ELSEVIER

PII: S0954-1810(96)00035-0

Artificial Intelligence in Engineering **12** (1998) 21–33

0954-1810/98/$19.00

Probabilistic reasoning and multiple-expert methodology for correlated objective data

Kwoh Chee-Keong

The Intelligent System Laboratory, School of Applied Science, Nanyang Technological University, Singapore 639798

&

D. F. Gillies

Department of Computing, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London, UK SW7 2BZ

(Received 19 June 1996; revised version received and accepted 12 September 1996)

In this paper, a numerical expert system using probabilistic reasoning with influence structure generated from the observed data is demonstrated. Instead of using an expert to encode the influence diagram, the system has the capability to construct it from the objective data. In cases where data are correlated, instead of compromising the performance by wrestling with different influence structures based on the assumption that all the environment variables are observed, we incorporated the flexibility of including unobservable variables in our system. The resulting methodology minimised the intervention of a domain expert during modelling and improved the system performance.

Global optimisation using all variables is often very difficult and unmanageable in probabilistic network construction. In our approach, we group all the variables into subsets and generate advice for these subsets of features using multiple small probabilistic networks, and then seek to aggregate these into a consensus output. We proposed a probabilistic aggregation using the joint probability of data and model approaches. In this approach, we avoided the very high-dimensional integration over all possible parameter configurations. The resulting system has the benefit of a multiple-expert system and is easily expandable when new information is to be added. © 1997 Elsevier Science Limited.

Key words: probabilistic network, bayesian inference, multiple-expert system, unobservable variables.

1 INTRODUCTION AND BACKGROUND

In recent years, there has been a change in the content and methodology of research in the area of artificial intelligence. It is now more common to build on existing theories to base claims on rigorous theorems or hard experimental evidence rather than on intuition.[1] The development of probabilistic reasoning is an example of the former and the latter is well represented in the field of neural networks. The probabilistic reasoning pioneered by Judea Pearl[2] marked a new acceptance of probability and decision theory in AI. The belief network formalism was invented to allow efficient reasoning about the combination of uncertain evidence, represented as probabilities. To date, the expert systems employing probabilistic reasoning are still dominant in medical related applications.[3,4] Applications of probabilistic reasoning in other areas, where domain experts are fewer or not available, are relatively limited.

The experimental framework of this research is to design and build a fully automated endoscopic navigation and advisory system. An endoscope is a medical instrument that is used for non-invasive observation of the inner surfaces of the human body. It is employed extensively in the diagnosis of colon and gastrointestinal tract diseases. Its main body has a flexible shaft with a manoeuverable tip which is usually inserted through a natural body opening. The orientation of the tip can be controlled by pull wires, or by some motorised controls. The tip has optical fibres to provide a cold light source for illumination and the visual feedback which is connected to a monitor screen and/or a frame grabber[5] for computer processing. Besides the viewing facility, there is a suction mechanism, an air-blowing

mechanism, a water jet and an extra 'operating' channel that allow the passage of flexible miniature operation instruments.

The first work on endoscopic navigation was done by Khan and Gillies[5-7] and it is set out in Khan's PhD thesis 'Machine vision for endoscope control and navigation'. Their main contribution was in the signal level processing which they implemented using one single dominant feature for lumen recognition. They proposed two methods of identifying the lumen. The first method used contour extraction. Contours are extracted by edge detection, thresholding and linking. This method requires images to be divided into overlapping squares with overlapping resolutions (8 by 8 and 4 by 4) where line segments are extracted by means of a Hough transform. Perceptual criteria, such as proximity, connectivity, similarity in orientation, contrast and edge pixel intensity, are used to group edges both strong and weak. This approach is called perceptual grouping. The second method they implemented is based on a region extraction and merging approach with spatial domain data. They used an N-level quadtree based pyramid structure to find the most homogeneous large dark region, which in most cases will correspond to the lumen (centre line of the colon). The algorithm processes the quadtree from the bottom (pixel level) up, recursively and computes the mean and variance of each region of the image corresponding to a quadtree node. On reaching the root, the largest uniform seed region, whose mean corresponds to a lumen is selected. The method works with 'local' pixel information using variance within a small region of the image to determine the most uniform seed region. Khan[5] concluded that the second method is the simplest of the two in determining the insertion direction and the easiest to implement for a real time application. He implemented the system with the simple crisp logic of if-then-else coupled with some experimental thresholds for the decision making.

In addition to Khan and Gillie's development, which is based on two-dimensional information, in the form of regions and contours, the use of three-dimensional shape could provide additional information that will enhance the system's capabilities. A two-dimensional colon image does not give direct information on the three-dimensional shape of the world. Shape or depth information from an image can be estimated by various methods. A technique particularly suitable for endoscopy is the shape from a shading algorithm developed by Rashid,[8] who assumes a point light source very close to the camera. This lighting model is a good approximation to a real endoscope. His shape from a shading algorithm obtains the relative depth of the colon surface in the image. It is simple and fast so it is suitable to be applied in real-time for navigation. The shape from the shading method reconstructs the surface normals $(p, q, -1)$ at a set of points in the image. The normals that we obtain from low-level processing consist of one vector (p, q) per pixel that gives the orientation of the surface at this point with respect to two orthogonal axes (x, y) that are perpendicular to the camera axis (z). If we assume that the colon has a shape similar to a tube and in the image only a section of the internal wall of this tube is observed, then a reasonable approximation of the position of the centre of the colon (lumen) will be a function of the direction in which the majority of the (p, q) vectors are pointing.[9] Although it was necessary to assume that the colon has Lambertian surfaces, the results could still be used to give a reasonable statistical estimate of the lumen position.

Sucar and Gillies[10] utilised both the above methods for signal level processing and implemented an advisory system for the control level based on Pearl's[2] probabilistic networks. In their design, they solicited the influence structures from the endoscopic expert, and the conditional probabilities are encoded by frequency count (the occurrence or observation of events given that some events have been observed). This created the first artificial intelligence system for navigation in the colon. Up to this point, there were two signal processing models in the system, which are analogous to two independent domain experts that provide independent opinions. Sucar and Gillies fused the posterior probabilities of these two outputs into a global output by utilising a naive Bayes structure.

In order to complement the short comings of the two models and to look into the problem from another perspective, we[11] devised and implemented a new Fourier domain method that uses global pixel information which is less sensitive to noise. As the Fourier transform extracts global information in the spatial domain, we also have an indication of the direction of search for the lumen even if it is out of the colon image. Since real time performance is required, we also simplified the mathematical equations by reducing the two-dimensional fast Fourier transform into two one-dimensional transforms for real-time processing. To minimise the edge discontinuity problem that leads to mis-classification we used a tapering window in the spatial domain. The results that we obtained demonstrate that the method is highly effective in identifying the lumen position and provides useful features for the advisory module.

The feature extraction modules described above form the first stage of the complete endoscopic navigation and advisory system. The next step is to devise the correct model for the 'control level' processing which will incorporate the information from the region segmentation model, shape from shading model and Fourier domain model. In the next section, we will outline the approach to construct the probabilistic networks from the objective data extracted from the above modules. We will also discuss our approach to modify the network when we find the recovered structure cannot represent the observed data faithfully. Finally, we will present our strategies for building smaller sub-networks

(branches of global net) to reduce the complexity of the network and the aggregation formulation.

2 PROBABILISTIC NETWORK

In his analysis of knowledge representation schemes for high level vision, Sucar[12] has shown that classical logic provides a theoretical framework for inference in deterministic systems but not for probabilistic systems. Provan[13] also argues that logic is an inadequate knowledge representation language for high level vision. In the research for representing and reasoning under uncertainty, there are not yet a general and widely accepted theorem and methodology. However, it has been accepted that numerical models should be used.

In real world problems (not restricted to high level vision), uncertainty in qualitative and quantitative information can arise from many sources: qualitative information based upon some procedures may be unreliable, it may be incomplete, ambiguous or inconsistent; quantitative values of variables are not exact as they are estimated from some mathematical model, and all information and data gathering processes incur some random elements and inaccuracies. Consequently, the implementation and the execution of the inference process, which depends on the application of a methodology, model selection criteria, and so on, also subject the system to further uncertainty (commonly known as the uncertainty about the model). Clark,[14] Sucar[12] and Ng and Abramson[15] had summarised and compared the concepts and approaches for the various techniques that have been devised for uncertainty management in artificial intelligence and expert systems. In real world problems, especially data oriented experiments, most evidence and hypotheses can be assigned to an exclusive and exhaustive list or approximated by virtual evidence.[2,16] Hence, the Bayesian approach is quite an adequate tool for a variety of modelling problems.[17–22]

The most well developed Bayesian approach is the probabilistic network, Which is also known as a Bayesian network, belief network or causal network. As the name implies, it consists of graphical structures (networks) used for representing the relationships and interactions between variables. It encodes and represents the conditionally independent information which is frequently derived from some subjective knowledge base. The second component of a probabilistic network is the matrices that store all the beliefs and likelihood information for every possible mutually exclusive and exhaustive state of the variables.

The most commonly used influence structure is singly connected. Given a singly connected probabilistic network, consider a general fragment of the network around a node, N, with multiple parents and multiple children as shown in Fig. 1. The set of all N's parents is denoted as $\mathbf{U} = \{U_1, U_2, \ldots U_{|U|}\}$ and the set of N's children is denoted as $\mathbf{Y} = \{Y_1, Y_2, \ldots Y_{|Y|}\}$. Let $\mathbf{E}$

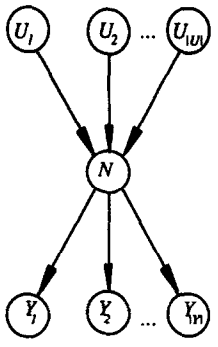

Fig. 1. The parents and children of a typical single connected node N.

represent the evidence set, then the posterior probability of a query node N, $Bel(N)$, is

$$P(N/\mathbf{E}) = P(N/\mathbf{E}^+, \mathbf{E}^-)$$

$$= \frac{P(\mathbf{E}^-/\mathbf{E}^+, N)P(N/\mathbf{E}^+)}{P(\mathbf{E}^-/\mathbf{E}^+)}$$

$$= P(\mathbf{E}^-/N)P(N/\mathbf{E}^+)\left[\frac{1}{P(\mathbf{E}^-/\mathbf{E}^+)}\right]$$

$$= \left[\frac{1}{\sum_N P(\mathbf{E}^-/N)P(N/\mathbf{E}^+)}\right]$$

$$\times P(\mathbf{E}^-/N)P(N/\mathbf{E}^+)$$

$$Bel(N) = \alpha\lambda(N)\pi(N) \tag{1}$$

using the Bayes's theorem and the conditional independence property.

Pearl (see Chapter 4 of Ref. 2) derived the propagation rule for each element, n, of the query node N, as follows

$$Bel(n) = \alpha\lambda(n)\pi(n) \tag{2}$$

The belief, posterior probability, of each element of N, n, is the product of the λ value, $\lambda(n)$, which consolidates all the evidence contribution from its children and the π value, $\pi(n)$, which consolidates all the evidence contribution from its parents; α is a normalising constant which is the joint probability of the total evidence.

From Pearl's derivation,

$$\lambda(n) = \prod_{i=1}^{|Y|} \lambda_{Y_i}(n) \tag{3}$$

where λ messages, $\lambda_{Y_i}(n)$, are the posterior probabilities of each child with the evidence from itself and its descendants reformatted to the dimension of node N.

$$\pi(n) = \sum_U P(n/\mathbf{u}) \prod_{i=1}^{|U|} \pi_N(u_i) \tag{4}$$

The π value is analogous to the posterior probability of node N with all evidence from its parents $\mathbf{U}$. The π value

involves the marginalisation over each dimension of a parent U_i together with its π message, $\pi_N(u_i)$, where each π message represents the posterior probability of each parent given the evidence from itself or its predecessors.

$$Bel(n) = \frac{P(n, \mathbf{e})}{P(\mathbf{e})} = \frac{bel(n)}{\sum_N bel(n)} \tag{5}$$

On the basis of the exhaustive property of the probabilistic approach, the posterior probability of node N, given all evidence, is the normalised product of the λ value and π value. The normalising constant represents the joint probability of evidence from all predecessors and successors.

After the node N has received a message from one of its children or parents, it has to send updating information to all other parents and children since it has revised its posterior probability. This information takes the form of π messages and λ messages away from N.

The λ message is derived as

$$\lambda_N(u_i) = \beta \sum_N \lambda(n) \sum_{U_{k,k\neq i}} P(n/\mathbf{u}) \prod_{k\neq i} \pi_N(u_k) \tag{6}$$

The λ message combines all the evidence from all other parents, in the summation over U_k, together with the λ value, which has fused all the evidential information from all the children of N, and represents the total evidential information that is required for U_i to update its posterior probability.

The π message is derived as

$$\pi_{Y_i}(n) = \alpha \frac{Bel(n)}{\lambda_{Y_i}(n)} = Bel(n)|_{\lambda_{Y_i}(n)=1} \tag{7}$$

The π message that is propagated to a particular child is the posterior probability of the node N without the evidential contribution from that particular branch. It can be understood intuitively as the updating message consolidating all the contributions from every other source of evidence except from the child node to which the message is sent.

The operating equations above deal with localised fusing and propagation of new evidence and beliefs through single connected probabilistic networks using messages so that each proposition (variable) will eventually be assigned a certainty measure consistent with the axioms of probability theory. Figure 2 shows the local variable for a node N and its links to its parents, $pr(N)$, and its children, $ch(N)$, and the messages going into and coming out of the processor. See page 168 of Ref. 2 for the detailed internal structure of a single processor of a probabilistic network.

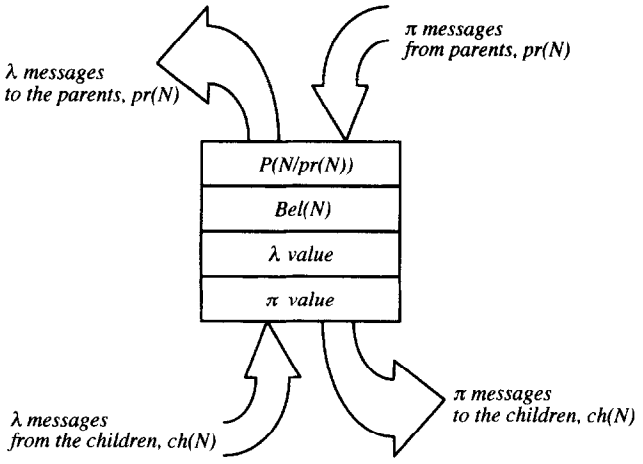

Fig. 2. The block diagram showing flow of signals in a generic node of the probabilistic network.

3 PROBABILISTIC REASONING FOR CONTINUOUS VARIABLES

The work in the above section addressed the problem of information fusion and propagation for discrete variables. These approaches are well suited for a situation where information is propositional and the conditional probability matrices are in the form of a contingency table which quantifies the association between the variables. However, in a natural environment there are quantities that are better thought of as continuous, linear or monotonic in nature. Measurements like time, weight and money, where the resolution may be fine, may require some compromises when handled by discrete variables. A possible solution is to quantise the continuous variables to discrete representation, and this has advantages in some situations. However it is often expensive both in memory use and computation time, especially if high precision is required. If the variables involved in the reasoning process are all continuous and can be represented by some known probabilistic density functions governed by some parameters, it will be better to use an efficient message propagation scheme for continuous variables. Such a scheme only sends parameters about the probabilistic density function to its neighbour given the observed information. In our work on the endoscopic navigation and advisory system, we have more than one mathematical model to estimate the lumen location. To combine all the estimates from the various models we proposed the use of a continuous probabilistic network where the uncertainties are approximated by Gaussian distributions. Figure 3 shows the information involved in the reasoning about the location of lumen.

Figure 4 shows some of the images where all the three models provide estimation of the lumen location. In these images, the cross is the estimated location of the lumen by the Fourier domain model with an ellipse indicating the estimated size of the lumen; the square is the large dark region estimated by the region based segmentation model and its centre will

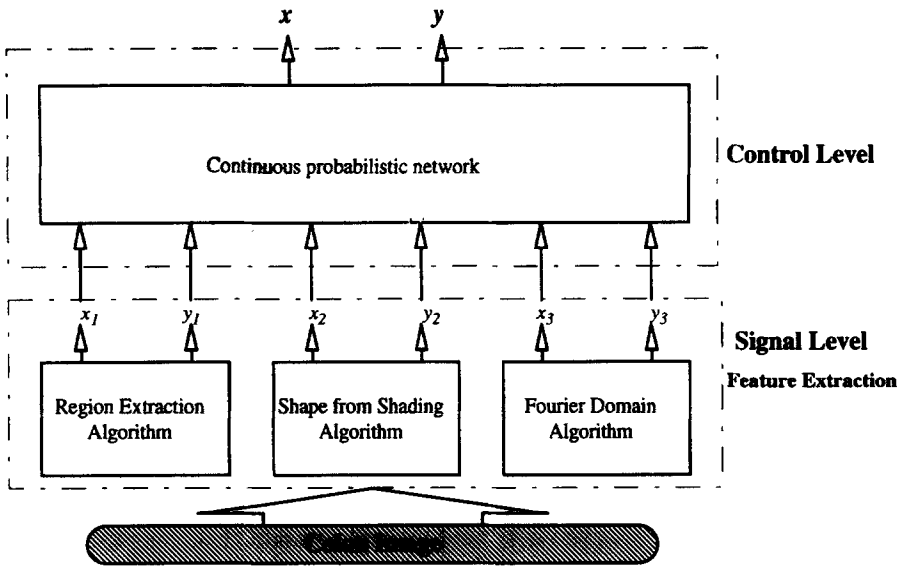

Fig. 3. Continuous probabilistic network for estimation of the lumen location in the navigation module.

be the estimated location of the lumen and the diamond in the image represents the estimated location of the lumen by the shape from shading algorithm.

Examples of the continuous variables to be decided are the system output, x and y, from individual model outputs $x_1, x_2, x_3, y_1, y_2, y_3$. If we assume the estimation in the x and y directions are independent, we can use one continuous probabilistic network for each set of variables, namely $\{x, x_1, x_2, x_3\}$ and $\{y, y_1, y_2, y_3\}$. The simplest probabilistic network that is adequate to model the pooled estimate and variance for x is shown in Fig. 5. A similar structure is also used to pool the estimates for y.

The propagation rule for continuous Gaussian variables is derived with the following assumptions:

(1) All interactions between variables are linear.
(2) The sources of uncertainty are normally distributed and are uncorrelated.
(3) The causal network is singly connected.

In the continuous probabilistic network, the π and λ messages are characterised by the means and variances of their Gaussian conditional densities. The derivation for a general single connected network can be found in Pearl.[2] The operating equations for data fusion for Fig. 5 are summarised as follows.

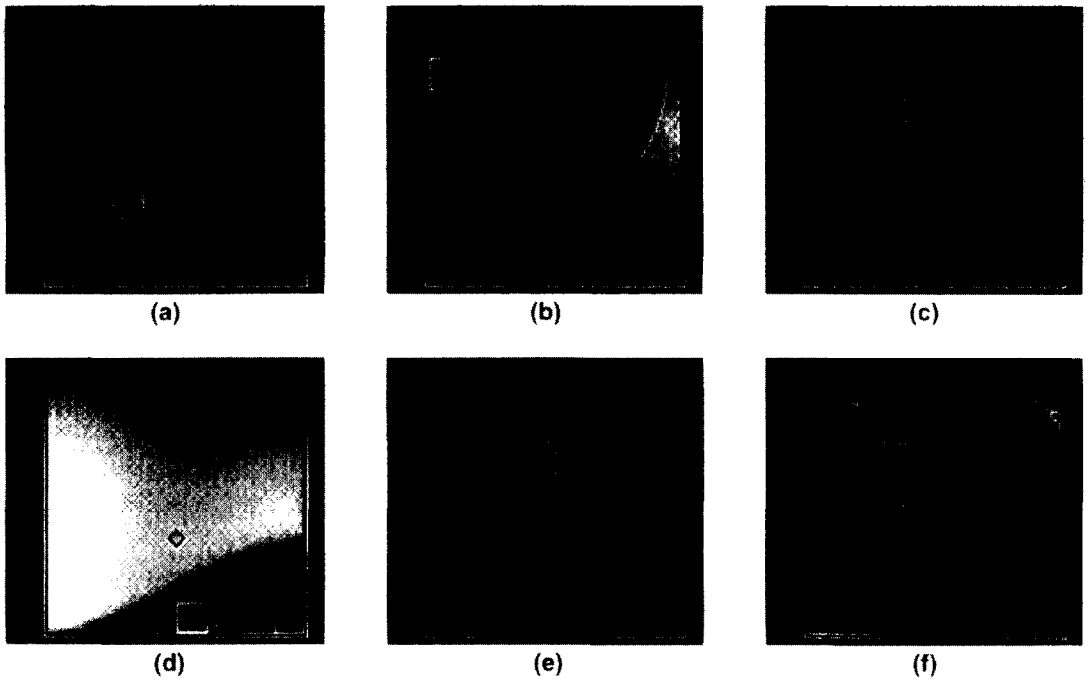

Fig. 4. Estimated lumen location by all three models.

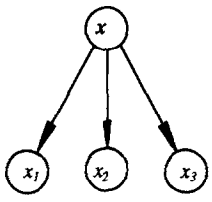

Fig. 5. Probabilistic network for estimation of x co-ordinate of lumen location.

The message from the parent of x is

$$\pi_x(u_i) = f(u_i/e_i^+) = N(u_i; \sigma_i^{2+}, \mu_i^+) \tag{8}$$

In eqn (8), it is said that the evidence from parent u_i, depends on the mean and the variance of the u_i governed by a Gaussian density function. These are the only two parameters that must be communicated to x. Since x is the root node in Fig. 5, we have $\pi_x(u_i)$ equal to a flat improper density function with unity mean and infinite variance. Hence the only contribution will come from its children.

$$\lambda_{x_j}(x) = f(e_j^-/x) = N(x; \sigma_j^{2-}, \mu_j^-) \tag{9}$$

The λ message from either of the children, say x_j, is also governed by the Gaussian density function. The two parameters that must be communicated are the mean and the variance from the child.

Knowing the parameters that node x receives, we have to update its π and λ values as in the discrete case. Similar to what it receives, its π and λ values are each quantified by their means and variances.

For the λ value,

$$\lambda(x) = f(e^-/x) = N(x; \sigma_\lambda^2, \mu_\lambda)$$

$$\sigma_\lambda^2 = \left[\sum_j \frac{1}{\sigma_j^{2-}}\right]^{-1}$$

$$\mu_\lambda = \sigma_\lambda^2 \sum_j \frac{\mu_j^-}{\sigma_j^{2-}} \tag{10}$$

which is exactly the pooled mean and variance as used in classical statistics involving blocked data.[23]

Since there is no parent, the π value will be

$$\sigma_\pi^2 = \infty$$

$$\mu_\pi = 1 \tag{11}$$

which guarantees a flat distribution that has no preference to any value at all.

The posterior probability density function for x is fully specified by

$$Bel(x) = f(x/\mathbf{e}) = N(x; \sigma_x^2, \mu_x)$$

$$\sigma_x^2 = \frac{\sigma_\pi^2 \sigma_\lambda^2}{\sigma_\pi^2 + \sigma_\lambda^2}$$

$$\mu_x = \sigma_\lambda^2 \left(\frac{\mu_\lambda}{\sigma_\lambda^2} + \frac{\mu_\pi}{\sigma_\pi^2}\right) \tag{12}$$

3.1 Estimation of parameters

Putting everything together in our system, we first estimated the variance for each model that provide estimates of the lumen location. In order to achieve that, we collected 100 images at random and used an expert's opinion to identify the centre of the colon (lumen), $[x_T, y_T]$. We then ran each model to find their estimated co-ordinates, namely $[x_1, y_1]$ for the region based segmentation estimate, $[x_2, y_2]$ for the shape from shading estimate and $[x_3, y_3]$ for the Fourier domain estimate. Variances of these methods were then computed by summing all the squares of differences and divided by the degree of freedom. Table 1 summarises the variance in the x and y co-ordinates (in pixels).

From the variance in Table 1, evidently the Fourier domain model has the most influence in the pooled estimate of the lumen co-ordinate in both the x and y directions, which can be explained by its use of 'global' information.

4 PROBABILISTIC REASONING FOR ADVISORY MODULE WITH OUTPUT FROM SIGNAL PROCESSING MODULES

Beside the navigation module, we have pictorial recognition and advisory module to generate advice for the endoscopist. This module is a numerical expert system that will aid an endoscopist in decision making by suggesting the right course of action with expert information. Figure 6 shows the information flow of probabilistic networks in our advisory module.

In the advisory module, the output from the system is usually propositional. For example in the application to endoscopy, it can be: push the endoscope; pause and search for lumen; pull back; inflate to open the intestine; suck to clear away the fluid, etc. Since these propositions represent alternatives that are mutually exclusive and further assumed to be exhaustive by including an 'others' proposition to replace the 'etc.' to complete the list, we can model the system with probabilistic network(s).

5 CONSTRUCTING A NETWORK FROM OBJECTIVE DATA

A numerical expert system often consists of many subsystems, each with an associated probabilistic network.

Table 1. Variance for each model in estimating the lumen locations in x and y co-ordinates (in pixels)

	Variance σ^2 (standard deviation)	
Model name	x	y
Region segmentation (1)	113·97 (10·67)	158·41 (12·59)
Shape from shading (2)	435·35 (20·86)	968·38 (31·11)
Fourier domain (3)	92·03 (9·59)	107·89 (10·38)

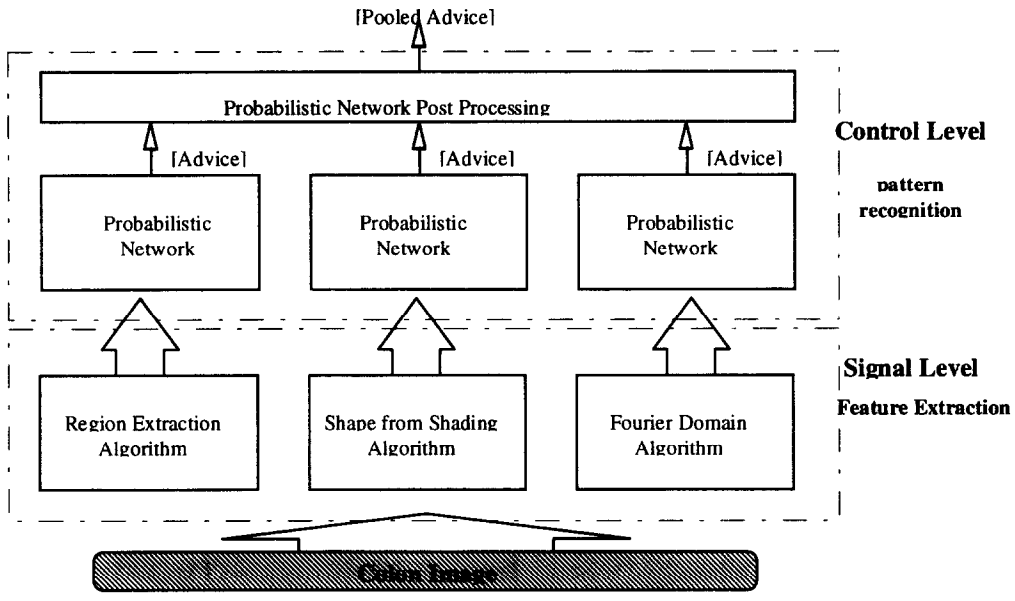

Fig. 6. Block diagram of various levels of processing in our endoscopic advisory and control system for advisory module.

They can be considered as branches in the total network. In the simplest approach, we assume that all the interacting variables are observed and we want to construct a probabilistic network for each sub-system, which we simply call a model, using all the observed variables. The observed data, together with the topology, derived from a knowledge base, will be translated into prior and conditional probabilities for each state of the variables. In order to determine the mapping from the problem to the solution space, the probabilistic network knowledge based system must be constructed from available data and information.

In the closed world definition, if **V** is the set of all interacting variables $\{V_1, V_2, V_3, \ldots\}$ for a model, then

$$P(\mathbf{V}) = \prod_{V_i \subset \mathbf{V}} P(V_i/pr(V_i)) \qquad V_i, pr(V_i) \subseteq V \tag{13}$$

where $pr(V_i)$ is the parent of the variable V_i. In other words, the joint probability of $P(\mathbf{V})$ can be expressed as the product of the conditional probabilities of each variable given the state of their parents and **V** is the set which contains all the possible combinations of V_i. (Throughout the text we use boldface to denote a collective of variables and italic to denote a variable. In general, lowercase letters are used to refer to an element of the corresponding uppercase variable.)

If we assume that the extracted features, denoted as variables **V**, are all that exist and are required to model a system, then we would expect to have observed data for all nodes in the desired network (maybe with occasional missing data for some elements). Many researchers have developed algorithms for constructing the network topology from empirical observations.[24–28] Most of their algorithms are improvements of the maximum-weighted spanning tree algorithm first formulated by Chow and Liu[29] which utilised a mutual information measure.

Chow and Liu[29] defined a divergence measurement between the true (measured) distribution P and the tree-dependent distribution P_t as

$$I(P, P_t) \sum_{\mathbf{x}} P(\mathbf{X}) \log \frac{P(\mathbf{X})}{P_t(\mathbf{X})} \tag{14}$$

which in fact is the cross-entropy measurement of the true distribution and the distribution encoded in the tree network. Using eqn (14) and only second order statistics, they choose the best approximated tree structure of the nth distribution by a $n-1$ second order distribution (the conditional probabilities) and one first order distribution (the prior probability of root). They derived the following equation using eqns (13) and (14)

$$I(P, P_t) = -\sum_{i=1}^{|X|} I(X_i, pr(X_i)) + \sum_{i=1}^{|X|} H(X_i) - H(\mathbf{X}) \tag{15}$$

In the above equation, $H()$ is the entropy measurement of a (marginal) distribution. $I()$ is the cross-entropy measurement between two distributions and $pr()$ represents the casual parent as in eqn (13). Since the last two terms of (15) are constants, minimising the divergence is equivalent to maximising the first term, the total branch weight. Hence, their algorithm is known as the maximum weighted spanning tree (MWST).

These types of maximum connection weight algorithms have the big advantage of not needing to consider all the possible trees that could be constructed from purely objective data. However, the possible ignorance of some interacting variables will generate many probabilistic networks that could closely approximate the given observed data.

During influence network construction, it is always

assumed that the variables starting from the same parent are conditionally independent. In practice, this assumption may not hold during validation, and ignorance of it will give rise to incorrect inferences.

Many reported works did not explain the validation process. In this section, we will look at the criteria to validate conditionally independence and we will look into the strategy of using hidden nodes as unobservable variables to model the dependency in the following section.

To depict the causal relationship graphically, the variables will be represented by vertices (nodes) and their relationships by directed edges (links). If A and B are two variables and the parent of B, $pr(B)$, is A, then A and B will be linked by a directed edge with an arrow pointing from A to B. In most modelling applications, the underlying structures for the observed variables are assumed to be tree-structured. Because of this assumption, we can investigate the conditional independence in the minimal case of three adjacent nodes, called a *triplet*. Figure 7 depicts the three possible types of adjacency for a triplet in a polytree. In types (a) and (b), variables B and C are both conditionally independent given A. Type (c) however, depicts the relationship that B and C are conditionally dependent given A. If we have a star structure, we will decompose it into triplets for verification.

In the probabilistic networks, the influence diagram represents purely qualitative relationships and the probability distributions, **P**, will encode the quantitative values of the distribution. Hence validating conditional independence, under the assumed causal relationship, is of utmost importance, and we will now discuss two of the criteria that we used in our system. One criterion is derived for continuous Gaussian variable triplets and the second criterion is derived for nominal variable triplets that represent two opposite types of distribution.

5.1 Conditional Pearson's test

The first is a classical statistical approach called the conditional Pearson's correlation coefficient test. It is best suited for continuous Gaussian variables. The test validates the conditional independence assumption for the children's distribution over the parent's distribution. If, for example, suppose that there are three variables A, B, C and we want to test whether B, C are conditionally independent given A, that is to say:

$$P(A,B,C) = P(A)P(B/A)P(C/A)$$

Let the covariance matrix of the variables set $\{A,B,C\}$ be:

$$COV(A,B,C) = \begin{bmatrix} \sigma_A^2 & \sigma_{AB} & \sigma_{AC} \\ \sigma_{AB} & \sigma_B^2 & \sigma_{BC} \\ \sigma_{AC} & \sigma_{BC} & \sigma_C^2 \end{bmatrix}$$

$$COV(B,C,A) = \begin{bmatrix} \sigma_B^2 & \sigma_{BC} \\ \sigma_{BC} & \sigma_C^2 \end{bmatrix} - \begin{bmatrix} \sigma_{AB} \\ \sigma_{AC} \end{bmatrix} [\sigma_A^2]^{-1}$$

$$\times [\sigma_{AB} \quad \sigma_{AC}] = \begin{bmatrix} Var(B/A) & Cov(B,C/A) \\ Cov(B,C/A) & Var(C/A) \end{bmatrix} \quad (16)$$

We use the correlation coefficient (ρ), which defined for any two variable sets $\{B,C\}$ given the assumed parent set $\{A\}$ as:

$$\rho(B,C/A) = \frac{Cov(B,C/A)}{\sqrt{Var(B/A)Var(C/A)}} \quad (17)$$

The $\rho(B,C/A)$ means that correlation coefficient (ρ) is obtained through conditional covariance and conditional variances. A high correlation coefficient between a pair of variables, given their parent(s), indicates that the conditional independence assumption is weak.[30]

5.2 Conditional mutual information criteria

The second test is based on information theory and shares several common properties with the contingency table statistics for nominal data. This approach uses the conditional mutual information criteria to validate the conditional independence assumption).[24] The conditional mutual information for variables $\{B,C\}$ given A, denoted by $I(B;C/A)$, is defined as:

$$I(B;C/A) = E\left[\sum_{B,C} P(b,c/a) \log \frac{P(b,c/a)}{P(b/a)P(c/a)}\right]$$

$$= \sum_{A,B,C} P(a,b,c) \log \frac{P(b,c/a)}{P(b/a)P(c/a)} \quad (18)$$

where lower case a, b, c represent the members of A, B, C.

5.3 Procedures to improve the influence structure

Our strategy for improving the network is as follows. First, we calculate the conditional dependency values using both criteria and a performance measure (which we will discuss shortly), and then we modify the influence diagram. Following that we recalculate the

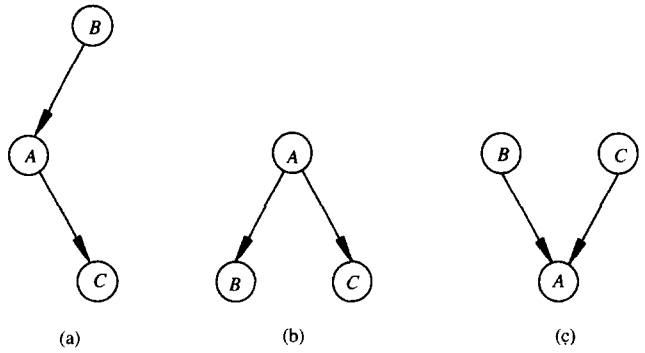

Fig. 7. Three possible types of adjacent triplets in a polytree.

conditional dependency values for both criteria and the new performance measure. We accept the modification if both criteria improve or one improves without worsening the other and the overall performance is improved. It is important to note that these criteria are derived for different distributions, and we cannot expect to have real life data that are completely conditionally independent according to both criteria. The data have either nominal independence or ordinal independence. We believe this qualitative approach is rational since we do not have the information to define an exact criterion for our data.

6 COPING WITH HIGHLY CORRELATED DATA

In the earlier work,[31] the advisory module was built with QUALQUANT methodology which assumes an expert who can identify all the conditional independence for the feature variables. However, during validation, there were dependencies that were not identified by the domain expert. To overcome the problem, Sucar and Gillies suggested a methodological solution, namely consultation with experts, to derive a node that makes the two independent variables conditionally independent. However, it will, in general, be a very difficult process for the expert to define a function that will combine the information from the two evidence variables into a coherent variable. Hence, we devise a way to create a hidden node[32] based on the statistical distribution of the two evidence variables and an objective function that satisfies the axioms of conditional independence in the framework of the probabilistic methodology. Our approach is to use the training data, without seeking expert opinion, to define a mapping which will fuse the dependent information.

6.1 Using hidden nodes as unobservable variables

In order to determine the conditional probability matrices for the hidden node, we use a gradient descent method. The objective function to be minimised is the squared-error between the measured and computed values of the instantiated nodes. Let the variable **A** have A states and a set of training data be denoted $\mathbf{E} \in \Im$ if we are interested in the posterior probability (belief) of **A** given some evidence **E**, denoted as $Bel(\mathbf{A})$. Then we express $Bel(\mathbf{A})$ as a non-linear function of all the evidence

$$Bel(\mathbf{A}) = f(\mathbf{E})$$

One formulation is the expectation of the squared-error cost function, Δ, and is

$$\Delta = E\{\xi\} = E\left\{\sum_{i=1}^{A}[D(a_i) - Bel(a_i)]^2\right\} \tag{19}$$

where $E\{.\}$ is the expectation operator, and $D(a_i)$ is the desired value of a_i. Using the joint probability of the input and the desired output, we[32] have shown that when the network parameters are chosen to minimise a squared-error cost function, the outputs are the conditional expectations of the desired outputs which minimise the mean-squared estimation error.

$$\min\{\Delta\} = \min\left\{E\left\{\sum_{i=1}^{A}[E\{D(a_i)/\mathbf{E}\} - (Bel(a_i))]^2\right\}\right\} \tag{20}$$

In eqn (20), $E\{D(a_i)/\mathbf{E}\}$ is the expected belief of **A** given the evidence **E**, and

$$E\{D(a_i)/\mathbf{E}\} = \sum_{j=1}^{A} D(a_i)P(a_j/\mathbf{E})$$

is the conditional probability of desired states given the evidence **E**, weighted by the instantiated value of the states vector.

In probabilistic networks, there is no strict direction of signal flow, and so queries can be made at any node in the network. Thus the training data set **X** is an unordered collective $\{X_1, X_2, X_3, \ldots, X_{|X|}\}$. Since probabilistic networks must be able to handle partial evidence, where some nodes remain un-instantiated, we must use a formulation to encompass situations:

$$\min_{\Delta}\left\{\Delta = \sum_{i=1}^{|X|}\sum_{Z\subseteq\mathbf{X}\setminus X_i} E\left\{\sum_{j=1}^{|X_i|}[Q(x_{ij}) - Bel(x_{ij}/\mathbf{Z})]^2\right\}\right\} \tag{21}$$

In our experiments we found that if we encompass both forward and backward propagations to compute the node probabilities, we can achieve very good results. When using a chained rule to perform partial derivatives, we have shown that the error gradients can be treated as updating messages and can be propagated in any direction throughout any singly-connected network. We use the simplest node-by-node creation approach for parents with more than two children. We tested our approach on different networks in an endoscope guidance system and, in all cases, demonstrated improved results.

6.2 Using orthogonal transformation for subsets of co-exist features

A hidden node approach is a very general method, however, it increases the complexity of the model,[30] and in the worst case, for n observables, we may need $(n-2)$ hidden nodes. Furthermore, with the introduction of hidden variables, there is a need for an iterative process (typically 100–1000 epochs) to search for a near optimal solution. The related issue in practical implementation of the searching strategy itself is a big area of research (operational research) and is covered extensively in Ref. 30. The use of an orthogonal transform need not involve

the same increase in network complexity, however, it maps a set of n observed data to a new set of data. The link matrices (conditional probabilities) can be constructed in one pass as the required topology is the simplest naive Bayes' structure.

The networks used in the endoscope control and advisory system have the following common characteristics: (1) the features for each sub-system always exist concurrently; (2) the query node is usually the root node; (3) the evidence nodes will never be queried in run time, where the most probable configuration is not of interest to the user. For example in the Fourier domain sub-system, we never terminate the software execution when one or two features are extracted. The difference in running time for extracting just the value [X-size] as opposed to all three values {[X-size], [Y-size] and [Energy]} is marginal, since these features are estimated from some common intermediate data. Knowing the behaviour of our feature extraction algorithms inspires us to perform a transformation with these data sets into a new orthogonal space by assuming that the data can be approximated by multi-variate normal distribution.[33]

The transformed variables for the Fourier domain sub-systems in the endoscope control and advisory system are as shown in Fig. 8. In order to avoid confusion with the probabilistic link of the probabilistic networks, we use solid lines and arrows to represent the causal links of probabilistic networks. Dashed lines and arrows indicate the direction of flow of data and information. For the sub-model, the residual sum-of squared error reduced from 0·1530 without orthogonal transformation to 0·1008 after transformation using a naive Bayes structure. This represents around 85% correct prediction if we use 1-of-m state discrete representation where the highest posterior probability is taken as the recommended state. Similar orders of improvement were observed in other sub-models. In Fig. 8, the mathematical model where the information is fused in the post-processing is another probabilistic network.

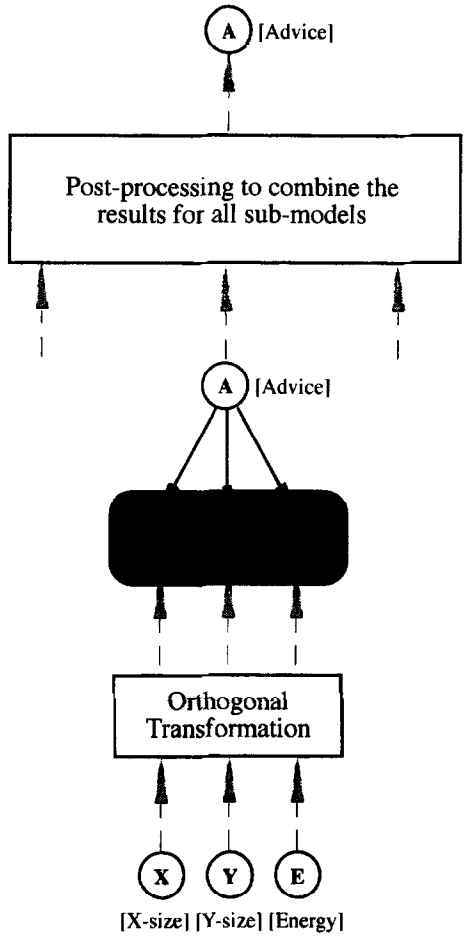

Fig. 8. Endoscope control and advisory system with orthogonal transformation. Solid lines and arrows represent the probabilistic networks. Dashed lines and arrows indicate the direction of flow of information.

7 MULTIPLE SUB-MODELS

In section 6, we discussed the strategies to cope with correlated data. In this section, we cover the macro-level of task diversion. Teams, groups, committees and panels play an important role in the modern world. The decision maker, chairman or leader should derive a consensus or select the best approach among all the contributions. In our system, we use multiple models at the micro-level, similar to the distributed system approach where each model represents the knowledge of an expert, to 'look' at the colon image from a different viewpoint. Each of these models offers its advice from the relevant findings. Since all the findings are derived from the colon image, the differing pieces of advice are dependent on each other, but we do not expect a consensus since each model utilises a different feature extraction model. Due to the lack of unanimous results, some co-ordination mechanism is necessary to arrive at the final decision. The easiest strategy is to discard all but one model that is supposed to give the best advice. Although this is a common strategy, many researchers feel it is only appropriate as a last resort. Another ambitious strategy is to build a global network for all the extracted information.

It is well known that global optimisation using all findings is usually very difficult and unmanageable. In our approach, we used multiple probabilistic networks for subsets of findings. The reasons are as follows. First, to build a global network that processes all the findings is a formidable task. There are many possible interactions between all observed variables and it is difficult to understand, encode, verify and modify the interaction between them. Secondly, the domain expert, when given all the observable variables, usually utilises only a few 'important' features for inference. Thus we propose to generate 'local' results from subsets of feature extraction models and use a posterior processing model to fuse these individual results into a global solution. This

approach has proved to exhibit the benefit of probabilistic multi-knowledge-base systems that combine different expert contributions.

7.1 Probabilistic aggregation using the joint probability of the data and model

Since the standard way of expressing the uncertainty of the model[26,34] in the posterior compromise approach involves a very high-dimensional integration of equation and direct evaluation of it can become impossible, we propose a method of expressing multiple models using the following pair of equations,

$$P(A/\mathbf{D}) = \sum_k P(A/M_k, \mathbf{D})P(M_k/\mathbf{D}) \tag{22}$$

$$P(M_k/\mathbf{D}) = P(\mathbf{D}/M_k)\frac{P(M_k)}{P(\mathbf{D})} \tag{23}$$

The first part of (23) is the likelihood of the observed data configuration to be generated by the given model and can be easily estimated by calculating the joint probability of the configuration of leaf nodes using only the prior distribution, without any instantiation of any node in the model. Putting the two equations together, we have

$$P(A/\mathbf{D}) = \frac{1}{P(\mathbf{D})}\sum_k P(A/M_k, \mathbf{D})P(\mathbf{D}/M_k)P(M_k) \tag{24}$$

Since $1/P(\mathbf{D})$ is common for all the models, it can be treated as the normalising constant and be removed from eqn (24). The next step to simplify this equation is to split the observed data $\mathbf{D}$ into two subsets $\{\mathbf{D}_k, \mathbf{D}_{-k}\}$ where the first subset $\mathbf{D}_k$ denotes all the findings that are relevant to the model k, and the second subset $\mathbf{D}_{-k}$ denotes all findings that are irrelevant to the model k.

$$\begin{aligned} P(A/\mathbf{D}) &= \frac{1}{\alpha}\sum_k P(A/M_k, \mathbf{D}_k, \mathbf{D}_{-k}) \\ &\quad P(\mathbf{D}_k, \mathbf{D}_{-k}/M_k)P(M_k) \\ &= \frac{1}{\alpha}\sum_k P(A/M_k, \mathbf{D}_k)P(\mathbf{D}_k/M_k) \\ &\quad P(\mathbf{D}_{-k}/M_k)P(M_k) \\ &= \frac{1}{\alpha}\sum_k P(A, \mathbf{D}_k/M_k)P(D_{-k})P(M_k) \end{aligned} \tag{25}$$

The first term of eqn (25) is the joint probability of advice with the relevant findings instantiated for the model. This is the non-normalised belief of the root node. The second term, $P(\mathbf{D}_{-k})$, is to account for the model predicting irrelevant findings. Since these nodes are independent from the model, such as [X-size] with respect to the region segmentation model, each of their probabilities will be taken as 1 over the number of states of the node. Doing this is equivalent to expressing total ignorance. This leaves us with the last term, which is the prior probability of the model. The simplest approach is to assume equal prior probabilities for each model or to set them using a function of performance statistics during training.

Another possible strategy uses the joint probability of the data and model is to select the result of the model that has the greatest joint probability. In doing so, we are putting our faith in the model that has the most experience with the particular configuration of data such that

$$P(A/\mathbf{D}) = \frac{1}{\alpha}\max_k[P(A, \mathbf{D}_k/M_k)P(\mathbf{D}_{-k}/M_k)P(M_k)] \tag{26}$$

When we put the maximum weight on the model that has the most prior experience with a particular configuration, we based our decision on the most reliable source. Just as a decision maker would like to ask an expert: 'How many cases of ... have you seen before?' and take that information into his decision process.

7.2 Experimental result

In our design for the endoscopic navigation and advisory system, we generate advice for subsets of features using multiple small probabilistic networks. We then seek to aggregate these different pieces of advice for a consensus output. In our experiment, we trained the network with 100 samples and tested the performance with another 290 data sets. The query node, which is constructed as the root node, has three possible states of advice and all the leaf nodes have six to ten ordinal states.

Table 2 summarises the different methods of choosing $P(M_k)$, which is the prior probability of the model, for the probabilistic posterior compromise method discussed in section 7.1. The choices that we considered are:

(a) The prior of the model is taken as one minus the expected sum-of-squared error of the model during training:

$$P(M_k) = 1 - \Delta_{\mathrm{SE}}(M_k)$$

(b) The prior of the model is taken as one minus the squared root of the expected sum-of-squared error of the model during training:

$$P(M_k) = 1 - \sqrt{\Delta_{\mathrm{SE}}(M_k)}$$

(c) The prior of the model is taken as the frequency of correct prediction during training.

(d) The prior of the model is taken as the correlation performance of the root node measured during training.

(e) Equal priors are assigned to each model

$$P(M_k) = 1$$

Table 2. Summary of the different choice of $P(M_k)$, the prior probability of the model, for the probabilistic posterior compromise approach

$P(M_k)$	Brier score (Δ_{SE})	No. correct
(a)	0·1641	0·6655
(b)	0·1646	0·6655
(c)	0·1644	0·6655
(d)	0·1635	0·6655
(e)	0·1652	0·6690
(f)	0·1615	0·7172

(f) Choose the prediction of model with the maximum joint probability of the data and model.

$$P(M_k) = \infty$$

All values are normalised to conform to axiom of probabilities.

For our experiment, a total ignorance system should have the prediction of 33% correct as there are three possible states of output. In our experimental results, we found that choosing the prediction of the model with the maximum joint probability of the data set and model produced the best performance. It is worth noting that the choice of prior for the posterior compromise approach has little influence on the performance in our system. This could be due to the fact that all our models perform equally well with some types of data and equally poorly with others.

8 CONCLUSION

Application of the methodology is demonstrated using an expert system for colon endoscopy. This provides a good test case because of the high degree of uncertainty in the knowledge and data, and the availability of real data from many different cases. Furthermore, with many independent researches to extract relevant visual information, we have a multiple expert knowledge based system with various probabilistic networks being implemented to infer from the main features for navigation. The system was tested with a large sample of real images from colonoscopy. The results show strong empirical evidence supporting our approach.

Furthermore, with the introduction of hidden nodes or orthogonal transformation, we do not have the same dilemma of choosing between different topological structures none of which fit the data coherently. When a probabilistic aggregation mechanism using the joint probability of the data and model is used for post-processing of results from all sub-models, we took into consideration of all advices from different 'experts' (sub-models) to come to the final inference. The overall system for the advisory module improved its performance from 80% correct to consistently above 90% for a trained data set, and above 75% for an untrained data set.

In this paper, we have presented our methods for handling correlation objective data in probabilistic networks. We have also discussed most of the relevant issues in constructing an objective probabilistic network and the methodology of multiple-expert systems. We verified all the algorithms in the endoscope navigation and advisory system. We believe our work has presented enough information to demonstrate the capabilities of a system built from objective data. We believe there are still numerous research areas that can be explored from our work.

REFERENCES

1. Russell, S. J. & Norvig, P., *Artificial Inteligence, A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 1995.
2. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
3. Anderson, K., Olesen, K. G., Jensen, F. V. & Jensen, F., HUGIN – a shell for building Bayesian belief universes for expert systems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, 20–25 August 1989, Vol. 2, Detroit, Michigan, Morgan Kaufmann, pp. 1080–5.
4. Heckerman, D., Probabilistic similarity networks. ACM Doctoral Dissertation Award, MIT Press, Cambridge, MA, 1990.
5. Khan, G. N., Machine vision for endoscope control and navigation PhD thesis, Imperial College, London, 1989.
6. Khan, G. N. & Gillies, D. F., A highly parallel shade image segmentation method. In *Proc. International Conference on Parallel Processing for Computer Vision and Display*, University of Leeds, 1988.
7. Khan, G. N. & Gillies, D. F., Extracting contours by perceptual grouping. *Image and Vision Computing*, 1992, **10** (2), 77–88.
8. Rashid, H., Shape from shading and motion parameter estimation under the near light source illumination. Ph D thesis, Imperial College of Science, Technology and Medicine, London, 1991.
9. Sucar, L. E., Gillies, D. F. & Rashid,, H., Integrating shape from shading in a gradient histogram and its application to endoscope navigation. In *5th International Conference on Artificial Intelligence (ICAI-V)*, Cancun, Mexico, 1992.
10. Sucar, L. E. & Gillies, D. F., Expressing relational and temporal knowledge in visual probabilistic networks. In *Uncertainty in Artificial Intelligence*, 1992, ed. D. Dubois, *et al.* North-Holland, 1992, pp. 303–309.
11. Kwoh, Chee Keong & Gillies, D. F., Using Fourier information for the detection of the lumen in endoscope images. In *IEEE Region 10's Ninth Annual International Conference, Proceedings of TENCON Conference*, Aug. 1994, Singapore, pp. 981–5.
12. Sucar, L. E., Probabilistic reasoning in knowledge-based vision systems. Ph D thesis, Imperial College, London, 1991.
13. Provan, G. M., An analysis of knowledge representation schemes for high level vision. In *Proc. First European Conference on Computer Vision (ECCV-90)*, 1990, pp. 537–41.
14. Clark, D. A., Numerical and symbolic approaches to uncertainty management in AI. *Artificial Intelligence Review*, 1990, **4**, 109–46.

15. Ng, Keung-Chi & Abramson, B., Uncertainty management in expert systems. *IEEE Expert*, April 1990, 29–47.
16. Neapolitan, R. E., *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley, New York, 1990.
17. Finn, V. J., Christensen, H. I. & Nielsen, J., Bayesian methods for interpretation and control in multi-agent vision systems. in *SPIE*, **1708**, *Application of Artificial Intelligence X: Machine Vision and Robotics*, 1992, 536–48.
18. Madigan, D. & Raftery, A. E., Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 1994, **89** (428), 1535–46.
19. Spiegelhalter, D. J., Dawid, A. P., Lauritzen S. L. & Cowell, R. G., Bayesian analysis in expert systems. *Statistical Science*, 1993, **8** (3), 219–83.
20. Spiegelhalter, D. J., Harris, N. L., Bull, K. & Franklin, C. G., Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. *Journal of the American Statistical Association*, June 1994, **89** (426), 435–43.
21. Ibarguengoytia, P. H., Sucar, L. E. & Vadera, S., Real time intelligent signal validation in power plants. In *International Federation of Automatic Control (IFAC)*, Mexico, December, 1995.
22. Sucar, L. E., Perez-Brito, J. & Ruiz-Suarez, J. C. Induction of dependency structures from data and its application to ozone predication. In *IEA/AIE-95*. Melbourne, Australia, 1995.
23. Montgomery, D. C., *Design and Analysis of Experiments*. John Wiley, New York, 1976.
24. Rebane, G. & Pearl, J., The recovery of causal poly-trees from statistical data. In *Uncertainty in Artificial Intelligence, 3*, ed. L. N. Kanal, T. S. Levitt & J. F. Lemmer. North-Holland, Amsterdam, 1989, pp. 175–82.
25. Geiger, D., An entropy-based learning algorithm of Bayesian conditional trees, In *Uncertainty in Artificial Intelligence*, ed. Dubois, Wellman, B. D. D'Ambrosio & P. Smerts. 1992, pp. 92–7.
26. Cooper, G. F. & Herskovits, E., A Bayesian method for constructing Bayesian belief networks for databases. In *7th Conference on Uncertainty in Artificial Intelligence*, UCLA, ed. B. D. D'Ambrosio, P. Smerts & P. P. Bonissone, Morgan Kaufmann, 1991, pp. 86–94.
27. Cooper, G. F. & Herskovits, E., A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 1992, **9**, 309–47.
28. Molina, R., de Campos, Luis M. & Mateos, J., Using Bayesian algorithms for learning causal networks in classification problems. In *Uncertainty in Intelligent Systems*, ed. B. BouchonMeunier *et al.* North-Holland 1993, pp. 49–58.
29. Chow, C. K. & Liu, C. N., Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 1986, **14** (3), 462–7.
30. Kwoh Chee Keong, Probabilistic reasoning from correlated objective data. Ph D thesis, Imperial College, London, 1995.
31. Sucar, E., Gillies, D. F. & Gillies, D. A., Objective probabilities in expert systems. *Artificial Intelligence*, 1993, **61**, 187–208.
32. Kwoh, Chee Keong & Gillies, D. F., Using hidden nodes in Bayesian networks. *Artificial Intelligence Journal* (in press).
33. Kwoh, Chee Keong, Ismaili, I. A. & Gillies, D. F., On the use of orthogonal transformations in probabilistic inference systems. *SIAM Journal on Computing* (submitted).
34. Madigan, D. & York, J., Bayesian graphical models for discrete data. *International Statistical Reivew*, 1993 (in press).