ ones. No variability over the 18 quality models has been detected whatever the performance measure assessed and the dataset considered.

This finding raises the question on the reliability of the image quality metrics performance evaluation tools based on correlation, accuracy and monotonicity and the impact of the structure of image quality databases. As a result of the conclusions of this research work, we propose a new research direction to investigate new benchmarking tools of image quality metrics in order to reliably measure the evolution of the field and to point out any shortcomings of each method.

CHAPTER

4

Chapter 4

Overview on Machine Learning and Artificial Neural Networks

4.1 Introduction

Learning is an inherent characteristic of humans and animals. Indeed, while executing similar tasks, the ability to improve performance is acquired via synthesizing different types of information. As regards machines, learning broadly refers to the enhancements of expected systems' performance while executing repeated tasks. Such learning tasks mainly involve density estimation, regression, and classification where the system attempts to predict the data density, a continuous target variable or a discrete target variable, respectively.

Machine learning (ML) is an active research field concerned with the development of algorithms able to generalize from their experience. This is possible by inferring rules from observing examples used for training the system and suitably making further predictions on future data. The importance of achieving learning in machines is manifold [49]:

- ✓ Important relationships and correlations might exist among large piles of data. ML methods can often be used to analyse and extract these relationships (data mining).
- ✓ The amount of knowledge available about certain tasks might be too large for explicit encoding by humans. Machines that learn this knowledge gradually might be able to capture which information is relevant to the actual task and which one is non-informative or redundant (dimensionality reduction / feature selection).

- ✓ Environments change over time and new knowledge about tasks is constantly being discovered by humans. Continuing systems redesign is impractical, but machines that can adapt to a changing environment would be able to track much of the new knowledge (evolutionary systems).
- ✓ Another important task of ML is classification, which is also referred to as pattern recognition, in which machines learn to automatically recognize complex patterns, to distinguish between them, and to make intelligent predictions on their class. When the classes are not known beforehand, machines might be able to assign a set of objects into groups called clusters so that the objects in the same cluster are more similar to each other than to those belonging to other clusters (clustering).
- ✓ Some tasks cannot be defined well except by example. That is, we might be able to specify data samples but not a concise relationship between inputs and desired outputs. There exist prediction techniques that allow approximate such an input/output function (regression).
- ✓ Learning in machines also might help us understand how humans learn. Biological phenomena are considered nonlinear by nature. In this thesis, we are interested in visual data quality appreciation which is a nonlinear natural cognitive task that evolves by time and personal experience. However, mechanisms leading to the visual data quality evaluation are still ill-understood. Machine learning techniques might help us determine how do humans make judgement about the quality of what do they regularly see on their screens and which factors mainly affect this process. In the following chapters (5 and 6), we develop machine learning based techniques for image quality assessment in an attempt to shed more light on this subject.

This chapter provides an overview of the principle of learning that can be adhered to machines to improve their performance. In the next section 4.2, we point out the major historical events that marked the ML research field. The different paradigms have also been briefly introduced in section 4.3 with a particular attention drawn on the supervised approach employed for further investigations in the thesis. The rest of the chapter is restricted to the artificial neural networks based techniques and more particularly to the feed-forward multi-layer perceptron employed in most contributions of our work.

In section 4.4, we address some basic issues related to the use of neural networks. We essentially focus on the generalization capabilities including the inconveniences presented by the backpropagation learning algorithm applied to multi-layer perceptrons (section 4.5) as well as the solutions provided to cope with these problems (section 4.6). The chapter ends with a brief discussion on the general research questions as well as the choice of the methods used in chapters 5 and 6.

4.2 Brief History

The learning problem is characterized by the following significant historical events [50]:

4.2.1 Construction of the first learning machines (the 60s)

The first model of machine learning called “perceptron” was suggested by Frank Rosenblatt in 1957 in [51]. The perceptron was not new concept, but the contribution was the definition of the model as a program for computers able to solve pattern recognition problems. The elementary perceptron has *N* input neurons and one output neuron. Rosenblatt has built later a model of several levels of neurons where outputs of neurons of the previous level (i-1) are inputs for neurons of the next one (i).

A. Novikoff proved the first theorem about the perceptron [52] that asserts that the perceptron algorithm converges after a finite number of iterations if the data set is linearly separable. Hence, the learning theory that aims at formalizing the automatic learning process is actually introduced, and the relationship between the generalization ability and the principle of minimizing the number of errors on the training set has been established.

4.2.2 Elaboration of the fundamentals of the learning theory (1960-1970s)

This period has known a proliferation of numerous learning models dedicated to solving real-life problems. The Madaline constructed by B. Widrow and M. E. Hoff [53], the learning matrices constructed by K. Steinbush [54], decision trees originally intended for experts systems [55], and hidden Markov models developed for speech recognition problems [56] are typical examples.

During this period, much work has been done in the statistical learning theory, therefore the structural risk minimization theory introduced by V. N. Vapnik and A. J. Chervonenkis [57] has became a popular subject of analysis. M. L. Minsky and S. Papert [58] showed in 1969 that a two layer feed-forward network can overcome many restrictions faced with a single-layer network, but did not present a solution to the problem of how to adjust the weights from input to hidden units.

4.2.3 Introduction of the neural networks (the 80s)

In 1986, Y. LeCun [59] and D. E. Rumelhart *et al.* [60] independently implemented methods for simultaneously setting the weights' values of all neurons of the perceptron using the so-called back-propagation algorithm. The central idea behind this solution is that the errors for the units of the hidden layer are determined by back-propagating the errors of the units of the output layer. For this reason the method is often called the back-propagation learning rule that can also be considered as a generalisation of the delta rule (also called the Least Mean Square (LMS) method) [53] for non-linear activation functions and multilayer networks.

Perhaps the most notable work in the field of machine learning in this period was the pioneering collective book (Volume I and II) by D. E. Rumelhart, J. L. McClelland, and the PDP research group on "Parallel Distributed Processing" [561, 62]. Thereby, interest in information processing models inspired from the nervous system was renewed and many earliest models particularly the Hebb learning rule [63] and the perceptron model [51] have been re-examined. The success of the book "Parallel Distributed Processing" is due to the introduction of new more efficient algorithms, and the emphasis made on the advantages and features of biological nervous systems (massive parallelism, learning capabilities and distributed memory).

4.2.4 Development of alternatives to neural networks (since the 1990s)

Big attention has been focused on alternatives to neural networks since the 1990s. During the time between constructing the perceptron (1957) and implementing the back-propagation technique (1986), the statistical learning theory introduced in the late 1960's

has been extremely developed from a theoretical point of view. Building on thirty years of analysis of learning processes, the synthesis of novel types of learning algorithms (called SVMs: Support Vector Machines) controlling generalization ability began in the middle of the 1990's. Many other kernel methods such as the Radial Basis Functions (RBF) have been developed later as well as the Bayesian neural networks.

Despite the proliferation of numerous powerful learning methods that generally share the same framework as the one depicted in figure 4.1 below, the artificial neural networks are still popular and employed in the most recent applications.

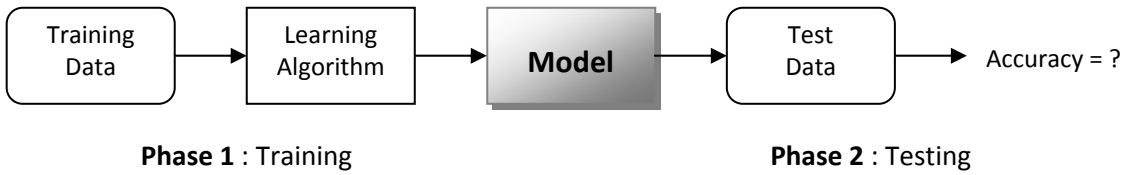

Figure 4.1: Framework of machine learning systems.

4.3 Machine Learning Paradigms

Learning algorithms fall into four categories with respect to the following factors: (i) the type of feedback available for the learning task, (ii) the representation of the learned information and (iii) the availability of the prior knowledge. The field of machine learning usually distinguishes supervised, unsupervised, semi-supervised and reinforcement learning. The key concepts and principles of each of the aforementioned learning paradigms are briefly introduced in this section.

4.3.1 Supervised learning

Supervised learning requires a trainer, who supplies the input-output training data in order to learn a model of it. The data is termed labelled and the learning system adapts its parameters by specific algorithms to generate the desired output patterns from a given input pattern. It is not always convenient to solve practical learning problems using the supervised techniques. In some applications like text processing, video-indexing and

bioinformatics, the labelled data are not available, expensive to generate or cannot be collected for the training process. In the context of classification problems, number of categories might not be available or known, as in the case of scene classification for example, or they might increase with more data as for objects identification. A possible solution to these problems is the use of unsupervised learning systems.

4.3.2 Unsupervised learning

In absence of labelled data or trainer, the desired output for a given input is not known and the system is provided with only unlabelled training data. Consequently, the learning system adapts its parameters autonomously to find a meaningful representation of complicated high dimensional data. Unsupervised learning is commonly used for dimensionality reduction and clustering applications. More details on these techniques can be found in [64].

Nevertheless, the unsupervised learning approach does not totally resolve problems encountered with its supervised counterparts. Clustering problems, for example, are often subjective in nature and the clusters generated by unsupervised models are difficult to evaluate and interpret. Two ways to overcome this issue have been suggested: (i) the first one is to supply limited labelled data to guide the unsupervised process (semi-supervised learning), (ii) the second solution is to incorporate the system's user suggestions and feedback (reinforcement learning).

4.3.3 Semi-supervised learning

As the name suggests, it is in between supervised and unsupervised learning techniques. The semi-supervised solution aims at making advantage of the strengths of both by reducing the amount of labelled data required for supervised learning as well as improving the results of unsupervised clustering to the expectations of the users. Many assumptions have to be made on the labelled/unlabelled data to achieve these goals. However, the difficulty of

verifying the semi-supervised assumptions or mathematically formalizing them is still an open issue [65].

4.3.4 Reinforcement learning

It constitutes another possible solution to the drawbacks of supervised and unsupervised learning techniques. Here, the emphasis is made on learning by the individual from direct interaction with its environment instead of relying on exemplary supervision or complete models of the environment.

In reinforcement learning, the system does not explicitly know the input-output examples, but it receives some form of feedback from its environment. The feedback signals help the learner to discover which actions yield the most reward by trying them. The learner thus adapts its parameters through trial-and-error interactions with a dynamic environment. Another challenging feature specific to reinforcement learning is called delayed reward where the selected actions may affect not only the immediate reward but also all subsequent rewards.

There are two main strategies for solving reinforcement-learning problems. The first is to search in the space of behaviours in order to find one that performs well in the environment. Genetic algorithms and genetic programming fall into this class of methods. The second is to use statistical techniques and dynamic programming methods to estimate the utility of taking one action or another.

4.4 Supervised Learning in Multilayer Neural Networks

The Artificial Neural Networks (ANNs) are variations of the parallel distributed processing idea. They are nonlinear statistical data modelling tools, and their architecture is based on interconnected computational building blocks inspired by biological nervous system which perform the processing in a parallel way as shown in the diagram of figure 4.2. Furthermore, a special interest in networks arises from their ability to perform nonlinear approximation. It is known that ANN with one hidden layer (and also higher layer networks) can interpolate

any multidimensional function with given accuracy and can exactly implement any arbitrary finite training set [66]. Indeed, artificial neural networks are commonly used to model complex relationships between inputs and outputs or to find patterns in data. In other words, they are capable to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. The above mentioned property of network mapping makes ANNs a powerful learning method with widespread applications like function fitting (also called regression or function approximation), pattern recognition, data clustering, and time series analysis. In this section, we focus on the function approximation learning capabilities of neural networks which we will exploit in further investigations related to image quality assessment.

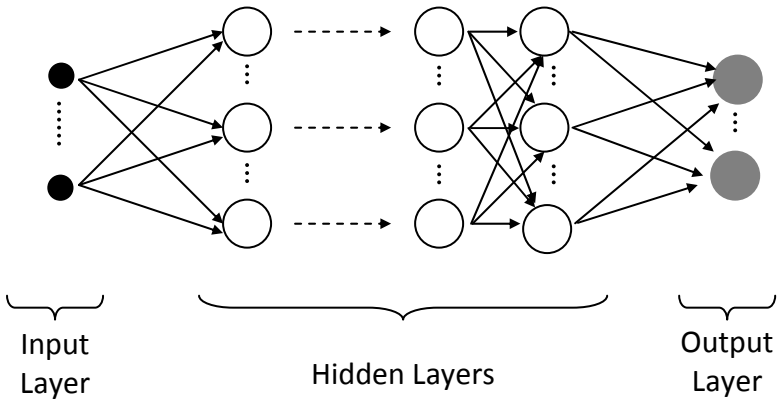

Figure 4.2: A multi-layer feed-forward perceptron generic diagram.

4.4.1 The Backpropagation Learning Algorithm

The backpropagation algorithm proposed in [59, 60] can be applied to networks with any number of layers. The universal approximation theorem has been shown; it states that only one layer of hidden units suffices to approximate any function with finitely many discontinuities to arbitrary precision, provided the activation functions of the hidden units are non-linear [67, 68]. In many applications a feed-forward network with a single layer of hidden units is used with a sigmoid logistic function (see figure 4.3) for the units.

The standard backpropagation is a gradient algorithm in which the weights of the units are computed by a forward pass through the network for each training case. Then, starting from the output units, the derivatives of the cost function are calculated in backward pass with respect to the weights of units. In function approximation applications with *n* outputs, the standard cost function (also called the performance function or the error rate) is the sum of the squared errors between the desired and the estimated outputs, y_i and $\hat{y}_i$ respectively is given by equation (4-1) below:

$$f(C) = SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (4\text{-}1)$$

Feed-forward multilayer perceptrons that have one or more logistic hidden units and are properly trained using backpropagation tend to give reasonable answers when presented with inputs that they have never seen. This generalization property makes it possible to emulate a large variety of tasks that humans perform provided a large enough pool of representative input-output samples.

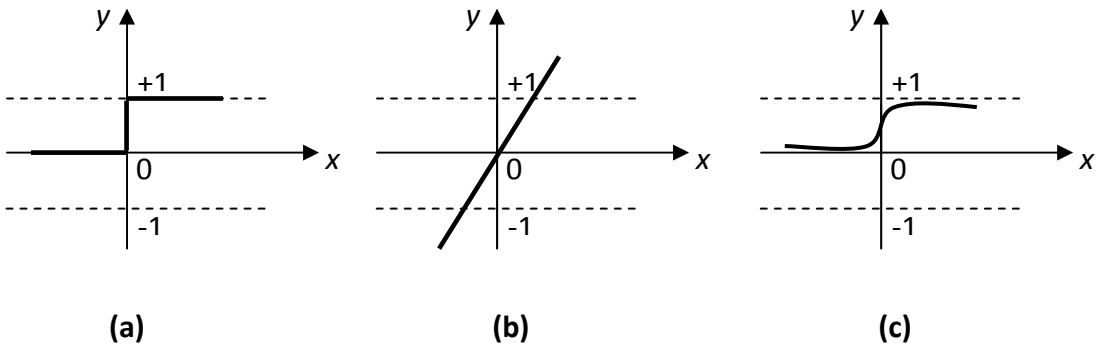

Figure 4.3: Different types of logistic functions: (a) hard-limit, (b) linear, and (c) sigmoid.

4.5 Problems Related to ANNs Generalization Capabilities

As evoked in the previous section, ANNs are often referred to as universal function approximators. According to the Kolmogorov's theorem, feed-forward backpropagation networks can exactly implement an arbitrary training set with a sufficient number of

neurons in the hidden layer [66]. Nevertheless, the capabilities of the nonlinear approximation of adaptive neural network systems can evolve during the training process for better or for worse. Unfortunately, they can result in bad generalization to new data that will lower the predictive ability of the network if it learns both investigated dependencies and noise. That is, cautions on two design issues must be considered: (i) size of ANN and (ii) time of ANN training.

The overfitting and overtraining problems refer to exceeding the optimal ANN size and the time of ANN training, respectively. They result in bad generalization of the system [69].

4.5.1 Overfitting

Overfitting occurs when the error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The network has memorized the training examples, but it has not learned to generalize to new situations. We say that it has learned by heart.

In addition, if we had access to an unlimited number of training data samples, the model that provides the lowest generalization error is theoretically the one which will be selected and the phenomenon is commonly referred to as the curse of dimensionality. However in real applications we only have access to a finite set of samples usually insufficient to directly solve the problem at hand. To cope with the lack of data, a complicated network with large number of weights is naïvely used for training which makes the model liable to overfit.

One method for avoiding overfitting is to find the optimal network architecture (number of hidden layers and number of neurons in layers) that provides the adequate solution with the optimal error. Unfortunately, it is difficult to know beforehand how large a network should be for a specific application. The larger network is used, the more complex the functions the network can create. Thereby, if the number of parameters in the network is much smaller than the total number of points in the training set, then there is little or no chance of overfitting. However, if the model is too simple an underfitting phenomenon can occur.

Knowing that the feed-forward multilayer neural networks do not require complex model to establish complex nonlinear relationships, overfitting does not have any influence on network generalization ability if overtraining is avoided [69] by the "split-sample" method evoked in the next sections.

4.5.2 Overtraining

The overtraining problem has not benefited from a big deal of attention in the literature. With time of training, the cost function of the network gradually decreases over learning set. It is optimal to stop net training before complete convergence has occurred, in other words the sum squared error should not reach zero.

A probable description of overtraining is that the network learns the gross structure first and then the fine structure that is generated by noise [69]. A stopping point should be determined for network training by testing error convergence on an additional test set. The early stopping method presented below is the most commonly proposed technique to avoid overtraining.

4.6 Solutions to Generalization Inconveniences

There are methods for improving neural network generalization: early stopping, cross-validation and Bayesian regularization described in the next subsections.

4.6.1 Early stopping

- The hold-out method

The hold-out method is the default technique used for early stopping in ANNs that aims at determining a stopping point for the backpropagation error. To do so, the set of available data is separated into two disjoint subsets: a training subset used to train the network as its name suggest and a test subset to estimate the generalization error of the obtained model.

It is crucial to realize that the test error is not a good estimate of the generalization error since it is periodically computed during training. One way for getting an unbiased estimate of the generalization error is to compute it on data samples that are not used at all during the training process: the three-way split-sample solution.

- The three-way split sample method

The three-way split-sample method is an alternative technique used for early stopping in ANNs that aims at improving their generalization capabilities. Here, the set of available data is separated into three disjoint subsets: training, validation and test. The training subset is applied to the ANN for computing the gradient and updating the network weights and biases. The validation subset is used to further validate the ANN parameters adjustments. The error on the validation set is monitored during the training process. It normally decreases during the initial phase of training. However, when the network begins to overfit the data, the error on the validation set begins to rise. Practically, when the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are retrieved. Finally, the test subset containing not already seen data is used only to assess the performance of the fully-trained neural network after the final model has been chosen. After evaluation of the final model on the test subset, no further model tuning would be done.

The division of data set into training, validation and test subsets is often performed randomly at different proportions. Throughout all our implementations, the input/output data is randomly split so that 80% of the samples are assigned to the training set, 10% to the validation set, and 10% to the test set.

- Advantages and shortcomings

Despite their simplicity and fastness, the hold-out and the three-way split-sample methods have two basic drawbacks. Firstly, in problems where we have a dataset of very limited size, we may not be able to afford the opportunity of setting aside a portion of the dataset for test (hold-out method) or even two portions for validation and test (three-way split-sample method). Secondly, since it is a single train-test (or train-validation-test) experiment, these methods estimate of generalization error will be misleading if the split has been

inappropriate. The limitations of the holdout can be overcome with the cross-validation family of methods at the expense of more computations. Figure 4.4 below shows the differences between the hold-out and the three-way split sample methods used for early stopping.

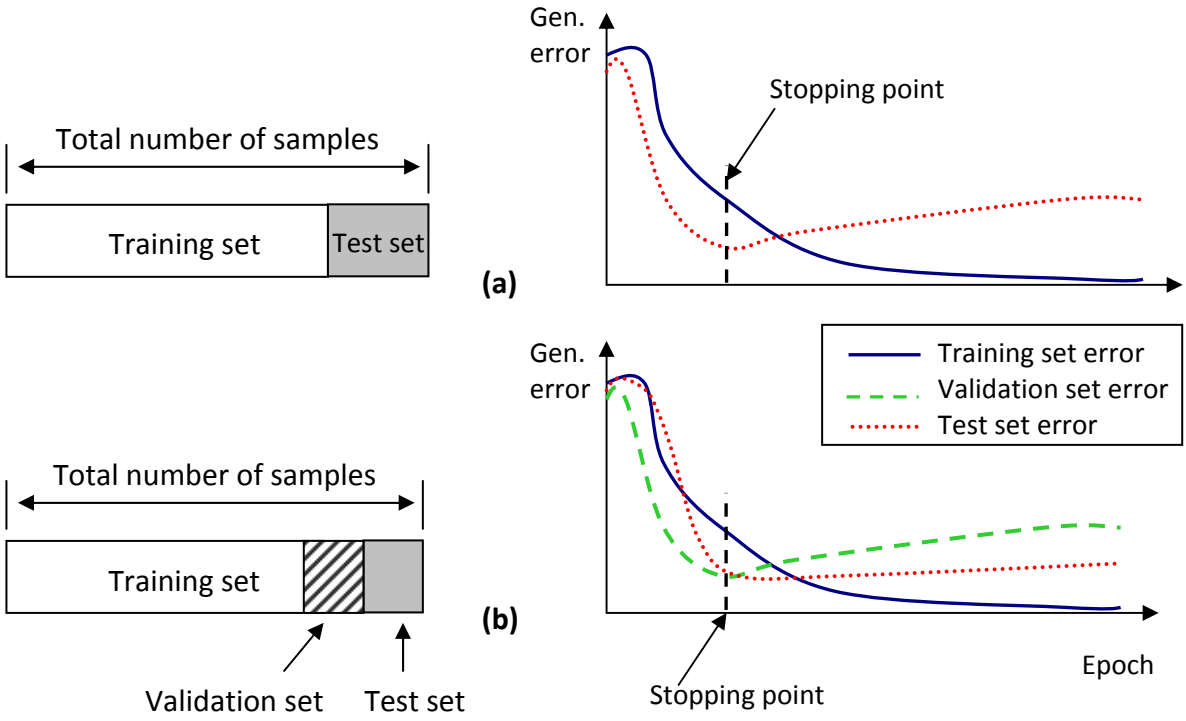

Figure 4.4: Splitting data samples and generalization error plots for (a) the hold-out method and (b) the three-way split sample method.

4.6.2 Cross-validation

The cross-validation method can be used either to evaluate the performance of a given model or to perform model selection. In the first case, the generalization error of the model is estimated for several architectures in order to choose the number of hidden units, for example. In the model selection case, the cross-validation is used to choose among several models with different inputs subsets the one that has the smallest estimated generalization error. The model selection problem is coped with in chapter 5. The two most popular cross-validation methods used in neural networks are K-fold and Leave-One-Out (LOO). Diagram of figure 4.5 allows make distinction between them.

- K-fold cross-validation

The K-fold divides all the samples in K groups of samples (called folds) of approximately equal sizes. For each of K experiments, the prediction function is learned using K – 1 folds, and the remaining fold is used for testing. The generalization error is estimated as the average error over the K test folds. A common choice for K-fold cross-validation is K=10 which we will adopt throughout all our implementations.

- Leave-One-Out cross-validation

For a sample set of size N, when K is chosen to be equal to N the K-fold methods is called Leave-One-Out (LOO). The Leave-One-Out (or LOO) is a simple cross-validation. Each learning set is created by taking all the samples except one, the test set being the sample left out. Thus, for N samples, we have N different learning sets and N different tests set. This cross-validation procedure does not waste much data as only one sample is removed from the learning set. As usual, the generalization error is estimated as the average error over the N test folds.

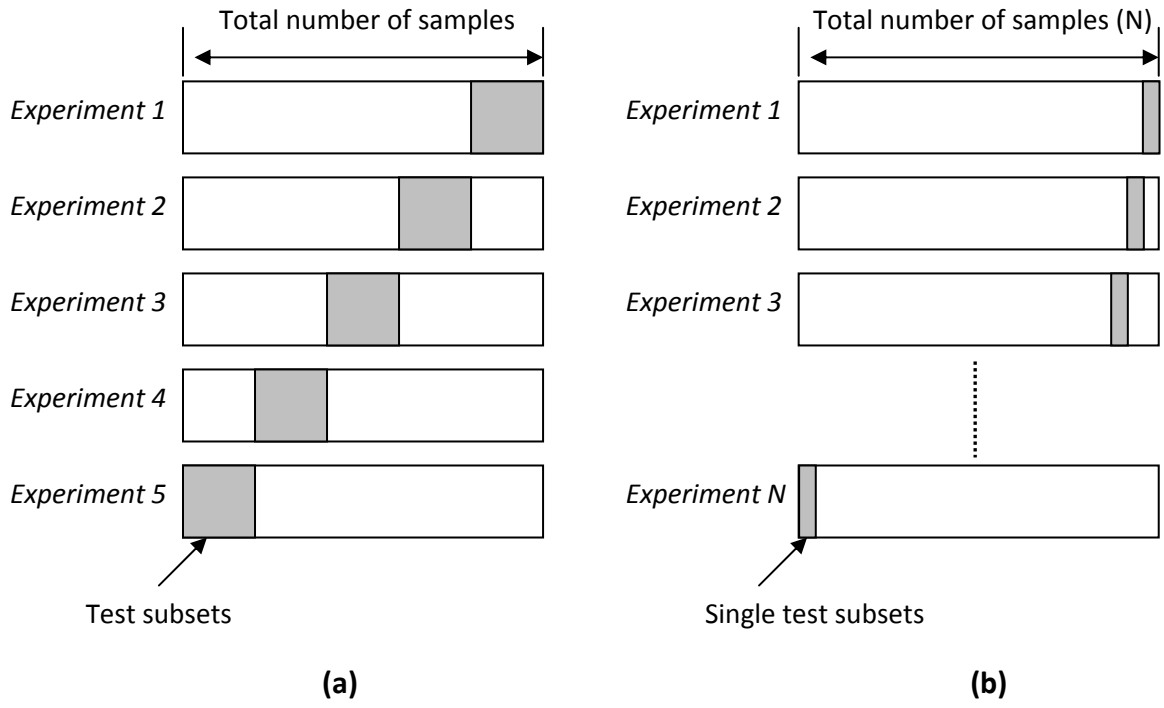

Figure 4.5: cross-validation experiments: (a) K-fold method (example K=5), and (b) LOO method (K=N).

➢ Advantages and shortcomings

The advantage of the K-fold cross-validation is that all the data samples are eventually used for both training and testing. However for limited data samples size, the $\frac{1}{K} \cdot 100\%$ portion set aside for testing can constitute a waste of data for training and give rise to underfitting problem. This can be overcome by leaving out only one sample at each experiment (leave-one-out cross-validation). With a large number of folds the generalization error is very accurately estimated since it is the average of the generalization errors computed over the K test folds. Nevertheless, the number of experiments and therefore computation time is increased.

For choosing subsets of inputs in regression, L. Breiman and P. Spector [70] found 10-fold and 5-fold cross-validation to work better than leave-one-out found to have some subtle deficiencies for model selection [71].

➢ The bias-variance dilemma

The bias-variance dilemma occurs when training is performed on different dataset samples equally representative as in the case of training with cross-validation. The number of folds needed to get good estimation results is the crucial issue while designing cross-validation based learning models. On the one hand, a learning algorithm is said to have a large bias for a given input vector V_i if, when trained on each of the different subset samples, it systematically gives biased predictions to the output vector V_o. On the other hand, a learning algorithm is said to have high variance for a given input vector V_i if it leads to completely different predictions of the output vector V_o on different training subsets.

Consequently, with a large number of folds:

- The estimator is very accurate, that is the bias of the optimal performance error is small,
- The variance of the estimator is large,
- And the computational cost is high due to the large number of experiments.

Inversely, with a small number of folds:

- The bias of the estimator is large,
- The variance of the model is reduced,
- And, the number of experiments and therefore computational cost are reduced.

As a matter of fact, it is commonly agreed that an unbiased estimator seems to be very interesting though it does not guarantee the lowest possible generalization error. The optimal model performance is attained when the correct tradeoff is found between the bias and the variance of the estimator. More details on the bias-variance dilemma can be found in [72].

4.6.3 Bayesian Regularization

Bayesian regularized artificial neural networks (BRANNs) are more robust for improving generalization than standard back-propagation nets and can reduce or eliminate the need for lengthy cross-validation. The BRANNs provide solutions to a number of problems such as choice of model, robustness of model, size of validation subset, and optimization of network architecture. This involves modifying the standard cost function, which is normally chosen to be the sum of squares of the network errors on the training set.

D. MacKay [73] has proposed a Bayesian framework which can be directly applied to the neural network learning problem. It also allows estimating the effective number of parameters (weights) actually used by the model to solve a particular problem. The cost function $f(C)$ is then expanded to search not only for the minimal error, but the optimal combination of sum squared errors and sum squared network parameters as shown in equation 4-2.

$$f(C) = \alpha \cdot E_e + \beta \cdot E_p \qquad (4\text{-}2)$$

where E_e is the sum of the squared errors, E_p is the sum of squared weights and biases, α and β are Bayesian hyper-parameters.

Bayesian regularized artificial neural networks are difficult to overtrain, since evidence procedures provide an objective Bayesian criterion for stopping training. They are also difficult to overfit, because the BRANN calculates and trains on a number of effective network parameters or weights, effectively turning off those that are not relevant. This effective number is usually considerably smaller than the number of weights in a standard fully connected back-propagation neural net [64].

4.6.4 Structural Risk Minimization Method

Structural risk minimization (SRM) introduced by Vapnik and Chervonekis in 1974 [57] is an inductive principle for model selection used for learning from finite training data sets. It describes a general model of capacity control and provides a trade-off between hypothesis space complexity (the VC dimension of approximating functions) and the quality of fitting the training data (empirical error). The procedure is outlined below.

1. Using a priori knowledge of the domain, choose a class of functions, such as polynomials of degree *n*, neural networks having *n* hidden layer neurons, a set of splines with *n* nodes or fuzzy logic models having *n* rules.
2. Divide the class of functions into a hierarchy of nested subsets in order of increasing complexity. For example, polynomials of increasing degree.
3. Perform empirical risk minimization on each subset (this is essentially parameter selection).
4. Select the model in the series whose sum of empirical risk and VC confidence is minimal.

4.7 Conclusion

The background principles of learning in machines have been exposed at the beginning of this chapter. A wide range of learning algorithms has been reviewed, each with its strengths and weaknesses. Without diving into the specifics of individual algorithms, interest has been focused on artificial neural networks (ANNs) with supervised learning that will be used for further research work in next chapters. The choice of ANNs, particularly the multilayer feed-

forward perceptrons, has been justified. Thereby, universality and powerful capabilities of this model has been explained.

The backpropagation learning algorithm is the standard method applied to the multilayer perceptrons. It is used to improve the generalization performances of machine learning systems; however it may suffer from overfitting and overtraining problems. These subtle inconveniences occur depending on how available data is representative, how amount of data is available and how much the problem to model is complex. Early stopping, cross-validation, Bayesian regularization, and structural risk minimization methods are among number of potential solutions to get better models' accuracy. They have been elucidated and adopted later in chapters 5 and 6.

CHAPTER

5

Chapter 5

Feature Selection for Image Quality Assessment

5.1 Introduction

In learning systems, the data provided may correspond to measurements performed on a physical system or to feature information gathered from observations on a phenomenon. Usually all features are not equally informative: some of them may be noisy, meaningless, redundant, correlated or irrelevant for the learning task [74]. As the name suggests, feature selection is the iterative process of selecting a subset of only pertinent features by removing interfering ones. Training after feature selection would be easier and better estimation performances would be achieved which leads to better model interpretability [75]. Nevertheless, it is not always obvious to find the subset of features at iteration (i) that contains the best subset of features at iteration (i+1). A feature can be informative when included within a feature subset but noisy or meaningless when associated with another different one.

Indeed, with the proliferation of large datasets within many domains and the development of new applications in recent years, new research topics of pressing needs have emerged. Hence, feature selection has been an active field of research for decades and is widely applied to data mining [76 -78] and ultra-high dimensional data [79].

The issue related to combining and selecting low-level features for image quality assessment is laid out in this chapter. The general process of feature selection is explained in section 5.2 and the different approaches are outlined in section 5.3. The section 5.4 gives more specific

details on how to select features using a neural network based approach that will be utilized for further experiments in this chapter. Experiments will also be carried out using another class of methods based on the polynomial regression. To do, an automated modelling software KXEN (Knowledge eXtraction ENgine) presented in section 5.5 has been employed in order to validate the results obtained with the neural networks. The experimental setup is described in section 5.6. It is worth noting that the same data samples have been used for low-level image quality indicators selection in both groups of experiments to fairly compare their results. The chapter ends with a discussion of the outcomes and a conclusion.

5.2 Feature Selection Process

As depicted in figure 5.1, feature selection is generally performed through two basic phases: feature selection, and model learning / performance evaluation. The feature selection phase is based on three interleaved functionalities: feature subset search, feature subset evaluation and stop criterion. Since the feature selection problem consists at identifying the optimal set of pertinent variables that lead to better recognition rate in the case of classification and to better prediction quality in the case of regression, a feature selection system would respond to the following questions:

1) How to measure the pertinence of a variable ? Here an evaluation criterion should be defined that measures the importance and the utility of a candidate feature or a candidate set of features.
2) How to generate the optimal subset ? Here a search strategy should be performed to look for a candidate set containing a subset of the pertinent features defined via the evaluation criterion.
3) Which optimality criterion to use ? The answer here consists at describing the condition(s) that when met, the search procedure is terminated [80].

Once a subset of features assumed to be the best is selected, the second phase consists at training the data based on the selected features to learn the model according to the task (prediction or classification). The performance achieved is computed on a test set of the pertinent features that has not been already used in the feature selection phase.

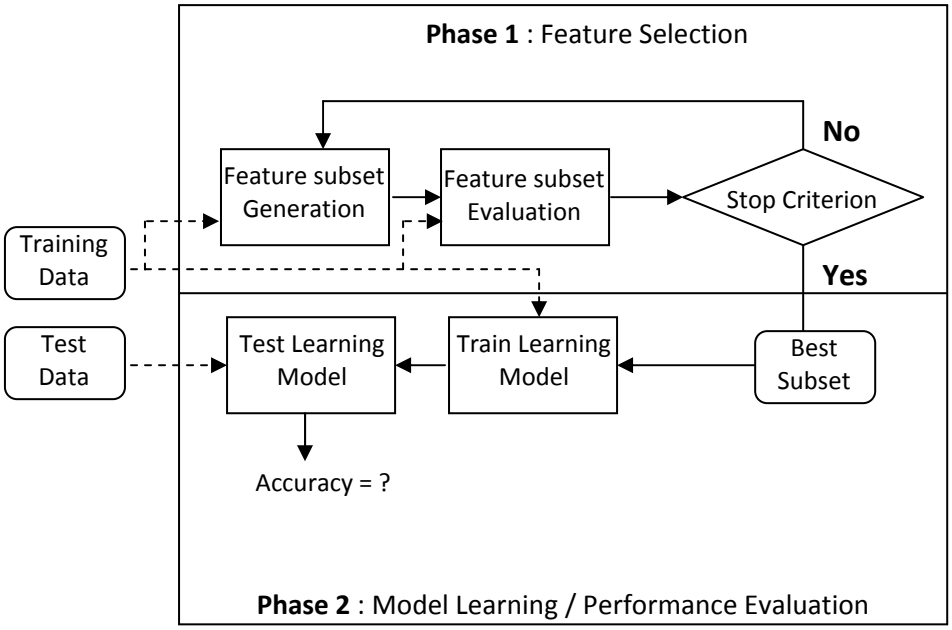

Figure 5.1: A generic scheme for the feature selection process [81].

5.2.1 Feature subset evaluation

The search problem is driven by a certain feature evaluation criterion which is used to assess the utility and the relevance of each feature subset. A variable is pertinent when "its suppression causes the deterioration of the learning system performances, i.e., the discrimination capabilities for classification or the prediction quality for regression" [81]. Furthermore, several feature evaluation criteria, based either on statistical grounds or heuristics, have been proposed for measuring the importance of variable subsets, comparing them and selecting one of them. For classification applications, classical criteria use probabilistic distances or entropy measures, often replaced in practice by simple interclass distance measures. Some statistical methods such as the error measures are also used for regression [82] and classification [83].

The feature evaluation criterion is sometimes assumed to be monotonous which means that the learning system's performance always improves each time a feature assumed to be non relevant is discarded from the subset. However, most of the existing and widely used feature evaluation criteria do not verify the monotonicity hypothesis.

5.2.2 Feature subset search

The feature selection problem can be stated as the search for an optimal number of features out of the total number of available ones without degrading the performance of the resulting learning model when using either set of features [84]. For a non monotonous evaluation criterion and for an initial set of *N* variables, there are 2^N-1 possible combinations of variables where "2" represents the two alternatives of whether to select a variable or not. In fact, comparison of feature subsets amounts to a NP-hard problem which becomes computationally unfeasible, even for moderate initial features set size.

A possible alternative is the use of the Branch and Bound exploration algorithm [76, 85] that allows reducing the search of optimal features subset for originally or assumed monotonous criteria. Nevertheless, the use of this technique is limited due to the non monotonicity of evaluation criteria in most cases. Sub-optimal search techniques that follow one of the following sequential search scenarios are then employed:

- Start with an empty set of features and iteratively add features that optimize the evaluation criterion to the already selected feature set (Forward Methods).
- Start with the full set of features and iteratively remove the less relevant features according to the evaluation criterion from the selected candidate set (Backward Methods).
- Start with an empty set and alternate forward and backward steps (Stepwise Methods). The number of forward and backtracking steps can be fixed before hand as in the "plus *l* take away *r*" algorithm [86] which alternates *l* forward selections and *r* backward deletions, or can be instead dynamically tuned using Floating Sequential Search Methods [87]. Approaches based on Genetic Algorithms [88, 89] constitute an attractive alternative to heuristic tree search methods.

It is worth noting that in the process of generating the candidate set and evaluating it, the feature selection algorithm may use the information from the training data, current selected features, target learning model, and given prior knowledge to guide the search and evaluation procedures [90]. Based on the search strategy, some features in the candidate set may be discarded or added to the selected feature set according to their relevance. Some methods rely only on the data for computing relevant variables and do not take into consideration the model which will then be used for processing these data after the selection step. They may rely on hypothesis about the data distribution (parametric methods) or not (non parametric methods). Other methods take into account simultaneously the model and the data; this is usually the case for neural network based variable selection.

5.2.3 Stop criterion

Let be given a feature evaluation criterion and a search procedure, the optimal number of pertinent features is not known a priori. The stopping criterion is needed to determine whether the current set of selected features is good enough when no more variable is significantly more informative. If it is, the feature selection algorithm will return the set of selected features, otherwise, it iterates until the stopping criterion is met. In most applications, an estimate of the generalization error is computed for the successive variable subsets generated by the search algorithm. The selected features will be those giving the best performances, i.e., minimizing the generalization error. This latter is obtained using a validation set to optimize the learning parameters in the data training step or the cross-validation strategy. When using a neural network, retraining the model is necessary for each subset.

5.2.4 Learning model and performance evaluation

Once a set of features is selected, it can be used to filter the training and test data for model fitting or classification. The performance achieved by a particular learning model on the test data can also be used for evaluating the effectiveness of the feature selection algorithm for that learning model [75].

Similarly to machine learning techniques presented in the previous chapter, the feature selection algorithms can be supervised, unsupervised or semi-supervised depending on whether the training data is labelled, unlabelled or partially labelled, respectively.

In the evaluation process, a supervised feature selection algorithm determines features' pertinence by evaluating their discrimination or predication power accuracy. Without labels, an unsupervised feature selection algorithm may exploit data variance or data distribution in its evaluation of features' relevance. A semi- supervised feature selection algorithm uses a small amount of labelled data as additional information to improve unsupervised feature selection [75].

5.3 Approaches to Feature Selection

Depending on how and when the evaluation of selected features is made, different approaches can be distinguished which broadly fall into three categories: filter, wrapper and embedded models.

- **Filters**: features evaluation and selection of filter model is performed independently of the learning algorithm that will use the selected variables. It relies essentially on analyzing the general characteristics of the data set which allows the algorithms to have very simple structure. Consequently, a straightforward and fast search strategy, such as forward selection or backward suppression is usually adequate. The performance assessment is defined using statistical tests. Compared to wrappers and embedded methods, this class of models has shown a relative robustness against overfitting, but may fail to select the most useful features.
- **Wrappers**: unlike the filter models that do not involve the learning model in the evaluation and selection procedure, the wrapper models take into consideration the influence of the selected variables on the performances of the learning algorithm that is used in the evaluation step as a measure of feature subsets usefulness. Wrappers adopt the exhaustive strategy that explores the space of all features subsets. They are more robust than filters in finding the most useful variables that

result in higher learning performance for a particular learning model. Unfortunately, they suffer from overfitting.

- **Embedded methods**: Algorithms of the embedded model incorporate feature selection as a part of the learning process (classification or regression), and variables are selected based on their usefulness in such a way to optimize the objective function of the learning model. Compared to wrappers, embedded models also employ cross-validation in the assessment step. Therefore, they are usually less computationally expensive and less prone to overfitting. Figure 5.2 shows up the differences between the three aforementioned approaches.

Artificial neural networks (ANNs) are a typical example of wrapper models capable to find the optimal global solution for any nonlinear and convex problem in a very efficient way.

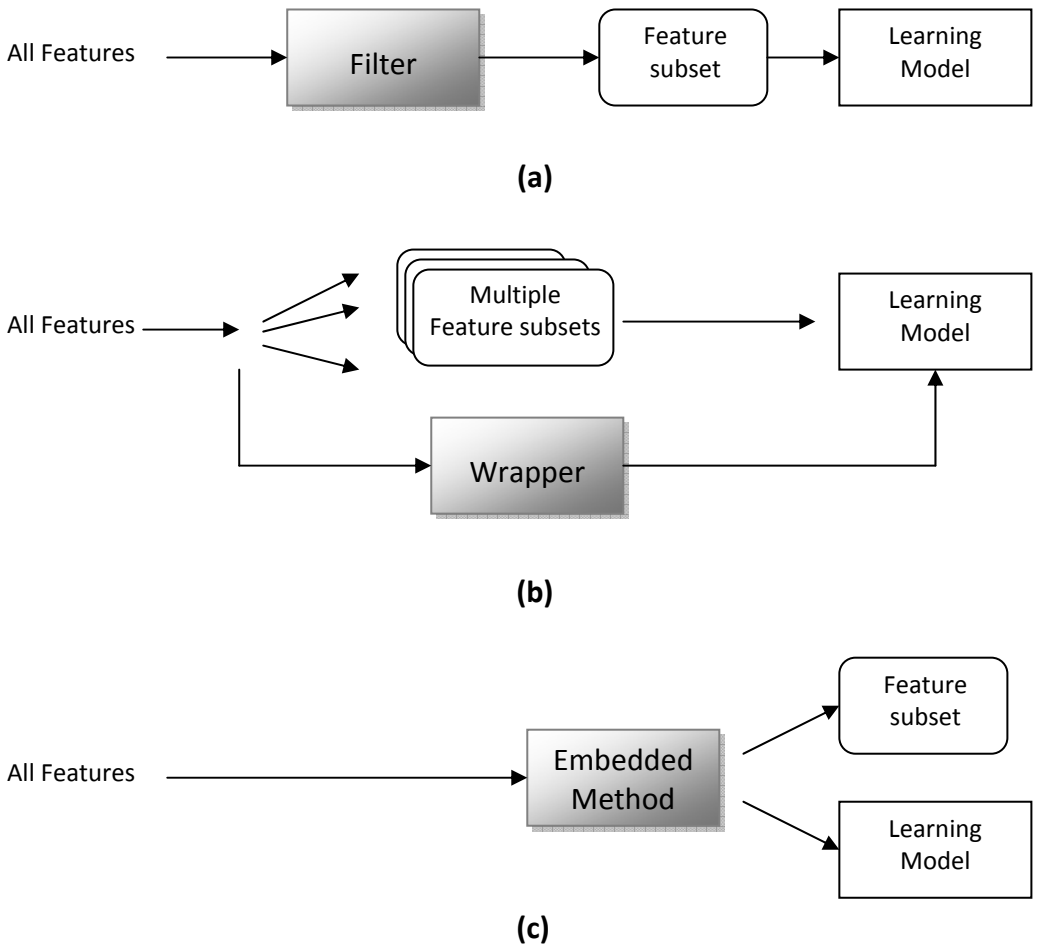

Figure 5.2: Flow charts of (a) filter, (b) wrapper, and (c) embedded methods for feature selection.

5.4 Feature Selection with Neural Networks

As explained above, feature selection with artificial neural networks is model dependent; the selection and the processing of the data are performed simultaneously. Unlike the model independent selection where the parameters optimization is tuned by the user, the neural network based selection is part of the training process and the model selection criterion is globally optimized depending on the model parameters. Here, the number of variables is directly related to the architecture of the neural network model as well as the complexity of the function being learned [80]. The main aspect used when deriving such algorithms is the nonlinearity of the ANN models. Hence, the ANN based approach avoids the linearity assumption on the input-output data that constrains the other existing model independent techniques.

Furthermore, no sophisticated search strategies are needed; the backward deletion technique is generally employed. The main concept is to make the ANN converge to a local minimum using the entire initial set of features, and then to make variable selection by suppressing the less pertinent ones. At each selection step, the neural network is retrained with the remaining features until the model parameters are globally optimized. The optimization relies on the importance of the architecture connexions measured by means of the first and second derivatives of the ANN cost function.

With the perspective to find the low-level image indicators that are the most relevant in the process of image quality evaluation, we apply the neural network based feature selection in order to estimate, at each experiment, the Mean Opinion Score (MOS) or the Difference Mean opinion Score (DMOS) based on the image features fed as input vectors. The experimental setup is described in detail in section 5.6. Another group of MOS/DMOS estimators is built using the KXEN software presented hereafter and results of both the two methods are reported in section 5.7 with concluding comments.

5.5 KXEN Statistical Modelling Software

A lot of statistical models can be used to build an estimator of the Mean Opinion Score (MOS) based on the image features. We chose to stick with a simple and efficient approach that consists in a regularized multi-dimensional polynomial estimator, implemented by KXEN (Knowledge eXtraction ENgine).

KXEN is composed of several modules that allow to automatically performing:

- Robust regression,
- Data standardization and recoding,
- Handling of missing data by inserting inferred values, etc.

The KXEN package K2R (Kxen Robust Regression) is the package responsible for regression modelling we have exploited in our work.

Like other statistical tools (for example: STATISTICA, SPSS, Weka, R, Rapid Miner, etc.), this model is an implementation of the statistical learning theory. Therefore, the K2R component is presented to be a very powerful tool for regression modelling. On the one hand, it can safely handle huge datasets (very high numbers of over one million input-output samples) in a highly automated manner. This makes it very easy to use since the amount of data preparation necessary before modeling is reduced which accelerates the modeling process. On the other hand, the prediction accuracy of the algorithm is high. It provides indicators and graphs to ensure that the quality and robustness of trained models can be easily assessed. It also gives indication of which attributes either contain no relevant information or are redundant with other attributes (http://kxen.com).

To optimize the models parameters and hyper-parameters, the KXEN algorithm relies on the Structural Risk Minimization (SRM) theory of V. N. Vapnik and A. J. Chervonenkis [57] introduced in the previous chapter. The most interesting feature of SRM is that the theoretical basis is much more generalized, and it does not make many of the assumptions that constrain some other approaches (normality, linearity, independence, etc.).

This kind of statistical modelling can supply an accurate estimation of the target variable (the MOS in our experiments) and can also estimate the contribution of the low-level image indicators that are basically ranked according to their weight in the polynomial expression. This allows to take into account eventual correlations between the features and cases where individual features are not correlated to the target variable, but their (non linear) combination carries valuable information.

5.6 Experimental setup

The objective of the experiments presented in this section is to determine the most relevant descriptors for the subjective image quality assessment. The first step is to extract the image features that may convey useful information to the process of image quality appreciation. In the second step, the models parameters are estimated on different sets of labelled images (with known MOS or DMOS) and with different input vectors. Thirdly, the feature selection is performed in order to choose the image features vector that gives best accuracy of the learning model [91, 92].

5.6.1 Step 1: image statistical features extraction

The results of the predictive performance study of eighteen objective image quality measures - carried out using the Friedman analysis in chapter 3 - show that the class of metrics based on statistical features such as the structural similarity index (SSIM) and its variants is an interesting class of measures in many cases of the investigation. Another motivation is that the statistical features are not computationally expensive; these are the arguments for which we have chosen to derive the statistical low-level indicators from images pairs (reference / test) after being converted into a grayscale representation as input vectors for the learning models developed in this chapter. The output vector contains the subjective quality ratings of the test images that consists either on the mean opinion scores (MOS) or the difference mean opinion scores (DMOS) depending on the data available with the image quality databases as shown below.

Table 5.1 summarizes the information related to the three image databases used for feature selection testing, namely the LIVE database (release 2), the TID and the CSIQ databases. However, there exist other subjective image quality databases that we could not use in the present experiments because of their limited size as can be seen in bold from Table 5.2. A thorough description of the six image quality databases is provided in chapter 2.

	LIVE	**TID**	**CSIQ**
Publisher	University of Texas at Austin	Tampere Univ. of Tech., Finland	Oklahoma State Univ., USA
Reference	[26]	[31]	[32]
Year of publication	2005	2008	2010
Number of reference images	29	25	30
Number of labelled images	982	1700	866
Number of observers	29	838	35
Subjective scores	DMOS	MOS	DMOS
Score's range	0 .. 100	0 .. 9	0 .. 9

Table 5.1: Summary of the image databases used in the modelling experiments.

	TOYAMA	**IVC**	**A57**
Publisher	University of Toyama, Japan	University of Nantes, France	Cornell Univ., Ithaca, NY, USA
Reference	[25]	[28]	[29]
Year of publication	2000	2006	2007
Number of reference images	14	10	3
Number of labelled images	**196**	**160**	**54**
Number of observers	16	15	7
Subjective scores	MOS	MOS	MOS
Score's range	1 .. 5	1 .. 5	0 .. 1

Table 5.2: Summary of the image databases excluded from the modelling experiments.

The extracted statistical descriptors including the mean, variance, covariance and mean squared error are listed and described in Table 5.3.

Sixty two (62) different input vectors have been generated by combining the statistical image attributes listed above. A battery of tests has been driven with the different input vectors using two distinct learning models including our neural network implementation and the polynomial regression using the KXEN algorithm.

Image indicator	Description
mu1, mu2	Mean pixels values of the reference and test images, respectively.
sigma1_sq, sigma2_sq	Variance of pixels values of the reference and test images pixels, respectively.
sigma12	Covariance between pixels values of the reference and test images.
mse_error	Mean squared error between reference and test images.

Table 5.3: List of statistical image indicators employed for feature selection.

5.6.2 Step2: the estimation models

At first stage, we use a multilayer perceptron (MLP) based model. The MLP is a one hidden layer neural network. The inputs are a features vector extracted from the pairs of images (reference and test). The output is a numerical real value that represents the MOS (or DMOS) corresponding to the test images. The transfer functions are the tangent sigmoid and the linear function for the hidden and the output layer respectively. The Levenberg-Marquardt learning algorithm is employed.

The feature selection phase is crucial; it allows finding the most informative combination of image features that will be used to develop objective image quality evaluation measures. It is useful to use more than one single estimation model in order to get the most reliable feature selection results.

The KXEN software has been used for his efficiency and ease of manipulation. It is important to understand that the variables are encoded using a non-linear procedure inside the Kxen Consistent Coder (K2C) component before being used by the polynomial regressor as depicted in figure 5.3.

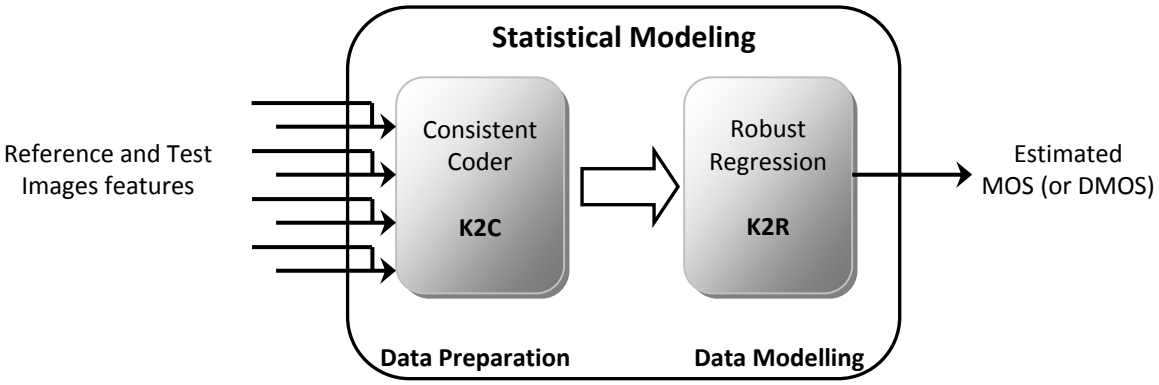

Figure 5.3: The estimation model involved by the KXEN algorithm.

5.6.3 Step3: the feature selection procedure

All presented results in next section are obtained by a 10-fold cross-validation procedure (described in section 4.6 of chapter 4). Because this step is essential for further investigations, it is necessary to be sure to avoid both the overfitting and overtraining. To do so, the image features set is randomly split into three parts for each combination: the training subset with 80% of images, the validation subset with 10% of the images, and a test subset with the remaining 10%.

On each of the ten runs, the models parameters are estimated on the training subset, and generalization (hyper-parameters) is controlled by observing the error on the validation one. The estimation models are trained 30 times and the minimum validation error is researched. When the validation error starts to go up, learning is stopped (see early stopping in section 4.6 in the previous chapter). Finally, the model that leads to this minimum error is validated and then applied on the test subset. It is worth noting that training and validation subsets are normalized according to the mean and the standard deviation of the training base values. Data processing is clarified on figure 5.4.

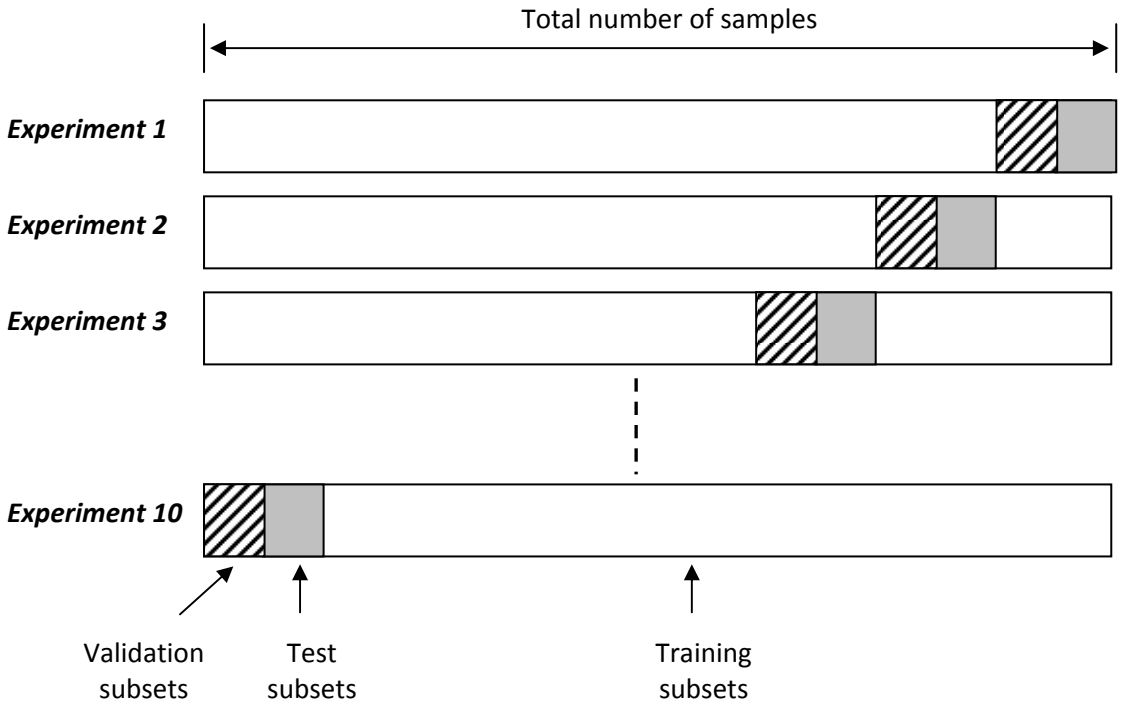

Figure 5.4: Data processing: the 10-fold cross-validation and three-way split-sample methods are applied to avoid overfitting and overtraining of the models.

Because the feature selection is performed to further develop image quality metrics using the selected image features, the performance measure is naturally the predictive performance measures including the Pearson's correlation coefficient (PCC) as indication on the correlation of the estimated MOS (or DMOS) and the real ones. The root mean squared error and the mean absolute error to gauge accuracy, and the Spearman's and Kendall's rank order correlation coefficients to check the monotonicity. These five measures are averaged on the ten distinct test subsets coming from cross-validation.

The model estimation procedure can be outlined as follows:

0. Data refers here to the image features vectors (inputs/outputs) corresponding to different combinations

1. For each combination, divide randomly the available data into 10 folds.

2. For each fold, split the data into training, validation and test subsets.

3. Select architecture and training parameters

4. For each fold, train the model using the training subsets

5. Evaluate the model using the validation subsets

6. Compute the validation error periodically during training

7. Stop training when the validation error starts to go up

8. Repeat steps 3 through 7 using different architectures and training parameters

9. Select the best model and train it using data from the training and validation subsets

10. Assess the final model using the test subsets.

5.7 Results and Discussion

As explained above, we have made multiple combinations of the five previously described statistical descriptors which resulted in 62 input vectors to the estimation models. After training, many combinations have been discarded and only 19 ones (C1 to C19) have been retained. They are listed in Table 5.4.

Performance measures corresponding to correlation, accuracy and monotonicity computed on the LIVE, TID and CSIQ image quality databases of the best estimation models for the 19 combinations are presented in Tables 5.5 to 5.9. For each database, results are obtained with the neural network based systems and with the KXEN modelling tool. The framed values correspond to the maximum performance achieved.

	Statistical image indicators combinations
C1	mu1, mu2
C2	mu1, mu2, mu1_sq, mu2_sq
C3	mu1, mu2, mu1_mu2
C4	mu1, mu2, sigma1_sq, sigma2_sq
C5	mu1, mu2, mu1_sq, mu2_sq, sigma1_sq, sigma2_sq
C6	mu1, mu2, mu1_mu2, sigma1_sq, sigma2_sq
C7	mu1, mu2, mu1_sq, mu2_sq, mu1_mu2, sigma1_sq, sigma2_sq
C8	mu1, mu2, sigma1_sq, sigma2_sq, sigma12
C9	mu1_mu2, sigma1_sq, sigma2_sq, sigma12
C10	mu1, mu2, mu1_sq, mu2_sq, sigma1_sq, sigma2_sq, sigma12
C11	mu1, mu2, mu1_mu2, sigma1_sq, sigma2_sq, sigma12
C12	mu1_sq, mu2_sq, mu1_mu2, sigma1_sq, sigma2_sq, sigma12
C13	mu1, mu2, mu1_sq, mu2_sq, mu1_mu2, sigma1_sq, sigma2_sq, sigma12
C14	mu1, mu2, sigma1_sq, sigma2_sq, sigma12, mse_error
C15	mu1_mu2, sigma1_sq, sigma2_sq, sigma12, mse_error
C16	mu1, mu2, mu1_sq, mu2_sq, sigma1_sq, sigma2_sq, sigma12, mse_error
C17	mu1, mu2, mu1_mu2, sigma1_sq, sigma2_sq, sigma12, mse_error
C18	mu1_sq, mu2_sq, mu1_mu2, sigma1_sq, sigma2_sq, sigma12, mse_error
C19	sigma1_sq, sigma2_sq, sigma12

Table 5.4: List of the retained statistical image indicators combinations.

	LIVE		TID		CSIQ	
	ANN	**KXEN**	**ANN**	**KXEN**	**ANN**	**KXEN**
C1	0,359	0,917	0,060	0,281	0,164	0,753
C2	0,383	0,917	0,087	0,281	0,154	0,753
C3	0,367	0,986	0,053	0,236	0,140	0,658
C4	0,663	0,917	0,387	0,281	0,325	0,753
C5	0,612	0,917	0,369	0,281	0,460	0,753
C6	0,630	0,986	0,394	0,236	0,373	0,658
C7	0,555	0,917	0,393	0,236	0,398	0,658
C8	0,954	0,986	0,798	0,281	0,894	0,894
C9	0,936	0,418	0,790	0,586	0,876	0,876
C10	0,947	0,917	0,785	0,299	0,889	0,753
C11	0,945	0,986	0,774	0,236	0,887	0,658
C12	0,949	0,418	0,770	0,586	0,893	0,893
C13	0,950	0,348	0,785	0,236	0,895	0,658
C14	0,944	0,906	0,769	0,287	0,890	0,690
C15	0,942	0,819	0,753	0,684	0,884	0,884
C16	0,948	0,906	0,774	0,287	0,895	0,690
C17	0,946	0,930	0,765	0,243	0,883	0,696
C18	0,951	0,468	0,786	0,399	0,887	0,887
C19	0,954	0,986	0,804	0,684	0,901	0,753

Table 5.5: Pearson's correlation coefficient (PCC) for the 19 features' combinations using the neural network approach and the KXEN software on the LIVE, TID and CSIQ databases.

	LIVE		TID		CSIQ	
	ANN	**KXEN**	**ANN**	**ANN**	**KXEN**	**ANN**
C1	28,942	31,130	1,340	1,447	0,258	0,272
C2	28,818	31,130	1,336	1,447	0,261	0,272
C3	28,923	30,708	1,339	1,461	0,261	0,277
C4	23,156	31,130	1,237	1,447	0,248	0,272
C5	24,271	31,130	1,244	1,447	0,229	0,272
C6	24,087	30,708	1,236	1,461	0,238	0,277
C7	26,109	30,708	1,235	1,461	0,239	0,277
C8	9,694	31,130	0,806	1,447	0,113	0,116
C9	10,890	34,829	0,814	1,636	0,125	0,125
C10	9,974	31,130	0,828	1,447	0,119	0,272
C11	10,064	36,522	0,842	1,461	0,119	0,277
C12	9,732	34,829	0,853	1,636	0,117	0,117
C13	9,304	30,708	0,823	1,461	0,117	0,277
C14	10,112	31,075	0,861	1,449	0,119	0,275
C15	10,427	33,776	0,880	1,657	0,123	0,123
C16	9,744	31,075	0,849	1,449	0,116	0,275
C17	9,926	30,694	0,867	1,464	0,123	0,274
C18	9,537	33,776	0,818	1,657	0,120	0,120
C19	9,649	30,708	0,794	1,447	0,116	0,272

Table 5.6: Root Mean Squared Error (RMSE) for the 19 features' combinations using the neural network approach and the KXEN software on the LIVE, TID and CSIQ databases.

	LIVE		TID		CSIQ	
	ANN	**KXEN**	**ANN**	**ANN**	**KXEN**	**ANN**
C1	24,493	25,821	1,084	1,161	0,218	0,224
C2	24,343	25,821	1,081	1,161	0,222	0,224
C3	24,601	25,717	1,089	1,176	0,222	0,227
C4	18,572	25,821	1,008	1,161	0,206	0,224
C5	19,553	25,821	1,016	1,161	0,185	0,224
C6	19,465	25,717	1,008	1,176	0,196	0,227
C7	21,578	25,717	1,004	1,176	0,196	0,227
C8	7,479	25,821	0,647	1,161	0,086	0,090
C9	8,410	28,967	0,655	1,292	0,096	0,096
C10	7,735	25,821	0,655	1,161	0,091	0,224
C11	7,804	25,717	0,673	1,176	0,091	0,227
C12	7,445	28,967	0,684	1,292	0,090	0,090
C13	7,390	30,267	0,662	1,176	0,090	0,227
C14	7,817	25,854	0,689	1,162	0,091	0,227
C15	8,251	28,306	0,706	1,307	0,094	0,094
C16	7,558	25,854	0,681	1,162	0,090	0,227
C17	7,681	25,790	0,682	1,179	0,095	0,227
C18	7,389	28,306	0,655	1,307	0,093	0,093
C19	7,120	25,717	0,633	1,161	0,089	0,224

Table 5.7: Mean Absolute Error (MAE) for the 19 features' combinations using the neural network approach and the KXEN software on the LIVE, TID and CSIQ databases.

	LIVE		TID		CSIQ	
	ANN	**KXEN**	**ANN**	**ANN**	**KXEN**	**ANN**
C1	0,289	0,668	0,075	0,220	0,219	0,844
C2	0,278	0,668	0,093	0,220	0,191	0,844
C3	0,277	0,648	0,086	0,278	0,171	0,787
C4	0,691	0,668	0,285	0,220	0,353	0,844
C5	0,679	0,668	0,274	0,220	0,514	0,844
C6	0,682	0,648	0,288	0,278	0,407	0,787
C7	0,591	0,648	0,294	0,278	0,432	0,787
C8	0,944	0,668	0,788	0,220	0,906	0,844
C9	0,934	0,478	0,788	0,547	0,894	0,894
C10	0,943	0,293	0,784	0,221	0,894	0,844
C11	0,942	0,648	0,776	0,278	0,897	0,787
C12	0,945	0,478	0,779	0,547	0,902	0,902
C13	0,948	0,648	0,794	0,278	0,905	0,787
C14	0,942	0,596	0,777	0,219	0,903	0,802
C15	0,940	0,554	0,776	0,714	0,897	0,897
C16	0,945	0,596	0,775	0,219	0,901	0,802
C17	0,942	0,576	0,770	0,279	0,897	0,794
C18	0,945	0,299	0,781	0,220	0,898	0,898
C19	0,946	0,668	0,793	0,714	0,905	0,905

Table 5.8: Spearman's Rank Order Correlation Coefficient (SROCC) for the 19 features' combinations using the ANN approach and the KXEN software on the LIVE, TID and CSIQ databases.

	LIVE		TID		CSIQ	
	ANN	**KXEN**	**ANN**	**ANN**	**KXEN**	**ANN**
C1	0,210	0,478	0,051	0,138	0,147	0,576
C2	0,202	0,478	0,065	0,138	0,131	0,576
C3	0,202	0,462	0,059	0,181	0,117	0,537
C4	0,522	0,478	0,192	0,138	0,249	0,576
C5	0,514	0,478	0,186	0,138	0,373	0,576
C6	0,514	0,462	0,194	0,181	0,294	0,537
C7	0,442	0,462	0,199	0,181	0,311	0,537
C8	0,804	0,478	0,595	0,138	0,735	0,576
C9	0,782	0,329	0,594	0,366	0,714	0,714
C10	0,800	0,204	0,593	0,138	0,718	0,576
C11	0,800	0,462	0,586	0,181	0,721	0,537
C12	0,804	0,329	0,586	0,366	0,727	0,727
C13	0,812	0,462	0,602	0,181	0,730	0,537
C14	0,799	0,425	0,583	0,144	0,729	0,554
C15	0,794	0,386	0,580	0,476	0,717	0,717
C16	0,805	0,425	0,585	0,144	0,725	0,554
C17	0,797	0,404	0,579	0,187	0,720	0,542
C18	0,804	0,208	0,590	0,138	0,723	0,723
C19	0,809	0,478	0,602	0,476	0,729	0,729

Table 5.9: Kendall's Rank Order Correlation Coefficient (KROCC) for the 19 features' combinations using the ANN approach and the KXEN software on the LIVE, TID and CSIQ databases.

5.8 Conclusion

Feature selection results presented in tables 5.5 through 5.9 show that combinations C8, C13 and C19 achieve better performances. This means that these input vectors provide good results in terms of correlation, accuracy and monotonicity on the LIVE, TID and CSIQ image quality databases. The C19 Combining the variance of both the reference and test images (σ_x^2, σ_y^2) and their covariance σ_{xy} generates the best results amongst the combinations mentioned above. It will be used to develop two reduced reference image quality measures based on multi-layer perceptron model, and are the object of the next chapter.

CHAPTER

6

Chapter 6

Reduced Reference Multilayer Perceptron based Metrics

6.1 Introduction

Objective image quality assessment is a complex problem due to the subjective nature of the human visual perception. Hence, image quality has been defined in several ways in the literature leading to different approaches that have been suggested to predict the human perception appreciation. In this chapter, we regard the image quality evaluation task in terms of a mapping function between qualitative and quantitative attributes. Thereby, developing a quality measure is to find a fitting function to the subjective human appreciation scores.

Feed-forward artificial neural networks are universal approximators; they are capable to fit any linear or non linear function. Designing and implementing multilayer perceptron is then a possible solution to the problem of objective image quality metrics design and development.

In this chapter, based on results of feature selection performed in chapter 5, we propose two reduced reference variance / covariance based image quality metrics using a neural network approach. The main contribution is that the proposed metrics are computationally simple and do not require the entire reference image to be calculated while still giving interesting performances when compared to eighteen full reference image quality metrics

available in the literature. The first metric called ECF (Error based Cost Function) is more accurate than most of its counterparts while the second measure called CCF (Correlation based Cost Function) outperforms the metrics to which it has been compared in terms of correlation and monotonicity. A comparative study has been conducted over three image quality databases including the LIVE (second release), TID2008 and CSIQ introduced in chapter 2.

Typically, image quality databases are constructed with labelled images where subjective quality scores are collected using thorough and expensive psycho-visual experiments. These individual scores are then transformed into either the Mean Opinion Score (MOS) or the Difference Mean Opinion Score (DMOS) with the reference images also rated.

The rest of the chapter is organized as follows; section 6.2 presents the details of the method and implementation. Explanations are provided about the design of the two different cost functions that have been used to develop the two image quality metrics. Section 6.3 includes the performance benchmarking methodology of image quality measures that is applied for evaluating the proposed metrics (ECF and CCF). A comparative study is then supplied where the superiority of the predictive capabilities of the proposed measures is shown. Finally, concluding remarks are made in section 6.4.

6.2 The Reduced Reference Image Quality Metrics Design

The contribution consists of using two different cost functions for the MLP and by performing a complete performance evaluation of the proposed image quality metrics [93, 94]. The block diagram of Fig. 6.1 shows the model components. We attempt to learn useful information about human image quality evaluation by mimicking the human reasoning when an observer proceeds to give his/her appreciation about the quality of an image.

The objective of implementing an MLP based model in the present work is threefold. The first one is for feature selection application that helps to determine the most relevant descriptors for the subjective image quality assessment among a set of feature vectors. This step has been laid out in chapter 5. The second goal is to map the subjective image quality

scores onto a mathematical formula that represents the objective image quality metric. The multilayer perceptron cost function is, at third stage, customized in such a way to optimize the image quality algorithms performance criteria including correlation, monotonicity and accuracy [94].

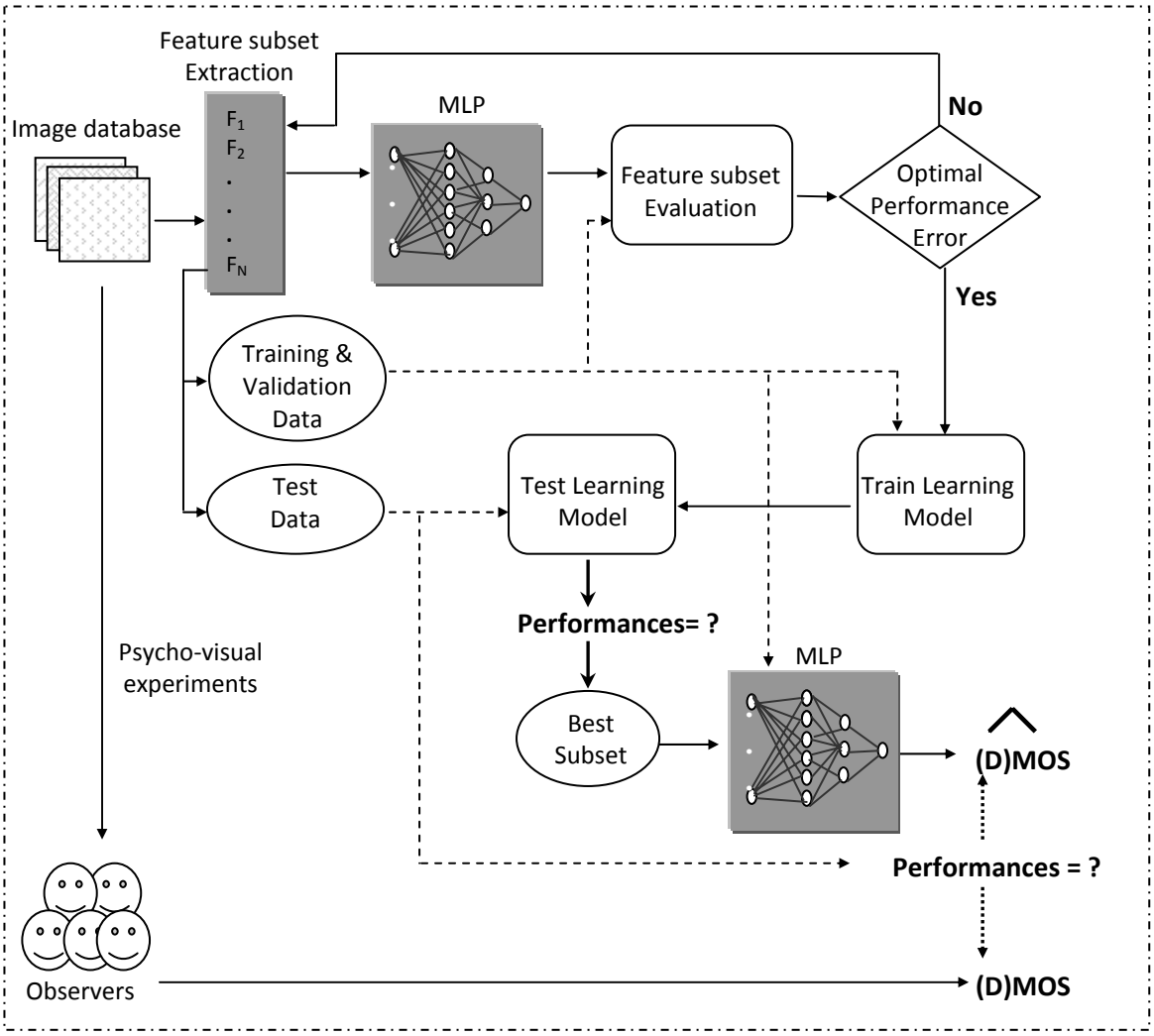

Figure 6.1: Systems' block diagram.

6.2.1 Standard error based cost function

In the first stage, we have used a fully connected multilayer perceptron with two hidden layers of 10 and 4 cells, respectively. The inputs are the features vector extracted from the pairs of images (reference and test) and the outputs are the known MOS (or DMOS) values associated with the test images. The transfer functions are the tangent sigmoid of the hidden layers and the linear function for the output layer. The neural network weights and biases are updated according to the Levenberg-Marquardt learning algorithm using the Bayesian regularization process of David MacKay [73].

The Bayesian regularization is adopted to improve the generalization capabilities of MLP networks. This consists of estimating the effective number of network parameters (weights and biases) actually needed to solve the problem at hand. The cost function is then expanded to search not only for the minimal error, but the optimal combination of sum squared errors and sum squared network parameters as shown in equation (6-1).

$$f(C) = \alpha \cdot E_e + \beta \cdot E_p \qquad (6\text{-}1)$$

where E_e is the sum of the squared errors, E_p is the sum of squared weights and biases, α and β are Bayesian hyper-parameters. Using this cost function causes the network to have smaller weights and biases, thus forcing the network's response to be smoother and less likely to overfit [95].

6.2.2 Correlation based cost function

In the second stage, we have modified the standard error based cost function to a correlation based one. This is due to the fact that we have noticed a significant and substantial decrease in both the root mean square error (RMSE) and the mean absolute error (MAE) between real MOS (DMOS) values and the computed neural network outputs as can be seen from results highlighted in bold in tables 6.1 and 6.2. Therefore, we propose to develop a cost function such as to increase the correlation between subjective and objective quality scores. We have used a one hidden layer feed-forward neural network with 10 cells. The transfer functions are the tangent sigmoid and the linear function for the hidden and the output layer, respectively. The Levenberg-Marquardt learning algorithm with early stopping paradigm is employed.

The proposed cost function is inspired from the one presented by Englisch *et al.* in [96] and can be written as:

$$r^2 = LCC^2 = \left(\frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \right)^2 \qquad (6\text{-}2)$$

r^2 is called coefficient of determination which is simply the square of the linear correlation coefficients between the observed quality values (vector X) and the predicted ones (vector Y). It is a convex and differentiable function whose values range from 0 to 1.

The backpropagation algorithm minimizes the cost function; hence the absolute value of the computed output of the MLP is then subtracted from 1 since we aim to maximize the correlation between calculated and desired outputs as given by equation (6-3) below:

$$\hat{Y} = 1 - |Y| \qquad (6\text{-}3)$$

6.3 Proposed image quality measures' performance

This is concerned with the results of a comparative performance investigation in terms of the correlation, accuracy and monotonicity. It is similar to the one conducted in chapter 3.

6.3.1 Accuracy

We first show the results related to the accuracy performance criterion of image quality measures. We have used a standard error based cost function described in section 6.2.1 to develop the ECF quality metric. It can be noticed from values highlighted in bold in tables 6.1 and 6.2 that both the root mean square error (RMSE) and the mean absolute error (MAE) have been drastically minimized with our proposed ECF metric over the three quality databases.

Error values are considerably inferior to the ones obtained for the 18 full reference state-of-the-art metrics (MSE, …, RFSIM) as well as for the newly developed reduced reference CCF metric before applying the logistic function described by eq. (3-7) and eq. (3-8) of chapter 3. After the logistic function has been carried out, a few exceptions of the RMSE and MAE results of MS-SSIM, VIF, VIFP, MSSIM and RFSIM on the LIVE database - and sometimes over TID and CSIQ datasets - are lower (so better) than our proposed metric (ECF) as underlined in tables 6.1 and 6.2.

	Root Mean Square Error – RMSE Before nonlinear regression				Root Mean Square Error – RMSE After nonlinear regression		
	LIVE	**TID**	**CSIQ**		**LIVE**	**TID**	**CSIQ**
MSE	1313,1	1332,7	394,9		36,695	1,341	0,233
PSNR	29796,5	22,9	30,3		73,589	1,129	0,172
SNR	29796,5	16,7	24,1		73,582	1,158	0,175
WSNR	29796,5	26,2	32,2		73,591	1,181	0,178
NQM	29796,5	18,8	27,3		73,585	1,098	0,182
UQI	49,3	4,0	0,6		10,834	1,089	0,147
SSIM	49,2	3,9	0,6		10,475	1,069	0,152
MS-SSIM	49,0	3,8	0,7		8,690	0,844	0,116
VIF	49,4	4,1	0,6		6,980	0,936	0,098
VIFP	49,4	4,2	0,5		6,956	0,997	0,113
IFC	60,4	21,7	9,8		10,400	1,038	0,213
M-SVD	24,5	65,5	38,8		14,507	1,320	0,218
PSNRHVS	45466,6	21,2	28,4		82,608	1,103	0,162
PSNRHVSM	45466,6	24,3	32,6		82,610	1,126	0,165
VSNR	61,1	1589,7	29,4		9,427	1,323	0,176
MSSIM	49,1	3,9	0,6		8,358	0,953	0,129
R-SVD	49,3	4,4	0,2		19,669	1,263	0,226
RFSIM	49,4	4,1	0,6		8,717	0,822	0,105
ECF	**9,595**	**0,790**	**0,119**		**9,587**	**0,789**	**0,114**
CCF	1073,349	15,287	0,517		36,449	1,341	0,263

Table 6.1: Root mean square error of ECF and CCF from the 18 image quality metrics over the LIVE, TID and CSIC databases.

Results also show that the nonlinear regression applied to the ECF metric has slightly changed the values compared to those obtained before the nonlinear mapping; which is not the case for all other image quality models considered in the present comparative study. This demonstrates that the nonlinearities of the human visual system have been taken into consideration and have been fitted using the multilayer perceptron (MLP) based model. This demonstrates that the MLP has learnt to reduce the error between the calculated output and the desired one which is the MOS (or DMOS) value of test images. This has led to considerably improve the metric's accuracy performance.

	Mean Absolute Error – MAE Before nonlinear regression				Mean Absolute Error – MAE After nonlinear regression		
	LIVE	**TID**	**CSIQ**		**LIVE**	**TID**	**CSIQ**
MSE	369,1	449,8	220,1		29,174	1,087	0,199
PSNR	13573,3	22,3	29,3		69,341	0,866	0,135
SNR	13575,5	15,8	22,7		69,333	0,894	0,139
WSNR	13573,9	24,3	30,0		69,347	0,899	0,137
NQM	13574,8	17,2	24,6		69,338	0,808	0,146
UQI	38,3	3,8	0,5		7,862	0,868	0,112
SSIM	38,2	3,7	0,5		7,579	0,847	0,116
MS-SSIM	38,1	3,6	0,6		6,608	0,616	0,088
VIF	38,4	3,9	0,5		5,366	0,680	0,074
VIFP	38,5	4,0	0,5		4,925	0,779	0,091
IFC	52,4	10,5	7,1		7,984	0,825	0,182
M-SVD	17,6	33,0	26,6		10,451	1,064	0,188
PSNRHVS	20699,4	20,1	26,8		76,747	0,780	0,124
PSNRHVSM	20699,7	22,7	30,2		76,752	0,800	0,127
VSNR	48,8	59,7	26,6		6,582	1,074	0,131
MSSIM	38,1	3,7	0,6		6,306	0,719	0,099
R-SVD	38,2	4,1	0,2		13,871	0,985	0,182
RFSIM	38,3	3,9	0,5		6,519	0,584	0,077
ECF	**6,936**	**0,631**	**0,092**		**6,964**	**0,629**	**0,086**
CCF	950,415	13,360	0,371		30,313	1,087	0,223

Table 6.2: mean absolute error of ECF and CCF from the 18 image quality metrics over the LIVE, TID and CSIC databases.

6.3.2 Correlation

Since correlation is also an important criterion for the predictive performance in image quality, we have proposed a neural network approach which is capable to improve its performance. Therefore, we have customized the cost function elucidated in section 6.2.2.

As highlighted in Table 6.3, the Pearson's correlation coefficient between our CCF (Correlation based Cost Function) quality metric and the subjective quality scores in the form of MOS (or DMOS) is visibly superior to the correlation results obtained for the other studied metrics. This includes the reduced reference metric (ECF) proposed in the present work for which the cost function is based on error as well as the 18 full reference quality measures except VIF and VIFP as underlined in the table below. This is true for the three image quality databases including LIVE (second release), TID2008 and CSIQ.

The results show that our CCF quality metric outperforms all other considered metrics in the present comparative study in terms of correlation but also in terms of monotonicity as explained in the next section.

	Pearson's correlation coeff. – PCC Before nonlinear regression				**Pearson's correlation coeff. – PCC After nonlinear regression**		
	LIVE	**TID**	**CSIQ**		**LIVE**	**TID**	**CSIQ**
MSE	0,478	0,042	0,462		0,589	0,042	0,461
PSNR	0,629	0,410	0,756		0,629	0,541	0,756
SNR	0,629	0,389	0,745		0,629	0,505	0,745
WSNR	0,629	0,399	0,736		0,629	0,505	0,736
NQM	0,629	0,499	0,719		0,629	0,576	0,719
UQI	0,936	0,554	0,829		0,938	0,585	0,829
SSIM	0,861	0,459	0,765		0,942	0,604	0,815
MS-SSIM	0,758	0,327	0,771		0,961	0,777	0,898
VIF	0,963	0,707	0,922		0,975	0,717	0,928
VIFP	0,945	0,588	0,881		0,975	0,669	0,903
IFC	0,685	0,213	0,582		0,944	0,634	0,582
M-SVD	0,778	0,180	0,556		0,886	0,180	0,556
PSNRHVS	0,629	0,481	0,785		0,629	0,570	0,785
PSNRHVSM	0,629	0,481	0,778		0,629	0,544	0,778
VSNR	0,723	0,043	0,743		0,954	0,247	0,743
MSSIM	0,841	0,457	0,804		0,964	0,704	0,871
R-SVD	0,777	0,307	0,476		0,778	0,337	0,509
RFSIM	0,960	0,757	0,913		0,960	0,790	0,916
ECF	0,952	0,808	0,892		0,952	0,809	0,900
CCF	**0,998**	**0,970**	**0,997**		**0,986**	**0,970**	**0,988**

Table 6.3: Pearson's correlation coefficient of ECF and CCF for 18 image quality metrics using the LIVE, TID and CSIC databases.

6.3.3 Monotonicity

Since the Spearman's and Kendall's rank order correlation coefficients are related to the Pearson's correlation coefficient, we are interested to assess the effect of the correlation based cost function on the monotonicity results. As shown in Table 6.4, the Spearman's Rank Order Correlation Coefficient (SROCC) and Kendall's Rank Order Correlation Coefficient (KROCC) values for our CCF quality metric are higher than those obtained for the 19 metrics (even VIF and VIFP) against which the predictive performance comparison is performed.

It is worth noting that the SROCC and KROCC values are not subject to the nonlinear mapping as suggested by the Video Quality Expert Group in [36].

	Spearman's Rank Order Correlation Coefficient - SROCC				**Kendall's Rank Order Correlation Coefficient – KROCC**		
	LIVE	**TID**	**CSIQ**		**LIVE**	**TID**	**CSIQ**
MSE	0,936	0,513	0,806		0,791	0,374	0,607
PSNR	0,936	0,513	0,806		0,791	0,374	0,607
SNR	0,931	0,483	0,799		0,783	0,346	0,600
WSNR	0,956	0,450	0,773		0,833	0,363	0,599
NQM	0,955	0,577	0,740		0,832	0,426	0,563
UQI	0,945	0,552	0,807		0,806	0,402	0,615
SSIM	0,954	0,585	0,837		0,821	0,424	0,632
MS-SSIM	0,972	0,789	0,914		0,862	0,605	0,739
VIF	0,976	0,702	0,919		0,866	0,552	0,753
VIFP	0,975	0,619	0,879		0,863	0,470	0,695
IFC	0,956	0,541	0,748		0,818	0,406	0,574
M-SVD	0,907	0,588	0,768		0,750	0,440	0,583
PSNRHVS	0,958	0,549	0,830		0,831	0,442	0,653
PSNRHVSM	0,964	0,518	0,822		0,844	0,417	0,653
VSNR	0,958	0,650	0,810		0,822	0,493	0,624
MSSIM	0,973	0,708	0,883		0,862	0,524	0,695
R-SVD	0,795	0,345	0,557		0,623	0,238	0,383
RFSIM	0,969	0,803	0,929		0,854	0,628	0,764
ECF	0,955	0,798	0,909		0,816	0,603	0,727
CCF	**0,989**	**0,984**	**0,998**		**0,992**	**0,990**	**0,998**

Table 6.4: Spearman and Kendall's correlation coefficients of ECF and CCF against those of the 18 image quality metrics using the LIVE, TID and CSIC databases.

6.4 Interpretation and Concluding Comments

In this chapter we have proposed two reduced reference variance/covariance based image quality metrics using a multilayer perceptron approach. Two different neural network cost functions have been employed, evaluated and compared in this chapter. In the first stage, we have used a standard error based cost function and the results obtained were very encouraging especially for the quality metric's accuracy. In addition, both the root mean square error and the mean absolute error values have been drastically minimized. Then, we have proposed a correlation based cost function and demonstrated that correlation between the subjective image quality scores and the objective measure values outperforms all the

quality measures presented in the literature. Pearson, Spearman and Kendall correlation coefficients are consequently maximized.

These results clearly show that the neural network based method is an efficient way to directly optimize performance features related to image quality problem. We suggest that neural network cost function has an important role for the fitting function outputs. We aim to design a multi-objective neural network where the two cost functions used in the present work are combined in order to simultaneously minimize the errors while still maximizing correlation. In so doing, we expect to obtain a quality metric that allows to considerably gaining in accuracy, correlation as well as monotonicity. We also plan to compare our metrics performance to the state-of-the art reduced reference measures [94].

It is worth noting that we did use Bayesian regularization method for the error based cost function metric but not for the correlation based one. We suggest employing it for future work and to make tests on different neural network architectures including the Support Vector Machine where the generalization capabilities are more controlled.

CHAPTER

7

Chapter 7

General Conclusion

7.1 Summary

In this thesis, the problem of image quality assessment has been tackled from different points of view. Initially, we have supplied a comprehensive overview of the most commonly used state-of-the-art objective image quality estimators. After that, we have studied the subjective quality measures and supplied a thorough and complete description of number of subjectively rated image quality databases that are publicly. We have also discussed the advantages and limitations of the objective and subjective quality assessment approaches, concluding that we cannot substitute for one approach or the other. Indeed, our interest must be simultaneously focused on these two ways of design since the ability of an objective model to predict the perceptual quality of an image should be established regarding the subjective ratings provided by human observers.

In this perspective, the predictive performance benchmarking problem of objective image quality measures has been addressed. We have examined eighteen standard, well-known and widely used full-reference image quality models. Their ability to automatically predict the human quality appreciation has been evaluated and compared over six image quality databases constructed using different standard protocols. Because it is robust and free-distribution, the Friedman statistical method has been suggested to analyse if there are significant differences between the performances of the investigated objective models on

the one hand, and between the behaviour of the existing subjective databases on the other hand. Performance evaluation focused on three standard measures: correlation, accuracy and monotonicity.

One of the major findings of the present work is that the performance of the objective image quality estimators as it is actually applied depends on the structure of the image database the algorithm is being assessed on. Structure here refers to the proportion of very good or very poor quality images. This questions the reliability of the standard performance evaluation methodology.

Another outcome of applying the Friedman test in the 95% confidence interval reveals that the quality algorithms that belong to the natural statistics class, such as MS-SSIM, MSSIM, VIF and VIFP introduced in chapter 2, perform better than their counterparts in terms of correlation, accuracy and monotonicity. However, the predictive performances of the most traditional image quality models that rely on simple mathematical formulas, - such as the MSE and PSNR also presented in chapter 2 -, fall into the same statistical range as many other more sophisticated measures inspired from the human visual system. The severe criticism that has been levelled to the classical quality algorithms in numerous research papers was based on counter-examples that show situations (i.e. image examples) where the PSNR or the MSE give objective scores that do not reflect the visual reality. In this way, whenever there is a new image quality measure, the metric developers tend to highlight the subjectivity of their own metric.

Nevertheless, there exists *theoretically* at least one counter-example to each quality measure that demonstrates its inconvenience since there does not exist one universal and 100% perfect image quality metric. In the case of PSNR and MSE, their mathematical formulas are simple, and it is easy to find many counter examples (i.e., image examples) where these models are shown to be poor quality predictors. Finding counter-examples for the existing image quality models just needs more efforts due to their relatively more complicated formulas.

Furthermore, the necessity to supplement objective measurements with subjective rating scores has led us to develop learning systems where the model is fed information about the image and its parameters are estimated according to the images subjective quality values.

7.2 Open Issues

Our work on the image quality assessment problem has lead to the production of several results, although some open problems remain.

The reliability of the image quality metrics performance evaluation tools based on correlation, accuracy and monotonicity is questioned since it was found that standard benchmarking methods are impacted by the structure of image quality databases. Looking for new processes for validating objective quality models and standardized image quality assessment solutions to make the perceptual quality measurement evaluation procedure more efficient are open research issues.

Concerning the reliability and robustness of the supervised learning algorithms developed and/or used in this research work, it is important to mention that there is no single learning algorithm that works best on all supervised learning problems as stipulated by the “No Free Lunch” theorem. It is then important to experiment the various options of machine learning approaches to find the best model that characterizes the problem at hand (feature selection integrated to image quality evaluation).

7.3 Perspectives

Like any research work, ours is far from to be complete. Some work still to be done in the same direction.

- ✓ The development of a taxonomic scheme for classification of the state-of-the-art objective image quality assessment algorithms will be of big help to sort out the different theoretic foundations upon which these estimators are built as well as the fundamental ideas behind their development.

- ✓ It would be interesting to develop an image quality measure that combines the predictive capabilities of the metrics proposed in chapter 6. The Error based Cost Function (ECF) quality model gives very attractive performances in terms of accuracy whereas the Correlation based Cost Function (CCF) quality metric works very well in terms of correlation and monotonicity. Designing a reduced-reference image quality algorithm that outperforms the existing both full-reference and reduced reference models is possible by customizing the perceptron's cost function.
- ✓ An image database independent objective measure would be a good improvement of ECF and CCF proposed metrics. Learning on one database and testing on a different one would be a possible alternative, as well as using second order image features.
- ✓ The learning models based on neural networks are sometimes criticised as they are considered as black box systems that do not allow understanding the mechanisms leading to learning the task being modelled. Nevertheless, it is possible to extract the neural network's formula knowing its architecture, the number of hidden layers and of units (input, hidden and output), the transfer functions at each layer, as well as the biases and weights values after estimation step.
- ✓ The neural networks have been selected to design image quality algorithms in this thesis to simulate the behaviour of the human visual system when giving judgment on the perceptual quality of a visual content. Other machine learning classes that do not owe their origins to the study of the human brain but that possess the ability to acquire and store knowledge can also be further experimented. Approaches based on statistical principles such as the Support Vector Machines would be interesting to customize for the current applications.
- ✓ Elaborate methods for Verification and Validation (V&V) of machine learning based systems are beginning to emerge. It would be thoughtful to apply the compatible V&V methods in order to validate the quality assessment metrics stemming from machine learning oriented models.

REFERENCES

References

[1] Z. Wang and A. C. Bovik, "Modern Image Quality Assessment", In *syntheses lectures on Image, Video and Multimedia Processing,* Morgan & Claypool Publishers, Mar. 2006.

[2] U. Engelke and H.-J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey", *3rd European Conference on Next Generation Internet Networks: Design Eng. Heterogeneity*, pp. 190–197*,* 2007.

[3] Z. Wang, A. C. Bovik, and E. P. Simoncelli, "Structural Approaches to image quality assessment," In *Handbook of Image and Video Processing* (Al Bovik, ed.), 2nd edition, Academic Press, June 2005.

[4] T. Mitsa and K. L. Varkur, "Evaluation of Contrast Sensitivity Functions for the Formulation of Quality Measures Incorporated in Halftoning Algorithms", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, MN, vol. 5, pp. 301-304, 1993.

[5] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective Picture Quality Scale (PQS) for image coding", *IEEE Trans. Communications*, vol. 46, no. 9, pp. 1215–1225, Sept. 1998.

[6] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image Quality Assessment Based on a Degradation Model", *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[8] Z. Wang and A. C. Bovik, "A Universal Image Quality Index", *IEEE Signal Processing Letters*, Vol. 9 No. 3, pp. 81-84, Mar. 2002.

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment", In *Proc. 37th IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 09-12, 2003.

[10] D. M. Rouse and S.S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM", In *Proc. SPIE*, vol. 6806, Human Vision and Electronic Imaging, 2008.

[11] Z. Wang and E.P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain", *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, pp. 573-576, Philadelphia, PA, Mar. 2005.

[12] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics", *IEEE Transactions on Image processing*, vol. 14, no. 12, pp. 2117-2128, Dec. 2005.

[13] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality", *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp.430 – 444, Feb. 2006.

[14] A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "An SVD-Based Grayscale Image Quality Measure for Local and Global Assessment", *IEEE Transactions on image processing*, vol. 15, no. 2, Feb. 2006.

[15] A. Mansouri, A. M. Aznaveh, F. Torkamani-Azar, and J. A. Jahanshahi, "Image Quality Assessment Using the Singular Value Decomposition Theorem*", Optical Review,* vol. 16, no. 2, pp. 49-53, 2009.

[16] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images", *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284-2298, 2007.

[17] Lin Zhang, Lei Zhang, and X. Mou, "RFSIM: a feature based image quality assessment metric using Riesz transforms", In *Proc. ICIP 2010*, Hong Kong, 26-29 Sept. 2010.

[18] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS", *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, Scottsdale, USA, 4 p, 2006.

[19] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions", *Proc. of the Third Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM-07*, Scottsdale, Arizona, USA, 4 p, 25-26 Jan. 2007.

[20] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” Geneva, 2002 (available at www.itu.org).

[21] ITU-R Recommendation BT.814-1, “Spécifications et méthodes de réglage de la brillance et du contraste des dispositifs de visualisation”, technical report, International Telecommunication Union, 1994.

[22] ITU-R Recommendation BT.815-1, “Spécification d’un signal de mesure du contraste des dispositifs de visualisation”, technical report, International Telecommunication Union, 1994.

[23] M. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies”, *Proceedings of SPIE*, vol. 5150, no. 3, pp. 573-582, 2003.

[24] A. Lahoulou, A. Bouridane, E. Viennet, M. Haddadi, “Full Reference Image Quality Metrics Performance Evaluation over Image Quality Databases”, *Accepted for publication in The Arabian Journal for Science and Engineering,* DOI: 10.1007/s13369-012-0509-6.

[25] Y. Horita, Y. kawayoke, and Z. M. Parvez Sazzad, “Image Quality Evaluation Database”, http://mict.eng.u-toyama.ac.jp/mictdb.html

[26] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “LIVE Image Quality Assessment Database Release 2”, http://live.ece.utexas.edu/research/quality

[27] H. R. Sheikh, “Image Quality Assessment Using Natural Scene Statistics”, Ph.D. dissertation, University of Texas at Austin, May 2004.

[28] P. Le Callet, and F. Autrusseau, “Subjective quality assessment IRCCyN/IVC database”, 2005, http://www.irccyn.ec-nantes.fr/ivcdb/

[29] A57 image database, 2007, available on: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[30] D. M. Chandler, K. H. S. Lim, and S. S. Hemami, “Effects of Spatial Correlations and Global Precedence on the Visual Fidelity of Distorted Images”, In *Proc. SPIE Human Vision and Electronic Imaging XI*, B. E. Rogowitz, T. N. Pappas, and S. Daly, Eds., San Jose, CA, 2006.

[31] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics”, *Advances of Modern Radio electronics*, vol. 10, pp. 30-45, 2009. www.ponomarenko.info/tid2008.htm

[32] E. C. Larson, and D. M. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy”, *Journal of Electronic Imaging*, vol. 19, no. 1, Mar. 2010. http://vision.okstate.edu/csiq/

[33] VQEG, “Final VQEG report on the validation of objective quality metrics for video quality assessment”, 2000. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/

[34] M. Gaubatz, “Metrix MUX Visual Quality Assessment Package”, 2007, http://foulard.ece.cornell.edu/gaubatz/metrix_mux

[35] S. Péchard, “Qualité d’usage en télévision haute définition : évaluations subjectives et métriques objectives”, Ph.D. dissertation, Ecole Polytechnique de l’Université de Nantes, Oct. 2008.

[36] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment – Phase II”, Rapport technique, Video Quality Experts Group, 2003, http://www.vqeg.org

[37] VQEG, “Multimedia Test Plan 1.19”, Video Quality Experts Group, 2007.

[38] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms”, *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, November 2006.

[39] J.C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, “Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions”, SIAM Journal of Optimization, vol. 9, no. 1, pp. 112-147, 1998.

[40] C.G. Broyden, “A new double-rank minimization algorithm”, Notices of the American Mathematical Society, vol. 16, 1969.

[41] R. Fletcher, “A New Approach to Variable Metric Algorithms”, Computer Journal, vol. 13, pp. 317-322, 1970.

[42] D. Goldfarb, “A Family of Variable-Metric Methods Derived by Variational Means”, Mathematics of Computation, vol. 24, pp. 23-26, 1970.

[43] D.F. Shanno, “Conditioning of Quasi-Newton Methods for Function Minimization”, Mathematics of Computation, vol. 24, pp. 647-656, 1970.

[44] M. Nauge, M.-C. Larabi, and C. Fernandez, “Benchmark de métriques de qualité sur bases de données d’images compressées”, *Conférence sur la Compression et Représentation des Signaux Audiovisuels, CORESA 2010*, Lyon, France, 4p, 26-27 Oct. 2010.

[45] K. Okarma, “Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Assessment”, *Lecture Notes in Computer Science*, vol. 6113, pp. 539-546, 2010.

[46] K. Okarma “Video Quality Assessment Using the Combined Full-Reference Approach”, *Advances in Intelligent and Soft Computing*, vol. 84, pp. 51-58, 2010.

[47] A. Lahoulou, E. Viennet, A. Bouridane, M. Haddadi, “A Complete Statistical Evaluation of State-of-the-art Image Quality Measures”, *The 7th International Workshop on Systems, Signal Processing and their Applications* (*Wosspa 2011*), ENP, Algiers, Algeria, pp. 219-222, 9-11 May 2011.

[48] A. Lahoulou, E. Viennet, A. Bouridane, M. Haddadi, “Technical Report: Full Numerical Results for image quality metrics performance benchmarking”, April 2011. Available on: http://www-l2ti.univ-paris13.fr/~lahoulou/tech_report.html

[49] N. J. Nilsson, “Introduction to Machine Learning”, Stanford: Stanford University, Nov. 1998.

[50] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, New York, 1996.

[51] F. Rosenblatt, “Principles of Neuro-dynamics”, Spartan Books, New York, 1962.

[52] A. Novikoff, “On convergence proofs of perceptrons”, In *Proceedings of the Symposium on the Mathematical Theory of Automata*, vol. 7, pp. 615–622, 1962.

[53] B. Widrow and M. E. Hoff, “Adaptive Switching Circuits”, In *IRE WESCON Convention Record*, part 4, pp. 96-104, 1960.

[54] K. Steinbush, “The learning matrix”, *Cybernetics*, pp. 36-45, 1961.

[55] E. Hunt, J. Marin, and P. Stone, “Experiments in Induction”, New York: Academic Press, 1966.

[56] J. K. Baker, “The DRAGON System - An Overview”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 24-29, 1975.

[57] V. N. Vapnik and A. Ja. Chervonenkis, “Ordered Risk Minimization (I and II), *Automation and Remote Control*, vol. 34, pp. 1226-1235 and 1403-1412, 1974.

[58] M. L. Minsky and S. Papert, “Perceptrons”, Cambridge, MIT Press, MA, 1969.

[59] Y. LeCun, “Learning Process in an asymmetric Threshold Network”, In F. Fogelman-Soulié E. Bienenstock and G. Weisbuch, editors, *Disordered systems and biological organization*, pp. 233-240, Les Houches, France, Springer-Verlag, 1986.

[60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by backpropagating errors”, Nature, vol. 323, pp. 533-536, 1986.

[61] D. E. Rumelhart, J. L. McClelland, and the PDP research group, “Parallel distributed processing: Explorations in the microstructure of cognition - Volume I”, Cambridge, MA: MIT Press, 1986.

[62] J. L. McClelland, D. E. Rumelhart, and the PDP research group, “Parallel distributed processing: Explorations in the microstructure of cognition - Volume II”, Cambridge, MA: MIT Press, 1986.

[63] D. O. Hebb, “The Organization of Behavior: A neuropsychological theory”, Wiley, New York, 1949.

[64] C. Bishop, “Neural Networks for Pattern Recognition”, Oxford University Press, Oxford, London, 1995.

[65] Tyler Lu, “Fundamental Limitations of Semi-Supervised Learning”, Master of Mathematics in Computer Science, Waterloo, Ontario, Canada, 2009.

[66] A. N. Kolmogorov, “On the Representations of Continuous Functions of Many Variables by Superposition of Continuous Functions of One Variable and Addition”, *Dokl. Akad. Nauk USSR*, vol. 214, pp. 953-956, 1957.

[67] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators”, *Neural Networks*, vol. 2, no. 5, pp. 359-366, 1989.

[68] E. J. Hartman, J. D. Keeler, and J. M. Kowalski, “Layered neural networks with Gaussian hidden units as universal approximations”, *Neural Computation*, vol. 2, no. 2, pp. 210-215, 1990.

[69] I. V. Tetko, D. J. Livingstone, and A. I. Luik, “Neural network studies. Comparison of overfitting and overtraining”, *J. Chem. Inf. Comput. Sci.,* vol. 35, pp. 826-833, 1995.

[70] L. Breiman, and P. Spector, “Sub-model selection and evaluation in regression: The X-random case”, *International Statistical Review*, vol. 60, pp. 291-319, 1992.

[71] J. Shao, “Linear model selection by cross-validation”, *Journal of the American Statistical Association*, vol. 88, pp. 486-494, 1993.

[72] S. Geman, E. Bienenstock, R. Doursat, “Neural Networks and the Bias/Variance Dilemma”, *Neural Computation*, vol. 4, no. 1, pp. 1-58, 1992.

[73] D.J.C. MacKay, "Bayesian interpolation", *Neural Computation*, vol. 4, no. 3, pp. 415-447, 1992.

[74] P. Leray and P. Gallinari, “Feature selection with neural networks”, *Behavior metrika (special issue on Analysis of Knowledge Representation in Neural Network Models)*, vol. 26, no. 1, pp. 145-166, 1999.

[75] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", *presented at Journal of Machine Learning Research - Proceedings Track*, pp. 4-13, 2010.

[76] H. Liu and H. Motoda, “Feature Selection for Knowledge Discovery and Data Mining”, Kluwer Academic Publishers, Boston, 1998.

[77] D. Cakmakov and Y. Bennani, “Feature Selection for Pattern Recognition”, Informa Press, Ed. 2002.

[78] I. Guyon, , S. Gunn, M. Nikravesh, and L. Zadeh, ”Feature Extraction, Foundations and Applications”, Editors. Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[79] J. Fan, R. Samworth, and Y. Wu, “Ultrahigh dimensional feature selection: Beyond the linear model”, *Journal of Machine Learning Research*, vol. 10, pp. 2013-2038, 2009.

[80] Y. Bennani, S. Guérif, and E. Viennet, “Réduction des dimensions des données en apprentissage artificiel”, *Revue des Nouvelles Technologies de l'Information (RNTI-A2)*, pp. 135-163, Mar. 2008.

[81] Y. Bennani, “Systèmes d’apprentissage connexionnistes : sélection de variables”, vol. 15 (3-4), Revue d’Intelligence Artiticielle, Paris, France : Hermes Science Publications, 2001).

[82] M.L. Thompson, “Selection of Variables in Multiple Regression. Part I: A Review and Evaluation”, *International Statistical Review*, vol. 46, pp. 1-19, “Selection of Variables in Multiple Regression. Part II”: *Chosen Procedures, Computations and Examples, in International Statistical Review*, vol. 46, pp. 129-146, 1978.

[83] G.J. McLachlan, “Discriminant Analysis and Statistical Pattern Recognition”, Wiley-Interscience publication, 1992.

[84] J. Kittler, “Feature Selection and Extraction”, Chapter 3 in *Handbook of Pattern Recognition and Image Processing,* Eds. Tzay Y. Young, King-Sun Fu, Academic Press, pp. 59-83, 1986.

[85] P.M. Narendra and K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection”, *IEEE Transactions on Computers*, vol. 26, no. 9, pp. 917-922, 1977.

[86] S.D. Stearns, “On Selecting Features for Pattern Classifiers”, In *3rd International Conference on Pattern Recognition*, Coronado, CA, pp. 71-75, 1976.

[87] P. Pudil, J. Novovicova, and S. Blaha, “Statistical Approach to Pattern Recognition: Theory and Practical Solution by Mean of PREDITAS System”, *Kybernetika*, vol. 27, pp. 1-78, 1991.

[88] F. J. Ferri, V. Kadirkamanathan, and J. Kittler, “Feature Subset Search using Genetic Algorithms”, In *IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Essex, 1993.

[89] W. Siedlecki and J. Sklansky, “A Note on Genetic Algorithm for Large-scale Feature Selection”, *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335-347, Nov. 1989.

[90] T. Helleputte and P. Dupont, “Partially supervised feature selection with regularized linear models”, In *ICML*, 2009.

[91] A. Lahouhou, E. Viennet, A. Beghdadi, “Combining and Selecting Indicators for Image Quality Assessment”, In *31st International Conference on Information Technology Interfaces*, Croatia, pp. 261-268, 22-25 June 2009.

[92] A. Lahouhou, E. Viennet, A. Beghdadi, “Selecting Low-level Features for Image Quality Assessment by Statistical Methods”, *Journal of Computing and Information Technology*, Vol. 18, No. 2, pp. 183-189, 2010.

[93] A. Lahoulou, E. Viennet, M. Haddadi, “Variable Selection for Image Quality Assessment using a Neural Network based Approach”, *IEEE European Workshop on Visual Information Processing, (EUVIP 2010),* Paris, France, pp. 45-49, 5-7 July 2010.

[94] A. Lahoulou, E. Viennet, A. Bouridane, M. Haddadi, “Customizing Cost Function for Optimizing Image Quality Measures Performances”, *International Journal of Electronics*, Vol. 99, No. 11, pp. 1533-1546, Nov. 2012.

[95] F.D. Foresee and M.T. Hagan, "Gauss-Newton approximation to Bayesian regularization", *Proceedings of the 1997 International Joint Conference on Neural Networks*, pp. 1930-1935, 1997.

[96] H. Englisch and Y. Hiemstra, “The Correlation as Cost Function in Neural Networks”, *IEEE World Congress on Computational Intelligence*, Orlando, FL, vol.7, pp. 4180 - 4185, 1994.

APPENDICES

A - E

Appendix A: Absolute values of the Pearson's Correlation Coefficient (PCC) after nonlinear regression

	TOY	LIVE	IVC	A57	TID	CSIQ
	All data set					
MSE	0,578	0,589	0,510	0,661	0,042	0,461
PSNR	**0,475**	**0,629**	0,720	0,634	0,541	0,756
SNR	**0,475**	**0,629**	0,652	0,574	0,505	0,745
WSNR	**0,475**	**0,629**	0,862	0,930	0,505	0,736
NQM	**0,475**	**0,629**	0,850	0,802	0,576	0,719
UQI	0,790	0,938	0,832	0,450	0,585	0,829
SSIM	0,841	0,942	0,792	0,419	0,604	0,815
MS-SSIM	0,909	0,961	0,887	0,848	0,777	0,898
VIF	0,933	0,975	0,903	0,621	0,717	0,928
VIFP	0,881	0,975	0,823	0,802	0,669	0,903
IFC	0,528	0,944	0,908	0,577	0,634	0,582
M-SVD	0,818	0,886	0,776	0,703	0,180	0,556
PSNRHVS	**0,475**	**0,629**	0,865	0,874	0,570	0,785
PSNRHVSM	**0,475**	**0,629**	0,891	0,916	0,544	0,778
VSNR	0,900	0,954	0,803	0,945	0,247	0,743
MSSIM	0,931	0,964	0,925	0,787	0,704	0,871
R-SVD	0,069	0,778	0,553	0,062	0,337	0,509
RFSIM	0,831	0,960	0,836	0,845	0,790	0,916
	Noised images set					
MSE	/	0,784	/	0,801	0,108	0,838
PSNR	/	**0,637**	/	0,935	0,909	0,953
SNR	/	**0,637**	/	0,987	0,806	0,925
WSNR	/	**0,637**	/	0,980	0,838	0,923
NQM	/	**0,637**	/	0,922	0,740	0,907
UQI	/	0,955	/	0,823	0,539	0,727
SSIM	/	0,970	/	0,862	0,788	0,895
MS-SSIM	/	0,989	/	0,874	0,790	0,941
VIF	/	0,990	/	0,896	0,880	0,957
VIFP	/	0,993	/	0,843	0,811	0,962
IFC	/	0,968	/	0,549	0,601	0,821
M-SVD	/	0,943	/	0,936	0,907	0,908
PSNRHVS	/	**0,637**	/	0,934	0,909	0,953
PSNRHVSM	/	**0,637**	/	0,921	0,891	0,956
VSNR	/	0,986	/	0,916	0,750	0,910
MSSIM	/	0,982	/	0,973	0,705	0,892
R-SVD	/	0,942	/	0,030	0,645	0,842
RFSIM	/	0,966	/	0,903	0,810	0,940

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG coded images set					
MSE	0,379	0,923	0,741	0,767	0,508	0,785
PSNR	**0,493**	**0,621**	0,739	0,700	0,830	0,891
SNR	**0,493**	**0,621**	0,660	0,557	0,756	0,805
WSNR	**0,493**	**0,621**	0,885	0,979	0,905	0,937
NQM	**0,493**	**0,621**	0,878	0,890	0,897	0,955
UQI	0,860	0,943	0,833	0,960	0,805	0,914
SSIM	0,746	0,970	0,832	0,971	0,888	0,940
MS-SSIM	0,879	0,987	0,925	0,956	0,916	0,981
VIF	0,923	0,990	0,939	0,953	0,918	0,988
VIFP	0,808	0,988	0,867	0,970	0,909	0,983
IFC	0,552	0,959	0,961	0,854	0,865	0,809
M-SVD	0,709	0,934	0,740	0,794	0,909	0,916
PSNRHVS	**0,493**	**0,621**	0,907	0,964	0,917	0,906
PSNRHVSM	**0,493**	**0,621**	0,941	0,959	0,922	0,931
VSNR	0,857	0,983	0,842	0,974	0,882	0,951
MSSIM	0,941	0,985	0,957	0,931	0,905	0,977
R-SVD	0,012	0,961	0,912	0,952	0,573	0,846
RFSIM	0,759	0,979	0,836	0,931	0,898	0,968
	Gaussian blurred images set					
MSE	/	0,728	0,782	0,852	0,657	0,878
PSNR	/	**0,711**	0,891	0,590	0,869	0,908
SNR	/	**0,711**	0,826	0,444	0,823	0,885
WSNR	/	**0,711**	0,935	0,911	0,866	0,958
NQM	/	**0,711**	0,981	0,894	0,823	0,970
UQI	/	0,974	0,942	0,955	0,833	0,949
SSIM	/	0,931	0,910	0,912	0,881	0,900
MS-SSIM	/	0,961	0,955	0,946	0,897	0,964
VIF	/	0,979	0,989	0,950	0,880	0,979
VIFP	/	0,980	0,975	0,919	0,883	0,965
IFC	/	0,976	0,978	0,844	0,844	0,901
M-SVD	/	0,827	0,805	0,866	0,864	0,898
PSNRHVS	/	**0,711**	0,968	0,838	0,889	0,938
PSNRHVSM	/	**0,711**	0,965	0,880	0,882	0,919
VSNR	/	0,964	0,981	0,915	0,873	0,915
MSSIM	/	0,967	0,956	0,957	0,880	0,959
R-SVD	/	0,973	0,944	0,895	0,254	0,870
RFSIM	/	0,955	0,962	0,902	0,889	0,956

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG 2000 coded images set					
MSE	0,820	0,929	0,844	0,775	0,800	0,871
PSNR	**0,458**	**0,658**	0,848	0,796	0,743	0,947
SNR	**0,458**	**0,658**	0,817	0,761	0,671	0,927
WSNR	**0,458**	**0,658**	0,915	0,979	0,818	0,961
NQM	**0,458**	**0,658**	0,859	0,906	0,817	0,969
UQI	0,735	0,925	0,809	0,987	0,774	0,903
SSIM	0,930	0,967	0,863	0,884	0,758	0,922
MS-SSIM	0,951	0,981	0,927	0,916	0,831	0,977
VIF	0,968	0,985	0,936	0,852	0,829	0,978
VIFP	0,950	0,981	0,883	0,943	0,816	0,978
IFC	0,506	0,949	0,935	0,766	0,799	0,825
M-SVD	0,928	0,946	0,930	0,808	0,663	0,937
PSNRHVS	**0,458**	**0,658**	0,906	0,887	0,816	0,966
PSNRHVSM	**0,458**	**0,658**	0,932	0,920	0,828	0,946
VSNR	0,943	0,978	0,862	0,959	0,800	0,932
MSSIM	0,957	0,981	0,934	0,910	0,818	0,974
R-SVD	0,155	0,955	0,785	0,594	0,720	0,919
RFSIM	0,904	0,965	0,830	0,896	0,807	0,968

Appendix B: Root Mean Squared Error (RMSE) after nonlinear regression

	TOY	LIVE	IVC	A57	TID	CSIQ
	All data set					
MSE	1,077	36,695	1,048	0,184	1,341	0,233
PSNR	**7,071**	**73,589**	0,846	0,190	1,129	0,172
SNR	**7,071**	**73,582**	0,924	0,201	1,158	0,175
WSNR	**7,072**	**73,591**	0,617	0,090	1,181	0,178
NQM	**7,071**	**73,585**	0,642	0,147	1,098	0,182
UQI	0,809	10,834	0,675	0,220	1,089	0,147
SSIM	0,714	10,475	0,743	0,223	1,069	0,152
MS-SSIM	0,551	8,690	0,563	0,130	0,844	0,116
VIF	0,477	6,980	0,524	0,193	0,936	0,098
VIFP	0,623	6,956	0,692	0,147	0,997	0,113
IFC	1,121	10,400	0,511	0,201	1,038	0,213
M-SVD	0,759	14,507	0,769	0,175	1,320	0,218
PSNRHVS	**10,843**	**82,608**	0,611	0,119	1,103	0,162
PSNRHVSM	**10,843**	**82,610**	0,554	0,099	1,126	0,165
VSNR	0,576	9,427	0,726	0,080	1,323	0,176
MSSIM	0,483	8,358	0,463	0,152	0,953	0,129
R-SVD	1,317	19,669	1,015	0,245	1,263	0,226
RFSIM	0,734	8,717	0,669	0,132	0,822	0,105
	Noised images set					
MSE	/	25,777	/	0,079	0,645	0,123
PSNR	/	**70,365**	/	0,046	0,255	0,069
SNR	/	**70,358**	/	0,021	0,362	0,086
WSNR	/	**70,370**	/	0,026	0,333	0,087
NQM	/	**70,362**	/	0,051	0,411	0,095
UQI	/	9,836	/	0,075	0,514	0,155
SSIM	/	8,058	/	0,066	0,376	0,101
MS-SSIM	/	4,905	/	0,064	0,374	0,077
VIF	/	4,697	/	0,058	0,291	0,066
VIFP	/	3,776	/	0,071	0,357	0,062
IFC	/	8,358	/	0,110	0,488	0,129
M-SVD	/	11,055	/	0,046	0,257	0,095
PSNRHVS	/	**83,250**	/	0,047	0,255	0,069
PSNRHVSM	/	**83,251**	/	0,051	0,277	0,067
VSNR	/	5,581	/	0,053	0,404	0,094
MSSIM	/	6,310	/	0,030	0,433	0,102
R-SVD	/	11,135	/	0,131	0,467	0,122
RFSIM	/	8,594	/	0,056	0,358	0,077

TOY	LIVE	IVC	A57	TID	CSIQ
JPEG coded images set					
1,221	13,551	0,782	0,165	1,479	0,189
7,060	**76,825**	0,784	0,183	0,949	0,139
7,060	**76,819**	0,874	0,213	1,116	0,181
7,060	**76,827**	0,543	0,052	0,724	0,107
7,060	**76,820**	0,557	0,117	0,752	0,090
0,674	11,702	0,645	0,072	1,010	0,124
0,879	8,601	0,645	0,061	0,785	0,104
0,629	5,743	0,441	0,075	0,681	0,059
0,507	4,845	0,401	0,078	0,674	0,047
0,777	5,501	0,580	0,063	0,711	0,056
1,101	9,984	0,321	0,134	0,854	0,180
0,930	12,620	0,782	0,156	0,709	0,123
10,831	**87,028**	0,490	0,068	0,679	0,130
10,831	**87,030**	0,394	0,073	0,659	0,111
0,679	6,522	0,627	0,058	0,802	0,095
0,445	6,175	0,338	0,093	0,726	0,065
1,319	9,680	0,477	0,078	1,396	0,163
0,859	7,126	0,639	0,095	0,748	0,077
Gaussian blurred images set					
/	16,432	0,712	0,105	0,920	0,137
/	**59,323**	0,518	0,162	0,580	0,120
/	**59,315**	0,644	0,180	0,666	0,133
/	**59,326**	0,405	0,083	0,587	0,082
/	**59,318**	0,221	0,090	0,667	0,070
/	5,386	0,383	0,059	0,649	0,090
/	8,750	0,473	0,082	0,556	0,125
/	6,613	0,338	0,065	0,519	0,077
/	4,901	0,171	0,063	0,557	0,058
/	4,809	0,253	0,079	0,551	0,075
/	5,193	0,238	0,107	0,630	0,124
/	13,508	0,677	0,100	0,591	0,126
/	**61,033**	0,287	0,109	0,537	0,099
/	**61,035**	0,299	0,095	0,553	0,113
/	6,369	0,222	0,081	0,573	0,116
/	6,219	0,334	0,058	0,557	0,081
/	5,495	0,376	0,089	1,135	0,141
/	7,211	0,311	0,087	0,537	0,084

TOY	LIVE	IVC	A57	TID	CSIQ
JPEG 2000 coded images set					
0,755	10,699	0,691	0,142	1,292	0,155
7,083	**67,747**	0,685	0,136	1,307	0,102
7,082	**67,743**	0,743	0,146	1,447	0,118
7,083	**67,748**	0,520	0,046	1,123	0,088
7,082	**67,744**	0,660	0,095	1,125	0,079
0,893	10,971	0,758	0,036	1,235	0,136
0,485	7,407	0,653	0,105	1,273	0,122
0,409	5,604	0,484	0,090	1,084	0,067
0,329	4,952	0,455	0,118	1,092	0,066
0,411	5,659	0,605	0,075	1,128	0,066
1,137	9,178	0,457	0,145	1,173	0,179
0,490	9,333	0,474	0,133	1,461	0,111
10,855	**81,196**	0,546	0,104	1,127	0,082
10,856	**81,197**	0,467	0,089	1,093	0,103
0,437	6,063	0,653	0,064	1,171	0,114
0,382	5,568	0,460	0,093	1,121	0,072
1,303	8,543	0,799	0,181	1,353	0,125
0,564	7,533	0,719	0,100	1,153	0,079

Appendix C: Mean Absolute Error (MAE) after nonlinear regression

	TOY	LIVE	IVC	A57	TID	CSIQ
	All data set					
MSE	0,942	29,174	0,914	0,150	1,087	0,199
PSNR	**4,410**	**69,341**	0,667	0,161	0,866	0,135
SNR	**4,408**	**69,333**	0,743	0,174	0,894	0,139
WSNR	**4,412**	**69,347**	0,484	0,076	0,899	0,137
NQM	**4,409**	**69,338**	0,524	0,117	0,808	0,146
UQI	0,617	7,862	0,523	0,170	0,868	0,112
SSIM	0,542	7,579	0,555	0,185	0,847	0,116
MS-SSIM	0,430	6,608	0,427	0,106	0,616	0,088
VIF	0,370	5,366	0,410	0,139	0,680	0,074
VIFP	0,453	4,925	0,519	0,115	0,779	0,091
IFC	0,959	7,984	0,396	0,161	0,825	0,182
M-SVD	0,593	10,451	0,600	0,138	1,064	0,188
PSNRHVS	**5,890**	**76,747**	0,461	0,090	0,780	0,124
PSNRHVSM	**5,891**	**76,752**	0,437	0,081	0,800	0,127
VSNR	0,422	6,582	0,559	0,062	1,074	0,131
MSSIM	0,374	6,306	0,363	0,126	0,719	0,099
R-SVD	1,197	13,871	0,831	0,206	0,985	0,182
RFSIM	0,565	6,519	0,501	0,112	0,584	0,077
	Noised images set					
MSE	/	20,298	/	0,071	0,538	0,103
PSNR	/	**65,574**	/	0,042	0,192	0,055
SNR	/	**65,566**	/	0,016	0,282	0,069
WSNR	/	**65,581**	/	0,018	0,263	0,070
NQM	/	**65,573**	/	0,043	0,313	0,075
UQI	/	7,629	/	0,060	0,423	0,124
SSIM	/	6,894	/	0,057	0,302	0,083
MS-SSIM	/	3,995	/	0,061	0,287	0,061
VIF	/	3,873	/	0,050	0,225	0,054
VIFP	/	2,865	/	0,062	0,277	0,052
IFC	/	6,655	/	0,097	0,392	0,103
M-SVD	/	8,802	/	0,040	0,193	0,078
PSNRHVS	/	**75,758**	/	0,042	0,191	0,055
PSNRHVSM	/	**75,762**	/	0,048	0,211	0,054
VSNR	/	4,213	/	0,040	0,317	0,072
MSSIM	/	4,905	/	0,027	0,348	0,083
R-SVD	/	7,410	/	0,111	0,381	0,100
RFSIM	/	7,056	/	0,050	0,274	0,064

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG coded images set					
MSE	1,100	10,275	0,605	0,124	1,231	0,151
PSNR	**4,413**	**71,494**	0,617	0,149	0,744	0,096
SNR	**4,411**	**71,486**	0,721	0,180	0,877	0,155
WSNR	**4,416**	**71,502**	0,425	0,043	0,489	0,088
NQM	**4,413**	**71,491**	0,461	0,090	0,529	0,073
UQI	0,501	8,189	0,514	0,066	0,779	0,089
SSIM	0,701	5,832	0,476	0,051	0,561	0,075
MS-SSIM	0,499	4,446	0,316	0,065	0,432	0,045
VIF	0,405	3,667	0,281	0,067	0,416	0,035
VIFP	0,586	3,647	0,395	0,047	0,487	0,042
IFC	0,936	7,596	0,256	0,121	0,649	0,155
M-SVD	0,762	8,324	0,624	0,116	0,465	0,097
PSNRHVS	**5,893**	**80,016**	0,332	0,055	0,446	0,109
PSNRHVSM	**5,895**	**80,022**	0,277	0,057	0,406	0,092
VSNR	0,507	4,567	0,512	0,048	0,536	0,070
MSSIM	0,349	4,866	0,244	0,085	0,487	0,048
R-SVD	1,192	6,692	0,366	0,060	1,115	0,110
RFSIM	0,683	5,128	0,485	0,085	0,534	0,054
	Gaussian blurred images set					
MSE	/	13,067	0,628	0,092	0,658	0,106
PSNR	/	**56,652**	0,378	0,144	0,415	0,094
SNR	/	**56,643**	0,473	0,154	0,513	0,107
WSNR	/	**56,658**	0,327	0,067	0,420	0,064
NQM	/	**56,649**	0,186	0,067	0,501	0,055
UQI	/	3,987	0,272	0,047	0,490	0,064
SSIM	/	6,563	0,342	0,065	0,397	0,085
MS-SSIM	/	5,678	0,260	0,056	0,356	0,058
VIF	/	3,563	0,139	0,046	0,371	0,042
VIFP	/	3,619	0,208	0,064	0,378	0,052
IFC	/	3,820	0,191	0,093	0,464	0,098
M-SVD	/	10,874	0,488	0,082	0,441	0,095
PSNRHVS	/	**58,071**	0,246	0,083	0,368	0,077
PSNRHVSM	/	**58,074**	0,257	0,074	0,387	0,092
VSNR	/	4,740	0,173	0,062	0,399	0,090
MSSIM	/	5,402	0,247	0,047	0,388	0,060
R-SVD	/	4,087	0,296	0,071	0,941	0,104
RFSIM	/	5,388	0,245	0,063	0,367	0,063

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG 2000 coded images set					
MSE	0,612	8,026	0,474	0,113	1,139	0,130
PSNR	**4,407**	**63,240**	0,474	0,099	1,009	0,073
SNR	**4,405**	**63,233**	0,538	0,109	1,138	0,086
WSNR	**4,408**	**63,244**	0,393	0,042	0,724	0,070
NQM	**4,406**	**63,236**	0,542	0,083	0,724	0,060
UQI	0,689	7,588	0,609	0,030	0,887	0,099
SSIM	0,372	5,195	0,454	0,089	0,935	0,088
MS-SSIM	0,323	4,394	0,332	0,065	0,645	0,051
VIF	0,259	3,631	0,336	0,090	0,643	0,048
VIFP	0,310	4,130	0,435	0,068	0,725	0,047
IFC	0,981	7,250	0,343	0,122	0,804	0,142
M-SVD	0,376	6,472	0,372	0,112	1,176	0,094
PSNRHVS	**5,886**	**73,139**	0,369	0,082	0,727	0,060
PSNRHVSM	**5,887**	**73,143**	0,322	0,080	0,666	0,085
VSNR	0,327	4,215	0,497	0,055	0,778	0,092
MSSIM	0,292	4,129	0,325	0,069	0,708	0,055
R-SVD	1,184	5,585	0,631	0,166	1,020	0,091
RFSIM	0,427	5,535	0,530	0,082	0,761	0,059

Appendix D: Spearman's Rank Order Correlation Coefficient (SROCC)

	TOY	LIVE	IVC	A57	TID	CSIQ
	All data set					
MSE	0,722	0,936	0,689	0,618	0,513	0,806
PSNR	0,722	0,936	0,689	0,618	0,513	0,806
SNR	0,695	0,931	0,636	0,572	0,483	0,799
WSNR	0,850	0,956	0,859	0,920	0,450	0,773
NQM	0,909	0,955	0,835	0,798	0,577	0,740
UQI	0,780	0,945	0,827	0,425	0,552	0,807
SSIM	0,840	0,954	0,779	0,407	0,585	0,837
MS-SSIM	0,906	0,972	0,885	0,840	0,789	0,914
VIF	0,916	0,976	0,897	0,622	0,702	0,919
VIFP	0,877	0,975	0,811	0,769	0,619	0,879
IFC	0,872	0,956	0,898	0,319	0,541	0,748
M-SVD	0,817	0,907	0,774	0,645	0,588	0,768
PSNRHVS	0,837	0,958	0,859	0,850	0,549	0,830
PSNRHVSM	0,884	0,964	0,884	0,896	0,518	0,822
VSNR	0,885	0,958	0,798	0,936	0,650	0,810
MSSIM	0,925	0,973	0,914	0,795	0,708	0,883
R-SVD	0,061	0,795	0,456	0,057	0,345	0,557
RFSIM	0,831	0,969	0,819	0,821	0,803	0,929
	Noised images set					
MSE	/	0,991	/	0,950	0,882	0,934
PSNR	/	0,991	/	0,950	0,882	0,934
SNR	/	0,982	/	0,967	0,809	0,925
WSNR	/	0,984	/	0,967	0,836	0,921
NQM	/	0,992	/	0,933	0,743	0,911
UQI	/	0,947	/	0,783	0,528	0,716
SSIM	/	0,978	/	0,817	0,785	0,894
MS-SSIM	/	0,984	/	0,917	0,791	0,933
VIF	/	0,989	/	0,950	0,877	0,951
VIFP	/	0,989	/	0,817	0,818	0,957
IFC	/	0,962	/	0,450	0,606	0,828
M-SVD	/	0,963	/	0,967	0,896	0,937
PSNRHVS	/	0,990	/	0,950	0,868	0,933
PSNRHVSM	/	0,992	/	0,950	0,881	0,943
VSNR	/	0,985	/	0,950	0,760	0,908
MSSIM	/	0,969	/	0,950	0,695	0,884
R-SVD	/	0,949	/	0,050	0,649	0,847
RFSIM	/	0,988	/	0,900	0,820	0,935

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG coded images set					
MSE	0,509	0,943	0,674	0,633	0,826	0,888
PSNR	0,509	0,943	0,674	0,633	0,826	0,888
SNR	0,480	0,937	0,608	0,550	0,758	0,878
WSNR	0,815	0,974	0,878	0,983	0,873	0,956
NQM	0,910	0,979	0,853	0,900	0,859	0,953
UQI	0,846	0,953	0,825	0,933	0,758	0,901
SSIM	0,739	0,970	0,807	0,950	0,871	0,922
MS-SSIM	0,875	0,985	0,918	0,950	0,887	0,962
VIF	0,917	0,980	0,924	0,900	0,878	0,970
VIFP	0,814	0,977	0,830	0,917	0,883	0,968
IFC	0,910	0,962	0,954	0,800	0,812	0,939
M-SVD	0,710	0,943	0,711	0,633	0,892	0,946
PSNRHVS	0,771	0,971	0,885	0,900	0,901	0,940
PSNRHVSM	0,894	0,979	0,924	0,950	0,882	0,952
VSNR	0,845	0,972	0,777	0,983	0,867	0,903
MSSIM	0,934	0,984	0,944	0,933	0,864	0,956
R-SVD	0,004	0,959	0,899	0,867	0,556	0,839
RFSIM	0,754	0,976	0,791	0,883	0,884	0,950
	Gaussian blurred images set					
MSE	/	0,874	0,805	0,467	0,863	0,929
PSNR	/	0,874	0,805	0,467	0,863	0,929
SNR	/	0,857	0,733	0,367	0,808	0,915
WSNR	/	0,950	0,882	0,800	0,862	0,965
NQM	/	0,933	0,956	0,917	0,817	0,958
UQI	/	0,967	0,941	0,867	0,829	0,944
SSIM	/	0,939	0,869	0,683	0,872	0,924
MS-SSIM	/	0,976	0,944	0,783	0,889	0,972
VIF	/	0,982	0,973	0,817	0,888	0,975
VIFP	/	0,974	0,953	0,667	0,876	0,968
IFC	/	0,977	0,952	0,883	0,831	0,959
M-SVD	/	0,802	0,738	0,233	0,850	0,910
PSNRHVS	/	0,935	0,923	0,600	0,886	0,961
PSNRHVSM	/	0,960	0,922	0,717	0,882	0,971
VSNR	/	0,964	0,968	0,900	0,872	0,945
MSSIM	/	0,978	0,947	0,783	0,873	0,967
R-SVD	/	0,963	0,908	0,900	0,214	0,875
RFSIM	/	0,946	0,942	0,633	0,882	0,963

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG 2000 coded images set					
MSE	0,884	0,956	0,850	0,800	0,650	0,936
PSNR	0,884	0,956	0,850	0,800	0,650	0,936
SNR	0,860	0,945	0,814	0,750	0,618	0,931
WSNR	0,900	0,966	0,918	0,917	0,772	0,970
NQM	0,917	0,977	0,859	0,933	0,769	0,963
UQI	0,723	0,941	0,796	0,983	0,731	0,881
SSIM	0,922	0,973	0,850	0,883	0,699	0,921
MS-SSIM	0,942	0,986	0,929	0,900	0,784	0,969
VIF	0,944	0,979	0,936	0,817	0,789	0,967
VIFP	0,935	0,975	0,883	0,933	0,770	0,970
IFC	0,853	0,954	0,934	0,717	0,755	0,926
M-SVD	0,912	0,960	0,932	0,883	0,775	0,974
PSNRHVS	0,909	0,972	0,907	0,883	0,775	0,962
PSNRHVSM	0,932	0,981	0,933	0,850	0,786	0,970
VSNR	0,925	0,971	0,868	0,967	0,762	0,948
MSSIM	0,945	0,984	0,925	0,850	0,775	0,963
R-SVD	0,131	0,957	0,784	0,433	0,681	0,908
RFSIM	0,900	0,972	0,832	0,850	0,760	0,964

Appendix E: Kendall's Rank Order Correlation Coefficient (KROCC)

	TOY	LIVE	IVC	A57	TID	CSIQ
	All data set					
MSE	0,540	0,791	0,522	0,430	0,374	0,607
PSNR	0,540	0,791	0,522	0,430	0,374	0,607
SNR	0,517	0,783	0,464	0,403	0,346	0,600
WSNR	0,656	0,833	0,665	0,753	0,363	0,599
NQM	0,739	0,832	0,634	0,592	0,426	0,563
UQI	0,600	0,806	0,627	0,332	0,402	0,615
SSIM	0,649	0,821	0,594	0,278	0,424	0,632
MS-SSIM	0,732	0,862	0,701	0,648	0,605	0,739
VIF	0,740	0,866	0,717	0,459	0,552	0,753
VIFP	0,691	0,863	0,631	0,564	0,470	0,695
IFC	0,678	0,818	0,719	0,238	0,406	0,574
M-SVD	0,626	0,750	0,588	0,471	0,440	0,583
PSNRHVS	0,642	0,831	0,672	0,675	0,442	0,653
PSNRHVSM	0,701	0,844	0,694	0,725	0,417	0,653
VSNR	0,698	0,822	0,604	0,803	0,493	0,624
MSSIM	0,764	0,862	0,738	0,595	0,524	0,695
R-SVD	0,045	0,623	0,330	0,029	0,238	0,383
RFSIM	0,637	0,854	0,645	0,631	0,628	0,764
	Noised images set					
MSE	/	0,924	/	0,833	0,694	0,761
PSNR	/	0,924	/	0,833	0,694	0,761
SNR	/	0,887	/	0,889	0,611	0,753
WSNR	/	0,898	/	0,889	0,641	0,747
NQM	/	0,928	/	0,833	0,542	0,731
UQI	/	0,803	/	0,611	0,368	0,520
SSIM	/	0,881	/	0,667	0,584	0,709
MS-SSIM	/	0,903	/	0,778	0,598	0,773
VIF	/	0,913	/	0,833	0,684	0,796
VIFP	/	0,914	/	0,667	0,622	0,818
IFC	/	0,830	/	0,333	0,430	0,629
M-SVD	/	0,851	/	0,889	0,721	0,770
PSNRHVS	/	0,920	/	0,833	0,661	0,760
PSNRHVSM	/	0,928	/	0,833	0,685	0,783
VSNR	/	0,895	/	0,889	0,556	0,731
MSSIM	/	0,854	/	0,833	0,510	0,692
R-SVD	/	0,812	/	0,111	0,466	0,650
RFSIM	/	0,914	/	0,778	0,622	0,770

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG coded images set					
MSE	0,362	0,801	0,519	0,500	0,650	0,692
PSNR	0,362	0,801	0,519	0,500	0,650	0,692
SNR	0,344	0,792	0,448	0,444	0,565	0,689
WSNR	0,620	0,876	0,696	0,944	0,690	0,813
NQM	0,750	0,889	0,666	0,722	0,679	0,812
UQI	0,672	0,822	0,625	0,833	0,541	0,719
SSIM	0,543	0,863	0,630	0,889	0,705	0,753
MS-SSIM	0,690	0,913	0,780	0,833	0,726	0,828
VIF	0,747	0,895	0,791	0,778	0,699	0,855
VIFP	0,621	0,880	0,676	0,778	0,726	0,839
IFC	0,740	0,831	0,816	0,611	0,595	0,795
M-SVD	0,522	0,804	0,537	0,500	0,743	0,790
PSNRHVS	0,573	0,863	0,734	0,833	0,770	0,781
PSNRHVSM	0,715	0,887	0,790	0,833	0,711	0,804
VSNR	0,654	0,861	0,583	0,944	0,695	0,715
MSSIM	0,780	0,905	0,821	0,778	0,674	0,815
R-SVD	0,010	0,824	0,722	0,722	0,372	0,652
RFSIM	0,559	0,879	0,622	0,722	0,728	0,797
	Gaussian blurred images set					
MSE	/	0,705	0,667	0,389	0,730	0,754
PSNR	/	0,705	0,667	0,389	0,730	0,754
SNR	/	0,693	0,582	0,333	0,643	0,740
WSNR	/	0,821	0,709	0,667	0,719	0,831
NQM	/	0,797	0,836	0,778	0,632	0,826
UQI	/	0,853	0,815	0,778	0,652	0,801
SSIM	/	0,797	0,709	0,611	0,743	0,765
MS-SSIM	/	0,872	0,825	0,611	0,774	0,851
VIF	/	0,887	0,899	0,667	0,774	0,866
VIFP	/	0,863	0,825	0,556	0,743	0,843
IFC	/	0,873	0,847	0,722	0,664	0,827
M-SVD	/	0,628	0,614	0,222	0,696	0,728
PSNRHVS	/	0,786	0,825	0,556	0,769	0,824
PSNRHVSM	/	0,836	0,794	0,611	0,756	0,849
VSNR	/	0,837	0,878	0,833	0,734	0,790
MSSIM	/	0,876	0,825	0,611	0,741	0,834
R-SVD	/	0,838	0,772	0,778	0,143	0,695
RFSIM	/	0,813	0,815	0,500	0,759	0,829

	TOY	LIVE	IVC	A57	TID	CSIQ
	JPEG 2000 coded images set					
MSE	0,713	0,829	0,726	0,667	0,484	0,766
PSNR	0,713	0,829	0,726	0,667	0,484	0,766
SNR	0,684	0,809	0,654	0,611	0,472	0,763
WSNR	0,734	0,851	0,766	0,778	0,696	0,845
NQM	0,755	0,874	0,680	0,778	0,684	0,840
UQI	0,546	0,800	0,606	0,944	0,621	0,707
SSIM	0,762	0,863	0,692	0,722	0,565	0,753
MS-SSIM	0,803	0,902	0,782	0,778	0,733	0,846
VIF	0,801	0,880	0,790	0,667	0,721	0,849
VIFP	0,777	0,865	0,723	0,833	0,689	0,849
IFC	0,648	0,808	0,779	0,556	0,658	0,781
M-SVD	0,752	0,841	0,787	0,722	0,697	0,856
PSNRHVS	0,742	0,862	0,761	0,722	0,698	0,822
PSNRHVSM	0,782	0,887	0,795	0,722	0,730	0,844
VSNR	0,757	0,859	0,695	0,889	0,671	0,792
MSSIM	0,809	0,896	0,767	0,722	0,706	0,831
R-SVD	0,103	0,820	0,588	0,278	0,548	0,740
RFSIM	0,727	0,861	0,665	0,667	0,664	0,831

APPENDIX

F

Appendix F: Friedman's analysis results of the variability of the 18 image quality metrics over the six databases (Toyama, LIVE, IVC, A57, TID and CSIQ) in the 95% Confidence Interval

	All data sets				
	SNR	MS-SSIM	VIF	MSSIM	RSVD
SNR		x	x	x	
MS-SSIM	x				x
VIF	x				x
MSSIM	x				x
R-SVD		x	x	x	

	JPEG coded images sets						
	MSE	PSNR	SNR	MS-SSIM	VIF	VIFP	MSSIM
MSE				x	x		
PSNR				x	x	x	x
SNR				x	x	x	x
MS-SSIM	x	x	x				
VIF	x	x	x				
VIFP	x		x				
MSSIM	x		x				

	JPEG 2000 coded images sets				
	SNR	MS-SSIM	VIF	MSSIM	RSVD
SNR		x	x	x	
MS-SSIM	x				x
VIF	x				x
MSSIM	x				x
R-SVD		x	x	x	

	Gaussian blurred images sets			
	MSE	SNR	VIF	VIFP
MSE			x	x
SNR			x	x
VIF	x	x		
VIFP	x	x		

Table F.1: Pearson's Correlation Coefficient (PCC) variability over the 18 quality metrics according to the Friedman test in the 95% CI.

	All data sets			
	MS-SSIM	VIF	MSSIM	R-SVD
MS-SSIM				x
VIF				x
MSSIM				x
R-SVD	x	x	x	

	JPEG coded images sets						
	MSE	PSNR	SNR	MS-SSIM	VIF	VIFP	MSSIM
MSE					x		
PSNR				x	x		
SNR				x	x	x	x
MS-SSIM		x	x				
VIF	x	x	x				
VIFP	x						
MSSIM	x						

	JPEG 2000 coded images sets				
	SNR	MS-SSIM	VIF	MSSIM	R-SVD
SNR		x	x	x	
MS-SSIM	x				x
VIF	x				x
MSSIM	x				x
R-SVD		x	x	x	

	Gaussian blurred images sets			
	MSE	SNR	VIF	VIFP
MSE			x	x
SNR			x	x
VIF	x	x		
VIFP	x	x		

Table F.2: Root Mean Square Error (RMSE) variability
over the 18 quality metrics according to the Friedman test in the 95% CI.

	All data sets		
	MSE	VIF	R-SVD
MSE		x	
VIF	x		x
R-SVD		x	

	JPEG coded images sets						
	MSE	PSNR	SNR	MS-SSIM	VIF	VIFP	MSSIM
MSE					x		
PSNR					x		
SNR				x	x	x	x
MS-SSIM			x				
VIF	x	x	x				
VIFP			x				
MSSIM			x				

	Gaussian blurred images sets					
	MSE	PSNR	SNR	VIF	VIFP	M-SVD
MSE				x		
PSNR				x		
SNR				x	x	
VIF	x	x	x			x
VIFP			x			
M-SVD				x		

Table F.3: Mean Absolute Error (MAE) variability
over the 18 quality metrics according to the Friedman test in the 95% CI.

	All data sets						
	SNR	MS-SSIM	VIF	VIFP	MSSIM	R-SVD	RFSIM
SNR		x	x		x		
MS-SSIM	x					x	
VIF	x					x	
VIFP						x	
MSSIM	x					x	
R-SVD		x	x	x	x		x
RFSIM						x	

	JPEG coded images sets						
	MSE	PSNR	SNR	MS-SSIM	VIF	PSNRHVSM	MSSIM
MSE				x	x		x
PSNR				x	x		x
SNR				x	x	x	x
MS-SSIM	x	x	x				
VIF	x	x	x				
PSNR-HVSM			x				
MSSIM	x	x	x				

	JPEG 2000 coded images sets						
	SNR	MS-SSIM	VIF	VIFP	PSNRHVSM	MSSIM	R-SVD
SNR		x	x		x	x	
MS-SSIM	x						x
VIF	x						x
VIFP							x
PSNRHVSM	x						x
MSSIM	x						x
R-SVD		x	x	x	x	x	

	Gaussian blurred images sets					
	MSE	PSNR	SNR	MS-SSIM	VIF	M-SVD
MSE					x	
PSNR					x	
SNR				x	x	
MS-SSIM			x			x
VIF	x	x	x			x
M-SVD				x	x	

Table F.4: Spearman's and Kendall's Rank Order Correlation Coefficients (SROCC & KROCC) variability over the 18 quality metrics according to the Friedman test in the 95% CI.

APPENDICES

G - L

Appendix G: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **TOYAMA** database.

	t(1)	t(2)	t(3)	t(4)		t(1)	t(2)	t(3)	t(4)		t(1)	t(2)	t(3)	t(4)
	All data set					JPEG coded images set					JPEG 2000 coded images set			
MSE	-3,87500	0,01615	50,02827	3,07079		-3,87500	0,02064	49,06286	3,01339		-3,87500	0,01366	50,99368	3,12819
PSNR	3,87500	0,00004	9390,39937	3,07079		3,87500	0,00004	9389,55481	3,01339		3,87500	0,00004	9391,24394	3,12819
SNR	3,87500	0,00004	9385,21226	3,07079		3,87500	0,00004	9384,36770	3,01339		3,87500	0,00004	9386,05682	3,12819
WSNR	3,87500	0,00004	9396,81980	3,07079		3,87500	0,00004	9397,60201	3,01339		3,87500	0,00004	9396,03759	3,12819
NQM	3,87500	0,00004	9388,87787	3,07079		3,87500	0,00004	9389,02318	3,01339		3,87500	0,00004	9388,73256	3,12819
UQI	3,87500	5,88620	0,72952	3,07079		3,87500	5,88048	0,73890	3,01339		3,87500	5,87973	0,72014	3,12819
SSIM	3,87500	11,80645	0,89926	3,07079		3,87500	15,80644	0,90388	3,01339		3,87500	9,81697	0,89464	3,12819
MS-SSIM	3,87500	44,90482	0,97893	3,07079		3,87500	69,15253	0,98235	3,01339		3,87500	36,17508	0,97550	3,12819
VIF	3,87500	4,28096	0,63064	3,07079		3,87500	4,80605	0,65133	3,01339		3,87500	3,90628	0,60995	3,12819
VIFP	3,87500	4,52716	0,56780	3,07079		3,87500	4,83941	0,55357	3,01339		3,87500	4,26484	0,58204	3,12819
IFC	3,87500	0,03767	15,21411	3,07079		3,87500	0,03770	15,44730	3,01339		3,87500	0,03746	14,98091	3,12819
M-SVD	-3,87500	0,11468	11,16983	3,07079		-3,87500	0,12934	12,13103	3,01339		-3,87500	0,10470	10,20863	3,12819
PSNRHVS	3,87500	0,00003	14313,68539	3,07079		3,87500	0,00003	14314,22020	3,01339		3,87500	0,00003	14313,15059	3,12819
PSNRHVSM	3,87500	0,00003	14318,00299	3,07079		3,87500	0,00003	14319,61736	3,01339		3,87500	0,00003	14316,38862	3,12819
VSNR	3,87500	0,03106	43,45760	3,07079		3,87500	0,03116	43,40589	3,01339		3,87500	0,03081	43,50931	3,12819
MSSIM	3,87500	21,87576	0,95358	3,07079		3,87500	30,11041	0,96203	3,01339		3,87500	18,40025	0,94513	3,12819
R-SVD	-3,87500	7,07319	0,30453	3,07079		3,87500	7,21534	0,29119	3,01339		-3,87500	6,96491	0,31787	3,12819
RFSIM	3,87500	5,22022	0,74800	3,07079		3,87500	6,34295	0,76494	3,01339		3,87500	4,54848	0,73106	3,12819

Appendix H: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **LIVE** database.

	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)
	All data set				JPEG coded images set				JPEG 2000 coded images set			
MSE	114,41000	0,00078	406,14947	38,53848	111,41000	0,00548	137,05398	37,99747	91,37200	0,00562	110,80341	32,46584
PSNR	-114,41000	0,00004	13569,27681	38,53848	-111,41000	0,00004	16334,99293	37,99747	-91,37200	0,00003	16767,16683	32,46584
SNR	-114,41000	0,00004	13564,43523	38,53848	-111,41000	0,00004	16330,36873	37,99747	-91,37200	0,00003	16762,64909	32,46584
WSNR	-114,41000	0,00004	13572,52338	38,53848	-111,41000	0,00004	16339,07774	37,99747	-91,37200	0,00003	16769,59265	32,46584
NQM	-114,41000	0,00004	13567,21482	38,53848	-111,41000	0,00004	16332,82521	37,99747	-91,37200	0,00003	16764,51088	32,46584
UQI	-114,41000	3,46005	0,65960	38,53848	-111,41000	3,59870	0,68623	37,99747	-91,37200	3,95309	0,69513	32,46584
SSIM	-114,41000	4,19940	0,78269	38,53848	-111,41000	5,95349	0,83253	37,99747	-91,37200	6,25206	0,84667	32,46584
MS-SSIM	-114,41000	6,71888	0,91148	38,53848	-111,41000	12,58539	0,93994	37,99747	-91,37200	14,13565	0,94746	32,46584
VIF	-114,41000	2,91652	0,53760	38,53848	-111,41000	2,81382	0,57194	37,99747	-91,37200	2,87506	0,54943	32,46584
VIFP	-114,41000	3,16192	0,50686	38,53848	-111,41000	3,09527	0,53235	37,99747	-91,37200	3,15930	0,55824	32,46584
IFC	-114,41000	0,03193	19,83364	38,53848	-111,41000	0,02984	23,04352	37,99747	-91,37200	0,02943	23,17655	32,46584
M-SVD	114,41000	0,03243	25,39984	38,53848	111,41000	0,05425	17,37517	37,99747	91,37200	0,05204	15,84758	32,46584
PSNRHVS	-114,41000	0,00002	20692,46619	38,53848	-111,41000	0,00002	24913,76305	37,99747	-91,37200	0,00002	25571,24597	32,46584
PSNRHVSM	-114,41000	0,00002	20695,06842	38,53848	-111,41000	0,00002	24917,17238	37,99747	-91,37200	0,00002	25573,53387	32,46584
VSNR	-114,41000	0,02899	39,40786	38,53848	-111,41000	0,02770	43,95902	37,99747	-91,37200	0,02767	44,84420	32,46584
MSSIM	-114,41000	4,99377	0,85499	38,53848	-111,41000	7,07389	0,88974	37,99747	-91,37200	8,19314	0,89926	32,46584
R-SVD	114,41000	4,40266	0,34442	38,53848	111,41000	4,69520	0,32530	37,99747	91,37200	4,56241	0,31530	32,46584
RFSIM	-114,41000	3,12062	0,61612	38,53848	-111,41000	3,10056	0,65128	37,99747	-91,37200	3,47929	0,67204	32,46584
	Noised images set				Gaussian blurred images set							
MSE	111,77000	0,00037	1533,25464	47,14174	93,40900	0,00451	185,42933	38,10032				
PSNR	-111,77000	0,00004	10941,74715	47,14174	-93,40900	0,00004	10944,79279	38,10032				
SNR	-111,77000	0,00004	10936,66773	47,14174	-93,40900	0,00004	10939,71337	38,10032				
WSNR	-111,77000	0,00004	10945,98016	47,14174	-93,40900	0,00004	10947,62453	38,10032				
NQM	-111,77000	0,00004	10940,43928	47,14174	-93,40900	0,00004	10942,82491	38,10032				
UQI	-111,77000	2,98446	0,53055	47,14174	-93,40900	3,97065	0,70304	38,10032				
SSIM	-111,77000	2,82603	0,57379	47,14174	-93,40900	6,56990	0,82551	38,10032				
MS-SSIM	-111,77000	3,99250	0,80186	47,14174	-93,40900	11,45514	0,94008	38,10032				
VIF	-111,77000	2,87454	0,49021	47,14174	-93,40900	3,31934	0,52537	38,10032				
VIFP	-111,77000	2,96226	0,42092	47,14174	-93,40900	3,64832	0,50315	38,10032				
IFC	-111,77000	0,03426	16,03972	47,14174	-93,40900	0,03505	17,66692	38,10032				
M-SVD	111,77000	0,02235	34,15496	47,14174	93,40900	0,03450	35,52911	38,10032				
PSNRHVS	-111,77000	0,00003	16685,54650	47,14174	-93,40900	0,00003	16686,56951	38,10032				
PSNRHVSM	-111,77000	0,00003	16688,08269	47,14174	-93,40900	0,00003	16689,01919	38,10032				
VSNR	-111,77000	0,03002	34,37642	47,14174	-93,40900	0,03123	34,28466	38,10032				
MSSIM	-111,77000	3,29153	0,71697	47,14174	-93,40900	7,44675	0,89410	38,10032				
R-SVD	111,77000	9,67348	0,22189	47,14174	93,40900	4,03297	0,45984	38,10032				
RFSIM	-111,77000	2,61431	0,49761	47,14174	-93,40900	3,79326	0,61807	38,10032				

Appendix I: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **IVC** database.

	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)
	All data set				JPEG coded images set				JPEG 2000 coded images set			
MSE	-3,88462	0,00959	111,22692	2,95655	-3,61538	0,01131	91,96977	2,76077	-3,88462	0,00939	87,54495	3,09462
PSNR	3,88462	0,26056	29,29060	2,95655	3,61538	0,30627	29,82619	2,76077	3,88462	0,22766	30,88586	3,09462
SNR	3,88462	0,23971	24,00789	2,95655	3,61538	0,27311	24,75899	2,76077	3,88462	0,21576	25,81866	3,09462
WSNR	3,88462	0,18760	32,89503	2,95655	3,61538	0,20997	34,18492	2,76077	3,88462	0,17611	33,43365	3,09462
NQM	3,88462	0,19428	26,05829	2,95655	3,61538	0,22319	25,97765	2,76077	3,88462	0,18178	26,03416	3,09462
UQI	3,88462	7,45915	0,60008	2,95655	3,61538	8,69831	0,60567	2,76077	3,88462	6,93538	0,60705	3,09462
SSIM	3,88462	9,81168	0,80612	2,95655	3,61538	12,34602	0,82323	2,76077	3,88462	8,81001	0,82829	3,09462
MS-SSIM	3,88462	23,04748	0,94547	2,95655	3,61538	34,64848	0,95469	2,76077	3,88462	19,78817	0,94836	3,09462
VIF	3,88462	6,21896	0,38407	2,95655	3,61538	7,39596	0,40076	2,76077	3,88462	5,32870	0,38336	3,09462
VIFP	3,88462	8,19939	0,39680	2,95655	3,61538	9,74760	0,40643	2,76077	3,88462	6,76340	0,42317	3,09462
IFC	3,88462	0,68107	2,84530	2,95655	3,61538	0,92373	2,74572	2,76077	3,88462	0,68340	2,74071	3,09462
M-SVD	-3,88462	0,09324	15,58082	2,95655	-3,61538	0,13380	13,27548	2,76077	-3,88462	0,10140	12,89837	3,09462
PSNRHVS	3,88462	0,22722	26,81286	2,95655	3,61538	0,26698	27,79448	2,76077	3,88462	0,19981	27,65437	3,09462
PSNRHVSM	3,88462	0,18030	29,72254	2,95655	3,61538	0,19626	31,33449	2,76077	3,88462	0,16866	30,26814	3,09462
VSNR	3,88462	0,16163	23,63299	2,95655	3,61538	0,19345	24,22597	2,76077	3,88462	0,13664	24,70046	3,09462
MSSIM	3,88462	13,23169	0,89128	2,95655	3,61538	17,81646	0,90108	2,76077	3,88462	11,47671	0,89160	3,09462
R-SVD	-3,88462	12,09973	0,50081	2,95655	-3,61538	10,86380	0,50011	2,76077	-3,88462	22,81586	0,49661	3,09462
RFSIM	3,88462	5,52770	0,59438	2,95655	3,61538	6,17788	0,59166	2,76077	3,88462	4,99423	0,59525	3,09462
	Noised images set				Gaussian blurred images set							
MSE					-3,50000	0,00952	171,85522	2,36346				
PSNR					3,50000	0,28282	26,86977	2,36346				
SNR					3,50000	0,25787	21,49393	2,36346				
WSNR					3,50000	0,14215	30,01189	2,36346				
NQM					3,50000	0,13943	26,95577	2,36346				
UQI					3,50000	5,57040	0,57012	2,36346				
SSIM					3,50000	8,43402	0,76022	2,36346				
MS-SSIM					3,50000	16,91468	0,91930	2,36346				
VIF					3,50000	4,53153	0,35822	2,36346				
VIFP					3,50000	6,85445	0,38076	2,36346				
IFC					3,50000	0,36324	3,53085	2,36346				
M-SVD					-3,50000	0,07463	31,21067	2,36346				
PSNRHVS					3,50000	0,20937	23,20008	2,36346				
PSNRHVSM					3,50000	0,16416	25,14843	2,36346				
VSNR					3,50000	0,17047	18,65126	2,36346				
MSSIM					3,50000	10,46189	0,85680	2,36346				
R-SVD					-3,50000	12,29608	0,62961	2,36346				
RFSIM					3,50000	5,72967	0,54138	2,36346				

Appendix J: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **A57** database.

	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)
	All data set				JPEG coded images set				JPEG 2000 coded images set			
MSE	0,91126	0,01208	104,68563	0,42153	0,76339	0,01185	103,61092	0,42482	0,68406	0,01094	107,29942	0,46549
PSNR	-0,91126	0,29642	29,19616	0,42153	-0,76339	0,28546	29,20825	0,42482	-0,68406	0,27853	29,13113	0,46549
SNR	-0,91126	0,24433	25,03052	0,42153	-0,76339	0,23471	25,04261	0,42482	-0,68406	0,23060	24,96549	0,46549
WSNR	-0,91126	0,17196	35,12519	0,42153	-0,76339	0,17420	36,82715	0,42482	-0,68406	0,23441	33,44610	0,46549
NQM	-0,91126	0,15993	25,83704	0,42153	-0,76339	0,20723	25,90169	0,42482	-0,68406	0,23080	24,19051	0,46549
UQI	-0,91126	8,07207	0,63982	0,42153	-0,76339	8,89553	0,58385	0,42482	-0,68406	9,29931	0,54283	0,46549
SSIM	-0,91126	10,21045	0,81418	0,42153	-0,76339	11,78730	0,81469	0,42482	-0,68406	11,12361	0,79437	0,46549
MS-SSIM	-0,91126	29,81371	0,95226	0,42153	-0,76339	28,87558	0,95562	0,42482	-0,68406	27,21765	0,94284	0,46549
VIF	-0,91126	6,43010	0,46798	0,42153	-0,76339	5,56319	0,43241	0,42482	-0,68406	8,33575	0,31549	0,46549
VIFP	-0,91126	10,66459	0,39902	0,42153	-0,76339	10,20271	0,37423	0,42482	-0,68406	10,59214	0,34573	0,46549
IFC	-0,91126	0,55909	3,85091	0,42153	-0,76339	0,63574	3,04236	0,42482	-0,68406	1,10479	2,15738	0,46549
M-SVD	0,91126	0,09658	17,40458	0,42153	0,76339	0,10877	15,43072	0,42482	0,68406	0,09366	16,05178	0,46549
PSNRHVS	-0,91126	0,25000	27,38595	0,42153	-0,76339	0,21176	28,82788	0,42482	-0,68406	0,27566	26,35621	0,46549
PSNRHVSM	-0,91126	0,18516	30,61693	0,42153	-0,76339	0,13860	33,47032	0,42482	-0,68406	0,24263	28,86637	0,46549
VSNR	-0,91126	0,19028	23,40258	0,42153	-0,76339	0,20195	24,29750	0,42482	-0,68406	0,24883	21,93632	0,46549
MSSIM	-0,91126	16,21330	0,90186	0,42153	-0,76339	14,50499	0,90578	0,42482	-0,68406	15,76032	0,88098	0,46549
R-SVD	-0,91126	8,66141	0,41777	0,42153	0,76339	32,89732	0,31087	0,42482	0,68406	14,26318	0,45757	0,46549
RFSIM	-0,91126	6,19119	0,59066	0,42153	-0,76339	6,07017	0,60262	0,42482	-0,68406	5,81491	0,53437	0,46549
	Noised images set				Gaussian blurred images set							
MSE	0,42151	0,01109	107,42054	0,30426	0,68490	0,01223	100,59463	0,37494				
PSNR	-0,42151	0,27854	29,11323	0,30426	-0,68490	0,28520	29,33477	0,37494				
SNR	-0,42151	0,23035	24,94760	0,30426	-0,68490	0,23407	25,16914	0,37494				
WSNR	-0,42151	0,22466	37,71460	0,30426	-0,68490	0,21577	37,39931	0,37494				
NQM	-0,42151	0,27047	27,77901	0,30426	-0,68490	0,22872	32,58922	0,37494				
UQI	-0,42151	14,24692	0,61595	0,30426	-0,68490	7,89945	0,72517	0,37494				
SSIM	-0,42151	10,13338	0,71007	0,30426	-0,68490	11,26377	0,85398	0,37494				
MS-SSIM	-0,42151	30,67147	0,94100	0,30426	-0,68490	40,03654	0,97273	0,37494				
VIF	-0,42151	9,00073	0,54982	0,30426	-0,68490	6,44719	0,54534	0,37494				
VIFP	-0,42151	12,98482	0,42559	0,30426	-0,68490	9,94988	0,45711	0,37494				
IFC	-0,42151	1,51108	3,69717	0,30426	-0,68490	0,48991	5,65918	0,37494				
M-SVD	0,42151	0,17506	10,98656	0,30426	0,68490	0,08003	23,91866	0,37494				
PSNRHVS	-0,42151	0,27732	29,08805	0,30426	-0,68490	0,26332	27,58524	0,37494				
PSNRHVSM	-0,42151	0,25759	32,71473	0,30426	-0,68490	0,19474	31,37267	0,37494				
VSNR	-0,42151	0,25965	25,57925	0,30426	-0,68490	0,24360	23,16782	0,37494				
MSSIM	-0,42151	14,88129	0,86703	0,30426	-0,68490	22,88095	0,94883	0,37494				
R-SVD	-0,42151	11,30714	0,46211	0,30426	0,68490	12,12341	0,50192	0,37494				
RFSIM	-0,42151	6,14345	0,64394	0,30426	-0,68490	7,56938	0,66902	0,37494				

Appendix K: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **TID** database.

	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)
	All data set				JPEG coded images set				JPEG 2000 coded images set			
MSE	-7,71430	0,00080	454,22416	4,47960	-5,55290	0,00083	375,91297	4,08799	6,44000	0,00079	464,84080	3,14254
PSNR	7,71430	0,17284	26,74428	4,47960	5,55290	0,18237	27,99072	4,08799	6,44000	0,21560	25,50371	3,14254
SNR	7,71430	0,16395	20,12782	4,47960	5,55290	0,17013	21,37426	4,08799	6,44000	0,19861	18,88725	3,14254
WSNR	7,71430	0,09732	28,72446	4,47960	5,55290	0,10136	32,62923	4,08799	6,44000	0,10220	28,16936	3,14254
NQM	7,71430	0,10238	20,90801	4,47960	5,55290	0,10608	23,07895	4,08799	6,44000	0,09426	19,47113	3,14254
UQI	7,71430	3,78208	0,65305	4,47960	5,55290	4,29783	0,53475	4,08799	6,44000	3,85867	0,41861	3,14254
SSIM	7,71430	4,83196	0,77183	4,47960	5,55290	5,64725	0,76067	4,08799	6,44000	5,66470	0,69780	3,14254
MS-SSIM	7,71430	5,63020	0,89569	4,47960	5,55290	5,76121	0,89961	4,08799	6,44000	5,21569	0,83981	3,14254
VIF	7,71430	3,56098	0,55467	4,47960	5,55290	4,11591	0,39077	4,08799	6,44000	4,81920	0,23442	3,14254
VIFP	7,71430	3,76367	0,50036	4,47960	5,55290	6,07741	0,34479	4,08799	6,44000	6,35924	0,25537	3,14254
IFC	7,71430	0,05001	13,44490	4,47960	5,55290	0,44300	2,99565	4,08799	6,44000	0,47790	1,97309	3,14254
M-SVD	-7,71430	0,01769	37,13071	4,47960	-5,55290	0,01816	33,73342	4,08799	-6,44000	0,01793	47,33449	3,14254
PSNRHVS	7,71430	0,13715	24,60270	4,47960	5,55290	0,13189	27,02617	4,08799	6,44000	0,16601	22,59517	3,14254
PSNRHVSM	7,71430	0,10711	27,16240	4,47960	5,55290	0,09857	31,12007	4,08799	6,44000	0,12799	25,03271	3,14254
VSNR	7,71430	0,00063	63,84381	4,47960	5,55290	0,10584	24,93174	4,08799	6,44000	0,09594	21,76092	3,14254
MSSIM	7,71430	4,86984	0,83611	4,47960	5,55290	5,07043	0,83229	4,08799	6,44000	4,34440	0,73387	3,14254
R-SVD	-7,71430	6,65229	0,35240	4,47960	-5,55290	14,67527	0,26972	4,08799	-6,44000	12,20020	0,24156	3,14254
RFSIM	7,71430	3,84116	0,56469	4,47960	5,55290	3,81561	0,51483	4,08799	6,44000	4,32104	0,37818	3,14254
	Noised images set				Gaussian blurred images set							
MSE	-2,46130	0,00077	368,77439	4,90393	-5,06060	0,00117	380,16441	3,94111				
PSNR	2,46130	0,19879	28,34892	4,90393	5,06060	0,19071	26,21491	3,94111				
SNR	2,46130	0,18294	21,73246	4,90393	5,06060	0,18436	19,59845	3,94111				
WSNR	2,46130	0,13937	34,75049	4,90393	5,06060	0,13019	27,85584	3,94111				
NQM	2,46130	0,13504	26,09736	4,90393	5,06060	0,12625	19,53482	3,94111				
UQI	2,46130	4,83294	0,58520	4,90393	5,06060	3,91683	0,57376	3,94111				
SSIM	2,46130	5,24487	0,68403	4,90393	5,06060	5,06958	0,74454	3,94111				
MS-SSIM	2,46130	5,83402	0,90803	4,90393	5,06060	5,76636	0,89284	3,94111				
VIF	2,46130	6,09800	0,52747	4,90393	5,06060	4,11223	0,39519	3,94111				
VIFP	2,46130	7,46502	0,40819	4,90393	5,06060	5,43918	0,37219	3,94111				
IFC	2,46130	0,55074	3,88038	4,90393	5,06060	0,34486	4,24173	3,94111				
M-SVD	-2,46130	0,01770	24,28429	4,90393	-5,06060	0,01922	50,75023	3,94111				
PSNRHVS	2,46130	0,17733	28,10309	4,90393	5,06060	0,16323	22,83852	3,94111				
PSNRHVSM	2,46130	0,15129	31,71447	4,90393	5,06060	0,13915	24,92483	3,94111				
VSNR	2,46130	0,13071	27,56039	4,90393	5,06060	0,12742	20,37970	3,94111				
MSSIM	2,46130	5,37626	0,84208	4,90393	5,06060	4,96406	0,82629	3,94111				
R-SVD	-2,46130	13,20471	0,32779	4,90393	-5,06060	14,02037	0,29357	3,94111				
RFSIM	2,46130	5,22465	0,60674	4,90393	5,06060	4,13290	0,50219	3,94111				

Appendix L: Coefficients values for each of the 18 image quality metrics over the five considered datasets for the **CSIQ** database.

	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)	t(1)	t(2)	t(3)	t(4)
	All data set				JPEG coded images set				JPEG 2000 coded images set			
MSE	1,00000	0,00305	220,41741	0,35074	0,91764	0,00690	130,41332	0,37005	0,99979	0,00475	186,38610	0,40531
PSNR	-1,00000	0,13058	29,60169	0,35074	-0,91764	0,11050	32,02640	0,37005	-0,99979	0,13809	29,73371	0,40531
SNR	-1,00000	0,12687	23,03935	0,35074	-0,91764	0,10955	25,46779	0,37005	-0,99979	0,13355	23,17510	0,40531
WSNR	-1,00000	0,08657	30,34404	0,35074	-0,91764	0,07811	36,48167	0,37005	-0,99979	0,09476	31,09866	0,40531
NQM	-1,00000	0,08551	24,97816	0,35074	-0,91764	0,09255	29,31941	0,37005	-0,99979	0,09387	25,73535	0,40531
UQI	-1,00000	3,93329	0,66861	0,35074	-0,91764	3,70138	0,63643	0,37005	-0,99979	3,68733	0,56288	0,40531
SSIM	-1,00000	5,87907	0,80651	0,35074	-0,91764	6,59988	0,82415	0,37005	-0,99979	5,61145	0,78630	0,40531
MS-SSIM	-1,00000	9,51296	0,91769	0,35074	-0,91764	14,79187	0,94101	0,37005	-0,99979	8,30981	0,90292	0,40531
VIF	-1,00000	3,45011	0,54719	0,35074	-0,91764	2,96358	0,51103	0,37005	-0,99979	3,23776	0,39032	0,40531
VIFP	-1,00000	4,00867	0,52199	0,35074	-0,91764	3,73643	0,47165	0,37005	-0,99979	4,07980	0,40671	0,40531
IFC	-1,00000	0,14972	7,39925	0,35074	-0,91764	0,16763	6,00551	0,37005	-0,99979	0,27532	3,75958	0,40531
M-SVD	1,00000	0,03533	26,99826	0,35074	0,91764	0,07082	15,67621	0,37005	0,99979	0,04511	22,20595	0,40531
PSNRHVS	-1,00000	0,10927	27,12021	0,35074	-0,91764	0,08710	31,49873	0,37005	-0,99979	0,11485	26,80599	0,40531
PSNRHVSM	-1,00000	0,08316	30,56177	0,35074	-0,91764	0,06604	37,05515	0,37005	-0,99979	0,09005	30,19246	0,40531
VSNR	-1,00000	0,08139	26,97878	0,35074	-0,91764	0,06407	30,94687	0,37005	-0,99979	0,07899	27,80378	0,40531
MSSIM	-1,00000	6,38539	0,85661	0,35074	-0,91764	7,95496	0,88484	0,37005	-0,99979	5,54863	0,83356	0,40531
R-SVD	1,00000	5,69780	0,38756	0,35074	0,91764	40,69038	0,25700	0,37005	0,99979	6,62560	0,43555	0,40531
RFSIM	-1,00000	3,41547	0,58937	0,35074	-0,91764	3,31483	0,62366	0,37005	-0,99979	3,28847	0,54000	0,40531
	Noised images set				Gaussian blurred images set							
MSE	0,81237	0,00910	100,97346	0,39411	0,98797	0,00439	211,44656	0,38995				
PSNR	-0,81237	0,15572	31,71788	0,39411	-0,98797	0,15460	28,52555	0,38995				
SNR	-0,81237	0,14865	25,15927	0,39411	-0,98797	0,14852	21,96694	0,38995				
WSNR	-0,81237	0,14818	26,67734	0,39411	-0,98797	0,09073	31,61332	0,38995				
NQM	-0,81237	0,16024	20,08702	0,39411	-0,98797	0,07401	30,35604	0,38995				
UQI	-0,81237	5,23394	0,75426	0,39411	-0,98797	3,30085	0,65625	0,38995				
SSIM	-0,81237	7,53398	0,85887	0,39411	-0,98797	5,37338	0,80619	0,38995				
MS-SSIM	-0,81237	12,02015	0,92651	0,39411	-0,98797	7,81222	0,90730	0,38995				
VIF	-0,81237	4,80324	0,61337	0,39411	-0,98797	3,16990	0,46756	0,38995				
VIFP	-0,81237	4,80752	0,57873	0,39411	-0,98797	3,99717	0,45549	0,38995				
IFC	-0,81237	0,31088	5,64399	0,39411	-0,98797	0,20562	6,03897	0,38995				
M-SVD	0,81237	0,06974	22,18878	0,39411	0,98797	0,04283	32,46851	0,38995				
PSNRHVS	-0,81237	0,15555	27,34557	0,39411	-0,98797	0,12567	25,66469	0,38995				
PSNRHVSM	-0,81237	0,14740	28,41241	0,39411	-0,98797	0,08117	30,60296	0,38995				
VSNR	-0,81237	0,12523	26,12798	0,39411	-0,98797	0,09711	24,74562	0,38995				
MSSIM	-0,81237	7,42861	0,86687	0,39411	-0,98797	5,45139	0,85708	0,38995				
R-SVD	0,81237	5,47580	0,49897	0,39411	0,98797	17,60429	0,28331	0,38995				
RFSIM	-0,81237	3,84595	0,54877	0,39411	-0,98797	3,40082	0,55846	0,38995				

ملخـص : يعتبر قياس جودة الصور عنصرا بالغ الأهمية في كل من الخدمات المرئية و كذا أنظمة معالجة الصور التي تستهدف مراقبين بشرا. الهدف الأول لهذه الأطروحة هو تقديم تقييم كمي كامل و شامل للأداء التنبؤي لمجموعة متنوعة من المعايير الموضوعية لجودة الصور كاملة المرجع تم اختيارها على أساس كونها الأكثر استعمالا، و قد طبق منهج التقييم على ستة قواعد للبيانات مخصصة لتقييم جودة الصور، وهي مصنفة حسب تقدير ذاتي غير موضوعي و يمكن تحميلها على الانترنت. الهدف الثاني هو تحديد مميزات الصور الأكثر ملائمة لتقييم جودتها، و قد استخدمت طريقتان لاختيار المميزات بما في ذلك طريقة التقليل من المخاطر الهيكلية و كذا المناهج القائمة على الشبكات العصبية. الهدف الثالث لهذا البحث هو استغلال تقنيات التعلم الآلي تحت الإشراف و لاسيما نموذج البرسبترون متعدد الطبقات للتقدير الآلي لجودة الصور. النظام المبتكر يتعلم من خلال درجات الجودة الذاتية و يمكنه بناء نموذج قادر على إنتاج قياسات موضوعية تتوافق مع تلك الذاتية لأي صورة كانت يتم تقديمها له. الهدف الرئيسي من هذا التطوير هو تحسين الأداء التنبؤي وفقا للترابط، الرتابة و الدقة، و لتحقيقه تم استخدام دالة التكلفة الافتراضية القائمة على أساس الخطأ لتصميم المعيار الأول (الذي سميناه ECF)، ثم قمنا بتخصيص هذه الدالة على أساس الترابط لتطوير قياس جديد ثان (سميناه CCF). بعد مقارنة هذين المقياسين بالنسبة لثمانية عشر خوارزميا آخر خلال ثلاثة قواعد بيانية تبين أن خوارزميي ECF وCCF يأخذان بعين الاعتبار الطبيعة اللاخطية للنظام البصري البشري.ECF هو أكثر دقة من معظم القياسات قيد الدراسة في حين أن CCF يتفوق على جميعها من حيث الترابط و بالتالي الرتابة.

Résumé : L'évaluation de la qualité d'image présente un intérêt substantiel pour les services ainsi que pour les systèmes de traitement d'images où le dernier maillon de la chaîne est l'observateur humain. Le premier objectif de cette thèse est de fournir une évaluation statistique complète et approfondie des performances prédictives d'une large variété de mesures objectives de qualité avec référence complète sur un certain nombre de bases de données étiquetées avec des scores indiquant la qualité subjective des images. Le second objectif consiste à définir les attributs de l'image les plus pertinents pour l'évaluation de sa qualité. Deux méthodes de sélection de caractéristiques ont été utilisées, à savoir la minimisation du risque structurel et l'approche basée sur le modèle connexionniste. Le troisième objectif de ce travail de recherche est d'exploiter les techniques d'apprentissage supervisé, en particulier le modèle du perceptron multicouche, pour l'estimation automatique de la qualité de l'image. Le système apprend à partir des étiquettes de la qualité subjective et construit un modèle capable de continuer à fournir une mesure objective toujours corresponde à l'avis de l'homme à toute image qui lui est présentée. Le but principal était d'optimiser la performance prédictive des mesures développées en fonction de la corrélation, la monotonicité et la précision. La fonction de coût par défaut basée sur l'erreur a été employée pour la première mesure développée (que nous avons appelé ECF) et une fonction de coût personnalisée basée sur la corrélation a été proposée pour concevoir la deuxième mesure (que nous avons appelé le CCF). L'étude comparative de ces deux nouvelles métriques à dix-huit autres algorithmes de qualité d'image avec référence complète sur trois bases de données de qualité d'image montre que les algorithmes d'ECF et CCF prennent en considération les non-linéarités du système visuel humain. L'ECF est plus précise que la majorité des mesures étudiées, tandis que la CCF améliore largement les résultats de toutes les métriques concurrentes en termes de corrélation et de monotonicité.

Abstract: Image quality assessment presents a substantial interest for image services that target human observers. The first objective of this thesis is to provide a complete and thorough statistical predictive performance assessment of a variety of full-reference objective quality measures over number of subjectively rated image quality databases. The second is to define the image attributes that are the most relevant to its quality evaluation. Two feature selection methods have been used including the structural risk minimization and the neural network based approaches. The third objective of this research work is to exploit the supervised machine learning techniques, especially the multilayer perceptron based model, for automatic image quality appreciation. The system learns from the subjective quality scores and builds a model capable to further provide an objective measure that continues to match with the human opinion to any other image. The main target was to optimize the predictive performance of the developed measures according to correlation, monotonicity and accuracy. The default cost function based on error was employed for the first developed measure (that we called ECF) and a customized cost function based on correlation was proposed to design the second metric (that we called CCF). The comparative investigation to eighteen other full-reference image quality algorithms over three image quality databases shows that both ECF and CCF take into consideration the nonlinearities of the human visual system. The ECF is more accurate than the majority of the metrics under study, while the CCF outperforms all its counterparts in terms of correlation and hence monotonicity.