Bayesian network model of overall print quality: construction and structural optimisation

Tuomas~Eerola[a,*], Lasse~Lensu[a], Joni-Kristian~Kamarainen[b], Tuomas~Leisti[c], Risto~Ritala[d], Göte~Nyman[c], Heikki~Kälviäinen[a]

[a] *Machine Vision and Pattern Recognition Laboratory (MVPR), Lappeenranta University of Technology, P.O. Box 20, FIN-53851 Lappeenranta, Finland*
[b] *MVPR/Computational Vision Group (Kouvola Unit)*
[c] *Department of Psychology, University of Helsinki, P.O. Box 9, FIN-00014 Helsinki, Finland*
[d] *Department of Automation Science and Engineering, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland*

Abstract

Prediction of overall visual quality based on instrumental measurements is a challenging task. Despite the several proposed models and methods, there exists a gap between the instrumental measurements of print and human visual assessment of natural images. In this work, a computational model for representing and quantifying the overall visual quality of prints is proposed. The computed overall quality should correspond to the human visual quality perception when viewing the printed images. The proposed model is a Bayesian network which connects the objective instrumental measurements to the subjective opinion distribution of human observers. This relationship can be used to score printed images, and additionally, to computationally study the connections of the attributes. A novel graphical learning approach using an iterative evolve-estimate-simulate loop learning the quality model based on psychometric data and instrumental measurements is suggested. The network structure is optimised by applying evolutionary computation (evolve). The estimation of the Bayesian network parameters is within the evolutionary loop. In this loop,

[*] Corresponding author. Tel.: +358 5 6212852; fax: +358 5 6212899
Email addresses: tuomas.eerola@lut.fi (Tuomas~Eerola), lasse.lensu@lut.fi (Lasse~Lensu), joni.kamarainen@lut.fi (Joni-Kristian~Kamarainen), tuomas.leisti@helsinki.fi (Tuomas~Leisti), risto.ritala@tut.fi (Risto~Ritala), gote.nyman@helsinki.fi (Göte~Nyman), heikki.kalviainen@lut.fi (Heikki~Kälviäinen)

the maximum likelihood approach is used (estimate). The stochastic learning process is guided by priors devised from the psychometric subjective experiments (performance through simulation). The model reveals and represents the explanatory factors between its elements providing insight to the psychophysical phenomenon of how observers perceive visual quality and which measurable entities affect the quality perception. By using true data, the design choices are demonstrated. It is also shown that the best-performing network establishes a clear and intuitively correct structure between the objective measurements and psychometric data.

Keywords: Print quality assessment, overall visual quality, Bayesian networks, structural optimisation

1. Introduction

The concept of visual quality describes the quality of a reproduction of visual information on some media, for example, a printed photograph. Ambiguity cannot be avoided when defining a visual quality measure because the ground truth is the subjective opinion of an "end-user" observing the reproduction. Numerous measures still do exist and new ones appear continuously, but their relation to the fundamental objective, prediction of the visual quality as perceived by humans, is often heuristically justified or covers only a narrow problem domain due to limited experimental data. Despite this, quality measurements are important for developing new products (paper grades, inks, printers, etc.) within the relevant industries, and therefore, are in active use in research laboratories. Popular measures are, for example, objective metrics focusing on artefacts in printed test patterns, such as print dot roundness and edge raggedness˜(Wolin et˜al., 1998b,a), micro-uniformity, macro-uniformity, colour rendition, text and line quality, gloss, sharpness, and spatial adjacency˜(Bouk et˜al., 2008), and unevenness of solid printed regions (e.g.,˜Armel and Wise (1998); Briggs et˜al. (1999); ISO/IEC (2001)). In their recent work, the authors showed how the full-reference measures developed for digital images can be transformed for use with prints~(Eerola et~al., 2010). It is also generally accepted, as demonstrated in~(Eerola et~al., 2007, 2008a,b), that the listed measures indeed contribute to the overall quality, but the nature of this combined perceptual, cognitive, and psychological process is not merely a weighted sum of the measures. The persistent problem remains: how the measurable properties and the recorded human observations should be connected, and is it feasible to explain the phenomenon of visual quality?

In realistic and challenging visual evaluation tasks involving aesthetic or even personal attributes, it is highly unlikely that the overall visual print quality can be measured with a single measurement and represented by a single scalar value~(Keelan, 2002). Even in restricted settings with artefactual or preferential attributes, human evaluators are likely to give different ratings for the same samples. With simple visual assessment tasks, the physiological cause for the variation near the just-noticeable difference (JND) of a visual attribute arises from differences between the observers and from the stochastic nature of perception. Despite the fact that photoreceptor cells in the human retina can detect even single photons~(Baylor et~al., 1979), the absorption of a photon by a photoreceptor pigment is a random process, detection thresholds are dependent on the adaptation of the visual system, and the photoreceptor functionality is affected by inherent noise~(Birge and Barlow, 1995). When the visual attributes composing a stimulus are well above the JND levels, only the consistency of an individual observer in her judgement (personal variation in the evaluation criteria) and the level of agreement in a jury of observers affect the variation of the combined result. In the case of a jury, also cultural factors and physiological differences can have an effect on the evaluation. When the visual assessment task is made more realistic by evaluating natural images of high quality, however, the subjective assessments are affected by the JNDs of related visual attributes, by the inconsistency of an individual observer, and by the level of agreement within a jury of observers. To keep the variation at a reasonable level, it is essential to carefully design both the visual stimuli and evaluation task for the subjective experiment. Also, a psychometric method is required that allows the subjects to express their visual experience in viewing a specific image. This can be accomplished in a suitable experimental set-up.

The stochastic nature of perception and interpretation of visual information motivates to treat the overall quality and its attributes as probability distributions. For this purpose, the Bayesian theory provides a natural tool for modelling and analysis. A Bayesian network is an attractive tool since it is a probabilistic model that represents a set of random variables (instrumental measurements and subjective attributes) and their conditional independences with a directed graph. The idea of using Bayesian networks for modelling visual quality is not completely new. In˜(de˜Freitas˜Zampolo and Seara, 2004) and˜(Pulla et˜al., 2008), Bayesian networks were used to describe the overall image quality. However, these studies were not complete. In˜(de˜Freitas˜Zampolo and Seara, 2004), a network was used to combine noise˜(Damera-Venkata et˜al., 2000) and distortion measures˜(de˜Freitas˜Zampolo and Seara, 2003). The work reported in˜(Pulla et˜al., 2008) was more similar to this work since the authors used the network to combine objective and subjective assessment data. The objective measurements were given as input values, and the overall image quality was viewed as a probability distribution of ratings. The previous works did not consider the problem of how to establish the network structure automatically based on true data. Instead, they showed the potential of Bayesian networks to model image quality and similar phenomena.

In this work, the idea in˜(Pulla et˜al., 2008) is further developed by proposing a method which automatically optimises the structure of the Bayesian network for using it as a model of visual print quality. This is done by making elementary hypotheses about the behaviour of the overall quality with respect to the objective measures (prior) and by computing the model fitness through simulation. The structure optimisation method is a genetic algorithm mainly due to the complexity of the optimisation problem and to the need for simulation to evaluate the fitness of a solution. The main contribution of this paper is an evolve-estimate-simulate optimisation loop, where the structure and connections are evolved using an evolutionary approach, network parameters estimated using the maximum likelihood rule, and network performance evaluated using simulation. The final network forms statistical dependencies between the collected psychometric data and instrumental measurements whos values were discretised to the range $[1, 5]$. The network can be used as a unified model representing and explaining the phenomenon, and also as a more practical tool producing just a single visual quality index (VQI) for any printed product.

The report is organised as follows: in Sec.˜2, the use of Bayesian network for the problem of estimating print quality is described. Structural optimisation of the network is presented in Sec.˜3. The experimental arrangements and results are given in Section˜4, and conclusions in Sec.˜5.

2. Visual print quality model

Bayesian networks˜(Pearl, 1988) can be used as a tool for decision making under uncertainty. A network is a graph with nodes and edges with parameters. The parameters represent conditional probabilities of node outputs given node inputs. Typical applications for Bayesian networks are, for example, medical diagnosis˜(Nikovski, 2000), troubleshooting˜(Breese and Heckerman, 1996), student modelling˜(Yang et˜al., 2007), complex genetic models˜(Friedman et˜al., 2000), and crime analysis˜(Oatley and Ewart, 2003). Related to the topic of this work, Bayesian networks have also been applied for controlling a complex printing system˜(Hommersom and Lucas, 2010). In the following, Bayesian networks as a model of visual print quality is formalised.

2.1. Main structure and variables

The first step to construct a Bayesian network is to select the nodes, that is, the random variables. In this work, the basic structure depicted in Fig.˜1 is used for the Bayesian network. On the left, the nodes represent the objective measures that are the essential external model inputs. In the figure, the rightmost node represents the overall quality, the network output in the form of a probability distribution within the range of possible values. The middle portion

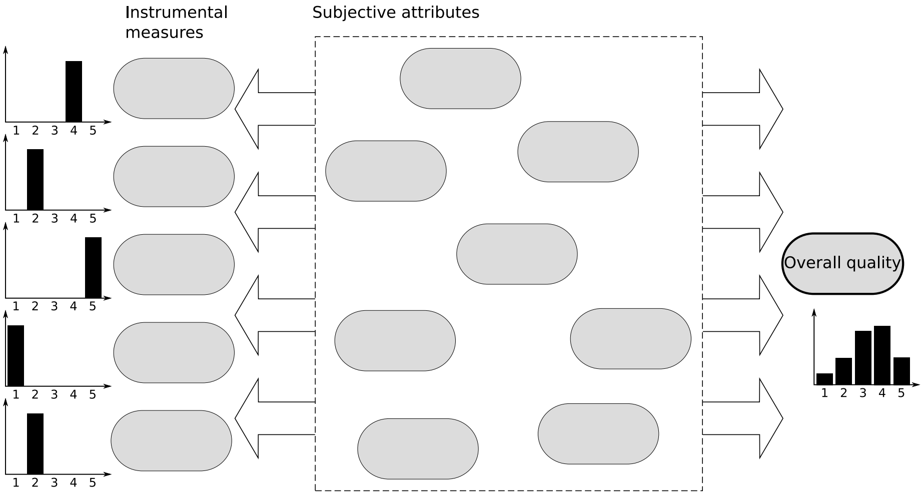


Figure 1: The basic Bayesian network structure with example distributions of random variables representing the inputs and output of the model.

consists of the subjective attributes which represent the abstract quality concepts shared and used by the individuals to form the basis of overall quality. These attributes were identified in the human experiments using well-defined psychometric tests described in Sec.˜2.2. The subjective attributes form an intermediate layer which transforms the objective measures into probabilistic overall quality. The arrows indicate the causality of the model. In the network, the subjective attributes are interpreted as “the reality” that is desired to be measured. On the one hand, the subjective attributes induce a certain measuring result (the objective measurements), and additionally, their combination forms the perceived overall quality. This is why the direction of the causality is from the subjective attributes to the objective measures as well as to the overall quality. However, constraining the direction of the causality towards the objective measures does not prevent the inference of subjective attributes and the overall quality based on the objective measures. It is important to notice that the objective measures on the left in Fig.˜1 can be accurately and repeatedly measured from printed photographs and test fields. The overall quality, or more precisely, its distributions from experiments with a jury of observers, can be estimated by carrying out psychometric experiments. From this viewpoint, the model produces the most likely distribution of the overall quality opinions of the observers, if the same material is physically presented to a number of them.

The best objective measures for the left portion in Fig.˜1 were selected according to the results of the prior works by the authors˜(Eerola et˜al., 2008a,b) where the most important instrumental and computational measures were surveyed and their relevance for explaining the overall quality was evaluated. The task was not possible by using the standard linear correlation; instead, nonlinear relationships were evaluated and ranked using the proposed cumulative match score histogram (CMSH)˜(Eerola et˜al., 2008a). The main idea behind the CMSH is the assumption that if two samples are visually perceived as being close to each other, they should be close to each other also based on the objective measurements. If a measure fails to meet the criterion, it was classified as irrelevant for subjective overall quality. Using the method, it was possible to rank the existing measures, and even exhaustively search for the optimal combinations of $N = 1, 2, \ldots, 6$ best measures. For digital printing (inkjet and electrophotography), the following six measures were selected: *computational mottling*˜(Sadovnikov et˜al., 2007), *colour gamut*, *mean colour density*, *print gloss*, *edge blurriness* and *edge raggedness*. This result is well in accordance with the current practises: these measures are commonly used in paper mill laboratories as well. Detailed descriptions of them can be found in˜(Eerola et˜al., 2008a).

The selection of subjective attributes was based on systematic interviews of observers during the far-reaching subjective experiments. As a standard psychological interview technique˜(Radun et˜al., 2008) the observers were asked to describe visual factors that affected their ratings after they had given a rating for overall quality for each image. Later, a common vocabulary was established from the factors by using manual search, frequency analysis, and term mappings, and it was revised in the next independent experiments. Specifically, the most common subjective attributes were as follows: *naturalness*, *clarity*, *colourfulness*, *subjective gloss*, *graininess*, *lightness*, *contrast*, and *sharpness*. It should be noted that these subjective attributes do not necessarily correspond to their physical analogues since the semantic meaning of a term varied between the observers. This is typical for the natural, fuzzy concepts that naïve observers use in their everyday speech in contrast to the well-defined concepts used by the professionals. This difference is not necessarily caused by the incorrect use of the terms, but by the fact that visual impressions of contrast, sharpness, naturalness etc. are not unambiguously related to any physical property of an image. For example, higher colour saturation may make the image look subjectively sharper, although the use of these concepts is separated among professionals. For this reason, the graph edges cannot be formed manually, but the relationships need to be learned.

2.2. Training and using subjective data

For training the network, the maximum likelihood estimation for complete data can be used since the numerical values of each node and for every sample exist. The instrumental and computational methods to perform the objective measures are described in˜(Eerola et˜al., 2008a). To obtain the numerical values of the subjective attributes and overall quality, several far-reaching psychometric experiments with interviews were conducted. These procedures with data have been described in detail in˜(Oittinen et˜al., 2008). In brief, samples from a specific set of natural images were placed on a table in a random order. Labels with the numbers from 1 to 5 were also presented on the table. The observer was asked to select the sample representing the lowest quality in the sample set and place it on the label with number 1. Then, the observer was asked to select the highest quality sample and place it on the label with number 5. After that, the observer's task was to place the remaining samples on the labels so that the quality increased steadily from 1 to 5. The subjective attributes were evaluated similarly, only now the observer was asked to label the samples based on a single attribute, such as sharpness or graininess. Information about the number of observers and paper samples are given in Section˜4.1. For the experiments in this study, the revised set of subjective attributes was selected using the previous psychometric experiments.

In this work, multi-dimensional arrays were used as the conditional probability distributions. Such arrays are suitable only if the data are discrete. The subjective data were readily discrete (1-5), but the objective variables needed to be discretised. The discretisation was carried out at the interval [1,5] with equal-width bins so that the largest value was given 5 and the lowest 1. When all the training data were collected and linearly discretised, the Bayesian network with a selected structure can be straightforwardly trained using the maximum likelihood parameter estimation˜(Holmes and Jain, 2008). Due to the potential problems with the low amount of training data, the Laplace correction was used.

2.3. Inference with Bayesian network

Inference with a Bayesian network can be understood as the computation of marginal distribution of one node (model output) based on evidence (model inputs) or finding of the most probable explanations, i.e., values for several nodes. The junction tree algorithm was used for the inference˜(Huang and Darwiche, 1996). The basic idea of the junction tree algorithm is to form a tree-structured Bayesian network equivalent to the initial multi-connected network. In the junction-tree, the nodes are cliques of the original nodes. For a tree structured Bayesian network, inference (computation of the conditional probabilities) can be performed in linear time. However, it should be noted that for large cliques the computation is exponential.

Let $G = (V, E)$ be a directed acyclic graph. Each node $i \in V$ corresponds to a random variable X_i with finite-set states, and pa_i is the set of parents of node i. Using the chain rule, a joint probability distribution represented by a Bayesian network is got

$$p(x) = \prod_{k=1}^{K} p(x_k|pa_k) \ . \tag{1}$$

A constructed network is an efficient tool for probabilistic inference. However, building such a network, especially finding its structure, remains the main problem. In the next section, the optimisation of a network structure in the case of print quality evaluation is discussed, and a method to find a working and general model even with a limited amount of training data is proposed.

3. Learning the structure

Learning the optimal structure for a Bayesian network has been shown to be NP-complete˜(Chickering, 1996). As a consequence, full search methods are infeasible. Several heuristic methods for structure learning exist (e.g., see˜(Neapolitan, 2004)). However, the laborious nature of collecting subjective data in the presented case severely limits the available amount of training data. Therefore, most heuristic methods, such as the PC algorithm˜(Spirtes and Glymour, 1991), are not applicable. The structure optimisation is, however, essential for solving the problem and needs to be implemented into the learning process.

In the case of print quality modelling, it is possible to form a number of hypotheses on how the model should behave. For example, if the undesired solid printed area unevenness (Mottling) increases while the other objective measures remain the same, the overall quality should decline. Similarly, if the colour gamut (a subset of colours a paper grade can reproduce with the available inks) expands, then the overall quality should improve. Using these heuristic and intuitively correct regulation rules, it is possible to produce a scalar value representing how logically correct a model is, that is, by randomly pruning how well model behaviour follows the hypotheses. This leads to a complex optimisation task: finding such a Bayesian network structure that the model behaves as logically as possible after its parameters have been estimated using the training data. In this learning scheme, a network is not evaluated according to how well it represents the training data, but how well it represents the prior knowledge after the estimation with the training data. Therefore, the prior knowledge of behaviour acts as a regularisation term which enables the optimisation process with a small number of data points.

3.1. Optimisation problem

The hypotheses related to the behaviour of the network are listed in Table˜1. These hypotheses are simply evaluated by computing the marginal distribution of overall quality with selected objective measure values set as evidence, and then, changing the value of a single objective measure and examining the sign change of the marginal distribution. To avoid the comparison of distributions, their expected values are used. The expected value of overall quality can be seen as an estimate for the mean opinion score (MOS).

Table 1: Prior rules as hypotheses for optimising the structure.

Id	Change on input	Effect on overall quality
H1	Mottling increases	decrease
H2	Colour densities increase	increase
H3	Colour gamut increases	increase
H4	Gloss increases	increase
H5	Edge blurriness increases	decrease
H6	Edge raggedness increases	decrease

The computation of marginal distributions is time consuming, and therefore, testing all prior hypotheses with all possible input combinations is infeasible. Instead, a number of random comparisons which provides statistical significance is selected. Let n be the sample size, that is, the number of tests per hypothesis, and X_i the test result (1 if a hypothesis is supported and 0 if not). Now,

$$p_n^* = \sum_{i=1}^{n} X_i/n \tag{2}$$

is the proportion of tests supporting the hypothesis: the higher the value, the better the evaluated model follows the hypothesis. Let $p = E[X_i]$ be the true value of p_n^* when all possible input combinations are tested. Let us select the sample size n from

$$P(|p_n^* - p| \leq \epsilon) \geq P_\epsilon, \tag{3}$$

that is, p_n^* should not differ from p more than ϵ with the probability P_ϵ. According to the central limit theorem,

$$p_n^* \sim N(p, \frac{p(1-p)}{n}), \tag{4}$$

and now,

$$P\left(|Z| \leq \frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}}\right) \geq P_\epsilon, \tag{5}$$

where $Z \sim N(0,1)$. It is apparent that the sample size depends on p, and $p(1-p)$ reaches its maximum value (1/4) when $p = 1/2$. Using (5), it is now straightforward to calculate the optimal sample sizes for different ϵ and P_ϵ (Table˜2).

Table 2: The minimum sample sizes (the number of hypothesis tests) for different ϵ and P_ϵ (the resulting error is smaller than ϵ with the probability P_ϵ).

P_ϵ	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
0.9	6764	271	68
0.95	9604	384	96

Next, the fitness function is formed based on the prior hypothesis evaluations. Let $X_{i,j}$ be the result of the jth hypothesis in the ith test. Now, the fitness function to be maximised can be formed as

$$f = \frac{1}{6}\sum_{j=1}^{6}\left(\sum_{i=1}^{n} X_{i,j}/n\right) \tag{6}$$

It should be noted that if the number of edges in the graph increases considerably, the situation may occur that the number of parents of some node is too large and the parameter estimation fails. This is due to the limited amount of training data, since the complexity of the joint distribution of a node grows exponentially in proportion to the number of its parents~(Bishop, 2006). This further causes hypothesis tests to fail and the fitness function value to decline. This, on the other hand, favours sparse networks where only the most important edges remain. Therefore, the presented fitness function also prevents over fitting to the existing data.

3.2. Genetic algorithms

The genetic/evolutionary approach was adopted mainly due to the complexity of the optimisation problem and the need of simulation to evaluate the fitness. Typical to all NP-complete problems, evolutionary approaches have been applied also to the structural optimisation of Bayesian networks~(Larrañaga et~al., 1996; van Dijk et~al., 2003) Also other stochastic metaheuristics such as ant colony optimisation~(de~Campos et~al., 2002), univariate marginal distribution algorithm~(Blanco et~al., 2003), and population-based incremental learning~(Blanco et~al., 2003) have been proposed. The earlier studies, however, typically optimise the structure directly for the given training data. This work differs from them in the sense that the training data is used only for parameter estimation, and the evolutionary strategy is applied only to guide the search towards solutions which are supported by the hypotheses. Moreover, with the limited data in our case, structural optimisation using only the data would be impossible.

A tailored genetic algorithm for the structure optimisation is presented in Algorithm~1. The Bayesian network structure is represented by adjacency matrices and converted to the population members by concatenating the matrix entries to vectors

$$a_{11}a_{21}...a_{n1}a_{12}a_{22}...a_{nn}, \text{ where } a_{ij} = \begin{cases} 1 & \text{if } j \text{ is a parent of } i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

In the basic structure (see Fig.~1), certain edges are not allowed (e.g., from overall quality to objective measures). The corresponding cells in the population member vector can be excluded from the search by fixing them to zero and this way speed up the optimisation process.

Algorithm 1. *Genetic algorithm*

```
Choose an initial population of size $N$
Estimate the model parameters for each population member
Evaluate the fitness of each initial population member using Eq. 6
repeat
    Randomly select pairs of individuals for crossover
    Generate a new population by uniform crossover and random mutations
    Estimate the model parameters for each population member
    Evaluate the fitness of each member of new population
    Combine the previous and new population, and select the $N$ best individuals for the next iteration
until termination
```

In Algorithm˜1, crossover means that some edges are swapped between two individuals with a fixed probability (crossover probability), and mutation that edges are removed or added randomly with a fixed probability (mutation probability). It should be noted that the crossover and mutation steps can produce offsprings which are not allowed, that is, the network structure becomes cyclic. This is handled by setting the fitness of each illegal population member to 0, which eliminates them before the next iteration.

4. Experiments

In this section, the data, selected paper types and printed test images are introduced, and the results are demonstrated discussing also the produced models.

4.1. Test set

The test set consisted of 21 paper grades (15 papers designed for electrophotography and 6 multi-purpose papers). The samples were printed using a Xerox DocuColor 6060 production-scale electrophotographic printer. Paper-specific colour management profiles were determined using a profiling target, spectrophotometer and profiling software, which is the current practise in high quality printing. The optimal print settings were used for each paper. The printing process corresponds to standard practises and is defined in more detail in˜(Oittinen et˜al., 2008). Two different collages of test images were printed on each paper grade (see Fig.˜2). The test images consisted of natural images and technical test fields. Ten copies of the images were printed, and the best prints were chosen for further study. In this way, it was made sure that the effect of media (papers on the final result) was studied, and not the printing process.

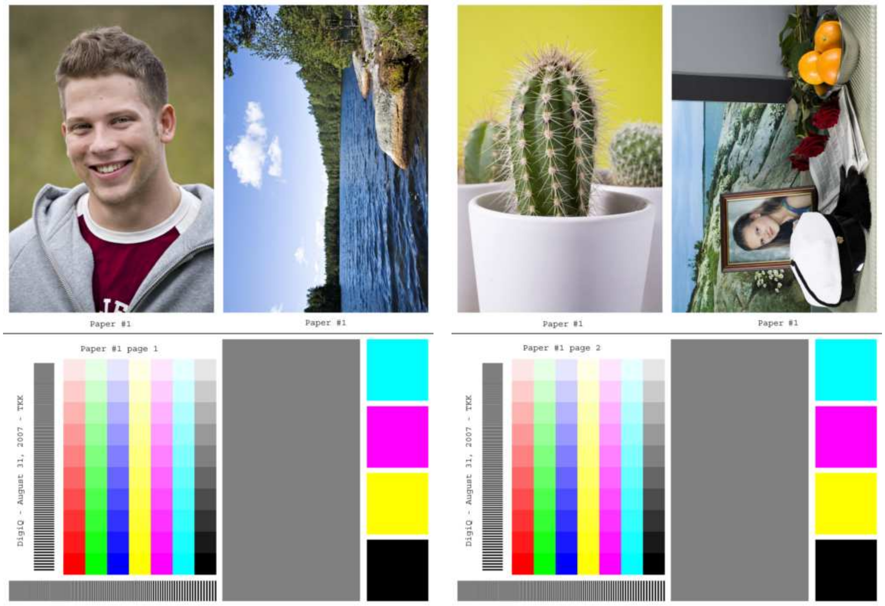


Figure 2: The used test images and technical fields.

The printed samples were scanned using a high quality scanner with 2500 dpi resolution and 48-bit RGB colours. A colour management profile was devised for the scanner before scanning, and colour correction, descreening and other automatic settings were disabled in the scanner software. The digitised images were saved using lossless compression.

The objective measures were computed from the technical test fields, and the subjective evaluation was carried out using three natural image contents (human portrait, landscape and cactus in Fig.~2). The subjective evaluation was conducted separately for each image content, and the number of observers was 29, so the number of training samples was $21 \times 3 \times 29 = 1827$. However, it should be noted that the objective measures were constant for each paper grade, and thus, the training data is extensive only for the subjective part of the model (only 21 different combinations of objective measures).

4.2. Optimising the structure

Table~3 shows the parameters used for the genetic algorithm. The initial population consisted of 20 educated guesses, 20 fully random networks, and 20 partly educated guesses (some of the edges manually selected). Due to the long computation time of marginal distributions, the number of hypothesis tests in the fitness function evaluation was set to small (100 per hypothesis), and thus, the error margin for the fitness function value was relatively large (see Table~2). Therefore, a list of the 100 best structures was maintained during the optimisation process, and evaluated with a larger number of hypothesis tests after the optimisation process. Due to the stochastic nature of genetic algorithm, the structure learning was repeated 10 times and the best structures were selected over all runs. The progress of the genetic algorithm is shown in Fig.~3

Table 3: Parameter values for the genetic algorithm.

Parameter	Value
Size of population	60
Number of iterations	1000
Crossover probability	0.5
Mutation probability	0.02

The best structures of the Bayesian network found in the optimisation process are shown in Fig.~4. As mentioned above, a list of the 100 best structures

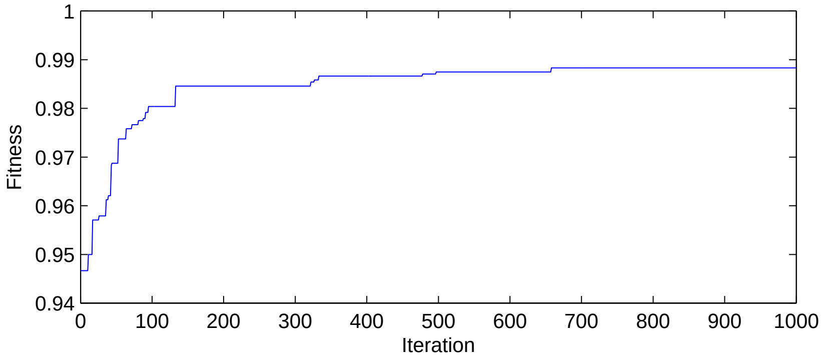


Figure 3: Progress of the genetic algorithm.

was maintained during the optimisation process and the optimisation was repeated 10 times resulting 1000 network structures. After the optimisation step, the fitness function was computed for the 1000 networks with a larger sample size (2000 per hypothesis). The best structure according to the fitness function is shown in Fig.~4(a). All the 1000 networks were evaluated against the subjective MOS using leave-one-out cross-validation. The best structure according to the correlation coefficient between the model output and MOS is shown in Fig.~4(b). In Fig.~5, the correlations against the subjective evaluation are plotted. The expectation values of the overall quality were used as the visual quality index (VQI). For a comparison, a tree-structured network learned using Chow-Liu algorithm~(Pearl, 1988) was also tested. With a correlation coefficient of 0.75 against MOS, the tree-structured network was outperformed by the networks found using proposed structure optimisation method. In Fig.~4(c) and Fig.~5(c), results are shown when, instead of optimising the fitness function presented in Sec.~3.1, the correlation coefficient to subjective evaluation was optimised. It is clear that the number of edges increases dramatically, and due to the small amount of training data, the generalisability of the model becomes weak.

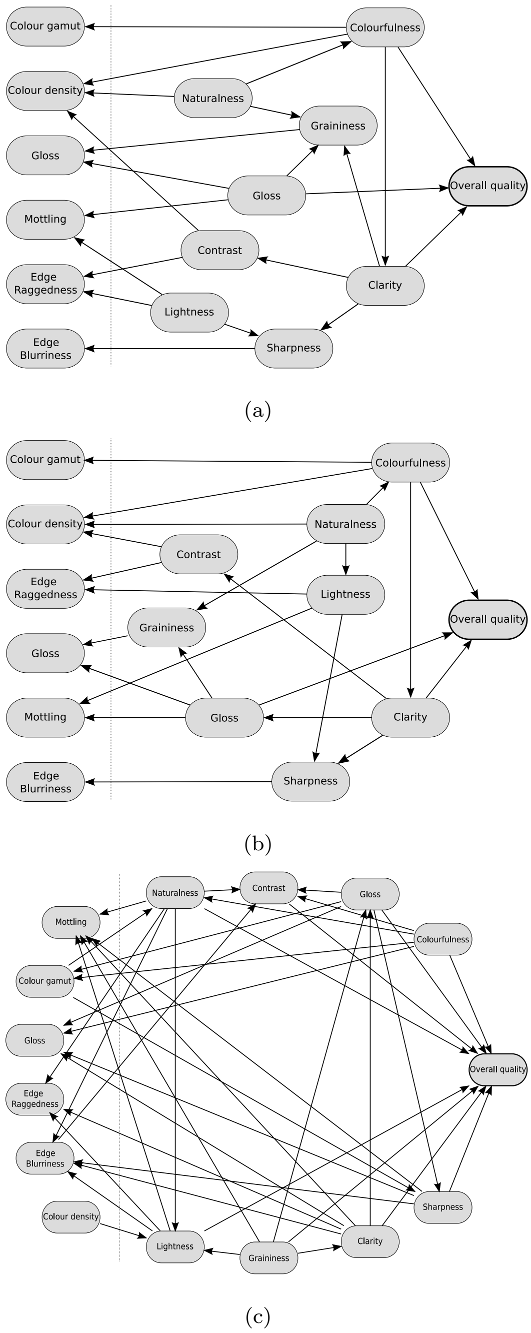
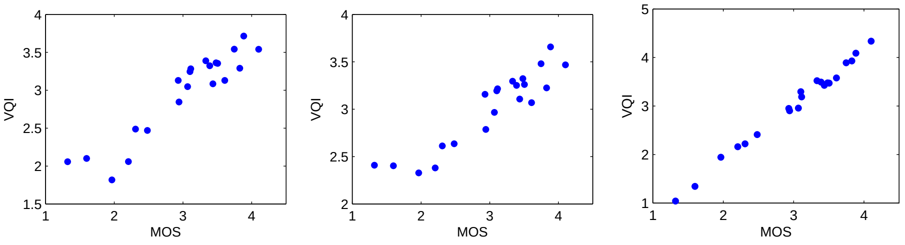


Figure 4: The best Bayesian network structures found: (a) Best simulation result; (b) Best correlation against the subjective evaluation (leave one out); (c) Correlation coefficient optimised instead of the fitness function.



(a) (b) (c)

Figure 5: Correlations between the model-produced visual quality index (distribution expectation) and subjective MOS. Correlations computed using the leave-one-out cross-validation: (a) Best fitness function value in Fig.˜4(a)(correlation: 0.924); (b) Best correlation against the subjective evaluation in Fig.˜4(b) (correlation: 0.933); (c) Correlation coefficient is optimised instead of the fitness function in Fig.˜4(c) (correlation: 0.994).

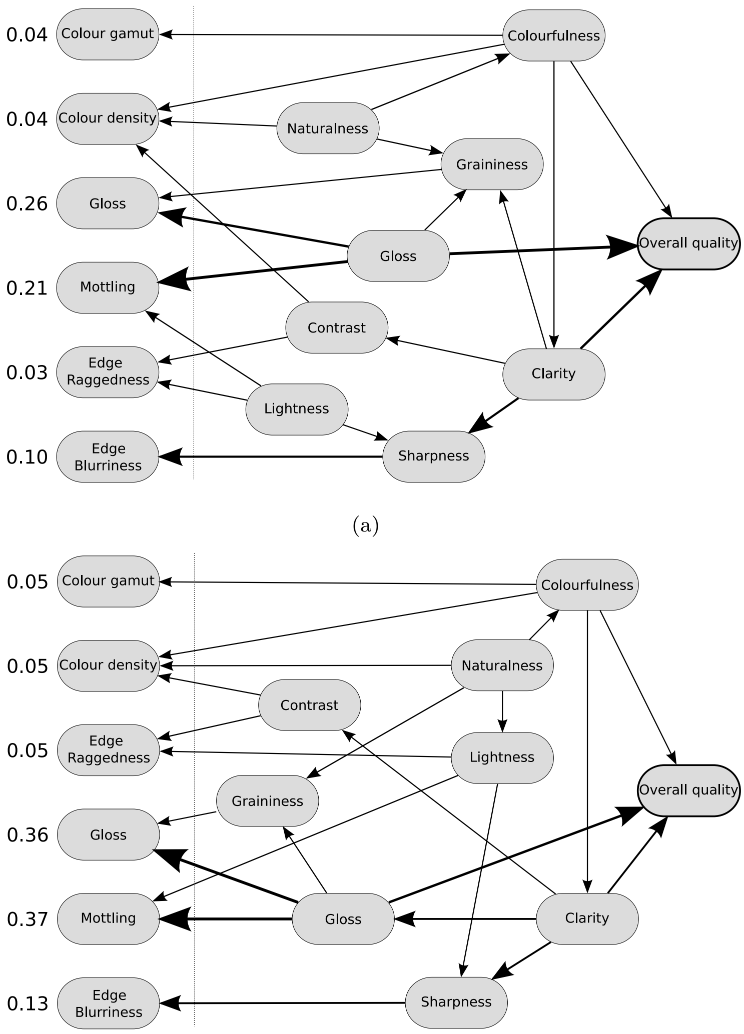
4.3. Model analysis

The significance of the edges in the produced networks was studied by removing them one at time and by computing the root mean square error (RMSE) between the output of the original model and the reduced model:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - b_i)^2}, \tag{8}$$

where a_i is the output of the original model (expected value of the marginal distribution) and b_i is the output of the reduced model. A large number ($N = 2000$) of random inputs were used in the computation. The results are shown in Fig.˜6. The line width represents the significance of an edge (large RMSE). Similarly, the significance of the inputs (objective measures) was studied and converted to scores (RMSE) in Fig.˜6 (larger score denotes higher significance).

The generality of the models was tested using the following procedure: (i) training the model with real subjective data, (ii) generating new data based on the model, (iii) re-training the model with the generated data, and (iv) validating the simulated model with real subjective data. For the generated model data, the inputs were randomly sampled from uniform distributions. Random inputs formed the evidence, and the marginal distributions for all other nodes

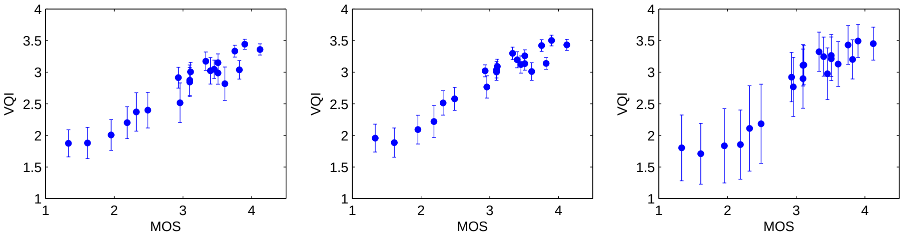


(a)

(b)

Figure 6: Sensitivity analysis of the best Bayesian network structures. The line width represents the significance of an edge, and the numbers represent the significance of the instrumental measures: (a) Best fitness function value; (b) Best correlation against the subjective evaluation (leave-one-out).

(subjective attributes and the overall quality) were computed. New data, that is, integer values for each node, were determined by sampling the computed marginal distributions. In this way, the generality of the models was tested in two stages: how well the generated data correspond to the real world data, and how well the simulated model predicts the real subjective data. The number of generated samples to train the simulated model was 2000 and the experiment was repeated 100 times. The results are shown in Fig.˜7. Error bars reperesent the standard deviations of the simulated model outputs (VQI). The mean correlation coefficients over 100 simulated model were as follows: 0.91 for network with best fitness function value (Fig.˜4(a)), 0.93 for network with best correlation to subjective evaluation (Fig.˜4(b)) and 0.80 when the correlation was optimised instead of the fitness function (Fig.˜4(c)). It can be seen that the correlation decreases significantly for the network structure found by optimising the correlation to subjective evaluation with the small training set (Fig.˜7(c)), but do not change considerably for the models where the behaviour of the model was optimised using hypothesis testing (Fig.˜7(a)-(b)). This confirms the previously mentioned assumption about generalisability becoming weak due to overfitting if only the correlation coefficient is optimised by using a small amount of subjective data.



(a)

(b)

(c)

Figure 7: Scatter plots between the model-produced visual quality index (distribution expectation) and subjective MOS when using a simulated model: (a) Best fitness function value (mean correlation: 0.91); (b) Best correlation against the subjective evaluation (mean correlation: 0.93); (c) Correlation coefficient optimised instead of the fitness function (mean correlation: 0.80).

5. Conclusions

In this paper, a Bayesian network model of overall visual print quality was presented. The learnt model in Fig.~6(a) is already an important result for further analysis of human visual quality perception and the value of subjective attribute data in modelling the perception. The second important contribution is the machine learning framework to search for the optimal structure of Bayesian network given an initial basic structure, a small amount of psychometric subjective data, and prior hypotheses concerning the model behaviour. The presented framework utilises a simulation-based fitness function and a tailored genetic algorithm to produce a set (population) of well-performing models. Selection of the best model depends on the application: is the objective to optimise the performance against the subjective evaluation in well-defined circumstances, or is there need for a model as general as possible? It should be noted, however, that if the printing method is changed, the structural optimisation needs to be re-run. The models presented here are valid only for electrophotography printing.

An interesting aspect of the presented model and its learning method is that the discovered connections and their significance help to understand the phenomenon of subjective quality evaluation. The discovered edges between the nodes (objective measures and subjective attributes) provide information about the abstract attributes that humans use for evaluating quality. The models also reveal what the nature of these subjective attributes is from the viewpoint of objective measures and how they contribute to each other. In addition, a full model, even with its limitations, is much more versatile than any traditional image quality measure. In the experiments, only the objective measures were used as evidence to infer the overall quality output, but similarly, the overall quality can be used as evidence and the distributions of the objective measures can be studied. This guides us to better understand the effects of different objective measures on the perceived quality. The result is useful in media technology as well as for psychophysical research.

Acknowledgement

The authors would like to thank Raisa Halonen from the Department of Media Technology in Helsinki University of Technology for providing the test material. The authors would like to thank also the Finnish Funding Agency for Technology and Innovation (TEKES) and partners of the DigiQ project (No. 40176/06) for support.

Armel, D., Wise, J., December 1998. An analytic method for quantifying mottle - part 1. Flexo, 70–79.

Baylor, D.~A., Lamb, T.~D., Yau, K.~W., 1979. Response of retinal rods to single photons. Journal of Physiology, London 288, 613–634.

Birge, R.~R., Barlow, R.~B., 1995. On the molecular origins of thermal noise in vertebrate and invertebrate photoreceptors. Biophysical Chemistry 55, 115–126.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Blanco, R., Inza, I., Larrañaga, P., 2003. Learning bayesian networks in the space of structures by estimation of distribution algorithms. International Journal of Intelligent Systems 18~(2), 205–220.

Bouk, T., Dalal, E.~N., Donohue, K.~D., Farnand, S., Gaykema, F., Dmitri, D., Haley, A., Jeran, P.~L., Kozak, D., Kress, W.~C., Martinezb, O., Mashtare, D., McCarthy, A., Ng, Y.~S., Rasmussen, D.~R., Robb, M., Shin, H., Quiroga, S.~M., Smith, E. H.~B., Tse, M.-K., Williams, D., Zeise, E., Zoltner, S., 2008. Recent progress in the development of incits w1.1, appearance-based image quality standards for printers. In: Proc. of SPIE IS&T Electronic Imaging. Vol. 6494.

Breese, J., Heckerman, D., 1996. Decision-theoretic case-based reasoning. IEEE Transactions on Systems, Man and Cybernetics, Part A 26~(6), 838–842.

Briggs, J., Forrest, D., Klein, A., Tse, M.-K., 1999. Living with ISO-13660: Pleasures and perils. In: IS&Ts NIP 15: 1999 International Conference on Digital Printing Technologies. IS&T, Springfield VA, pp. 421–425.

Chickering, D.˜M., 1996. Learning from Data: Artificial Intelligence and Statistics V. Springer, Ch. Learning Bayesian networks is NP-complete, pp. 121–130.

Damera-Venkata, N., Kite, T.˜D., Geisler, W.˜S., Evans, B.˜L., Bovik, A.˜C., April 2000. Image quality assessment based on a degradation model. IEEE Transactions On Image Processing 9˜(4), 636–650.

de˜Campos, L.˜M., Fernandez-Luna, J.˜M., Gamez, J.˜A., Puerta, J.˜M., 2002. Ant colony optimization for learning bayesian networks. International Journal of Approximate Reasoning 31˜(3), 291–311.

de˜Freitas˜Zampolo, R., Seara, R., 2003. A measure for perceptual image quality assessment. In: In proc. of the International Conference on Image Processing (ICIP'03). pp. 433–436.

de˜Freitas˜Zampolo, R., Seara, R., 2004. Perceptual image quality assessment based on bayesian networks. In: In proc. of the International Conference on Image Processing (ICIP'04). pp. 329–332.

Eerola, T., Kamarainen, J.-K., Leisti, T., Halonen, R., Lensu, L., Kälviäinen, H., Nyman, G., Oittinen, P., 2008a. Is there hope for predicting human visual quality experience? In: Proc. of the IEEE International Conference on Systems, Man, and Cybernetics. Singapore.

Eerola, T., Kamarainen, J.-K., Leisti, T., Halonen, R., Lensu, L., Kälviäinen, H., Oittinen, P., Nyman, G., 2008b. Finding best measurable quantities for predicting human visual quality experience. In: Proc. of the IEEE International Conference on Systems, Man, and Cybernetics. Singapore.

Eerola, T., Kamarainen, J.-K., Lensu, L., Kälviäinen, H., 2007. Visual print quality evaluation using computational features. In: International Symposium on Visual Computing. Lake Tahoe, USA, pp. 403–413.

Eerola, T., Lensu, L., Kälviäinen, H., Kamarainen, J.-K., Leisti, T., Nyman, G., Halonen, R., Oittinen, P., January/February 2010. Full reference printed image quality: Measurement framework and statistical evaluation. Journal of Imaging Science and Technology 54˜(1), 1–13.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian networks to analyze expression data. Journal of Computational Biology 7˜(3–4), 601–620.

Holmes, D., Jain, L. (Eds.), 2008. Innovations in Bayesian Networks: Theory and Applications. Springer.

Hommersom, A., Lucas, P., 2010. Using bayesian networks in an industrial setting: Making printing systems adaptive. In: Proceedings of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal.

Huang, C., Darwiche, A., 1996. Inference in belief networks: A procedural guide. International Journal of Approximate R 15˜(3), 225–263.

ISO/IEC, 2001. 13660:2001(e) standard. information technology - office equipment - measurement of image quality attributes for hardcopy output - binary monochrome text and graphic images. ISO/IEC.

Keelan, B.˜W., 2002. Handbook of Image Quality: Characterization and Prediction. Marcel Dekker, New York.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R.˜H., Kuijpers, C. M.˜H., 1996. Structure learning of bayesian networks by genetic algorithms: Performance analysis of control parameters. IEEE Transactions on Pattern Analysis And Machine Intelligence 18˜(9), 912–926.

Neapolitan, R.˜E., 2004. Learning Bayesian Networks. Prentice Hall.

Nikovski, D., 2000. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. IEEE Transactions on Knowledge and Data Engineering 12˜(4), 509–516.

Oatley, G., Ewart, B., 2003. Crimes analysis software: 'pins in maps', clustering and bayes net prediction. Expert Systems with Applications 25˜(4), 569–588.

Oittinen, P., Halonen, R., Kokkonen, A., Leisti, T., Nyman, G., Eerola, T., Lensu, L., Kälviäinen, H., Ritala, R., Pulla, J., Mettänen, M., 2008. Framework for modelling visual printed image quality from paper perspective. In: Image Quality and System Performance V. San Jose, USA.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann.

Pulla, J., Mettänen, M., Konkarikoski, K., Ritala, R., 2008. Bayesian network model as an overall image quality measurement system. In: Proc. of the 12th IMEKO TC1-TC7 Joint Symposium on Man Science and Measurement. Annecy, France.

Radun, J., Leisti, T., Häkkinen, J., Ojanen, H., Olives, J.-L., Vuori, T., Nyman, G., 2008. Content and quality: Interpretation-based estimation of image quality. ACM Trans. Appl. Percpt. 4˜(4).

Sadovnikov, A., Lensu, L., Kälviäinen, H., 2007. Automated mottling assessment of colored printed areas. In: Proc. of the 15th Scandinavian Conference on Image Analysis. Aalborg, Denmark, pp. 621–630.

Spirtes, P., Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review 9˜(1), 62–72.

van Dijk, S., Thierens, D., van˜der Gaag, L.˜C., 2003. Building a ga from design principles for learning bayesian networks. In: Proceedings of Genetic and Evolutionary Computation Conference. Chicago, USA, pp. 886–897.

Wolin, D., Johnson, K., Kipman, Y., 1998a. Automating image quality analysis. In: IS&T's NIP 14: International Conference on Digital Printing Technologies. Toronto, Ontario, Canada, pp. 627–630.

Wolin, D., Johnson, K., Kipman, Y., 1998b. The importance of objective analysis in image quality evaluation. In: IS&T's NIP 14: International Conference on Digital Printing Technologies. Toronto, Ontario, Canada, pp. 603–606.

Yang, Q., Wang, X., Huang, Z., Zheng, S., 2007. Research of student model based on bayesian network. In: Proceedings of the First IEEE International Symposium on Information Technologies and Applications in Education. Kunming, China, pp. 514–519.