

19

Bayesian Inference in Vision

Computer Vision © Department of Computing, Imperial College
GZ Yang and DF Gillies • <http://www.doc.ic.ac.uk/~gzy>

A frequent problem in vision is the incorporation of prior knowledge into the vision process. Usually, there is uncertainty in the features that we extract from the image, and this poses a problem in how to associate the prior knowledge with what we extract. When we looked at active contours, we saw that the segmentation could be regarded as an optimisation on the position of the boundary, and that by setting the heuristic constants in the energy equation, we could make the snake behave differently, either following the contour in detail or smoothing it out and filling gaps. To strengthen this idea we now note that really we have two different types of information that we are using in the decision process. These we will call:

1. **Prior Information**
2. **Likelihood Information.**

The prior information, in the case of snakes is determined by some of the terms used in the objective function that we are minimising, for example the curvature term. In essence they encapsulate what we expect to find in the image before we do any processing whatever. The likelihood information refers to the things that we find in the image. Information of this kind is the gradient and gradient direction of the pixels on the contour. In the last lecture we introduced another kind of deformable template which we said could be used as a measure of how well the information fitted our concept of a bicycle. Here again we see that the template can be considered our prior knowledge. With the springs un-stretched we assert that the data conforms exactly to what we expect a bicycle to be. The more they are stretched to fit the data the less we believe that it is a bicycle that we have extracted from the image. In the same example we noted that we might put a penalty on missing information, and this again will be seen to be prior information. If a wheel is missing from our template or belief that we have extracted a bicycle is again reduced. As with the snakes we can also associate likelihood information with the objects extracted from the image. For example, in extracting a wheel using the Hough transform, we could use summed gradient magnitudes of the voting pixels (normalised to the circumference) as a measure of how good the wheel really was. The problem is how do we now combine all this information to make an inference. One way of achieving this is to use probability theory, often generally known as Bayesian Inference.

The fundamental idea is incorporated in Bayes theorem (which was proved over 300 years ago). The proof of Bayes theorem is simple. Suppose that we want to estimate the probability of a double event, which we write as $P(D\&S)$ where, for illustrative purposes we think of D standing for "disease" and S standing for "symptom". Suppose that the symptom is a headache and the disease is an in growing toe nail, then we can safely say that the two are independent and that:

$$P(D\&S) = P(D)P(S)$$

That is the probability of any one member of the populace having both a headache and an in growing toe nail is simply the product of the individual probabilities. We can measure these probabilities experimentally. No suppose that the symptom is again a headache, but this time the disease is a "hangover". We know (mostly from personal experience) that these two events are strongly related. Thus the knowledge that a person has a hangover will strongly suggest to us that he or she will have a headache as well. Thus we write:

$$P(D\&S)=P(D)P(S|D)$$

where P(S|D) the probability of the symptom given the disease. We can again measure this by testing a population of hangover sufferers to determine how many of them have headaches. (You could carry out this experiment one Sunday morning, but don't expect much co-operation.) Now, since the and operator is symmetric we can also write:

$$P(D\&S)=P(S)P(D|S)$$

and so by eliminating P(D&S) we get:

$$P(S)P(D|S)=P(D)P(S|D)$$

$$P(D|S)=\frac{P(D)P(S|D)}{P(S)} \quad \text{Bayes Theorem (1763)}$$

Bayes theorem can now be seen as an inference mechanism since, in our illustrative example, the left hand side is an inference (or diagnosis), namely the probability of a disease given a symptom, and everything on the left hand side is a measurable quantity. The values P(s) and P(D) are referred to as "prior" probabilities. The value P(S|D) is known as the conditional probability. To tie this up with our previous ideas, we now note that, since we are making inferences about a disease we need no longer worry about the prior probability P(S). We can replace it with a normalising constant, and thus write:

$$P(D|S)=\alpha P(D)P(S|D)$$

The value of α can always be found since probabilities always sum to 1. In the simple case of the disease being present or not we can write $P(DIS) + P(\text{not DIS}) = 1$. Now we can interpret P(D) as our prior knowledge about the disease, and P(DIS) as the likelihood information, given that some measurement (the symptoms) have been made. Bayes theorem tells us that we simply multiply these together. Thus to apply Bayesian inference to vision we need to characterise both our prior knowledge and our likelihood information as probabilities. These we then combine by simply multiplying them together.

We are now faced with the problem of determining our probabilities, and there are two schools of thought as to how this may be done: the subjective and the objective. The objective approach (sometimes called the empirical Bayes method) relies on deriving our probabilities from experiment. For example, if our prior measurement was the summed spring tensions in fitting our template to the data, then we could conduct an experiment in which a large number of images containing bicycles or similar objects are processed and the spring deformations are measured. We could only use cases where the likelihood information was near to perfect for this purpose, for example by segmenting our images by hand. We would then be able to build a distribution of probability over extension, as illustrated in diagram 19.1. In practice this would be done discretely by dividing the x axis into a number of bands, and finding the proportion of the images, in each band, in which a bicycle was correctly identified. We can use objective probabilities if:

1. We can make measurements of the input data
 2. We can (possibly later) identify the correct classification for the data
- Notice that this is a form of learning by experience, updating after the decision is made (a posteriori), and is sometimes referred to as training.**

The diagram of 19.1 illustrates an important point about objective probabilities, namely that the distribution may not conform to what we originally expected. The fact that the probability

does not monotonically decrease with deformation may indicate that there are classes of bicycles that do not conform with the spring model. On the other hand it may be an indication that our experimental method was flawed in some way, for example by not having sufficiently diverse examples of bicycles. Thus the subjective approach is to express our knowledge independently of the data, for example by assuming some distribution. A half way house could be to fit a known mathematical distribution (for example the Gaussian) to the measured data.

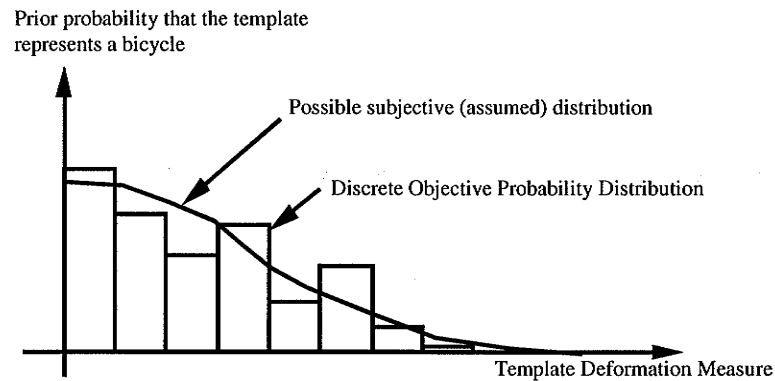


Diagram 19.1: Prior Probability Distribution

The likelihood information can be estimated similarly to the prior information. In this case, having fixed the prior information we now do a further series of experiments, this time with less than perfect images segmented by our computer vision algorithms, and determine the distribution of the probability of successful identification over the measure that we use for likelihood. For example the average gradient magnitude of the pixels that fit our template. Again we could adopt purely subjective methods to turn the measurements into probabilities.

Looking back at the course we can see that the division of information into prior and likelihood is very clear, and in the light of the above discussion it should be clear that they are both forms of uncertain knowledge.

Likelihood Information	Prior Information
Gradient Magnitude	Mean/Variance segmentation models
Gradient Directions	Co-occurrence matrices
Probabilities from relaxation labelling	Curvature terms in active contours
Magnitudes in the Hough array	Moments

Although probability now gives us a good theoretical approach to manage inference, it does not tell us how to combine all this information together. For example if we extend our likelihood in the above example to include the magnitudes of the peaks in the Hough transform for extracting the wheels, then how do we combine this with the gradient information? We could possibly build an objective probability function of two variables, but that would require a huge amount of data, and for more than two likelihood measures the objective approach becomes even less feasible. Conversely, if we choose the subjective approach, we are back to techniques such as combining the measures into one using heuristic constants. In other words the use of probability has added nothing to our existing inference mechanisms. Both approaches have been tried but the most successful method of fusing information of this kind is the Bayesian Network, which we will study in the next lecture.

Bayesian Networks in Vision

A Bayesian Network expresses a causal structure through which different information (prior and likelihood) can be combined to make an inference. For example, our prior knowledge for bicycle recognition might be expressed by rules of the kind: "a bicycle is identified by a frame two large wheels and a small wheel between them". The causal tree is shown in Diagram 19.2.

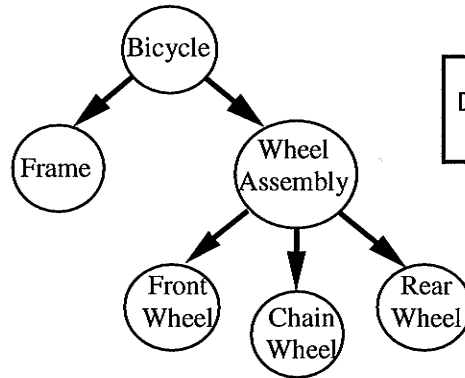
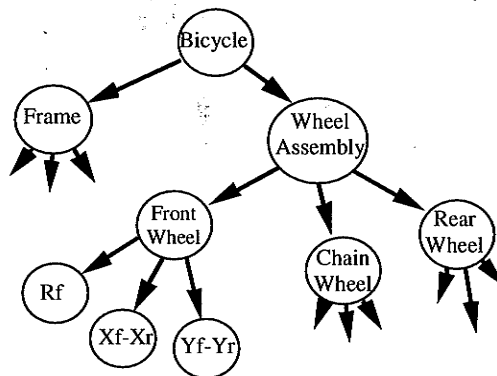


Diagram 19.2
Decision (or causal) tree for recognising a bicycle

We can expand the tree at each leaf node with measurable properties of the image as shown in Diagram 19.3. The problem is to estimate the probabilities of the non leaf nodes: ie given a certain radius and centre difference of two circles what is the probability of having found a front wheel &c:



Notation:

Rf : Radius of front wheel
(Xf, Yf): Centre of front wheel

Diagram 19.3
Bayesian Network with inputs from the image

In the usual case where we make an inference from many pieces of information we write:

$$P(D|S_1 \& S_2 \& \dots S_n) = \frac{P(S_1 \& S_2 \& \dots S_n)P(D)}{P(S_1 \& S_2 \& \dots S_n)}$$

if S1 S2 and Sn are conditionally independent

$$P(S_1 \& S_2 \& \dots S_n) = P(S_1)P(S_2) \dots P(S_n)$$

and

$$P(S_1 \& S_2 \& \dots S_n | D) = P(S_1 | D) P(S_2 | D) \dots P(S_n | D)$$

So

$$P(D | S_1 \& S_2 \& \dots S_n) = \frac{P(S_1 | D) P(S_2 | D) \dots P(S_n | D) P(D)}{P(S_1) P(S_2) \dots P(S_n)}$$

Consider in our example just the recognition of a front wheel:

$$P(FW | Radius \& X_{diff} \& Y_{diff}) = \frac{P(Radius | FW) P(X_{diff} | FW) P(Y_{diff} | FW) P(FW)}{P(Radius) P(X_{diff}) P(Y_{diff})}$$

$P(FW)$, $P(Radius)$, $P(X_{diff})$ and $P(Y_{diff})$ are all measurable from the set of images that we process, and, as before, can be considered constants for the population (given a set of measurements). So in effect we have:

$$P(FW | Radius \& X_{diff} \& Y_{diff}) = \alpha P(Radius | FW) P(X_{diff} | FW) P(Y_{diff} | FW) P(FW)$$

As before we can build a distribution, either subjectively or objectively, that shows the probability of finding that radius in a bicycle front wheel. The distribution might be something like diagram 19.4. If estimates are also taken of $P(Radius | FW)$, $P(X_{diff} | LDR)$ and $P(Y_{diff} | LDR)$, a probability can be found relating to whether a front wheel has been identified. This can be further propagated up the decision tree.

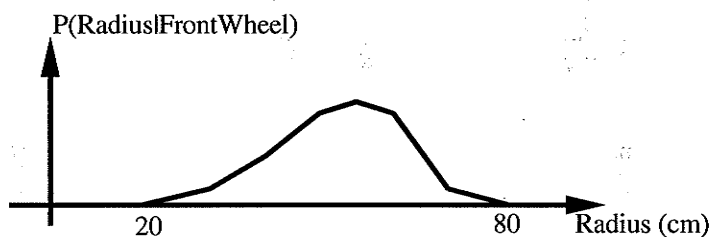


Diagram 19.4: Subjective distribution for front wheel radius.

Notice that, as long as our nodes do not have multiple parents, on each arc we have a matrix of conditional probabilities, as shown in diagram 19.5. "Front wheel" has only two states ("Front wheel found" or "Front wheel not found"). If we quantise our radius measurements into say ten levels we have a ten by two matrix on the arc. Nodes may have more than two states, for example "wheel assembly" could have say three states ("wheel assembly found", "wheel assembly not found" and "possible wheel assembly found").

Instantiation

In a Bayesian Network some of the nodes (typically the leaf nodes such as radius) represent measured values. When a measurement is made they are "instantiated". We can either instantiate them directly, equivalent to supplying for example a measurement for a radius, or we can admit an uncertainty in our likelihood information by supplying a probability distribution

over a set of radii.

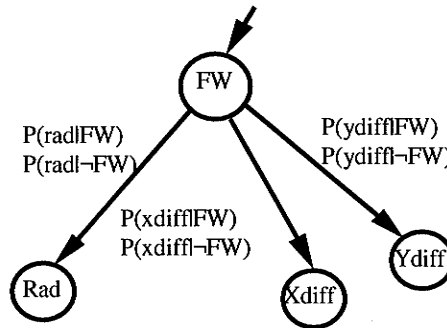


Diagram 19.5: Conditional probabilities on Bayesian tree

Probability Propagation

For a given instantiation of the measured variables (evidence) we need to compute the probability of each possible value of one or more of the other nodes. The probabilities then allow us to choose between competing hypotheses (eg bicycle or motorbike).

Evidence

Consider first the usual case of a tree where the leaf nodes are instantiated and the root node represents the hypothesis we are testing. Using Bayes theorem with assumed conditional independence we saw that there was a simple multiplicative rule for propagating the evidence:

$$P(FW | Radius \ \& X_{diff} \ \& Y_{diff}) = \alpha P(Radius|FW)P(X_{diff}|FW)P(Y_{diff}|FW)P(FW)$$

we call the term:

$$P(Radius|FW)P(X_{diff}|FW)P(Y_{diff}|FW)P(FW)$$

the evidence for a node taking a particular value in general:

$$\lambda(B_i) = \prod_{(sons)} P(D_j|B_i)$$

The evidence can be considered an un-normalised probability value for that node.

Conditioning

In general, if the sons are not leaf nodes then we have to condition the node, that is sum the total evidence at that node:

$$\lambda(B_i) = \prod_{(sons)} \sum_j \lambda(D_j)P(D_j|B_i)$$

and we can consider instantiated nodes as those that have $\lambda(N_i)=1$ for one value and zero for all others.

The elimination of prior probabilities

Consider the tree shown in diagram 19.6.

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

$$P(B|D\&E\&F) = \frac{P(B)P(D|B)P(E|B)P(F|B)}{P(D)P(E)P(F)}$$

we can eliminate P(B) to get

$$P(A|B\&C) = \frac{P(A)P(B|A)P(C|A)P(D|B)P(E|B)P(F|B)}{P(C)P(D)P(E)P(F)P(B|D\&E\&F)}$$

since the bottom is independent of A we write $\alpha = 1/(P(C) P(D) P(E) P(F) P(B|D\&E\&F))$

$$P(A|B\&C) = \alpha P(A)P(B|A)P(C|A)P(D|B)P(E|B)P(F|B)$$

so all we need to know is the prior probabilities of the values of A, P(A), and α can be eliminated since P(A|B&C) must sum to 1 over the possible values of A.

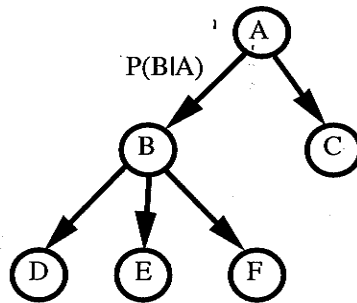


Diagram 20.5 Eliminating Prior Probabilities

Generalisation of Bayesian Networks

In general any node in a tree can be instantiated. If a non leaf node is instantiated a mechanism must be introduced to propagate evidence down the tree as well as up it. Instantiating a node with children is somewhat similar to determining a prior probability for the children, and this is expressed as a p value (unnormalised prior probability), similar to the evidence that was used to express the unnormalised likelihood being communicated by the children to the parent. Note also that it is not even necessary to restrict Bayesian Networks to be trees. The simplest extension is to allow multiple parents, in which exact calculations of the posterior probabilities can always be made. If loops are introduced in the causal structure (like the balance of nature: foxes rabbits and long grass) different techniques are required to avoid infinite recursion in propagating probabilities. (See Pearl, J. [1988], Probabilistic Reasoning in Intelligent Systems, Morgan-Kaufmann, San Mateo, Calif., USA.)

Conditional Independence

An important assumption in propagation of probabilities is that the children nodes are conditionally independent. It may be necessary to check this assumption. Considering the estimate of FrontWheel we have used: $P(\text{Radius}|\text{FW}) P(\text{XdifffFW}) P(\text{YdifffFW})$, but maybe in the population, these are not independent, i.e. we could write say: $\text{Xdifff} = f(\text{Radius})$.

If such a relationship exists then Bayes theorem has not been applied correctly. The relationship can be found by correlating the input data, and if the correlation coefficient is high, we know that the structure is incorrect, and we must correct it for example by:

- **Incorporating the known functional relationship**
- **Removal of correlated variables**
- **Altering the network structure**