Shape Matching with Belief Propagation: Using Dynamic Quantization to Accomodate Occlusion and Clutter

James Coughlan and Huiying Shen
Smith-Kettlewell Institute
San Francisco, CA 94115
coughlan@ski.org, *hshen@ski.org*

Abstract

Graphical models provide an attractive framework for shape matching because they are well-suited to formulating Bayesian models of deformable templates. In addition, the advent of powerful inference techniques such as belief propagation (BP) has recently made these models tractable. However, the enormous size of the state spaces involved in these applications (about the size of the pixel lattice) has restricted their use to models drawing on sparse feature maps (e.g. edges), which are typically unable to cope with missing or occluded features since the locations of missing features are not represented in the state space.

We propose a novel method for allowing BP to handle partial occlusions in the presence of clutter, which we call dynamic quantization (DQ). DQ is an extension of standard pruning techniques which allows BP to adaptively add as well as subtract states as needed. Since DQ allows BP to focus on more probable regions of the image, the state space can be adaptively enlarged to include locations where features are occluded, without the computational burden of representing all possible pixel locations. The combination of BP and DQ yields deformable templates that are both fast and robust to significant occlusions, without requiring any user initialization. Experimental results are shown on deformable templates of planar shapes.

1. Introduction

A variety of problems relating to the detection and matching of shapes have been formulated in terms of deformable template models [23], which explicitly model the shape and appearance of flexible objects. A common property of many deformable template models is the specification of a global shape in terms of parts, with geometric relationships among the parts enforcing constraints on the global shape. To find the best match of such a deformable template model with an image, candidates for the model parts are extracted from the image by a feature selection process (or in some cases drawn from a quantized space of values aligned to the image pixel grid), and the candidates that best satisfy the desired geometric relationships among the parts are chosen as the best-fitting solution.

An early example of this paradigm is the “pictorial structures” framework of [9], which describes a face template composed of elementary parts (eyes, nose, etc.) with spring-like spatial interactions between them, and a technique similar to dynamic programming (DP) used to find a nearly globally optimal solution. Similar matching procedures based on DP include landmark matching [2] and work addressing the related problem of finding generic smooth image contours [3]. Recently, [8] has introduced a deformable template which exploits the decomposition of a large class of shapes into triangulated polygons, giving rise to tree-shaped graphical models (without loops) that can be solved exactly with DP. One of the major advantages of deformable templates which can be matched by DP is that the globally optimal match is found without the need for an initial approximate guess of where the target is located in the image. By contrast, local gradient-based optimization procedures, which need a suitable initial condition to converge to the correct solution, are used to match many other types of deformable templates (e.g. [4]).

Dynamic programming is a useful optimization tool for these types of deformable templates, but it is subsumed by a more powerful technique, belief propagation (BP) [16], which is appropriate when a deformable template is cast as a graphical model (Markov random field). BP is exact on graphical models without loops (i.e. chains or trees) – the same models for which DP is guaranteed to work – but has also been shown empirically to provide good approximate solutions on a variety of loopy graphs [14]. Moreover, the graphical model formulation is attractive because a graphical model specifies a statistical model of the shape and appearance variability, rather than simply specifying a global fitness function to evaluate the quality of a match between a template and an image. The shape (prior) and appearance (likelihood) models are specified separately, which makes it easier to understand the behavior of the deformable template, and both the prior and likelihood models can be learned from training data.

0-7695-2158-4/04 $20.00 (C) 2004 IEEE

Graphical models used to represent shape deformations and matching processes have recently been applied to deformable templates in tracking applications [17, 10] and to dense stereo matching [21], all of which use BP to perform inference on the models. Work by [20, 12] to extend the use of BP to the continuous variable domain demonstrates simple graphical models for recovering facial appearance under partial occlusion and for detecting articulated objects in clutter. More recently, an optimization technique related to BP is used in the graphical model-based shape matching work by [18]. All of these algorithms require the entire target to be visible, except for [20, 12, 19], which are too computationally demanding for real-time use. Finally, current research on 3-D registration by [1] using BP to match 3-D non-rigid surfaces defined by range data can handle significant amounts of missing data (range occlusions), but requires that there be no clutter.

In previous work [6] we devised the first deformable template we are aware of based on BP. This template used a graphical model to represent the contour of a deformable planar shape such as a hand or a letter. The algorithm used a pre-pruning step to eliminate unlikely pixel locations from subsequent processing by BP. Although this pre-pruning stage greatly sped up the algorithm, it did so at the expense of requiring the entire target to be visible. Earlier work with a DP-based deformable template [5] removed this restriction by eliminating the pre-pruning step and allowing the possibility of features to be visible or occluded at *every* pixel. This technique was robust to occlusions but computationally very expensive.

As an alternative, we propose a novel technique for speeding up BP, called dynamic quantization (DQ), which is a compromise between pre-pruning unlikely candidates (and removing them from subsequent consideration) and representing all possible pixel locations. The main idea of DQ is to combine standard pruning techniques, which remove states that are sufficiently unlikely, with a technique for augmenting state spaces by adding sufficiently promising states. In this way, the number of allowed states increases or decreases over time as needed, allowing BP to focus on the more probable regions of state space (corresponding to more important regions of the image). As a result, DQ allows BP to perform template matching even with partial occlusions in the presence of considerable clutter, while remaining computationally tractable. We illustrate the DQ modification of BP with experimental results on planar deformable templates, demonstrating robustness to partial occlusions.

2. Generative Model

In this section we describe in detail the graphical models we use to define our deformable templates, including the shape prior, appearance likelihood model and a discussion of the resulting posterior.

2.1. Graphical Model Shape Prior

This subsection summarizes the basic graphical model used in our previous work [6], which models the shape of the boundary of a planar object as a sequence of points with associated orientations. The graphical model enforces spring-like geometric relationships between neighboring points to ensure that the overall shape is similar to a reference shape used to define the template, assigning high probabilities to configurations of points that are most similar to the reference. We will begin this section by specializing to the case of the letter A; generalization to models of other shapes follows naturally, which we discuss in the next subsection and demonstrate in our experimental results.

The variability of the template shape is modelled by the shape prior, which assigns a probability to each possible deformation of the shape. The shape is represented by a set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ in the plane which trace the contours of the shape, and by an associated chain $\theta_1, \theta_2, \cdots, \theta_N$ of tangent directions which describe the orientation of the tangent to the boundary at each point. In the case of the model of the letter A, $N = 20$ (see Figure 1). (Our notation is different from that used in [6], in which θ_i represented the orientation of the normal to the boundary rather than the tangent.) Each point $\mathbf{x}_i$ has two components x_i and y_i. For brevity we define variable $\mathbf{q}_i = (\mathbf{x}_i, \theta_i)$, which we also refer to as "node" i. The *configuration* $\mathbf{Q}$, defined as $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N)$, completely defines the shape.

The shape prior is defined relative to a *reference shape* so as to assign high probability to configurations $\mathbf{Q}$ which are similar to the reference configuration $\tilde{\mathbf{Q}} = (\tilde{\mathbf{q}}_1, \tilde{\mathbf{q}}_2, \cdots, \tilde{\mathbf{q}}_N)$ and low probability to configurations that are not. This is achieved using a graphical model which penalizes the amount of deviation in shape between $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ in a way that is invariant to global rotation and translation. (The scale of the shape prior is fixed and we assume knowledge of this scale when we execute our algorithm.)

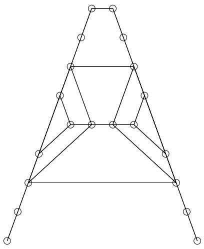

Figure 1: Letter A template. Nodes drawn as circles, with lines indicating graph connectivity. All nodes are shown in their reference shape positions $(\tilde{x}_i, \tilde{y}_i)$.

Deviations in shape are measured by the geometric relationships of connected pairs of points $\mathbf{q}_i$ and $\mathbf{q}_j$ on the template (see Figure 1 for the connectivity), and are expressed in terms of *pairwise interaction potentials* $\psi_{ij}(\mathbf{q}_i, \mathbf{q}_j)$. High potential values occur for highly probable shape configurations, for which the geometric relationships of pairs of points are most faithful to the reference shape $\tilde{\mathbf{Q}}$. We first outline the main properties of the potentials before defining them precisely.

Two kinds of geometric relationships, both of which are invariant to global translations and translations, are used to define the potentials. First, we expect that the relative orientations of tangent directions at nearby points on the boundary should be roughly invariant to local deformations. In other words, we expect $\theta_j - \theta_i \approx \tilde{\theta}_j - \tilde{\theta}_i$ for connected nodes i and j. Second, we note that the location of point $\mathbf{x}_j$ can be expressed relative to the location $\mathbf{x}_i$ and tangent θ_i; this relationship should also be roughly invariant to local deformations. Just as θ_j must be roughly consistent with θ_i, the location $\mathbf{x}_j$ must also be roughly consistent with the location and tangent direction of $\mathbf{q}_i$. (The reciprocal relationship between the location $\mathbf{x}_i$ and the location and tangent direction of $\mathbf{q}_j$ also holds.)

More precisely, we can express these geometric relationships in terms of *interaction energies* $U_{ij}(\mathbf{q}_i, \mathbf{q}_j)$. Low interaction energies occur for highly probable shape configurations, for which the geometric relationships of pairs of points tend to be faithful to the reference shape $\tilde{\mathbf{Q}}$, and high interaction energies are obtained for improbable configurations; the precise connection to probabilities is formulated in Equation (5).

The soft constraint that $\theta_j - \theta_i \approx \tilde{\theta}_j - \tilde{\theta}_i$ is expressed in the following interaction energy $U_{ij}(\mathbf{q}_i, \mathbf{q}_j)$:

$$U_{ij}^C(\mathbf{q}_i, \mathbf{q}_j) = \sin^2\left(\frac{\theta_j - \theta_i - C_{ij}}{2}\right) \qquad (1)$$

where $C_{ij} = \tilde{\theta}_j - \tilde{\theta}_i$. This energy attains a minimum when $\theta_j - \theta_i = C_{ij}$ (and a maximum when $\theta_j - \theta_i = C_{ij} + \pi$).

Next we define the relationship between the location of point $\mathbf{x}_j$ relative to $\mathbf{q}_i$. $\tilde{\mathbf{x}}_i$ and $\tilde{\theta}_i$ define a local coordinate system, and the coordinates of $\tilde{\mathbf{x}}_j$ in that coordinate system are invariant to global translation and rotation. If we define the unit tangent vectors $\mathbf{t}_i = (\cos\theta_i, \sin\theta_i)$ and $\tilde{\mathbf{t}}_i = (\cos\tilde{\theta}_i, \sin\tilde{\theta}_i)$ and vectors perpendicular to them $\mathbf{t}_i^\perp = (-\sin\theta_i, \cos\theta_i)$ and $\tilde{\mathbf{t}}_i^\perp = (-\sin\theta_i, \cos\theta_i)$, then the dot product of $\mathbf{x}_j - \mathbf{x}_i$ with $\mathbf{t}_i$ and $\mathbf{t}_i^\perp$ should have values similar to the corresponding values for the reference shape: $(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{t}_i \approx (\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i) \cdot \tilde{\mathbf{t}}_i$ and $(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{t}_i^\perp \approx (\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i) \cdot \tilde{\mathbf{t}}_i^\perp$. Now we can define the remaining two terms in $U_{ij}(\mathbf{q}_i, \mathbf{q}_j)$, the energies

$$U_{ij}^A(\mathbf{q}_i, \mathbf{q}_j) = [(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{t}_i - A_{ij}]^2 \qquad (2)$$

and

$$U_{ij}^B(\mathbf{q}_i, \mathbf{q}_j) = [(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{t}_i^\perp - B_{ij}]^2 \qquad (3)$$

where $A_{ij} = (\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i) \cdot \tilde{\mathbf{t}}_i$ and $B_{ij} = (\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i) \cdot \tilde{\mathbf{t}}_i^\perp$. The full interaction energy is then given as (omitting arguments $(\mathbf{q}_i, \mathbf{q}_j)$ for brevity):

$$U_{ij} = \frac{1}{2}\{K_{ij}^A U_{ij}^A + K_{ij}^B U_{ij}^B + K_{ij}^C U_{ij}^C\} \qquad (4)$$

where the non-negative coefficients K_{ij}^A, K_{ij}^B and K_{ij}^C define the strengths of the interactions and are set to 0 for those pairs i and j with no direct interactions (the majority of pairs). Higher values of K_{ij}^A, K_{ij}^B and K_{ij}^C produce a stiffer (less deformable) template.

Noting that in general $U_{ij}(\mathbf{q}_i, \mathbf{q}_j) \neq U_{ji}(\mathbf{q}_j, \mathbf{q}_i)$, we symmetrize the interaction energy as follows: $U_{ij}^{sym}(\mathbf{q}_i, \mathbf{q}_j) = U_{ij}(\mathbf{q}_i, \mathbf{q}_j) + U_{ji}(\mathbf{q}_j, \mathbf{q}_i)$. We use the symmetrized energy to define the shape prior:

$$P(\mathbf{Q}) = \frac{1}{Z}\prod_{i<j}\exp(-U_{ij}^{sym}(\mathbf{q}_i, \mathbf{q}_j)) \qquad (5)$$

where Z is a normalization constant and the product is over all pairs i and j, with the restriction $i < j$ to eliminate double-counting and self-interactions. Note that the prior is a Markov random field or graphical model that has pairwise connections between all pairs i and j whose coefficients are non-zero. The graph connectivity and the values of the coefficients were chosen experimentally by stochastically sampling the prior using a Metropolis MCMC sampler, i.e. generating samples from the prior distribution to illustrate what shapes have high probability (as in [6]).

2.1.1 Constructing Shape Priors

Once we constructed and tested our letter A template, we used a simple procedure to construct other deformable template prior models. A clear, representative image of the shape was taken to define the reference shape. The points used to define the reference shape contour were chosen manually by clicking on the image with a mouse at roughly equal intervals along the contour. (The associated tangent orientations were also extracted by defining them to be perpendicular to the image gradient direction at each point, assuming an idealized version of the appearance model described in the next section.)

In our experiments it sufficed to use the same coefficient values for all the shapes we tried. The connectivity was initially chosen so most nodes had no more than two nearest neighbors (as in a Markov chain), with more nearest neighbors chosen near junctions and high-curvature points. Longer-range connections were added as needed to preserve the continuity of the shape estimates determined by BP (i.e. so that adjacent parts of the shape would not be matched to non-adjacent regions of the image). Learning techniques such as maximum likelihood estimation could be employed to determine more optimal coefficient values and graph connectivities.

2.2. Appearance Model

The appearance (i.e. imaging, or likelihood) model explains what image data may be expected given a specific shape configuration. Rather than model raw image pixel data, we extract information from the image gradient.

The first type of image data derived from the raw grayscale image $I(\mathbf{x})$ is an edge map $I_e(\mathbf{x})$, and an associated edge orientation map $\phi(\mathbf{x})$ to provide estimates of the orientation of edges throughout the image. $I_e(\mathbf{x})$ is defined as the magnitude of the image gradient: $I_e(\mathbf{x}) = |G \star \nabla I(\mathbf{x})|$ where $G(.)$ is a smoothing Gaussian. The orientation map $\phi(\mathbf{x})$ is calculated as $\arctan(g_y/g_x)$ where $(g_x, g_y) = G \star \nabla I(\mathbf{x})$.

The edge strength part of the likelihood model quantifies the tendency for edge strength values to be high *on* edge and low *off* edge. As in our previous work [6], we use the approach of Geman and Jedynak [11] and define two conditional distributions of edge strength that are empirically measured: $P_{on}(I_e(\mathbf{x})) = P(I_e(\mathbf{x})|\mathbf{x}\, ON\ edge)$ and $P_{off}(I_e(\mathbf{x})) = P(I_e(\mathbf{x})|\mathbf{x}\, OFF\ edge)$. The most important property of these two distributions is that the log likelihood ratio $\log P_{on}(I_e)/P_{off}(I_e)$ increases monotonically with edge strength, meaning that higher edge strengths correspond to greater evidence for edges.

As shown in [6], *the overall likelihood of the entire template is proportional to terms depending only on these log likelihood ratios,* rather than the individual distributions $P_{on}(I_e)$ and $P_{off}(I_e)$. Rather than learning these individual distributions, we will propose simple models of the likelihood ratios themselves. More specifically, we will approximate the likelihood ratio $P_{on}(I_e)/P_{off}(I_e)$ as a step function having a value of 1 for edge strengths above a certain threshold and a value of 0.1 below threshold. In other words, *we are binarizing the edge strength map.* However, this thresholding procedure is done just for simplicity and is not a necessary step of our algorithm. The likelihood models could be extended to continuous (or more finely quantized) values, but we found the thresholding procedure to suffice for our deformable template model.

Next we turn to the orientation map. We expect that on a true object boundary the direction of ∇I should point roughly *perpendicular* to the tangent of the boundary. Denoting the true normal tangent direction of the boundary as θ, we then expect that $\phi(\mathbf{x})$ is approximately equal to either $\theta + \pi/2$ or $\theta - \pi/2$ (corresponding to the two possible edge polarities). This relationship between θ and $\phi(\mathbf{x})$ may also be quantified as a conditional distribution $P_{ang}(\phi|\theta)$, which was assumed to be of the form $P_{ang}(\phi - \theta)$ and measured [6] to have sharp peaks at $\pm\pi/2$. If $\mathbf{x}$ is not on an edge then we may assume that the distribution of $\phi(\mathbf{x})$ is uniform in all directions: $U(\phi) = 1/2\pi$. Again, we approximate the likelihood ratio $P_{ang}(\phi - \theta)/U(\phi - \theta)$ as a step function having a value of 1 when ϕ and θ are aligned within 10° of perpendicular and a value of 0.1 otherwise.

The complete imaging model is a distribution of all the image gradient data across the entire image, conditioned on the template shape configuration $\mathbf{Q}$. We can express the likelihood model as a distribution that factors over every pixel. We define $\mathbf{d}(\mathbf{x}) = (I_e(\mathbf{x}), \phi(\mathbf{x}))$ and let $\mathbf{D}$ denote the values of $\mathbf{d}(\mathbf{x})$ across the entire image. Defining the likelihood ratio

$$R(\mathbf{q}_i) = \frac{P_{on}(I_e(\mathbf{x}_i)}{P_{off}(I_e(\mathbf{x}_i)} \frac{P_{ang}(\phi(\mathbf{x}_i) - \theta_i)}{U(\phi(\mathbf{x}_i) - \theta_i)} \quad (6)$$

we obtain (after some manipulation [6]):

$$P(\mathbf{D}|\mathbf{Q}) \propto [\prod_{i=1}^{N} R(\mathbf{q}_i)] \quad (7)$$

where the constant of proportionality is a function of $\mathbf{D}$ only and does not depend on $\mathbf{Q}$. (This dependence on $\mathbf{D}$ will not matter for estimating the template configuration, described in the next subsection.)

2.3. Posterior Distribution

The shape configuration $\mathbf{Q}$ is determined by the posterior distribution $P(\mathbf{Q}|\mathbf{D}) = P(\mathbf{Q})P(\mathbf{D}|\mathbf{Q})/P(\mathbf{D})$. Multiplying the likelihood Equation (7) by the prior yields an expression for the posterior of the following form:

$$P(\mathbf{q}_1, \cdots, \mathbf{q}_N|\mathbf{D}) = \frac{1}{Z} \prod_i \psi_i(\mathbf{q}_i) \prod_{i<j} \psi_{ij}(\mathbf{q}_i, \mathbf{q}_j) \quad (8)$$

where $\psi_i(\mathbf{q}_i)$ is the local evidence for $\mathbf{q}_i$ (from the likelihood ratios in Equation (7)) and $\psi_{ij}(\mathbf{q}_i, \mathbf{q}_j)$ is the compatibility (or pairwise potential) between $\mathbf{q}_i$ and $\mathbf{q}_j$ (from the shape prior, Equation (5)).

Given the posterior distribution on shape configurations, a natural decision rule for choosing the most representative configuration is the MAP (maximum a posterior) estimate. Instead we use the output of BP – estimates of marginals for each variable – to calculate the MPM (Marginal A Posteriori Modes), i.e. the MAP estimate applied separately to the marginal posterior of each variable $\mathbf{q}_i$. More precisely, the MPM is given by $\mathbf{q}_i^* = \arg\max_{\mathbf{q}_i} P(\mathbf{q}_i|\mathbf{D})$ for all i. If the posterior is strongly peaked about its mode, as it should be when there is sufficient evidence for one correct match in the image, we expect the MPM and MAP to be similar. (Techniques similar to those used to find multiple targets in [5] may be used in the case of multiple matches, i.e. multiple targets in one image.) Empirically we find the MPM

0-7695-2158-4/04 $20.00 (C) 2004 IEEE

to be a satisfactory estimator, as shown in our experimental results.

3. Methods: Dynamic Quantization

Now that we have defined our graphical model of a deformable template, we will use a modified version of BP to perform inference with it. Although BP could be straightforwardly applied once the variables $\mathbf{q}_i$ are quantized to a sufficiently fine lattice (e.g. (x_i, y_i) lying on the input image pixel lattice and θ_i drawn from a discrete set of equally spaced angles from 0 to 2π), a huge number of allowed states would result, making BP prohibitively slow. Instead, we propose a dynamic quantization technique for efficiently quantizing the states of the variables in the graphical model.

3.1. Motivation

BP was originally formulated for use with graphical models having discrete variables, whereas our graphical models are most naturally represented using continuous variables. Recent work [20, 12] building on ideas from particle filtering has made it possible to perform BP on graphs with continuous variables, using stochastic particles to represent messages. However, a major drawback of the particle representation is that the multiplication of these particle-based messages – a fundamental component of each BP iteration – requires a computationally intensive Gibbs sampling procedure. As mentioned above, it is also possible to finely quantize the continous variables to obviate the need for the particle representation, at the expense of performing BP on very large (discrete) state spaces.

We chose instead to selectively quantize the variables in our graphical models, based on our intuition that only a fairly small number of "hot spots" in the variable state spaces are important. Therefore, it should suffice to quantize the space only in the neighborhood of these hot spots.

Our new dynamic quantization technique builds on our previous work [6], which used two forms of state pruning. The first form of pruning, which we called pre-pruning or adaptive quantization, initializes the allowed states for each variable $\mathbf{q}_i = (\mathbf{x}_i, \theta_i)$ to encompass those values most consistent with image evidence. More specifically, only those locations $\mathbf{x}_i$ are allowed that correspond to pixels with edge strengths above a certain threshold. At these locations only two possible orientation values of θ_i are allowed, $\phi \pm \pi/2$, where ϕ denotes the image gradient direction. (These two values correspond to assuming that the image gradient direction is exactly equal to the normal orientation of the edge boundary, and allowing for two possible polarities of the edge.) We use the same procedure to initialize the state spaces in BP.

The second form of pruning, called belief pruning, consists of monitoring the beliefs of each variable at each iteration of BP and discarding any states whose beliefs dropped below a certain threshold. (This is very similar to the "beam search" technique used to prune states in hidden Markov models (HMM's) in speech recognition [13].) While we have also retained this second form of pruning in our current work, we have added a modification that allows for new states to be created as well as old states to be destroyed.

The new ingredient that DQ adds to the two existing pruning techniques is a procedure for deciding when there is a possible deficiency of important states in a variable's state space, and a method for determining which states to add to correct such deficiencies. The intuition for this new procedure can be illustrated by considering the case of a simple graphical model deformable template representing a generic smooth curve. (For simplicity no orientation variables are used in this model, only location variables). If the variable state spaces are initialized very conservatively, so that no pixels with true edges are omitted, then BP will detect the correct target curve. The cost of such a conservative initialization will be to slow down BP with a lot of superfluous states corresponding to false positives and background clutter in the edge detection.

If the variable state spaces are initialized to include only pixels with edge strengths above a moderate threshold, then a small fraction of the true edge pixels will be omitted from the state spaces, and BP will find an incorrect solution because of their absence. However, a simple criterion will allow BP to consider previously disallowed states: if an edge pixel candidate lacks any suitable continuations, then all pixel locations that would make reasonable continuations should be added to the appropriate state space for consideration.

In the next subsection we formalize this procedure for "resurrecting" previously disallowed states.

3.2. Definition

We can formalize the DQ procedure by considering the general form of the BP message update equation for the message from node i to node j:

$$m_{ij}(\mathbf{q}_j) \mapsto \frac{1}{Z_{ij}} \sum_{\mathbf{q}_i} \psi_{ij}(\mathbf{q}_i, \mathbf{q}_j) \psi_i(\mathbf{q}_i) \prod_{k \in N(i) \backslash j} m_{ki}(\mathbf{q}_i) \tag{9}$$

where Z_{ij} is a normalization factor and the neighborhood $N(i)$ denotes the set of nodes directly coupled to i (excluding i itself).

We assume that, for each variable $\mathbf{q}_i$, every possible state space S_i contains states belonging to some lattice L_i (e.g. the cross product of the pixel lattice and a set of regularly spaced orientations from 0 to 2π), i.e. $S_i \subset L_i$. Messages can then be represented as weighted lists of states belonging to these lattices. A message list corresponding to $m_{ij}(\mathbf{q}_i)$ can be thought of as a sparse representation of the entire message $m_{ij}(.)$; states absent from the list imply corresponding message values of zero. Thus, in the message product term $\prod_{k \in N(i) \setminus j} m_{ki}(\mathbf{q}_i)$, if there is a state with non-zero message value coming from one neighboring node that is absent from the message(s) coming from one or more neighboring nodes, the product at that state is set to zero.

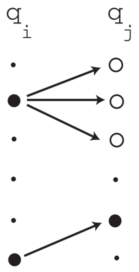

Figure 2: Illustration of DQ in the update of the message $m_{ij}(.)$ from node i to node j. Dots, filled circles and empty circles represent the states of the lattices L_i and L_j (six states in each). Filled circles represent allowed states, i.e. members of S_i and S_j (which initially correspond to edge locations). Given the value of $\mathbf{q}_i$ at the lower left, at least one allowed state of $\mathbf{q}_j$ exists that is compatible with it (indicated by lowest arrow). However, given the value of $\mathbf{q}_i$ near the top left, no states $\mathbf{q}_j$ are compatible. DQ will therefore add the states in the fan-out $F_{ij}(\mathbf{q}_i)$, represented by empty circles, to the state space S_j.

Next we define the *fan-out* of a state $\mathbf{q}_i$ to node j to be all states $\mathbf{q}_j$ (on the lattice) that are consistent according to the pairwise potential ψ_{ij}:

$$F_{ij}(\mathbf{q}_i) = \{\forall \mathbf{q}_j \in L_j | \psi_{ij}(\mathbf{q}_i, \mathbf{q}_j) > \epsilon\}$$

where ϵ is a small constant.

The "active ingredient" of DQ (see Figure (2)) is to check that, for each allowed state $\mathbf{q}_i$, there is at least one allowed state $\mathbf{q}_j$ in its fan-out (a continuation of the edge at $\mathbf{q}_i$). If not, then the fan-out states $F_{ij}(\mathbf{q}_i)$ are added to the state space S_j. After this procedure has been applied to every possible allowed state $\mathbf{q}_i$, the message update proceeds as usual (except that the state space S_j may now have been modified). The usual belief pruning procedure is also used to discard very improbable states, which offsets the growth of states from the adaptive quantization.

DQ allows BP to begin with modest-sized state spaces, corresponding to strong edges in the image. It adds new states as needed in order to "fill in" features which are either faint or entirely missing because of occlusions. Much like the DDMCMC (data-driven Monte Carlo Markov chain) technique [22] for searching posterior distributions with the help of simpler data-driven distributions, DQ draws on candidate states suggested by the image data, but is not limited by the choice of candidates. We feel that this technique preserves one of the most attractive properties of continuous BP – the ability to dynamically allocate resources (in the form of particles) only to the more important regions of state space – without requiring costly sampling procedures to perform message multiplications.

4. Results

We tested the DQ modification to BP on our graphical deformable template models applied to real images. Four template shapes were tested: the letters A and B, a car shape and a cat shape. For the first two templates the images were grayscale images of a whiteboard with handwritten characters; street scene images were used for the car template and close-up images were used for the cat template. The original images, which ranged from about 400 x 300 to 2000 x 1500, were decimated by a factor of 3 in both dimensions before the image gradient information was computed. Aside from the scale of the target shape, which was chosen manually for each image, no other form of user initialization was required.

Figure (3) shows typical detection results obtainable for the letter A template *without* DQ, demonstrating the deformable template's ability to automatically find a correct match even in the presence of substantial clutter, as well as its rotation invariance and robustness to local shape deformations.

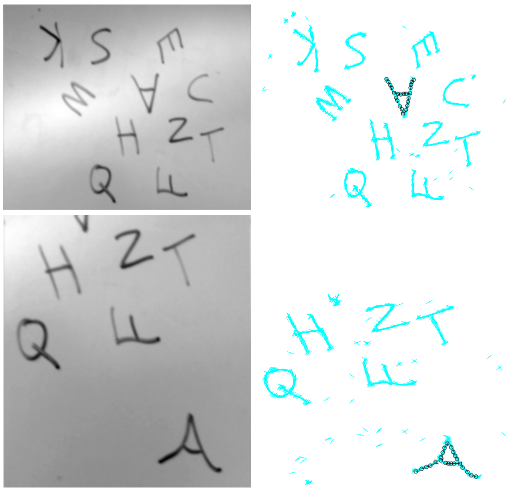

Figure 3: Typical detection results without DQ. Original images on left, solutions superimposed in black on right. Top row example demonstrates rotation invariance of deformable template, and bottom row example shows robustness to local shape deformations.

We tested the ability of DQ to allow the algorithm to recover from gaps in the target shapes. An example of this capability is shown in Figure (4). Pixel locations in the gap are not represented in the state spaces $\{S_i\}$ at the beginning of BP, but the DQ procedure is able to "fill" the gap using the prior knowledge of the shape prior. Although the locations in the gap have substantially weaker edge evidence than locations along the existing target edges, these locations are favorable as continuations of the existing edges near them. Once states corresponding to these gap locations are resurrected in the course of BP, subsequent message updates establish that they are probable locations given the context of the entire target shape.

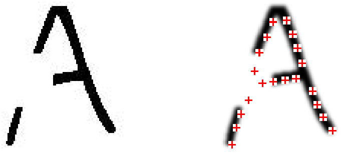

Figure 4: Robustness of template to partial occlusion. (a) Image contains an A missing its middle left corner and portions of adjoining segments. (b) DQ is able to recover the missing points (red and white crosses), enabling a successful detection.

We also demonstrate that DQ is able to fill in gaps on targets surrounded by clutter, as shown in Figure (5). When clutter is present, many spurious states corresponding to non-target locations are added by DQ in attempts to explore possible continuations of features resembling letter A parts. (E.g. parts of the letter X that resemble the sides of the letter A are likely to spawn states corresponding to a non-existent centerline between the sides of the A.) However, subsequent BP message updates rule out such states as improbable, since they correspond to matches with even less edge support than the true target.

Additional results are demonstrated in Figures (6), (7), (8), and (9). The letter B example in Figure (6b) demonstrates a successful match for a partially occluded target in clutter; note that the solution deviates from some of the visible edges of the target, reflecting the influence of the prior (reference) shape. The car example in Figure (7) is based on a simple contour model of the tops of both wheels and the chassis between them. The algorithm finds a satisfactory match in the presence of considerable clutter (green/blue pixels in Figure (7b) show the edges selected by pre-pruning) and occlusions (see Figure (8)). Finally, we demonstrate a simple head-on cat head template based on the contour of the two ears and the portion of the head between them. An edge-based contour representation is less appropriate for this object than for the others we modelled, since it is difficult to extract clean edges from the cat's furry silhouette. However, Figure (9) shows some successful matches, including one with an occlusion.

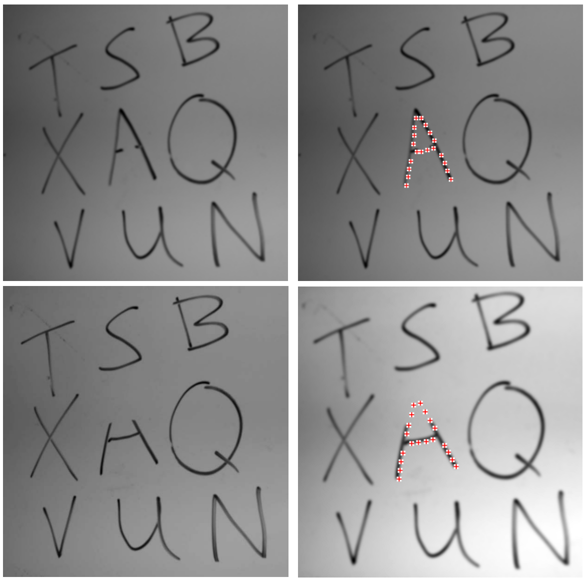

Figure 5: Robustness of template to partial occlusions in presence of clutter. Top row: (a) Image contains an A missing part of its centerline, with solution in (b). Bottom row: (a) Image contains an A missing part of its top, with solution in (b).

Execution times were on the order of tens of seconds on a standard desktop PC running a C++ implementation of the algorithm. We emphasize that there is no need for the user to initialize the template near the target shape, since BP considers all edge pixels across the entire image to be equally likely a priori, and the template is invariant to global rotation and translation. (However, the scale of the target is assumed to be known.)

5. Conclusions

DQ is a promising new enhancement of BP for deformable template matching. It extends standard pruning techniques, allowing BP to adaptively add as well as subtract states as needed. Since DQ allows BP to focus on the more probable regions of the image, state spaces can be adaptively enlarged to include locations where features are occluded, without the computational burden of representing all possible pixel locations. As a result, deformable template matching by BP is able to fill in gaps in target shapes in reasonable amounts of time. Although DQ is presented in the context of deformable template matching, we note that the technique should apply to inference on any graphical model in which the pairwise potentials are sparse (i.e. the state of one node strongly constrains the possible states of neighboring

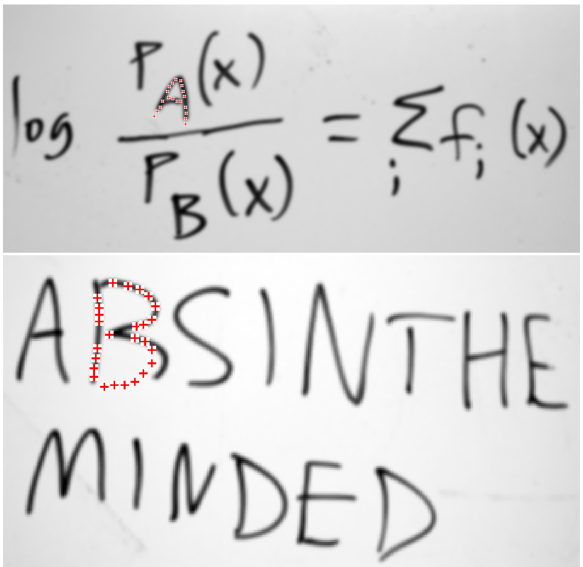

Figure 6: Additional A result, and sample result for B template.

nodes) and the states are fairly low in dimension (perhaps three or lower).

It is desirable to extend the current framework to include features which are stronger (i.e. more distinctive) than edges, such as corners. Because such strong features are comparatively rare in the background (and along the contour of the target itself), their presence allows the model to "hone in" on the target more quickly. Our preliminary experiments with deformable templates that include corner evidence show significant speed-ups compared to models that rely on image gradient information alone. In addition, more structured features combining regional and edge properties of the image (such as eyes for face models or wheels for car models), can be used to speed up the search even further, as in [7].

The use of structured features also motivates the use of hierarchical graphical models to represent objects, in which multiple levels of representation are integrated in one graphical model. For example, a letter model could represent strokes at the top level, and the edges they are composed of at the bottom level. Any combination of edge, corner and stroke detectors could naturally be combined as evidence for the model. We will investigate the use of hierarchical models and structured features in future research.

Acknowledgements

We would like to thank Drago Anguelov for his useful comments and suggestions, and Fusilli Campbell for graciously agreeing to pose for us. Both authors were supported by the National Institute on Disability and Rehabilitation Research grant number H133G030080 and the Smith-Kettlewell Eye Research Institute RERC grant.

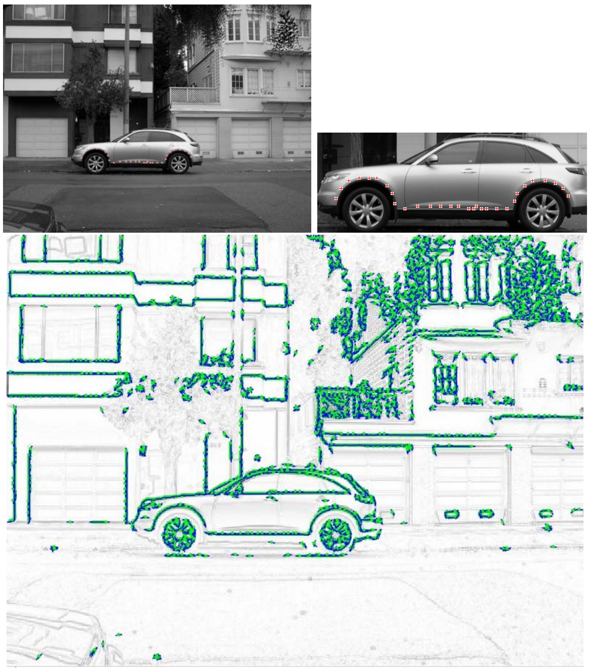

Figure 7: Car template. Top: solution shown superimposed on original image on left, and zoomed in on right. Bottom: edge pixels in green/blue.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, H. Pang and J. Davis. "The Correlated Correspondence Algorithm for Unsupervised Registration of Nonrigid Surfaces." Technical Report TR-SAIL-2004-100. Stanford University, 2004.

[2] Y. Amit and A. Kong. "Graphical Templates for Image Matching." PAMI, Vol 18, No. 3, pp. 225-236. 1996.

[3] M. Barzohar and D. B. Cooper, "Automatic Finding of Main Roads in Aerial Images by Using Geometric-Stochastic Models and Estimation," CVPR 1993. pp. 459-464.

[4] T. F. Cootes and C. J. Taylor, "Active Shape Models - 'Smart Snakes'," *British Machine Vision Conference,* pp. 266-275, Leeds, UK, September 1992.

[5] J. Coughlan, A.L. Yuille, C. English, D. Snow. "Efficient Deformable Template Detection and Localization without User Initialization." Computer Vision and Image Understanding, Vol. 78, No. 3, pp. 303-319. June 2000.

[6] J. Coughlan and S. Ferreira. "Finding Deformable Shapes using Loopy Belief Propagation." The Seventh European Conference on Computer Vision (ECCV '02). pp. 453-468. Copenhagen, Denmark. May 2002.

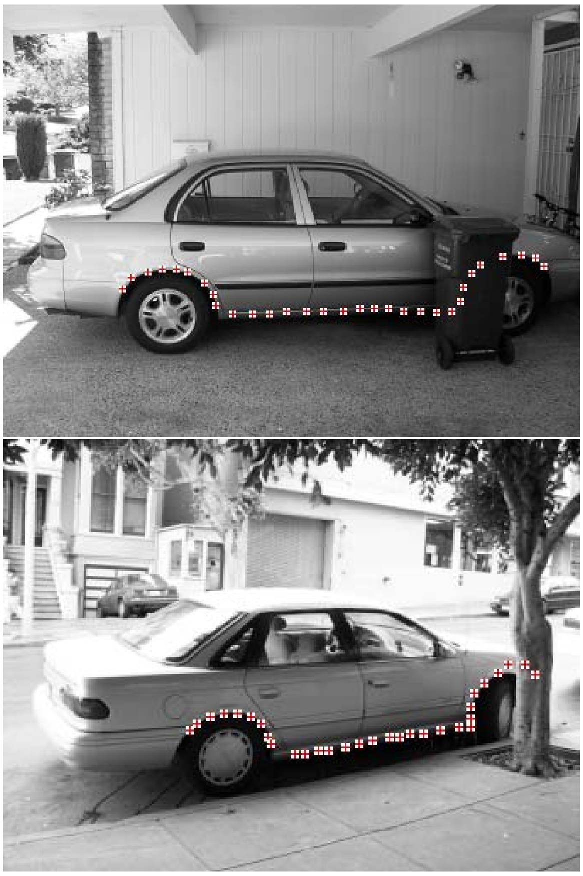

Figure 8: Car template results with occlusions.

[7] P. Felzenszwalb and D. Huttenlocher. “Efficient Matching of Pictorial Structures.” CVPR 2000. pp. 66-73.

[8] P. Felzenszwalb. “Representation and Detection of Deformable Shapes.” CVPR 2003. pp. 102-108.

[9] M.A. Fischler and R.A. Erschlager. ”The Representation and Matching of Pictorial Structures.” IEEE Trans. Computers. C-22. 1973.

[10] J. Gao and J. Shi. “Inferring Human Upper Body Motion Using Belief Propagation.” Tech. report CMU-RI-TR-03-06, Robotics Institute, Carnegie Mellon University, June, 2003.

[11] D. Geman and B. Jedynak. “An active testing model for tracking roads in satellite images”. PAMI. Vol. 18. No. 1, pp 1-14. January 1996.

[12] M. Isard. “Pampas: Real-Valued Graphical Models for Computer Vision.” In CVPR, 2003. pp. 613-620.

[13] B. Lowerre and R. Reddy, “The Harpy Speech Understanding System.” Trends in Speech Recognition. Ed. W. Lea. Prentice Hall. 1980.

[14] K.P. Murphy,Y. Weiss and M.I. Jordan. ”Loopy belief propagation for approximate inference: an empirical study”. In Proceedings of Uncertainty in AI. 1999.

[15] K. Murphy, A. Torralba, W. Freeman. ”Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes.” NIPS’03.

[16] J. Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufman. 1988.

[17] D. Ramanan and D. Forsyth. ”Finding and Tracking People from the Bottom Up.” CVPR 2003. pp. 467-474.

[18] A. Rangarajan, J. Coughlan and A. L. Yuille. A Bayesian network framework for relational shape matching. ICCV 2003. Nice, France. pp. 671-678. October 2003.

[19] L. Sigal, M. Isard, B. H. Sigelman, M. J. Black. “Attractive People: Assembling Loose-Limbed Models using Non-parametric Belief Propagation.” Advances in Neural Information Processing Systems 16, NIPS 2003.

[20] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky. Nonparametric belief propagation. In CVPR 2003. pp. 605-612.

[21] J. Sun, H. Shum, and N. Zheng. Stereo matching using belief propagation. ECCV 2002. pp. 510-524, 2002.

[22] Z.W. Tu and S.C. Zhu. ”Image segmentation by Data-driven Markov chain Monte Carlo.” PAMI, May 2002.

[23] A. L. Yuille, “Deformable Templates for Face Recognition”. *Journal of Cognitive Neuroscience.* Vol 3, Number 1. 1991.

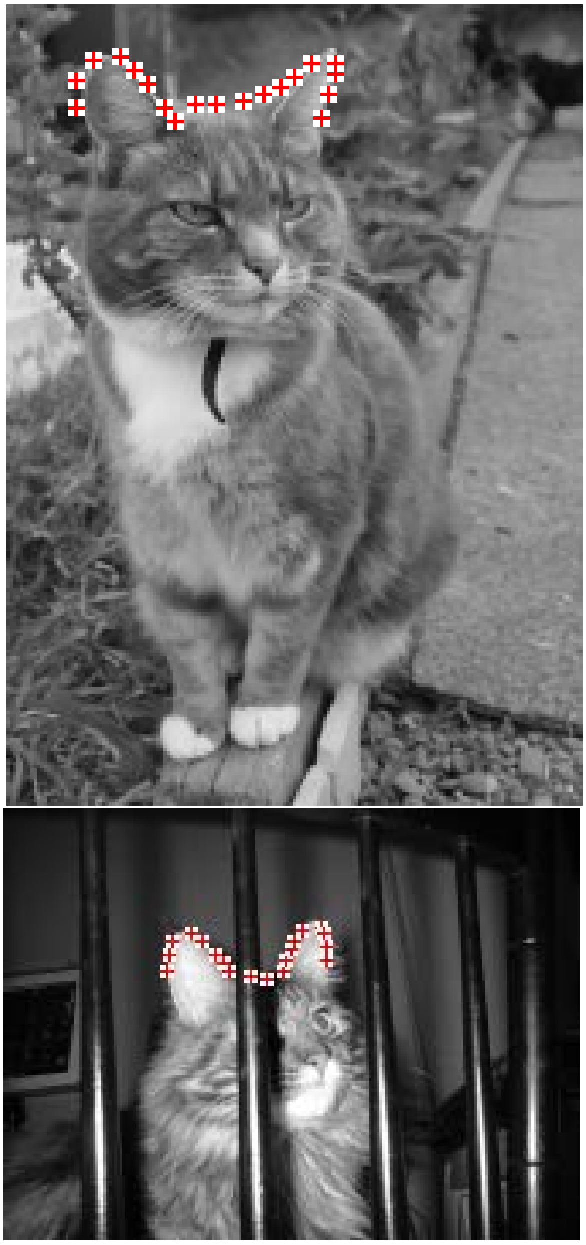

Figure 9: Cat results. Note occlusion on bottom panel.