TAMPERE UNIVERSITY OF TECHNOLOGY
Degree Programme in Information Technology

Md. Asadul Haque

Human Jury Assessment of Image Quality as a Measurement: Modeling with Bayes Network

Master of Science Thesis

Examiners: Prof. Risto Ritala
and Heimo Ihalainen
Examiners and topic approved in the
Computing and Electrical Engineering
Faculty Council meeting on 19th August, 2009

We don't live in a world of reality
we live in a world of perceptions.

Gerald J. Simmons

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Information Technology
Haque, Md. Asadul: Human Jury Assessment of Image Quality as a Measurement: Modeling with Bayes Network
Master of Science Thesis, 44 pages, 6 Appendix pages
August 2009
Major: Signal Processing
Examiner: Professor Risto Ritala, Heimo Ihalainen

Keywords: Human Jury Assessment, Bayes Network, Image Quality, Instrumental Measurement

Image quality assessment has been done previously manually by human jury assessment as reference. Due to lack of rationality in human jury voting and its high costs it is desirable to replace it with instrumental measurements that can predict jury assessment reliably. But high uncertainty in jury assessments and sensitivity of image context make it cumbersome for the instrumental measurements. Previous research has shown that modeling with a Bayesian network can resolve some of the problems.

A Bayesian network is a belief network of causal model representation of multivariate probabilistic distributions that describes the relationships between the interacting nodes in the form of conditional independency. By conditioning and marginalization operations we can estimate the conditional probabilities of unmeasured elements and their uncertainty in Bayes network. In this thesis we have considered a four-layer pre-existing Bayes network consisting of both qualitative and quantitative component and we have tried to assess probabilities of quality elements assessed by jurors based on instrumental measurement values. To analyze and to quantify the relationship between perceptual quality elements and instrumental measurements, we have calculated mutual information from our provided data set. Based on mutual information calculation and Kullback-Leibler distance measure we have investigated the sensitivity of the network, and we have tried to validate a feasible network model where network parameters have been selected such a way that it minimizes the uncertainties of our chosen Bayes network.

PREFACE

This Master of Science thesis work was carried out in the Tampere University of Technology (TUT) at the Department of Automation Science and Engineering. The thesis is a part of the DigiQ research project, which was a collaboration of four universities, TUT, Department of Psychology in University of Helsinki, the Laboratory of Media Technology in the Helsinki University of Technology and the Laboratory of Information Processing in Lappeenranta University of Technology.

I'm immensely grateful to my supervisor Prof. Risto Ritala for giving me the opportunity to work on such an interesting topic, for his constant support and his patience due to my slowness in research progress. I'm very much indebted to PhD. researcher Marja Mettänen who supported me and guided me throughout my work. I'm also very grateful to Heimo Ihalainen for his useful advices and providing me necessary books. Since it was my first independent research literally, I could never complete this research work without their help.

Finally I express my gratitude to all the staff of the Department for providing me such nice working environment.

Asadul Haque

Tekniikankatu 14 F 272

33720 Tampere, Finland

CONTENTS

LIST OF FIGURES

LIST OF TABLES

TERMS AND ABBREVIATIONS

HVS	Human visual system
MI	Mutual information
BN	Bayes network
VQI	Visual quality index
MOS	Mean opinion score
JND	Just noticeable difference
CDF	Cumulative distribution function
NSRPJ	Normalized sum of rational partial juries
ρ (rho)	Spearman's rank correlation coefficient or Spearman's rho
MSE	Mean squared error
PSNR	Peak signal to noise ratio
RGB	Red, green and blue colors
CMYK	Cyan, Magenta, Yellow and Black colors
HSI	Hue, saturation and intensity
HSV	Hue-saturation-value color model
DAG	Directed acyclic graph
PQE	Perceptual quality elements
KL	Kullback-Leibler divergence
PDF	Probability density function
MLE	Maximum Likelihood Estimation
CIELAB	Commission Internationale de l'Éclairage L*a*b*

1. Introduction

1.1. Background

Digital images are a powerful and efficient means for communicating information. The quality of images is affected by attributes such as noise, color, resolution and sharpness. That is why, it is important to develop effective image quality assessment models, which will enable us to monitor the quality of images.

Assessment of image quality conceptually is a subjective matter. It is rooted in both the objective properties of an image and the psychological processes of perception. Perception is the construction of an internal representation of the image using primarily low-level knowledge of the visual world. It largely depends on the state of mind of evaluator, age, cultural background etc. It also depends on how intensely the person's eyes perceive the colors of the image. Due to the multidimensional quantity of image quality the evaluator faces difficulty in choosing the quality attributes that need to be considered while assessing the image quality.
Human visual system responds to a limited range of spatiotemporal frequencies of colors. Another fact is, there is no perfect human vision model until now proposed by scientists, which can be used in constructing image quality model.

Hence, measuring image quality is a laborious task in general. Researchers have been working continuously to evaluate the printed image quality for the imaging industries. Until recently assessment of image quality has been based on human jury assessment, which is done by a group of evaluators. This group is called a jury. Since jury assessment is costly, cumbersome and finally time consuming, replacing the jury assessment with a set of instrumental measurements that can produce some numerical values of image quality might be worthy. Furthermore, those numerical values can be used by machine for evaluating the quality of images or visual print quality automatically and faster than human jury.

Due to the complex form of dependency between the instrumental measurement and the jury assessment, previous research has suggested modeling both measurement and jury assessment

with a Bayesian network. Bayesian networks is nowadays a widely used method for analyzing knowledge with uncertainty and efficient reasoning. The grading of instrumental measurements is directly connected with the grading of perceptual quality elements. In other words, it implies the objective instrumental measurements can be used to predict image quality performance by the evaluators.

This research is the continuation of Mr. Johannes Pulla's Master of Science thesis work about Jury assessment as Reference for Instrumental Measurements of Image Quality [1].

1.2 Goals of this Research

The main objectives of this thesis are expressed by the following:

- Estimation of unmeasured elements and their uncertainty in the form of probabilities in Bayes network.
- Identifying the parameters or the conditional probability estimates in the network submodels.
- Selecting the network structure or, from which submodels the network model is composed of.

To fulfill those objectives we have chosen an existing Bayes network structure with certain parameters and have tried to find mutual information (MI) between instrumental measurements and each of the quality elements separately. We have also tried to find alternative Bayes network structure using that mutual information. We have synthesized our existing network to understand statistical dependency based on the Mutual Information (MI) measurement between instrumental measurements and perceptual quality elements.

After that we have simulated the Bayes network to generate network input elements (in this case perceptual quality elements in the network) of image quality assessment and instrumental measurements so that we can understand how those elements are affecting the network. The ultimate goal of this research is to develop a robust model, which can predict the overall quality from the perception of the individual attributes in the multivariate environment.

An image quality estimator can be constructed by means of instrumental measurements, low level attributes and high level attributes of an image. In the estimator the Jury assessment test case data is collected from the low level or perceptual quality elements and high level attributes of the image. Based on this model we can identify essential parameters, which correspond to image quality assessment. Later a Bayesian model can be constructed according to the mutual information results computed from the test case data. The following figure 1.1 shows a simple estimator for visual image quality index.

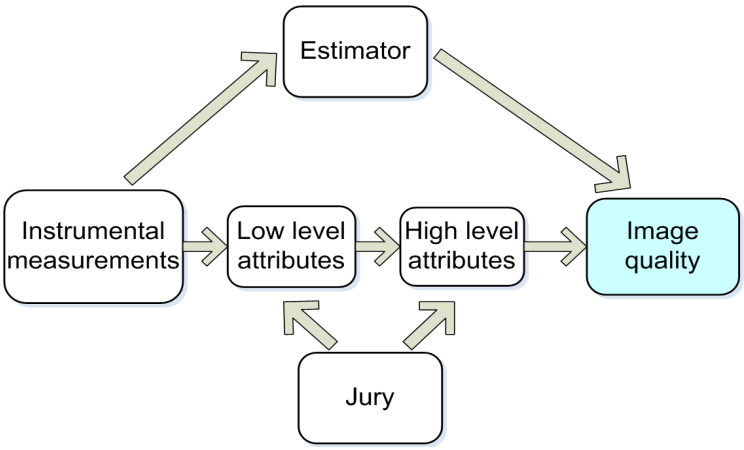

Figure 1.1: The estimator of visual quality index.

1.3 Structure of this Thesis

This thesis is organized as follows. Chapter 2 gives some details about various methods for image quality assessment. Chapter 3 discusses the modeling of image quality estimator, where Bayes Networks, Mutual information and Sensitivity of Networks have been discussed. In chapter 4, case study has been considered with the manipulation of network model and few experiments were carried out on it. Chapter 5 examines the method and analyzes the mutual information obtained during the experiment with proposed Bayes network. Chapter 6 concludes the thesis with some discussion about the results and future works can be done on it.

2. Methods for Image Quality Assessment

2.1 Subjective Image Quality Measurements

Human beings are the end users of all images. The quality of images is assessed by people looking at images, not by electronic measurements or other machine determined parameters. For this reason, humans are the effective means for assessing the quality of images. Paired comparison, categorical judgment, anchored scaling are some of subjective evaluation methods often used in assessing image quality. Pair comparison is the evaluation of determining small quality differences between pairs of sample images. Categorical judgment is most suitable for the large sample sets of images.

Rank ordering is the method of assessing image quality where evaluator arranges the test samples into a certain order according to the ordering criteria. Mean opinion score (MOS) gives a numerical indication of image quality where human observers are required to evaluate the subjective image quality. The group of evaluators is called jury. The jury can consist of a homogeneous group of people to minimize the variation in answers while assessing the image quality.

In spite of choosing homogenous group of people as jury, this method has several shortcomings, like lack of rationality in jury voting and being not cost-effective in real-world applications.

2.1.1 Just Noticeable Differences

The concept of just noticeable difference (JND) is very useful in image quality characterization and prediction. It is a measure of the perceptual continuum and can be presented as a probability distribution. Since paired comparison is performed between the test image and its analogous reference image, it is crucial to understand the significance of quality difference between them. Just noticeable differences are small units of change that can be used to construct calibrated numerical scales that quantify wide ranges of quality of images. JNDs are natural which provides a natural unit for quality scale calibration and making multivariate quality predictions. [2]

2.1.2 Low-level Attributes

Low-level attributes of an image are those attributes that are perceived by a human evaluator from the image at first sight. They are concrete subjective attributes that relate to simpler physical properties of an image. For example, sharpness is a low-level attribute of an image, which enables viewers to render fine details from it. Clarity, graininess, brightness, colorfulness are other low-level attributes which are considered in the model we have experimented here. These low-level attributes we have used to manipulate our images. The low-level attributes are subjective and cannot be measured instrumentally with only one measurement. [3]

2.1.3 High-level Attributes

The quality of an image is determined by the degree to which the image is both useful and natural from the observation point of view. High-level quality consists of the attributes naturalness and usefulness. They both are abstract aspects of the image and together compose the overall quality.

The usefulness of an image refers to the precision of the internal representation of the image and naturalness refers to the degree of apparent similarity between the internal references and the reproduced image color, environment and their perception without any distortions. Naturalness can be assessed by the mental recollection of the colors of familiar objects in color reproduction. These higher-level subjective attributes are considered as the abstract level characteristics of an image. They are the psychological attributes that largely depend on the lower level attributes and very subjective in nature. High-level attributes are used to reason meaning of the low level attributes for the quality rating, their frequencies, description and image quality concepts that resemble those attributes. [4]

2.1.4 Visual Quality Index

There are many attributes that influence the overall quality of images in the network. Finding significant dependency among the attributes is crucial, because it will help us to construct a feasible network model that combines all necessary image quality attributes. To meet this goal, it might be easy to understand the interrelation between each pair of attributes. But in the case of many attribute variables, it becomes burdensome to realize how all those attributes interact with each other and how they contribute to overall image quality. In the network the higher-level attributes usefulness and naturalness have formed the overall image quality, which is regarded as visual quality index (VQI). It is simply an ordering index. Visual quality index is derived from the instrumentally measured image quality value and the context of the image. It utilizes the knowledge of the human visual system (HVS) to a lesser or higher extent in order to increase the correlation with human judgment. Visual quality index (VQI) is the ultimate evaluation of an image, which can be used for quality control systems. [5]

2.1.5 Jury Rationality

Empirical evidence shows that the conception of human rationality is somewhat inaccurate. Human made decision is heavily influenced by emotions such as attention, elation, grief, lust, sympathy, anxiety etc. Emotion is largely cognitive in origin and it affects a goal of the human agent. So decisions caused by emotional states can be irrational.

Human jury rationality is seen as the consistency of the assessment made by a jury. The reason in using jury as a reference is rationality, although the ambiguity of different image quality attributes, subjectivity of perceived quality and lack of consistency are inherent in the image quality assessments made by a jury. Jury rationality can be measured by the fraction of rational partial juries, which is normalized sum of rational partial juries (NSRPJ). For a jury of *N* evaluators and the partial jury size n_p evaluators ($n_p \leq N$) the number of partial jury combinations is

$$S_p = \binom{N}{n_p} = \frac{N!}{n_p!(N-n_p)!}$$

With the increase of number of evaluators, the computational complexity increases too.

The result of instrumental measurements can be predicted based on jury assessment results. Jury assessment data can be interpreted in terms of discrete probability distributions and from there we can find how much of predicted jury assessment changes when we make small change in instrumental data. [6]

2.1.6 Overall Quality prediction from Image Quality Attributes

The term image quality describes the overall visual impression of an image that is derived from multiple components of human eye's perception, for example sharpness, brightness, noisiness, contrast, colorfulness etc. Overall image quality can be defined by various quality related perceptions, and each quality related perception is referred to as an image quality attribute. Since image quality model falls within a perceptual framework, while developing a method for predicting the overall image quality, it is essential to take into account all those important perceptual quality attributes. Prediction of multivariate image quality from associated quality attributes is quite cumbersome and requires understanding well the nature of interactions between attributes.

The overall quality of an image is affected by a set of attributes. Moreover, the presence of one attribute influences the perception of another attribute. It is difficult to develop an underlying theory explaining all interactions, since interactions between attributes may be expected to be specific to each pair of attributes or in some cases even to higher-order interactions. Although as a result of interactions between attributes there may be changes in the appearance of images, often such changes do not affect image quality. In reality, the interaction of perceptual attributes is less pervasive and a model can be constructed based on experiments involving combinations of attributes when an interaction is significant. Another fact is, especially among the artifactual attributes significant interactions between carefully defined attributes are quite uncommon. The

results of interaction among the image quality attributes from the different investigations may be integrated into a unified model for predicting overall image quality with a reasonable expectation. [7]

2.1.7 Rank Correlation

Rank correlation is a means to measure the relationship between rankings of different ordinal variables where a ranking is assigned as “first”, “second”, “third” etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and tells us how significant their relation is.

Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter ρ (rho), is a nonparametric measure of statistical dependence between two variables

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where d is the difference between ranks and n is the sample size.

It assesses how well the relationship between two variables can be described using a monotonic function. A monotonic function is a function which is always either entirely non-increasing or non-decreasing. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Spearman's coefficient is appropriate for both continuous and discrete variables, including ordinal variables. The difference between the Pearson correlation and the Spearman correlation is that the Pearson is most appropriate for measurements taken from an interval scale, it measures the strength of a linear relationship between paired data, on the other hand, the Spearman is more appropriate for measurements taken from ordinal scales. [8]

2.2 Objective Image Quality Measurements

The goal of the objective image quality measurement is to measure the errors or signal difference between the distorted and reference images. Objective measure provides an analytical result by identifying the sources of artifacts in the images. Designing a computational model that can predict perceived image quality accurately is not an easy task. There exists no proper psycho-visual model that can model numerically the visual error sensitivity features of human visual system. One easily computable measure of images is mean squared error (MSE) or peak signal to noise ratio (PSNR). If x is an original image and y is a distorted image, then the MSE and PSNR are respectively:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

$$\text{PSNR} = 10\ \log_{10}\frac{L^2}{MSE}$$

Here L is the dynamic range of image pixel intensities. In many cases, PSNR is a poor indicator of subjective image quality and does not correlate well with perceived quality measurement. MSE and PSNR are two common methods used to assess the quality of the distorted image. [9]

2.2.1 Full-Reference, No-Reference and Reduced-Reference Image Quality Measures

Three types of knowledge can be used for the design of image quality measure: knowledge based on the original image, knowledge from the distortion process of the image, and knowledge about the Human Visual System (HVS). Those classifications are called full-reference, no-reference and reduced-reference.

Measures that require both the original image and the distorted image are called full-reference methods. The availability of an original image is considered to be distortion-free or perfect quality and it can be used as a reference in evaluating a distorted image. Due to compression, acquisition process, transmission through noisy channels image can suffer distortion. Digital images can also undergo quality improvement processes, like enhancement, restoration

techniques. In each step it is necessary to quantify the quality of the resulting image. One easy way to do it is by using a full-reference image to carry out this task. In most of the proposed objective quality measures we assume that the undistorted original reference image exists and is fully available. [10]

Measures that do not require the original image are called no-reference methods. Sometimes it is necessary to develop objective quality assessment that correlates well with human perception without the reference image or no-reference. In many practical applications an image quality assessment system does not have access to the reference images. So in no-reference objective image quality assessment we try to construct a computational model that can predict the human-perceived quality of distorted images accurately and automatically without any prior knowledge of reference images. In some cases no-reference image quality assessment can be difficult. [11]

Measures that require both the distorted image and the partial information about the original image are called reduced-reference methods. Reduced-reference image quality assessment can be considered a solution between full-reference methods and no-reference methods. In reduced-reference methods the reference image is not fully available. Instead, the system includes certain features at the sender side that are extracted from the reference image and a feature extraction, comparison process and quality analysis process at the receiver side to evaluate the quality of the distorted image. The extracted features describing the reference image are transmitted to the receiver as side information through an ancillary channel. At the receiver side we compare distorted image and extracted features by using reduced-reference quality analysis method. Available bandwidth for transmitting the side information is an important parameter in a reduced-reference system. The reduced-reference system must select the most effective and efficient features to optimize image quality prediction accuracy under the constraint of the available bandwidth. The following figure shows the framework used for reduced reference image quality assessment metric. [12]

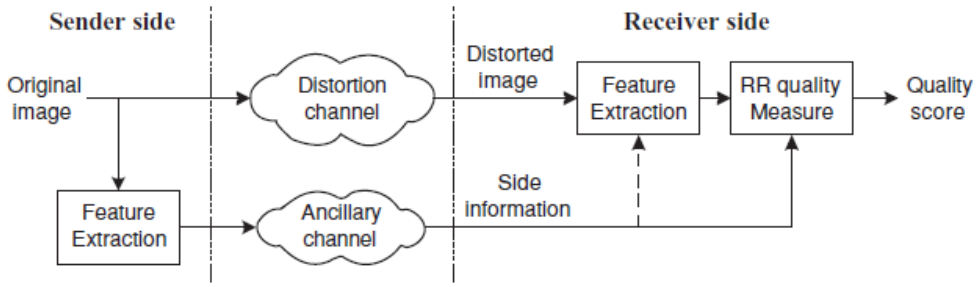

Figure 2.1: Diagram of a reduced-reference image quality assessment system

2.2.2 RGB and CMYK Color Spaces

RGB stands for red, green and blue colors. Merging these three primary colors can produce all other human-perceivable colors of the visible spectrum. CMYK stand for Cyan, Magenta, Yellow and Black colors. RGB and CMYK are two different color spaces.

Most scanners, digital cameras, and video camera save files as RGB and the conversion of RGB files to CMYK can be done in many ways.

Printed colors are produced from cyan, magenta and yellow printing inks by subtracting varying degrees of red, green and blue from white light to produce a selective gamut of spectral colors. RGB gamut is larger than the CMYK gamut. Conversion between RGB and CMY is performed by the following equation:

$$R = 255 \times (1-C) \times (1-K)$$

$$G = 255 \times (1-M) \times (1-K)$$

$$B = 255 \times (1-Y) \times (1-K)$$

Printing inks are not perfect reflectors of RGB colors. So as a result, this subtractive system can't duplicate all colors displayed on a computer screen. In other words, printers can't print pure red, green and blue using the CMYK system. [13]

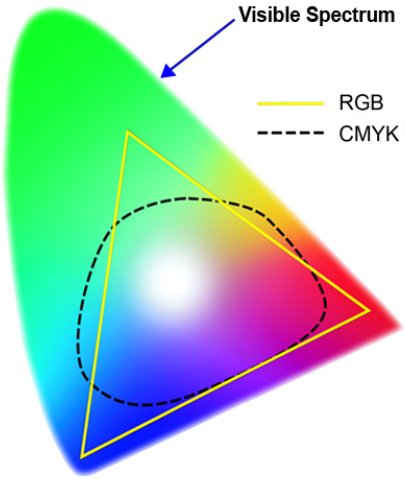

Figure 2.2: RGB and CMYK Color Models.

2.2.3 HSI and CIELAB Color Spaces

RGB and CMY are not well suited for describing colors that are practical for human interpretation. Hue, saturation, and brightness or intensity (HSI) are aspects of color in the red, green, and blue scheme. All possible colors can be specified according to hue, saturation, and brightness. Hue is a color attribute that described a pure color. Saturation gives a measure of the degree to which a pure color is diluted by white light. It is an expression for the relative bandwidth of the visible output from a light source.

Brightness is a relative expression of the intensity of the energy output of a visible light source. It can be expressed as a total energy value where the intensity is greatest. It is a key factor in describing color sensation.

CIELAB is a color space that scientifically describes how the average human eye sees color. It is proposed by the CIE (Commission Internationale de I'Éclairage L*a*b*). It is a three-dimensional uniform color space, which describes all the colors visible to the human eye.

In the CIELAB color space depicted below the L* axis runs from top to bottom. The maximum for L* is 100, which represents a perfect reflecting diffuser. The minimum for L* is zero, which represents black. The a* and b* axes have no specific numerical limits. Positive a* is red, negative a* is green, positive b* is yellow and negative b* is blue. [14]

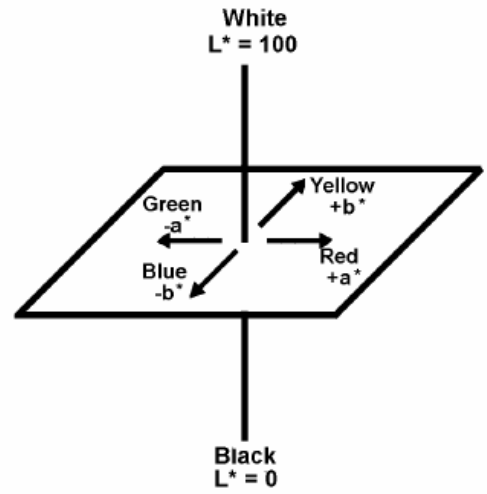

Figure 2.3: A diagram representing the CIELAB color space.

2.2.4 Instrumental Measurement Quantities

The visual appearance of an image is aesthetic characteristic, which is judged by humans based on its gloss, color, brightness, smoothness and some other characteristics. The appearance of an image is usually a property of the surface i.e. smoothness, reflectivity, noisiness, color etc. which can be measured instrumentally. Instrument measurements of appearance are objective and have several advantages. It produces a quantifiable measure of performance or same numeric values every time, which can be used for designing statistical model.

Dot quality, line quality, mottling and color quality are some of the elements of instrumental measurement quantities.

Quantification of dot quality refers to its shape, size, position etc. Quantifying line quality refers to its average width, width variation, raggedness, sharpness etc. Mottling and color quality refer to the measurement of luminance variations and CIE L* a* b* measurements respectively.

In the experimental model, we have considered low-pass filtering, noise and HSV saturation as instrumental measurement quantities.

3. Modeling of Image Quality Estimator

The ultimate goal of modeling an image estimator is to make prediction of final image quality. The goal is achieved in terms of components and subsystem properties described in the system. Since subjective and objective measures of images are the only principle components used constructing such estimator, both measures should be chosen correctly, so the overall system is viable. A practical mathematical method need to be applied that can quantify accurately the propagation of quality from component and subsystem properties through the network. We investigate the robustness of Bayesian networks against parameter changes. In Bayesian experimental analysis our results are continually revised in light of new evidence on the basis of Bayes theorem.

3.1 Bayes Networks

A Bayesian network or belief network is a causal probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph. The network consists of two distinct parts: a directed acyclic graph (DAG) and a set of parameters for the DAG. The DAG in a Bayesian network can be used to represent causal relationships among a set of random variables. Its nodes represent random variables and arcs represent direct probabilistic dependencies on its predecessors. Nodes with no predecessors are described by prior probability distributions.

A directed acyclic graph offers a simple and unique rule for expanding the joint probability in terms of simple conditional probabilities. Graphical models can represent conditional independence relationships efficiently among the variables that are directly related with each other. These parameter variables relationships can be estimated in the form of joint probability distribution easily later on. Because of causal structure, it gives a useful, modular insight into the interactions among variables and allows for prediction of the effects of external manipulation. Already there are many learning algorithms for automatically building Bayesian networks from a data set and some of them are based on testing conditional independences. [15]

Under the condition of uncertainty Bayesian network tool can be quite useful. This network can be used efficiently for modeling image quality estimator and so far the result seems promising.

Figure 3.1 presents the graphical scheme of Bayes network for assessing the visual quality of images we have been using in the experiment.

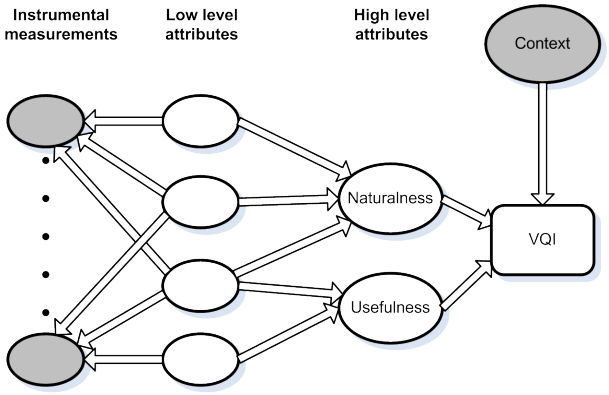

Figure 3.1: The scheme of graphical Bayes network model.

3.2 The Likelihood Function and Probability Density Function

The Likelihood function is the joint probability density function of observable random variables, which are regarded as the function of the parameters given the realized random variables.
Let F(x) be the distribution function for a continuous random variable X. The probability density function (PDF) for X is given by

$$f(x) = \frac{dF(x)}{dx}$$

wherever the derivative of F(x) exists.
F(x) is a non-decreasing function of x. That means its derivative is f(x) is always nonnegative.

The probability density function is discrete, whereas the likelihood function is continuous. Probability density function is a function of the data where the value of the parameter is fixed; on the other hand the likelihood function is a function of the parameter where the data is fixed. [16]

3.3 The Maximum Likelihood Estimation

The maximum likelihood estimate is the procedure of finding values of the parameter that maximize the sample likelihood or makes the observed data most likely. It provides a consistent approach to parameter estimation problems and can be developed for a large variety of estimation situations. Maximum likelihood methods have reasonable intuitive statistical and optimality properties. They become minimum variance unbiased estimators as the sample size of data increases. Once we have derived maximum-likelihood estimator, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference. This method is very widely applicable and is simple to apply. [17]

3.4 The Prior Probability and Posterior Probability

In Bayesian statistical inference a prior probability distribution is an initial probability value originally obtained before any additional information is obtained or some evidence is taken into account. The prior distribution is the probability distribution that the person has before observing.

A posterior probability is a probability value that has been revised by using additional information that is later obtained. Prior probabilities are the original probabilities of an outcome, which we can update with new information to create posterior probabilities.

We calculate posterior probability by updating the prior probability using Bayes theorem. In Bayes theorem, the posterior probability is the probability of event A occurring given that event B has occurred. The formula is as follows:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

where:

P(A) is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.

P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

P(B|A) is the conditional probability of B given A. It is also called the likelihood.

P(B) is the prior or marginal probability of B, and acts as a normalizing constant. [18]

3.5 Mutual Information

The value of information analysis in Bayesian network is based on the concept of mutual information, entropy and information gains. Mutual information is a quantity that measures the mutual dependency of two variables. It is a dimensionless quantity, which reduces the uncertainty about one random variable given knowledge of another. High mutual information means a large reduction in uncertainty, low mutual information indicates a small reduction and zero mutual information between two random variables means the variables are independent.

The mathematical representation for mutual information of the random variables X and Y are as follows:

$$I(X;Y) = \sum_{y\epsilon Y} \sum_{x\epsilon X} p(x,y) \log(\frac{p(x,y)}{p_1(x)p_2(y)})$$

Where, $p(x,y)$ is the joint probability distribution function of X and Y respectively. $p_1(x)$ is the marginal probability distribution function of X and $p_2(y)$ is the marginal probability distribution function of Y. Mutual information and Pearson correlation are used for choosing edges between nodes in a Bayesian model. [19]

3.6 Kullback-Leibler divergence

Kullback-Leibler divergence is a natural measure of the distance between two probability distributions. This is a discriminant function, which is intimately related to mutual information. For any two distributions $P(z)$ and $Q(z)$, it is defined as follows:

$$D_{KL}(P(z)\|Q(z)) \equiv \sum_z P(z) \log\left[\frac{P(z)}{Q(z)}\right]$$

Kullback-Leibler divergence has two essential properties:

$D_{KL}(P,Q) \geq 0$ for all distributions of P and Q

$D_{KL}(P,Q) = 0$ if and only if P = Q

We measure the closeness of the two distributions with Kullback-Leibler (KL) divergence.

Kullback-Leibler can be used to determine how far away a probability distribution P is from another distribution Q. That means Kullback-Leibler divergence can be used as a measure of the information gain in moving from a prior distribution to a posterior distribution in modeling Bayesian estimator. The goal in modeling Bayesian estimator is to maximize the expected Kullback-Leibler divergence between the prior and the posterior.

But this divergence is not a real distance measure, because it is not symmetric. Another property of Kullback-Leibler divergence is it is always non-negative and it does not satisfy the triangle inequality. [20]

3.7 Sensitivity of Networks

One goal of our present study of the Bayesian network is to quantify experimentally how much can be learned from the least amount of data available to construct a feasible network. The Bayesian model we have experimented here, instrumental measurement nodes have three states and all other nodes have five states. Sensitivity analysis refers to identifying the most important parameters, which have maximum impact in the Bayes network so that unimportant parameters can be discarded in the network. Sensitivity analysis in Bayesian networks is broadly concerned with understanding the relationship between local network parameters and global conclusions drawn based on the network.

One way of performing sensitivity analysis of Bayesian networks is to compute the conditional probabilities of a target node in the network when some evidence values of nodes are available and to observe how sensitive these conditional probabilities are to small changes in the parameters or evidence values. Of course not all parameters are equally sensitive since they all have different effects on the network's performance. In some cases a network can be very sensitive to small parameter changes.

In a one-way sensitivity analysis, the values of a single parameter are changed one at a time to compute the conditional probabilities of a target node, keeping the values of all other parameters fixed. Single parameter changes are easy to compute and visualize the effects in the network. [21]

In a two-way sensitivity analysis of a probabilistic network, two parameters are varied simultaneously to see the joint effect of their variation on a probability of interest. It is also possible to change more than two parameters at the same time, though it is hard to interpret such manipulation in sensitivity analysis. Multiple parameter changes can be more meaningful, and may disturb the probability distribution less significantly than single parameter changes. [22]

In a Bayesian network identifying most important parameters from the huge number of probability parameters is cumbersome and may require quite large data sets in order to learn accurate parameter estimates. An exponential number of conditional independence tests are required in most dependency-analysis based Bayesian network algorithms. In data-mining applications it is very common to have hundreds of variables in the data sets. A mathematical

function, namely, sensitivity function, can be used to express the sensitive change in posterior probability of the target query due to the variation of a Bayesian network's probability parameters. [23]

3.8 An Example of Bayes Network

The directed edges of a Bayesian network describe the probabilistic relations between the nodes. For doing the experiment of sensitivity analysis, here we have considered a probability distribution of several variables. The table below is a conditional probability of X1, given the prior of X2.

X2
0.1
0.3
0.4
0.2

X1 / X2	1	2	3	4
1	0.9	0.05	0.05	0
2	0.05	0.9	0.05	0
3	0	0.05	0.9	0.05
4	0	0.05	0.05	0.9

Table 3.1: The conditional probability of X1, given the prior of X2.

X3 / X2	1	2	3	4
1	0.9	0	0.05	0.05
2	0.05	0.05	0.9	0
3	0.05	0	0.9	0.05
4	0.05	0	0.05	0.9

Table 3.2: The conditional probability of X3, given X2.

X4 / X3	1	2	3	4
1	0	0.05	0.9	0.05
2	0.05	0	0.05	0.9
3	0.05	0.9	0.05	0
4	0.9	0	0.05	0.05

Table 3.3: The conditional probability of X4, given X3.

We have considered four nodes X1, X2, X3 and X4 where X1, X3 and X4 are conditioned on X2 for a given prior of X2. The joint probability of all four variables is:

$$P(X1,X2,X3,X4) = P(X1|X2)P(X2)P(X3|X2)P(X4|X3)$$

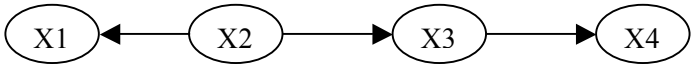

Figure 3.2: The four nodes conditional probability distribution model.

We have calculated conditional probability distribution of X1 and X2, given the probability of X2 and conditional probability distribution of X4, given the probability of X3. After that we have calculated all pair-wise mutual information. The calculated results are tabled below.

	X1 and X2	X1 and X3	X1 and X4
Mutual Information	0.8935 (nats)	0.3912 (nats)	0.2838 (nats)

Table 3.4: Mutual information values for Sensitivity Analysis Example

From the table we can see that the mutual information between X1 and X3 or X1 and X4 is smaller than the mutual information between X1 and X2. It is natural because of the structure of the network. It also shows causal conditional probabilities are easier to estimate.

4. Experiments and Case Study

To construct a probabilistic belief system, it is necessary to define a set of random variables, which will represent nodes in our Bayes network. For this we have collected some jury assessment data from a case study carried out. The idea behind the case study is to take a digital image and modify it in different ways by digital image processing. The digital image was modified in such a way that it can cover the subjective and objective quality factors as much as possible. The visual assessment test of those modified images was carried out using a monitor display. The nodes of the Bayesian networks are presented as probability distributions of image quality elements and instrumental measurements. Based on prior knowledge, a Bayes network proposed is shown in figure 4.1

4.1 Setup: manipulation

A digital image was modified by three different methods in Matlab by using image processing toolbox: low-pass filtering, noise addition and HSV saturation. These three versions of modifications have been simulated as instrumental measurements in our Bayesian model. There were three distinct degrees of modification for each method: no modification, mild, and moderate level. The combination of all modified images was used in the subjective assessment test as jury assessment data collected in trial sessions. The total number of images was twenty-seven and each image was assessed with respect to eight attributes (which have been embedded in our Bayesian model as PQE, low level and high level) on a scale from 1 to 5. The evaluators assessed one attribute at a time. That means each subject was asked to label the twenty-seven images with the grades 1-5 eight times. The evaluation data and original images are presented in Appendix A and in Appendix B.

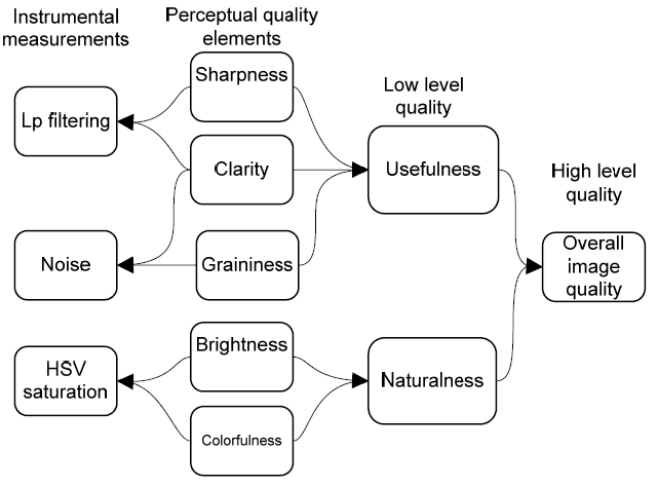

Figure 4.1. The Bayesian model based on prior knowledge.

4.2 Experiment

4.2.1 Case study I

With one image content of different modified variations the visual assessment test was conducted by six human evaluators. They were asked to give grades 1 to 5 of all those modified twenty-seven images in terms of eight attributes. So we got six sets of evaluations data from six different people for the purpose to utilize those data in our Bayesian networks as probability distributions. After that we have tried to find the causal relationship between those six sets of evaluations data and three versions of modified instrumental measurements.

For this purpose, we have computed all pair-wise mutual information between the three instrumental measurements and all the quality elements. The computed Pearson correlation values and mutual information between the attributes related in the model are listed below in the tables.

Overall Quality	Usefulness	Naturalness	Sharpness	Brightness	Colorfulness	Graininess	Clarity	Lp filtering	Noise	HSV Saturation
1.0000	0.5899	0.4252	0.5288	0.0309	0.2029	0.4330	0.5799	0.0751	-0.4160	-0.5374
0.5899	1.0000	0.5691	0.7660	0.1911	0.2069	0.3081	0.7374	0.0238	-0.7256	-0.3925
0.4252	0.5691	1.0000	0.4973	0.2648	0.4776	0.3027	0.5943	-0.0409	-0.4035	-0.3742
0.5288	0.7660	0.4973	1.0000	0.2297	0.2379	0.2251	0.6412	0.0284	-0.8392	-0.2438
0.0309	0.1911	0.2648	0.2297	1.0000	0.0665	-0.0180	0.2632	0.0172	-0.1547	-0.1203
0.2029	0.2069	0.4776	0.2379	0.0665	1.0000	0.1820	0.2365	-0.2256	-0.1722	-0.0891
0.4330	0.3081	0.3027	0.2251	-0.0180	0.1820	1.0000	0.4126	0.0208	-0.0728	-0.6085
0.5799	0.7374	0.5943	0.6412	0.2632	0.2365	0.4126	1.0000	-0.0407	-0.5349	-0.4942
0.0751	0.0238	-0.0409	0.0284	0.0172	-0.2256	0.0208	-0.0407	1.0000	0	0.0000
-0.4160	-0.7256	-0.4035	-0.8392	-0.1547	-0.1722	-0.0728	-0.5349	0	1.0000	0
-0.5374	-0.3925	-0.3742	-0.2438	-0.1203	-0.0891	-0.6085	-0.4942	0.0000	0	1.0000

Table 4.1: Pearson correlation values for Bayesian network

	Low-pass filtering (MI in nat)	Noise (MI in nat)	HSV Saturation (MI in nat)
Sharpness	0.0027	0.5633	0.0949
Clarity	0.0150	0.2101	0.1989
Graininess	0.0304	0.0319	0.3343
Brightness	0.0755	0.0441	0.0377
Colorfulness	0.1825	0.0404	0.0144
Usefulness	0.0290	0.3918	0.1157
Naturalness	0.1234	0.0999	0.1219
Overall quality	0.0156	0.1807	0.2167

Table 4.2: Mutual information values between Instrumental measurements and quality elements in Bayesian network

In the table 4.1 of Pearson correlation values we can see strong correlation (> 0.5) among Usefulness, Overall Quality, Naturalness, Sharpness and Clarity. This means that changes in one variable are strongly correlated with changes in the second variable. If one of the variables increases in value, the second variable also increases in value. Perceptual quality element Colorfulness has moderately strong correlation (> 0.2) with Overall Quality, Usefulness, Naturalness and Sharpness. In the table we can also see negative correlation values in some variables that implies if one variable increases in value, another variable decreases in value.

From the table 4.2 we can see instrumental measurement Low-pass filtering has higher mutual information (> 0.1) with Colorfulness and Naturalness. That means Colorfulness is affected by Low-pass filtering significantly. Noise has higher mutual information (> 0.2) with the Sharpness, Usefulness and Clarity. Instrumental measurement HSV saturation has higher mutual information with Graininess and Clarity. Based on the results of Pearson correlation and mutual information values we have initially proposed a Bayesian model, which is shown in figure 4.2. Compared to initial model, figure 4.1, brightness appears statistically independent from the other variables and is left out.

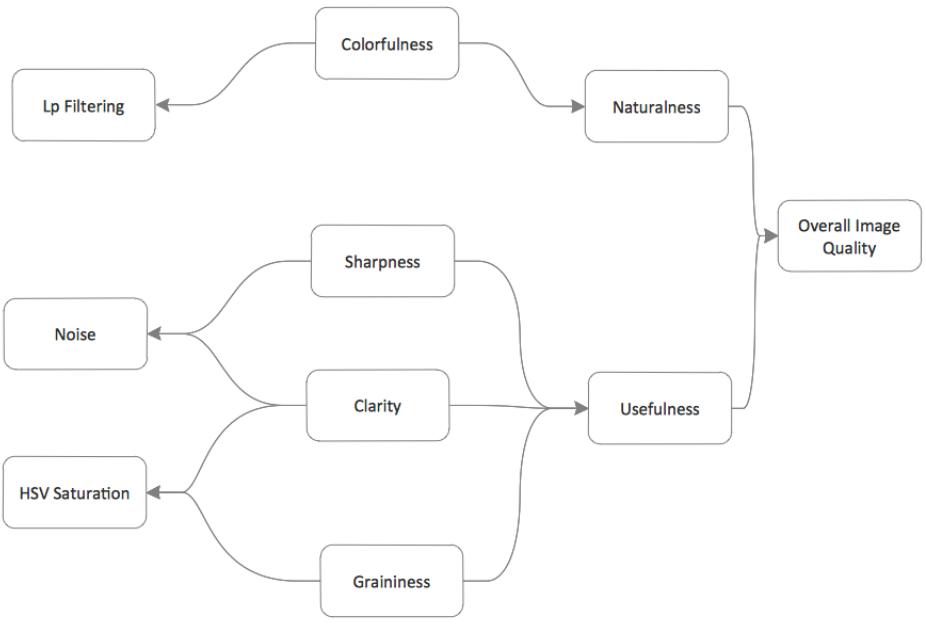

Figure 4.2. The Bayesian model constructed according to the mutual information results computed from the test case data.

However, we have not succeeded in constructing the probability models due to insufficient data and three conditioning variables of usefulness. Lack of insufficient data and more than two conditioning variables create no-model cases by means of zero probability which increases the uncertainty of our Bayesian model. No-model case means that the data does not contains all possible combinations of values of conditioning variables.

So we have decided to simplify the model in such a way where number of no-model cases is limited. For this we have decided to omit the quality variable Clarity as it's impacts are mostly covered by Sharpness and Graininess and the initial experimental results look promising.
The constructed two Bayes network models are shown in figure 4.3 and figure 4.4.

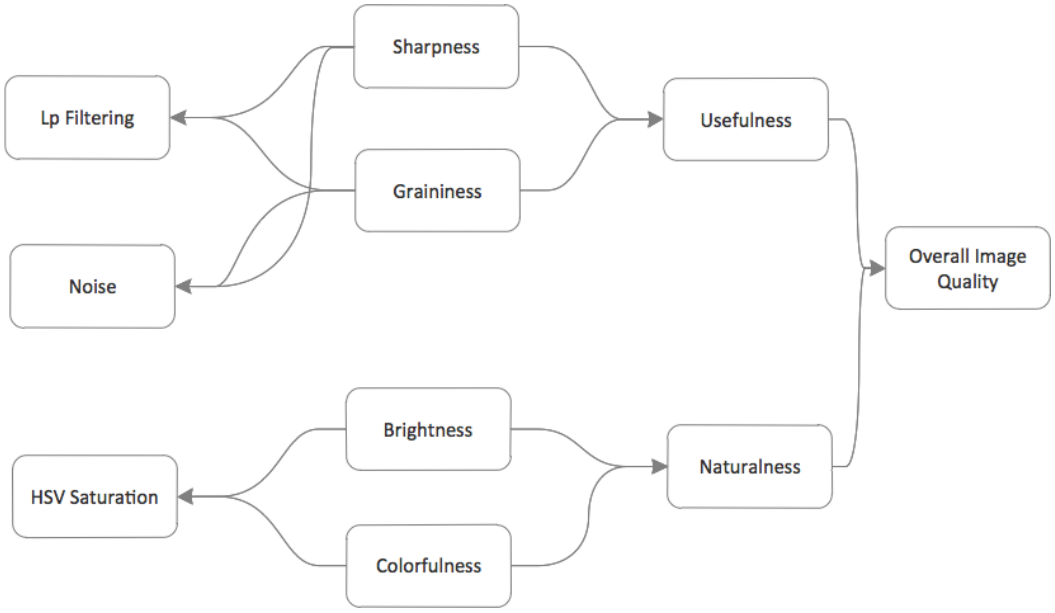

Figure 4.3. First simplified Bayesian model

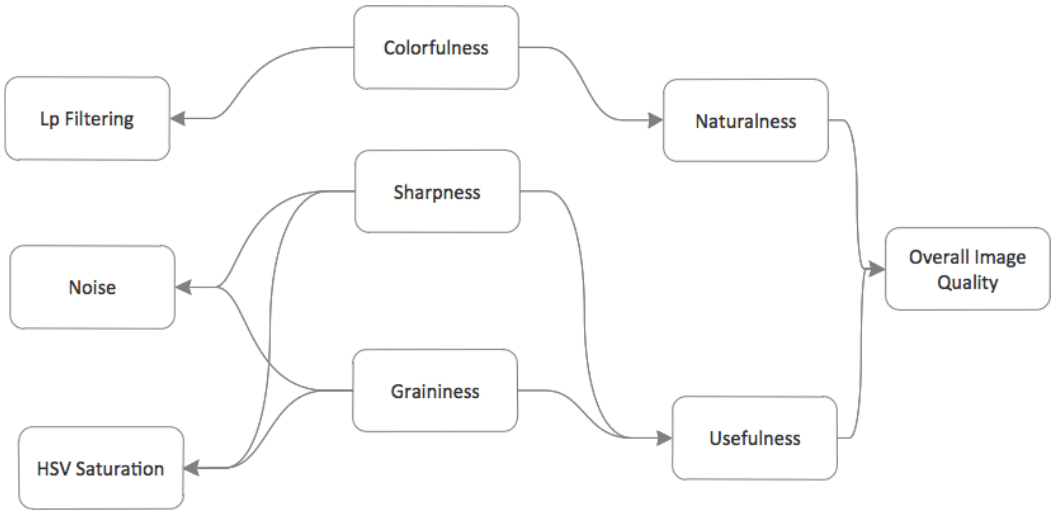

Figure 4.4. Second simplified Bayesian model

4.2.2 Case study II

After constructing two simplified Bayesian models, on these cases of models no-models cases are limited to the last one, Usefulness, Naturalness ⇒ Overall quality; the missing cases occur when either Usefulness is high but Naturalness is low and vice versa. For the missing models in both Bayesian networks we have decided to replace following probability values in the following positions (Prob_model_1_Overall(i,j,k) means probability of overall quality being i, given naturalness is j and usefulness is k)

Prob_model_1_Overall(:,1,5) = [0.6 0.4 0 0 0]
Prob_model_1_Overall(:,2,5) = [0 0.25 0.5 0.25 0]
Prob_model_1_Overall(:,5,1) = [0 0.25 0.5 0.25 0]

Prob_model_2_Overall(:,1,5) = [0.6 0.4 0 0 0]
Prob_model_2_Overall(:,2,5) = [0 0.25 0.5 0.25 0]
Prob_model_2_Overall(:,5,1) = [0 0.25 0.5 0.25 0]

After we have constructed simulator based on two simplified models, we have trained the model with the real subjective data, we have randomly generated perceptual quality elements data based on the models, then we have trained the model again with the generated data and we have validated the simulated models by means of mutual information and Pearson correlation value comparisons.

We have provided randomly generated perceptual quality elements data (Sharpness, Graininess, Brightness, Colorfulness for the first model and Colorfulness, Sharpness, Graininess for the 2nd model), where simulators generate with probability models a data vector as [Lpf Noise HSV usefulness naturalness Overall]. This two simulators provide artificial "evaluation vectors" and tries to mimic a human evaluation.

How good are those models? To identify it we have generated a large number (1000 to10000) of synthetic evaluation data sets with the simulators to mimic as if a large number of evaluation data set is generated by human evaluators. Later we have computed mutual information and correlation so that we can compare those results with the results produced earlier by original data set. By generating similar size as original data (162 samples) we have simulated our models multiple times (20 times), and then we have calculated average Pearson correlation, mutual information and standard deviation values. Our human evaluation data are in many places illogical, which are mainly due to that there are so few data vectors.
The acquired results from synthetic evaluation data allow us to compare them with original data we collected from human evaluators. The calculated values from the both models are listed in tables 4.3 - 4.10.

Overall Quality	Usefulness	Naturalness	Sharpness	Brightness	Colorfulness	Graininess	Clarity	Lp filtering	Noise	HSV Saturation
1.0000	0.5179	0.2681	0.4145	0.0969	0.1258	0.0620		-0.0010	-0.3664	-0.0422
0.5179	1.0000	0.5179	0.8114	0.0437	-0.0179	0.1347		0.0163	-0.6969	-0.0166
0.2681	-0.0017	1.0000	0.0090	0.2917	0.5009	0.0121		-0.0165	-0.0204	-0.1785
0.4145	0.8114	0.0090	1.0000	0.0324	-0.0031	0.0137		0.0254	-0.8660	-0.0099
0.0969	0.0437	0.2917	0.0324	1.0000	0.0969	0.0052		-0.0212	-0.0406	-0.1967
0.1258	-0.0179	0.5009	-0.0031	-0.0008	1.0000	0.1258		-0.0163	0.0037	-0.1023
0.0620	0.1347	0.0121	0.0137	0.0052	0.0200	1.0000		-0.0065	0.0825	-0.0252
-0.0010	0.0163	-0.0165	0.0254	-0.0212	-0.0163	-0.0065		1.0000	-0.0071	-0.0004
-0.3664	-0.6969	-0.0204	-0.8660	-0.0406	0.0037	0.0825		-0.0071	1.0000	0.0190
-0.0422	-0.0166	-0.1785	-0.0099	-0.1967	-0.1023	-0.0252		-0.0004	0.0190	1.0000

Table 4.3: Pearson correlation mean values from 20 simulations for first model

Overall Quality	Usefulness	Naturalness	Sharpness	Brightness	Colorfulness	Graininess	Clarity	Lp filtering	Noise	HSV Saturation
0	0.0584	0.0579	0.0724	0.0848	0.0643	0.0826		0.0596	0.0779	0.0534
0.0584	0	0.0641	0.0218	0.0626	0.0748	0.0894		0.0744	0.0263	0.0843
0.0579	0.0641	0	0.0754	0.0791	0.0497	0.0696		0.0691	0.0720	0.0897
0.0724	0.0218	0.0754	0	0.0765	0.0783	0.0740		0.0679	0.0152	0.0969
0.0848	0.0626	0.0791	0.0765	0	0.0643	0.0899		0.0859	0.0653	0.0686
0.0643	0.0748	0.0497	0.0783	0.0643	0	0.0774		0.0622	0.0869	0.0764
0.0826	0.0894	0.0696	0.0740	0.0899	0.0774	0		0.0727	0.0637	0.0897
0.0596	0.0744	0.0691	0.0679	0.0859	0.0622	0.0727		0	0.0763	0.1117
0.0779	0.0263	0.0720	0.0152	0.0653	0.0869	0.0637		0.0763	0	0.0925
0.0534	0.0843	0.0897	0.0969	0.0686	0.0764	0.0897		0.1117	0.0925	0

Table 4.4: Standard Deviation of Pearson correlation for first model

	Low-pass filtering (MI in nat)	Noise (MI in nat)	HSV Saturation (MI in nat)
Sharpness	0.0248	0.6279	0.0256
Clarity			
Graininess	0.0836	0.0422	0.0268
Brightness	0.0278	0.0256	0.0700
Colorfulness	0.0245	0.0310	0.0512
Usefulness	0.0261	0.3568	0.0219
Naturalness	5.1500e-04	0.0244	0.0508
Overall quality	0.0276	0.1067	0.0237

Table 4.5: Mean mutual information values between Instrumental measurement and quality elements for first model

	Low-pass filtering	Noise	HSV Saturation
Sharpness	0.0124	0.0372	0.0100
Clarity			
Graininess	0.0277	0.0239	0.0196
Brightness	0.0122	0.0144	0.0202
Colorfulness	0.0111	0.0174	0.0191
Usefulness	0.0142	0.0332	0.0094
Naturalness	0.0144	0.0099	0.0239
Overall quality	0.0137	0.0342	0.0081

Table 4.6: Standard Deviation of mean mutual information values between Instrumental measurement and quality elements for first model

Overall Quality	Usefulness	Naturalness	Sharpness	Brightness	Colorfulness	Graininess	Clarity	Lp filtering	Noise	HSV Saturation
1.0000	0.2387	0.6069	0.2137		0.3043	0.0520		-0.0774	-0.1816	-0.1139
0.2387	1.0000	-0.0212	0.8253		0.0317	0.0785		-0.0279	-0.7113	-0.3879
0.6069	-0.0212	1.0000	-0.0092		0.5196	0.0048		-0.1052	0.0000	0.0166
0.2137	0.8253	0.5196	1.0000		0.0350	-0.0297		-0.0210	-0.8708	-0.3076
0.3043	0.0317	0.5196	0.0350		1.0000	0.0037		-0.2078	-0.0269	-0.0059
0.0520	0.0785	0.0048	-0.0297		0.0037	1.0000		-0.0216	0.1339	-0.4611
-0.0774	-0.0279	-0.1052	-0.0210		-0.2078	-0.0216		1.0000	0.0040	0.0104
-0.1816	-0.7113	0.0000	-0.8708		-0.0269	0.1339		0.0040	1.0000	0.1416
-0.1139	-0.3879	0.0166	-0.3076		-0.0059	-0.4611		0.0104	0.1416	1.0000

Table 4.7: Pearson correlation mean values from 20 simulations for 2nd model

Overall Quality	Usefulness	Naturalness	Sharpness	Brightness	Colorfulness	Graininess	Clarity	Lp filtering	Noise	HSV Saturation
0	0.0791	0.0605	0.0726		0.0878	0.0729		0.0987	0.0648	0.0669
0.0791	0	0.0780	0.0274		0.0711	0.0933		0.0720	0.0429	0.0782
0.0605	0.0780	0	0.0670		0.0619	0.0626		0.0857	0.0679	0.0726
0.0726	0.0274	0.0670	0		0.0679	0.0585		0.0754	0.0190	0.0628
0.0878	0.0711	0.0619	0.0679		0	0.0910		0.0657	0.0668	0.0840
0.0729	0.0933	0.0626	0.0585		0.0910	0		0.0838	0.0599	0.0906
0.0987	0.0720	0.0857	0.0754		0.0657	0.0838		0	0.0733	0.0908
0.0648	0.0429	0.0679	0.0190		0.0668	0.0599		0.0733	0	0.0632
0.0669	0.0782	0.0726	0.0628		0.0840	0.0906		0.0908	0.0632	0

Table 4.8: Standard Deviation of Pearson correlation for 2nd model

	Low-pass filtering (MI in nat)	Noise (MI in nat)	HSV Saturation (MI in nat)
Sharpness	0.0206	0.6422	0.1358
Clarity			
Graininess	0.0228	0.0469	0.2901
Brightness			
Colorfulness	0.2091	0.0283	0.0220
Usefulness	0.0226	0.3790	0.1356
Naturalness	0.0646	0.0247	0.0233
Overall quality	0.0435	0.0510	0.0349

Table 4.9: Mean mutual information values between Instrumental measurement and quality elements for 2nd model

	Low-pass filtering	Noise	HSV Saturation
Sharpness	0.0091	0.0440	0.0321
Clarity			
Graininess	0.0156	0.0145	0.0650
Brightness			
Colorfulness	0.0340	0.0102	0.0088
Usefulness	0.0100	0.0558	0.0403
Naturalness	0.0209	0.0101	0.0119
Overall quality	0.0166	0.0163	0.0199

Table 4.10: Standard Deviation of mean mutual information values between Instrumental measurement and quality elements for 2nd model

Our next task is to examine and compare those Pearson correlation and mutual information values obtained from the two models so that we can validate and possibly propose a Bayesian model for the future.

5. Analysis method

5.1 Bayesian Model Identification

A Bayesian network (BN) is a graphical representation of a causal probabilistic relationship and it consists of two components: a directed acyclic graph and a set of conditional probability distributions. The conditional probability distribution of a random variable node is defined for every possible outcome of the preceding causal node. Edges between pairs of nodes are representing the causal relationship of these nodes, and a conditional probability distribution in each of the nodes. If there exists a causal probabilistic dependence between two random variables in the graph, the corresponding two nodes are connected by a directed edge. As the number of edges increase, the model becomes more complex. The complexity of the joint distribution of a node and its parent nodes grows exponentially in proportion to the number of parent nodes. Greater complexity means that a larger jury data is needed for the model identification. [24]

Our reference Bayesian network has eleven nodes initially, 3 instrumental, 5 low-level attributes, 2 high-level attributes, and the overall image quality. The instrumental measurement nodes have three possible discrete states and the rest of the nodes have five discrete states. Every node has maximum three parents. From the state probabilities in the reference Bayesian model we have simulated evaluation data for each attribute and new model parameters were estimated from the simulated data. In our first modified Bayesian model we decided to remove one perceptual quality attribute, which is brightness, based on the mutual information and Pearson correlation values calculation. We also exchanged the positions of high-level attributes usefulness and naturalness, where colorfulness is the parent node of naturalness, and sharpness, clarity and graininess are the parent nodes of usefulness. Due to three conditioning variables in perceptual quality elements the number of no-model cases increase significantly which prohibit us to validate our first model.

So further simplification of the model persuades us to construct two Bayesian models where sharpness and graininess become the parent nodes for usefulness and brightness and colorfulness become the parent nodes for naturalness in the first model. In the second simplified model we decided to keep only three perceptual quality elements colorfulness, sharpness and graininess where colorfulness is the parent node for naturalness and sharpness, graininess are the parent nodes for usefulness. In our both models the number of no-model cases were limited that facilitate us to reduce the uncertainty of the networks.

Simulations were performed on both models by using original jury assessments data and synthetically generated data to assess the overall reliability of the networks. The two models were simulated simultaneously multiple times with different data set to take account of the variation between the sampled data sets. Then the original and simulated models were compared through evaluating the Mutual information values, Pearson correlation values and Standard deviation values.

5.2 MI Analysis

If we want to construct a Bayesian network, we need to have prior expert knowledge of our model. To facilitate this, we need to understand the interaction between the nodes in instrumental measurement and nodes in image quality elements. Calculating the mutual information between them helps us to understand this conditional probability of nodes and tell us the overall reliability of the network's output. In Bayesian networks, if two nodes are dependent, knowing the value of one node will give us some information about the value of the other node. Hence, the mutual information between two nodes can tell us if the two nodes are dependent and if so, how close their relationship is. In our Bayesian network each node takes a finite set of discrete values. Mutual information calculation from those discrete values probability distribution gives us a clue how nodes in instrumental measurement and nodes in perceptual quality elements are interconnected. [25]

In our experiment the four-layer Bayes network comprises both qualitative and quantitative part. We have analyzed our chosen network by performing the statistical dependency test based on the Mutual Information (MI) measurement between instrumental measurements and perceptual quality elements. After calculating mutual information for different grading of quality elements we are able to investigate how sensitive our Bayesian network is. When the mutual information is higher with a perceptual quality element, we assume it has higher correlation with the nodes in instrumental measurement data. If the mutual information is low with a perceptual quality element, we assume it has lower correlation with the nodes in instrumental measurement data. Those perceptual quality element nodes have lower or insignificant mutual information values we discard them from the network and try to minimize the uncertainty.

In this way, we make parameter tuning in our Bayesian network and we construct a network model that is convenient from a practical point of view.

5.3 Bayes Network From Case Study

Sensitivity analysis refers to identifying the most important parameters so that unimportant parameters may be discarded from the model. If small changes are made in the parameters of the input evidence values, sensitivity analysis of the identified model reveals how much the output probability distribution changes.

Our objective was to calculate the posterior conditional probability distribution of each of the possible unobserved causes, in our network, which are Instrumental measurements, given the set of observed evidence, which are perceptual image quality elements. The Bayesian model structure and parameters should be chosen such that the uncertainties due to them are minimized. For this purpose, we have performed conditional independency tests in order to measure the degree of interaction between unmeasured nodes Instrumental measurements and evidence nodes perceptual quality elements. In order to do that, we have calculated mutual information between Instrumental measurements and each of the quality elements separately. Calculating Mutual information is a measure of knowing the dependence between two random variables. Also these information measures are easy to compute using probabilistic inference. Moreover we have

observed how different grading (from 1 to 5) in perceptual quality elements correlate with the grading of instrumental measurements (Low-pass filtering, Noise and HSV saturation).

Based on our corresponding obtained results and quality of the correlation we have initially proposed an alternative Bayes network. In our initially proposed Bayes network we have decided to cut away Brightness from the perceptual quality elements layer. Since we had not succeeded with the model, we decided to simplify the model further by constructing two models. In the first model Clarity has been replaced by Brightness and in our 2nd proposed Bayes network we have kept only three perceptual quality elements Colorfulness, Sharpness and Graininess.
Since removing the quality elements Clarity and Brightness from the network models doesn't affect the overall posterior probabilities significantly and the remaining quality parameters are sufficient to minimize the uncertainties of our network, the selection of our proposed networks can be considered as feasible models in our case study.

Finally for comparison purpose we have intuitively assumed if the Pearson correlation value between two quality elements is more than 0.4, we have calculated the differences with the Pearson correlation values of our initial Bayesian model. In the same way we have performed such test for Mutual information values for both models where mutual information value is more than 0.1 between two quality elements. Based on the results we have tried to propose which simplified Bayesian model is best for the assessment of image quality.

	Pearson Correlation data	Case1	Case2	STD 1	STD 2	Values for Case1 & Case2	Remarks
Overall & Usefulness	0.5899	0.5179	0.2387	0.0584	0.0791	1.2329 < 2 4.4399 > 3	Good Poor
Overall & Naturalness	0.4252	0.2681	0.6069	0.0579	0.0605	2.7133 > 2 3.0033 > 3	Perhaps Poor
Overall & Sharpness	0.5288	0.4145	0.2137	0.0724	0.0726	1.5787 < 2 4.3402 > 3	Good Poor
Overall & Graininess	0.4330	0.0620	0.0520	0.0826	0.0729	4.4915 > 3 5.2263 > 3	Poor Poor
Usefulness & Naturalness	0.5691	-0.0017	-0.0212	0.0641	0.0780	8.9048 > 3 7.5679 > 3	Poor Poor
Usefulness & Sharpness	0.7660	0.8114	0.8253	0.0218	0.0274	2.0826 > 2 2.1642 > 2	Perhaps Perhaps
Naturalness & Sharpness	0.4973	0.0090	0.5196	0.0754	0.0670	6.4761 > 3 0.3328 < 2	Poor Good
Naturalness & Colorfulness	0.4776	0.5009	0.5196	0.0497	0.0619	0.4688 < 2 0.6785 < 2	Good Good

Table 5.1: Pearson Correlation comparison for Bayesian models identification

	Mutual Information Data	Case1	Case2	STD 1	STD 2	Values for Case1 & Case2	Remarks
Lp filtering & Colorfulness	0.1825	0.0245	0.2091	0.0111	0.0340	14.2342 > 3 0.7824 < 2	Poor Good
Lp filtering & Naturalness	0.1234	0.0005	0.0646	0.0144	0.0209	8.5347 > 3 2.8134 > 2	Poor Perhaps
Noise & Sharpness	0.5633	0.6279	0.6422	0.0372	0.0440	1.7366 < 2 1.7932 < 2	Good Good
Noise & Usefulness	0.3918	0.3568	0.3790	0.0332	0.0558	1.0542 < 2 0.2294 < 2	Good Good
Noise & Overall Quality	0.1807	0.1067	0.0510	0.0342	0.0163	2.1637 > 2 7.9571 > 3	Perhaps Poor
HSV & Graininess	0.3343	0.0268	0.2901	0.0196	0.0650	15.6888 > 3 0.6800 < 2	Poor Good
HSV & Usefulness	0.1157	0.0219	0.1356	0.0094	0.0403	9.9787 > 3 0.4938 < 2	Poor Good
HSV & Naturalness	0.1219	0.0508	0.0233	0.0239	0.0119	2.9749 > 2 8.2857 > 3	Perhaps Poor
HSV & Overall Quality	0.2167	0.0237	0.0349	0.0081	0.0199	23.8272 > 3 9.1357 > 3	Poor Poor

Table 5.2: Mutual information comparison for Bayesian models identification

In Pearson correlation values from the table 5.1 we can see our first model tends to perform well over the 2nd model in terms of correlation values performances. For the case of mutual information performance we see the different scenario. From the table 5.2 we notice our 2nd model clearly outperforms over our first simplified Bayesian model.

So at the end we can say, when we have simulated the both Bayes networks considering specific grades (generated by both human evaluator and synthetically) of perceptual quality elements, comparing the table 5.1 and table 5.2 we come to a conclusion that our 2nd modified alternative model can be our chosen Bayesian model where the uncertainties are significantly minimized.

6. Conclusion

In Bayesian approach the principle goal is to construct the posterior probability distribution for the unknown entities from the given data sample in a model. To identify important parameters we have used mutual information and marginal distribution of a given data sample set which gives a feasible end model of the network. In reality, we have tried to find some relation between posterior probabilities and prior probabilities. One limitation of this approach is to take into account all required parameters of prior probabilities.

The overall value of image quality is often measured by summing up all the measurable perceptual image quality attribute values. When we are talking about perceptual quality elements, it may not be sufficient enough to comprise only such attribute variables as sharpness, clarity, graininess, brightness and colorfulness. In this case, it demands explicit definition of all required attributes of perceptual quality elements. On the other hand, if the more sophisticated image quality attributes are included, fruitful evaluation becomes extremely difficult for human evaluators and it will produce huge number of conditional probabilities.

Image quality is a visual or aesthetic characteristic such as color, smoothness, reflectivity, light scatter etc. that contradicts with the quantifiable instrumental values, although it will produce the same numeric values every time. We should keep in mind that an objective measure should have a consistent result with the subjective measure or the perceived quality of an image. It is necessary to understand how well the synthetically generated evaluation data correspond to the real world data, and how well the simulated model predicts the real subjective data.

Another limitation of it is the difficulty to acquire the plenty of data samples, which can validate the final model of the network in a pragmatic way. Many Bayesian network-learning algorithms require additional information, which is not always available. The uncertainty effects due to finite size of jury assessment data are common in validating a Bayesian model. In our effort a Bayesian networks is developed based on mutual Information calculation, conditional independence among the variables. But a feasible Bayesian network model can’t be achieved unless correct and reliable data are provided to us. Also it might be cumbersome and a

challenging task collecting the sufficient amount of data for identifying a robust network model. When sufficient amount of data are available, a Bayesian network may be built automatically straight from the databases using algorithms reliable estimates of conditional probability distributions. It is important that we develop more analytic tools to understand and explain the sensitivity of certain parameter changes for state-of-the-art in image quality assessment.

References

[1] Johannes Pulla. Jury Assessment as Reference for Instrumental Measurements of Image Quality, Master of Science Thesis, Tampere University of Technology, (2007)

[2] Brian Keelan. Handbook of Image Quality: Characterization and Prediction, 1st Edition, p.35-37, (2002)

[3] Raisa Halonen, Stina Westman, Pirkko Oittinen. Naturalness and Interestingness of Test Images for Visual Quality Evaluation, p.1-3, (2011)

[4] Halonen, R., Nuutinen, M., Oittinen, P., Leisti, T., Nyman, G., Mettänen, M., Ritala, R., Eerola,T., Lensu, L., Kämäräinen, J-K., Kälviäinen, H. Fusion of Digital and Visual Print Quality – Final Report of DigiQ Project, p.15-16, (2010)

[5] Heinz Hofbauer and Andreas Uhl. An Effective and Efficient Visual Quality Index Based On Local Edge Gradients, p.1, (2011)

[6] Johannes Pulla. Jury Assessment as Reference for Instrumental Measurements of Image Quality, Master of Science Thesis, p.16, Tampere University of Technology, (2007)

[7] Brian W. Keelan. Handbook of Image Quality: Characterization and Prediction, 1st Edition, p.149-168, (2002)

[8] Rank Correlation: https://en.wikipedia.org/wiki/Rank_correlation

[9] Sonja Grgic, Mislav Grgic and Marta Mrak. Reliability of Objective Picture Quality Measures, p.1-3, (2004)

[10] Santiago Aja-Fernandez, Raul San Jose Estepar and Carlos Alberola-Lopez. Full Reference Image Quality Assessment based on Local Statistics, p.1, (2014)

[11] Peng Ye and David Doermann. No-Reference Image Quality Assessment Using Visual Codebooks, IEEE Transactions on Image Processing, vol. 21, no. 7, (2012)

[12] Zhou Wang, Alan C. Bovik. Modern Image Quality Assessment, p.13, (2006)

[13] Color conversion, available: http://www.rapidtables.com/convert/color/cmyk-to-rgb.htm

[14] Measuring Color using Hunter L, a, b versus CIE 1976 L*a*b*. available: http://www.hunterlab.com/an-1005b.pdf

[15] Luis M. de Campos. A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests, Journal of Machine Learning Research 7, p.1, (2006)

[16] Lecture note: The Likelihood Function, Maximum Likelihood Estimator (MLE), Logit & Probit, Count Data Regression Models, available: http://econweb.rutgers.edu/tsurumi/liklihood.pdf

[17] Maximum Likelihood: Engineering Statistics Handbook, available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3652.htm

[18] Bayes Theorem, available: https://en.wikipedia.org/wiki/Bayes%27_theorem

[19] Mutual information and Kullback-Leibler divergence. available: http://cis.legacy.ics.tkk.fi/aapo/papers/NCS99web/node26.html

[20] Kullback-Leibler divergence. available: http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

[21] Hei Chan and Adnan Darwiche. Sensitivity Analysis in Bayesian Networks: From Single to Multiple Parameters, (2004)

[22] Hei Chan, PhD dissertation: Sensitivity Analysis of Probabilistic Graphical Models, University of California, Los Angeles, p.77, (2005)

[23] Enrique Castillo, Jose Manuel Gutierrez and Ali S. Hadi. Sensitivity Analysis in Discrete Bayesian Networks, IEEE Transactions on Man, Cybernetics and Systems, Vol. 27, 412-424, (1997).

[24] Halonen, R., Nuutinen, M., Oittinen, P., Leisti, T., Nyman, G., Mettänen, M., Ritala, R., Eerola,T., Lensu, L., Kämäräinen, J-K., Kälviäinen, H. Fusion of Digital and Visual Print Quality – Final Report of DigiQ Project, p.24, (2010)

[25] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, Weiru Liu. Learning Bayesian networks from data: An information-theory based approach, p.48, (2002)

Appendix A

Jury Simulation Data

This jury assessment data was obtained in a trail session where six people were asked to give grade in terms of eight attributes of image quality on a scale from 1 to 5.

Column 1: Overall quality
Column 2: Usefulness
Column 3: Naturalness
Column 4: Sharpness
Column 5: Brightness
Column 6: Colorfulness
Column 7: Graininess
Column 8: Clarity

Evaluator 1

5	5	5	5	2	5	5	5
4	4	5	4	4	5	2	4
2	2	2	4	2	4	1	2
3	3	5	3	2	5	5	3
3	3	5	3	3	5	2	4
2	1	3	2	3	4	1	1
1	2	3	2	2	5	4	2
2	2	2	1	3	2	3	1
1	1	1	1	2	2	1	1
4	5	4	5	5	3	5	5
2	4	2	4	4	2	3	4
2	2	1	4	5	1	1	2
3	4	4	3	4	3	4	4
2	3	2	3	5	3	2	2
1	1	1	2	4	2	2	2
1	2	2	1	4	2	4	1
1	1	2	1	4	2	2	1
1	1	1	1	5	1	2	1
5	5	4	5	1	3	5	5
4	5	4	5	1	4	3	4
1	2	1	4	1	1	1	2
4	4	1	3	1	1	5	4
3	3	3	3	1	3	3	3
2	1	1	2	2	3	1	1

2	1	2	1	1	1	4	2
2	2	1	1	1	1	3	1
1	1	1	1	2	2	1	1

Evaluator 2

5	5	4	5	2	4	4	4
3	5	4	5	4	4	3	4
3	4	2	4	3	2	2	3
3	3	3	3	4	5	5	2
3	2	3	2	3	3	2	3
4	2	2	3	3	3	2	3
1	1	4	1	3	4	5	2
2	2	2	1	2	3	2	2
1	1	3	2	2	4	1	1
5	5	3	5	1	1	4	4
3	4	2	4	2	1	3	2
3	4	1	4	2	2	2	2
5	4	2	2	2	1	4	4
4	3	1	3	1	1	2	1
3	2	2	3	1	2	2	3
1	1	2	2	1	2	5	1
1	2	1	1	2	2	2	2
1	1	2	1	2	1	1	1
5	5	5	4	5	5	5	5
5	4	5	4	4	4	3	2
2	3	4	3	4	2	1	2
4	3	3	4	3	3	4	5
4	3	4	3	3	3	2	2
2	3	3	2	4	2	1	3
2	1	1	1	5	2	4	2
2	2	3	2	4	3	2	1
1	1	3	2	4	3	1	1

Evaluator 3

5	5	4	5	2	5	5	5
2	5	5	4	3	2	4	4
2	4	3	3	4	4	1	1
4	4	4	2	3	2	4	4
3	2	4	1	2	2	3	3
2	3	1	1	4	3	1	4
2	2	4	1	2	4	5	2
2	2	4	2	4	4	2	2
1	2	3	2	3	3	1	1
4	4	3	5	5	3	5	5
2	5	2	4	5	3	4	4
2	2	1	3	5	1	1	3
3	3	2	5	5	2	5	3
3	4	3	4	4	2	2	3

2	3	2	2	4	2	1	2
3	2	3	1	5	2	3	3
1	3	2	2	4	1	2	2
2	3	1	2	5	2	1	1
5	3	5	5	4	5	5	5
4	5	5	5	4	5	3	5
2	4	3	2	4	3	1	3
5	4	4	4	1	5	5	3
3	3	5	3	3	5	2	3
2	2	2	4	2	4	1	1
3	2	3	2	4	4	5	2
3	3	3	1	1	3	3	2
1	1	2	2	2	4	1	1

Evaluator 4

5	5	5	5	5	4	5	5
4	4	3	3	4	4	3	4
2	4	4	3	4	4	1	3
4	4	5	3	2	4	4	4
3	3	2	2	1	2	3	3
2	2	2	2	3	3	1	5
2	3	4	1	1	4	4	4
2	2	3	1	1	3	2	2
1	1	3	1	4	3	1	1
4	5	4	5	2	2	5	5
4	4	3	4	4	2	3	3
2	4	2	3	2	2	1	2
4	4	1	2	1	3	2	3
3	4	2	2	3	1	1	3
2	2	2	2	1	2	4	1
1	2	1	1	5	2	2	2
2	1	1	2	4	2	1	2
1	2	2	1	2	5	5	1
5	5	5	5	2	5	3	5
4	5	5	4	3	4	1	4
3	3	3	3	2	5	5	4
4	4	4	3	3	5	3	3
3	4	4	2	1	4	2	3
2	3	2	2	3	5	4	3
2	3	2	1	4	4	3	2
2	3	3	1	2	4	1	1
1	1	2	1	3	2	2	2

Evaluator 5

5	4	4	5	5	5	3	5
3	4	3	4	3	4	3	3
1	3	3	4	4	4	1	3
5	4	4	4	3	4	5	5
4	4	4	3	4	5	3	4
2	3	4	2	3	4	1	3
3	3	4	2	4	4	5	5
2	3	2	2	4	4	3	4
2	2	4	2	3	4	2	3

5	5	4	4	2	2	4	4
2	3	3	4	2	2	3	4
1	3	3	3	1	3	2	3
4	4	2	3	2	2	4	5
4	3	2	3	2	2	3	4
2	2	3	4	4	5	1	3
3	4	2	3	3	2	5	4
2	2	2	2	2	4	3	4
1	3	2	2	1	1	2	3
5	5	5	5	5	3	3	4
4	4	4	4	5	3	3	4
1	3	4	3	4	3	1	3
5	4	5	3	5	2	5	5
3	3	5	2	5	3	3	4
1	3	3	3	4	3	2	3
3	3	5	2	5	3	5	5
3	4	3	2	4	3	3	4
1	3	3	2	4	2	2	3

Evaluator 6

1	5	5	5	5	2	2	5
1	5	5	5	5	5	5	4
1	4	3	4	2	5	2	3
2	4	4	3	3	4	3	4
5	3	4	3	4	3	5	3
1	2	3	2	2	4	4	2
2	2	2	2	3	3	1	3
3	2	2	1	3	3	1	2
2	1	1	1	1	2	1	1
4	5	4	5	4	2	2	4
5	5	5	5	4	4	3	5
1	4	1	3	1	3	1	2
4	4	2	4	5	5	2	4
3	3	4	2	2	2	2	3
4	2	1	2	1	5	4	1
2	2	1	1	2	2	4	2
2	1	1	1	2	3	5	2
2	1	1	1	1	1	1	1
1	5	5	5	5	1	1	5
3	5	5	5	5	1	5	5
2	4	2	4	3	3	3	3
4	4	3	3	4	1	5	3
3	3	4	3	3	2	1	3
1	3	3	2	2	2	5	2
5	2	2	2	3	1	1	2
1	1	2	1	2	1	2	2
3	1	1	1	2	1	5	1

Appendix B

Experimental Data

There were three distinct degrees of modification for each instrumental measurement method: no modification, mild, and moderate level.

Column 1: Lp filtering

Column 2: Noise

Column 3: HSV Saturation

1	1	1
1	1	2
1	1	3
1	2	1
1	2	2
1	2	3
1	3	1
1	3	2
1	3	3
2	1	1
2	1	2
2	1	3
2	2	1
2	2	2
2	2	3
2	3	1
2	3	2
2	3	3
3	1	1
3	1	2
3	1	3
3	2	1
3	2	2
3	2	3
3	3	1
3	3	2
3	3	3

Figure B.1. Studio image used in the Bayesian network identification test.