Preserving Modes and Messages via Diverse Particle Selection

Jason Pacheco[*] PACHECOJ@CS.BROWN.EDU
Department of Computer Science, Brown University, Providence, RI 02912, USA

Silvia Zuffi[*] SILVIA.ZUFFI@TUE.MPG.DE
Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany; ITC-CNR, 20133 Milan, Italy

Michael J. Black BLACK@TUE.MPG.DE
Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

Erik B. Sudderth SUDDERTH@CS.BROWN.EDU
Department of Computer Science, Brown University, Providence, RI 02912, USA

Abstract

In applications of graphical models arising in domains such as computer vision and signal processing, we often seek the most likely configurations of high-dimensional, continuous variables. We develop a particle-based max-product algorithm which maintains a diverse set of posterior mode hypotheses, and is robust to initialization. At each iteration, the set of hypotheses at each node is augmented via stochastic proposals, and then reduced via an efficient selection algorithm. The integer program underlying our optimization-based particle selection minimizes errors in subsequent max-product message updates. This objective automatically encourages diversity in the maintained hypotheses, without requiring tuning of application-specific distances among hypotheses. By avoiding the stochastic resampling steps underlying particle sum-product algorithms, we also avoid common degeneracies where particles collapse onto a single hypothesis. Our approach significantly outperforms previous particle-based algorithms in experiments focusing on the estimation of human pose from single images.

1. Introduction

Algorithms for computing most likely configurations, or *modes*, of posterior distributions play a key role in many applications of probabilistic graphical models. The *max-product* variant of the *belief propagation* (BP) message-passing algorithm can efficiently identify these modes for many discrete models (Wainwright & Jordan, 2008). However, the dynamic programming message updates underlying max-product have cost that grows quadratically with the number of discrete states. In domains such as computer vision and signal processing, we often need to estimate high-dimensional continuous variables for which exact message updates are intractable, and accurate discretization is infeasible. Monte Carlo methods like simulated annealing provide one common alternative (Geman & Geman, 1984; Andrieu et al., 2003), but in many applications they are impractically slow to converge.

Inspired by work on *particle filters* and *sequential Monte Carlo* methods (Cappé et al., 2007), several algorithms employ particle-based approximations of continuous BP messages. In these approaches, a non-uniform discretization adapts and evolves across many message-passing iterations (Koller et al., 1999; Sudderth et al., 2003; Isard, 2003; Ihler & McAllester, 2009). This literature focuses on the sum-product BP algorithm for computing marginal distributions, and corresponding *importance sampling* methods are used to update particle locations and weights. These stochastic resampling steps may lead to instabilities and degeneracies unless the number of particles is large.

Motivated by complementary families of *maximum a posteriori* (MAP) inference problems, we instead develop a *diverse particle max-product* (D-PMP) algorithm. We view the problem of approximating continuous max-product BP messages from an optimization perspective, and treat each particle as a hypothesized solution. Particle sets are kept to a computationally tractable size not by stochastic resampling, but by an optimization algorithm which directly min-

[*] J. Pacheco and S. Zuffi contributed equally to this work.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

imizes errors in the max-product messages. We show that the D-PMP algorithm implicitly seeks to maintain all significant posterior modes, and is substantially more robust to initialization than previous particle max-product methods.

We begin in Section 2 by reviewing prior particle BP algorithms, and contrast the MAP objective of max-product BP with the more widely studied marginalization problem of sum-product BP. We develop the D-PMP particle selection criterion and algorithm in Section 3. Section 4 provides an extensive validation on the challenging problem of articulated human pose estimation from single images, demonstrating state-of-the-art performance and significant improvements over other particle max-product algorithms.

2. Particle-Based Message Approximations

Consider a pairwise *Markov random field* (MRF), in which edges $(s,t) \in \mathcal{E}$ link pairs of nodes, and each node $s \in \mathcal{V}$ is associated with a continuous random variable x_s:

$$p(x) \propto \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t). \tag{1}$$

BP algorithms (Wainwright & Jordan, 2008) focus on a pair of canonical inference problems: sum-product computation of marginal distributions $p_s(x_s)$, or max-product computation of modes $\hat{x} = \arg\max_x p(x)$. Exact inference is intractable for most non-Gaussian continuous x, and numerical approximations based on a fixed discretization are only feasible for low-dimensional models. Particle-based inference algorithms instead aim to dynamically find a good, non-uniform discretization for high-dimensional models.

2.1. Sum-Product Particle Belief Propagation

For all BP algorithms, the local *belief* $\mu_s(x_s)$ is determined by multiplying the local potential $\psi_s(x_s)$ with messages $m_{ts}(x_s)$ from neighbors $\Gamma(s) = \{t \mid (s,t) \in \mathcal{E}\}$:

$$\mu_s(x_s) \propto \psi_s(x_s) \prod_{t \in \Gamma(s)} m_{ts}(x_s). \tag{2}$$

For sum-product BP, this belief is an estimate of the marginal distribution $p_s(x_s)$, and messages are defined as

$$m_{ts}(x_s) \propto \int_{\mathcal{X}_t} \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{k \in \Gamma(t) \setminus s} m_{kt}(x_t)\, dx_t, \tag{3}$$

where $\mathcal{X}_t$ is the continuous domain of x_t. However, this continuous BP update does not directly provide a realizable algorithm: the integral over $\mathcal{X}_t$ may be intractable, and the message function $m_{ts}(x_s)$ may not have an analytic form.

Importance Sampling Because BP messages are non-negative and (typically) normalizable, the BP message update can be viewed as an expectation of the pairwise potential function $\psi_{st}(x_s, x_t)$. Importance sampling methods (Andrieu et al., 2003) provide a general framework for approximating such expectations via weighted samples:

$$\mathbb{E}[g(x)] = \int_{\mathcal{X}} g(x) p(x)\, dx \approx \sum_{i=1}^{N} g(x^{(i)}) w(x^{(i)}),$$

$$x^{(i)} \sim q(x), \quad w(x) \propto \frac{p(x)}{q(x)}, \quad \sum_{i=1}^{N} w(x^{(i)}) = 1. \tag{4}$$

The proposal distribution $q(x)$ is used to approximate the expectation of $g(x)$ with respect to the target $p(x)$. Under fairly general conditions, this estimator is asymptotically unbiased and consistent (Andrieu et al., 2003).

Particle BP Returning to the message update of Eq. (3), let $M_{ts}(x_t) = \psi_t(x_t) \prod_{k \in \Gamma(t) \setminus s} m_{kt}(x_t)$ denote the *message foundation*. Given N particles $\mathbb{X}_t = \{x_t^{(1)}, \dots, x_t^{(N)}\}$ sampled from some proposal distribution $x_t^{(i)} \sim q_t(x_t)$, let

$$\hat{m}_{ts}(x_s) = \sum_{i=1}^{N} \psi_{st}(x_s, x_t^{(i)}) w_t(x_t^{(i)}), \tag{5}$$

where $w_t(x_t) \propto M_{ts}(x_t)/q_t(x_t)$. We can then construct a belief estimate $\hat{\mu}_s(x_s)$ by substituting the message approximation $\hat{m}_{ts}(x_s)$ in Eq. (2). Koller et al. (1999) and Ihler & McAllester (2009) take $q_t(x_t) = \hat{\mu}_t(x_t)$ so that particles are sampled from the approximate marginals, but other proposal distributions are also possible. In some cases, Metropolis-Hastings MCMC methods are used to iteratively draw these proposals (Kothapa et al. (2011)).

For junction tree representations of Bayesian networks, Koller et al. (1999) describe a general framework for approximating clique marginals given appropriate marginalization and multiplication operations. The nonparametric BP (Sudderth et al., 2003) and PAMPAS (Isard, 2003) algorithms approximate continuous BP messages with kernel density estimates, and use Gibbs samplers (Ihler et al., 2004) to propose particles from belief distributions. The sum-product *particle belief propagation* (PBP) algorithm of Ihler & McAllester (2009) associates particles with nodes rather than messages or cliques, and thus avoids the need for explicit marginal density estimates.

2.2. Max-Product Particle Belief Propagation

Rather than approximating marginal expectations, the max-product algorithm solves the optimization problem of finding posterior modes. The standard max-product algorithm is similar to sum-product BP, but the integration in Eq. (3) is replaced by a maximization over all $x_t \in \mathcal{X}_t$. The beliefs $\mu_s(x_s)$ then become *max-marginal* distributions (Wainwright & Jordan, 2008) encoding the probability of the most likely joint configuration with any fixed x_s.

While max-product message updates are sometimes simpler than sum-product, this optimization remains intractable for many continuous graphical models. However, given any set of N particles $\mathbb{X}_t$, we may approximate the

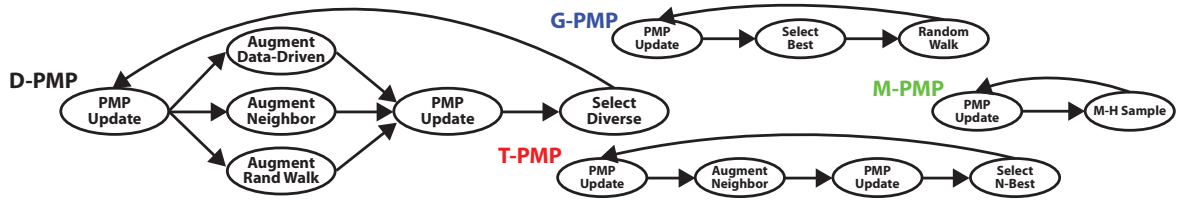


Figure 1. **PMP flowcharts** We compare the Metropolis PMP (M-PMP) of Kothapa et al. (2011), the Greedy PMP (G-PMP) of Trinh & McAllester (2009), the PatchMatch BP (T-PMP) of Besse et al. (2012), and our novel Diverse PMP (D-PMP) algorithm.

true continuous max-product messages as follows:

$$\hat{m}_{ts}(x_s) = \max_{x_t \in \mathbb{X}_t} \psi_{st}(x_s, x_t)\psi_t(x_t) \prod_{k \in \Gamma(t) \setminus s} \hat{m}_{kt}(x_t). \quad (6)$$

Because we do not seek an unbiased estimator of an integral, there is no need for the importance weighting used in the particle sum-product message updates of Eq. (5), or even for knowledge of the distribution from which particles were drawn. Since $\mathbb{X} \subset \mathcal{X}$ for any candidate particle set, particle max-product updates lower-bound the true mode:

$$\max_{x \in \mathbb{X}} \log p(x) \leq \max_{x \in \mathcal{X}} \log p(x). \quad (7)$$

The bound is tight whenever $\mathbb{X}$ contains the true MAP configuration. Various *particle max-product* (PMP) algorithms (see Fig. 1) have been devised for optimizing this bound.

Metropolis Particle Max-Product (M-PMP) Building directly on the sum-product PBP algorithm of Ihler & McAllester (2009), Kothapa et al. (2011) approximately sample particles $x_t^{(i)}$ from the current max-marginal estimate $\hat{\mu}_t(x_t)$ using a Metropolis sampler with Gaussian random walk proposals. Because the entire particle set is replaced at each iteration, discovered modes may be lost and the bound of Eq. (7) decrease. While drawing particles from max-marginals does explore important parts of the state space, the Metropolis acceptance-ratio computation requires an expensive $\mathcal{O}(N^2)$ message update.

Greedy Particle Max-Product (G-PMP) Rather than using conventional resampling rules, Trinh & McAllester (2009) employ a greedy approach which selects the single particle with highest max-marginal value at each iteration, $x_s^* = \arg\max_{x_s \in \mathbb{X}_s} \hat{\mu}_s(x_s)$. New particles are then generated by adding Gaussian noise, $x_s^{(i)} \sim N(x_s^*, \Sigma)$. This approach can be guaranteed to monotonically increase the MAP objective by retaining x_s^* in the particle set, but discards all non-maximal modes after each iteration, and thus is fundamentally local in its exploration of hypotheses.

PatchMatch & Top-N Particle Max Product (T-PMP) Besse et al. (2012) recently proposed a *PatchMatch BP* algorithm specialized to models arising in low-level computer vision. At each iteration, the particle sets at each node are augmented with samples generated from their neighbors. Max-marginals $\hat{\mu}_s(x_s)$ are computed on the augmented set, and the N particles with largest max-marginal are retained. This approach produces monotonically increasing MAP estimates, but their proposal distribution is specialized to pairwise MRFs in which potentials prefer neighboring nodes to take identical values. To provide a baseline for the more sophisticated selection rules defined in Sec. 3, we define a T-PMP method which employs the PatchMatch particle selection rule, but employs neighbor-based proposals appropriate for arbitrary potentials.

2.3. Discovering Diverse Solutions

Statistical models of complex phenomena are often approximate, and the most probable hypotheses may not be the most accurate (Meltzer et al., 2005; Szeliski et al., 2008). However, given many candidate solutions (modes) from an approximate model, one can then rerank them based on more complex non-local features. This approach has lead to state-of-the-art results in natural language parsing (Charniak & Johnson, 2005), image segmentation (Yadollahpour et al., 2013), and computational biology problems like protein design (Fromer & Yanover, 2009).

Several algorithms for finding the "M-Best" joint state configurations have been developed (Nilsson, 1998; Yanover & Weiss, 2003; Fromer & Globerson, 2009). However, especially for models derived from some discretization of underlying continuous variables, the top solutions are typically slight perturbations of the same hypothesis. Batra et al. (2012) propose an alternative algorithm for selecting the "Diverse M-Best" modes of a discrete model. Given the global MAP, they iteratively find the next best solution which differs from all previous solutions according to some externally provided dissimilarity metric. However, for general models it may be difficult to define and tune such a metric, and it is unclear how to solve the corresponding optimization problems for graphs with continuous variables. We develop an alternative approach which leverages the model potentials to implicitly encourage diversity, at a scale automatically tuned to the graphical model of interest, with no need for an explicit dissimilarity measure.

3. Diverse Particle Max-Product (D-PMP)

Our *diverse particle max-product* (D-PMP) algorithm replaces the typical *resample-update* paradigm employed by

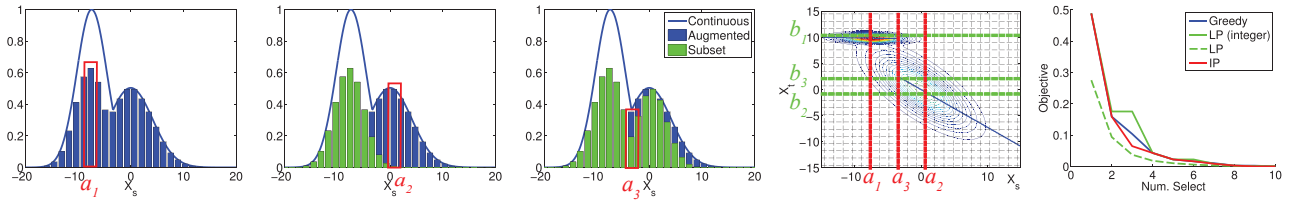


Figure 2. **Greedy diverse selection of particles** at x_t to preserve the message $m_{ts}(x_s)$. The model is a two-node correlated Gaussian pairwise potential $\psi_{st}(x_s, x_t) = N(x \mid \mu_{st}, \Sigma_{st})$, and unary potentials of evenly-weighted mixtures of two Gaussians. To aid visualization, particles are arranged on a regular grid (dashed lines). The first three plots show successive message approximations $\hat{m}_t^{(0)}, \hat{m}_t^{(1)}, \hat{m}_t^{(2)}$ (green, Eq. (12)), which lower-bound the true message m_t (blue, Eq. (10)). Indicies a_1, a_2, a_3 denote locations of maximum approximation error (red, Eq. (13)). The message foundation matrix M (second from right) shows particles selected with indicies b_1, b_2, b_3 as horizontal lines (green, Eq. (14)). The objective function values are shown (right, Eq. (11)) versus the number of particles selected by our greedy algorithm, the optimal IP solution (via exhaustive enumeration), a standard linear programming (LP) relaxation lower bound, and the LP solution rounded to a feasible integer one.

NBP and particle BP, and more broadly by most particle filters, with an optimization-guided stochastic search. As shown in Figure 1, we first use stochastic proposals to *augment* the set of particles with a candidate set of hypotheses. We then *update* the messages over the new particle set using the PMP message updates of Eq. (6). Finally, we *select* the subset of particles (or hypotheses) which preserve the current message values.

For simplicity, we formulate D-PMP for a pairwise MRF, however this should not be viewed as a limitation of the algorithm. We allocate N particles to each node at the start of each iteration. Stochastic proposals augment these sets to contain αN particles for some $\alpha > 1$; in our later experiments, $\alpha = 2$. Updated max-product messages are then used to select N particles for the subsequent iteration.

3.1. Augmentation step

Given N particles $\mathbb{X}_s$ from the preceding iteration, we draw $(\alpha - 1)N$ new particles $\mathbb{X}_s^{\text{prop}}$ by independently sampling from some proposal distribution $q_s(x_s \mid \mathbb{X}_s)$. Rather than discarding current particles as in typical resampling rules, we define an augmented set $\mathbb{X}_s^{\text{aug}} = \mathbb{X}_s \bigcup \mathbb{X}_s^{\text{prop}}$ containing αN particles. Various proposal distributions can be randomly or deterministically interleaved across iterations.

Data-Driven A distribution proportional to the observation potential, $q_s^{\text{data}}(x_s) \propto \psi_s(x_s)$, can often be either exactly or approximately sampled from. These *data-driven* proposals explore modes of the local likelihood function.

Neighbor-Based By conditioning on a single particle $\bar{x}_t$ for each neighboring node, we define a local conditional distribution $q_s^{\text{nbr}}(x_s) \propto \prod_{t \in \Gamma(s)} \psi_{st}(x_s, \bar{x}_t)$ as in Gibbs samplers. Such proposals can lead to global propagation of good local hypotheses, resulting in high-probability global modes. As sampling from a product of pairwise potentials is not generally tractable, we propose an approximation based on the *mixture importance sampler* of Ihler et al. (2004). For each new particle, we first sample some neighboring particle $\bar{x}_t \sim \mu_t(x_t)$ according to the current max-marginal estimate, and then take $q_s^{\text{nbr}}(x_s) \propto \psi_{st}(x_s, \bar{x}_t)$. Several samples are drawn with respect to each $t \in \Gamma(s)$.

Random-walk We also utilize a Gaussian random walk proposal $q_s^{\text{walk}}(x_s) = N(x_s \mid x_s^{(i)}, \Sigma)$, where proposals are sampled with respect to various $x_s^{(i)} \in \mathbb{X}_s$. The proposal covariance matrix Σ can be tuned to favor refinement of existing hypotheses, or exploration of new hypotheses.

3.2. Particle selection step

For each node $t \in \mathcal{V}$ we now have an augmented particle set $\mathbb{X}_t^{\text{aug}}$ containing αN particles. While we would prefer to never discard hypotheses, storage and computational constraints force us to reduce to only N important particles $\mathbb{X}_t^{\text{new}} \subset \mathbb{X}_t^{\text{aug}}$. We propose to do this by *minimizing the maximum error* between max-product messages computed on the augmented and reduced sets. The resulting *integer program* (IP) encourages diversity among the selected particles without any need for explicit distance constraints. Instead, the goal of message preservation automatically allocates particles near each non-trivial max-marginal mode.

IP formulation The particle message approximation $\hat{m}_{ts}(x_s)$ in Eq. (6) is a continuous function of x_s, which we seek to preserve on the augmented particle set $\mathbb{X}_s^{\text{aug}}$. Letting $a = 1, \ldots, \alpha N$ index particles at node s, and b index particles at node t, define a *message foundation matrix* as

$$M_{st}(a, b) = \psi_{st}(x_s^{(a)}, x_t^{(b)}) \psi_t(x_t^{(b)}) \prod_{k \in \Gamma(t) \setminus s} \hat{m}_{kt}(x_t^{(b)}), \quad (8)$$

where $M_{st} \in \mathbb{R}^{\alpha N \times \alpha N}$. The message foundation gives a compact representation for computing the message between two nodes over the augmented particle set.

Because node t sends messages to all $d = |\Gamma(t)|$ of its neighbors, we construct a "stacked" matrix of the message foundations for all neighbors $\Gamma(t) = \{s_1, \ldots, s_d\}$:

$$M_t = \left[M_{s_1 t}^T, \ldots, M_{s_d t}^T\right]^T \in \mathbb{R}^{d\alpha N \times \alpha N}. \quad (9)$$

The maximal values of the rows of the message foundation matrix are then the max-product messages $\hat{m}_{ts}(x_s^{(a)})$ sent

to all particles a at all neighbors $s \in \Gamma(t)$:

$$m_t(a) = \max_{1 \le b \le \alpha N} M_t(a, b). \tag{10}$$

Selecting a set of particles corresponds to choosing a subset of the message foundation columns. We let $z_{tb} = 1$ if column b (particle $x_t^{(b)}$) is selected, and $z_{tb} = 0$ otherwise. Particles are selected to *minimize* the *maximum* absolute error in the message approximations to *all* neighbors:

$$\arg\min_{z_t} \left[\max_{1 \le a \le d\alpha N} \left(m_t(a) - \max_{1 \le b \le \alpha N} z_{tb} M_t(a, b) \right) \right]$$
$$\text{subject to} \quad \sum_{b=1}^{\alpha N} z_{tb} = N, \qquad z_t \in \{0, 1\}^{\alpha N}. \tag{11}$$

The solution vector z_t directly provides a new set of N particles $\mathbb{X}_t^{\text{new}}$. The IP of Eq. (11) is likely NP hard in general, and so we develop an approximation algorithm.

Greedy approximation algorithm We begin with an empty particle set, and at each iteration $k = 1, \ldots, N$ a single particle with index b_k is selected to produce an improved approximation $\hat{m}_t^{(k)} \in \mathbb{R}^{d\alpha N}$ of the outgoing messages m_t of Eq. (10). Because maximization is associative,

$$\hat{m}_t^{(k)}(a) = \max \left\{ \hat{m}_t^{(k-1)}(a),\ M_t(a, b_k) \right\}, \tag{12}$$

where $\hat{m}_t^{(0)} = \vec{0}$ for the empty initial particle set. To choose a particle to add, we first identify the neighboring particle index $a_k \in \{1, \ldots, d\alpha N\}$ with the largest message approximation error, and greedily select the particle index $b_k \in \{1, \ldots, \alpha N\}$ which minimizes this error:

$$a_k = \arg\max_{1 \le a \le d\alpha N} \ m_t(a) - \hat{m}_t^{(k-1)}(a), \tag{13}$$

$$b_k = \arg\max_{1 \le b \le \alpha N} \ M_t(a_k, b). \tag{14}$$

The particle selection of Eq. (14) always eliminates errors in the max-product message for particle a_k, because

$$\hat{m}_t(a_k) = M_t(a_k, b_k) = \max_b M_t(a_k, b) = m_t(a_k).$$

It may also reduce or eliminate errors in messages for particles a where $\psi_{st}(x_s^{(a)}, x_t^{(b_k)})$ is large.

See Figure 2 for a graphical depiction of the greedy selection procedure on a Gaussian mixture model. Each step of Eq. (12,13,14) requires $\mathcal{O}(d\alpha N)$ time, so the overall cost of selecting N particles is $\mathcal{O}(d\alpha N^2)$. This quadratic cost is comparable to the max-product message updates of Eq. (6). While our experiments treat N as a fixed parameter trading off accuracy with computational cost, it may be useful to vary the number of selected particles across nodes or iterations of D-PMP, for example by selecting particles until some target error level is reached.

4. Application to Human Pose Estimation

D-PMP is particularly well suited to applications in computer vision, where the unknown quantities are typically high-dimensional continuous random variables, and weak likelihoods lead to multimodal density functions with many local optima. In addition, D-PMP places no restrictions on the parametric form of the model potentials, allowing for complex likelihood functions. In this section we apply D-PMP to human pose and shape estimation from single images. We employ the *deformable structures* (DS) model (Zuffi et al., 2012), an articulated part-based human body representation which models pose and shape variation. Discrete approximations are infeasible due to the high-dimensionality of the DS latent variables, making it an ideal model for demonstrating D-PMP inference.

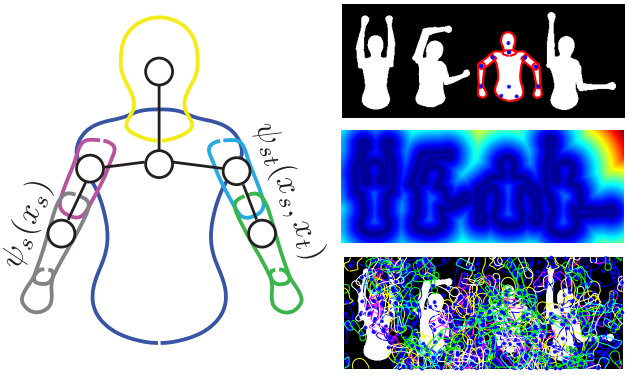


Figure 3. **Human pose estimation** *Left:* The DS upper body model, encoded as a tree-structured 6-node MRF (circles). *Right:* Human silhouettes posed to spell "ICML" with the MAP pose in red (top), the edge-based distance map likelihood (middle, small distances in blue), and an uninformative initialization based on 200 particles per node sampled uniformly at random (bottom).

4.1. Deformable Structures

The DS model specifies a pairwise MRF with nodes $s \in \mathcal{V}$ for each body part, and links kinematic neighbors with edges $(s, t) \in \mathcal{E}$ (Figure 3). Shape is represented by learned PCA coefficients z_s. With global rotation θ_s, scale d_s, and center o_s, the state of part s is

$$x_s = (z_s, o_s, \sin(\theta_s), \cos(\theta_s), d_s)^T. \tag{15}$$

A pair (s, t) of neighboring body parts is connected by joints with locations p_{st} and p_{ts}, respectively. To model their relationships, we first capture the parts' relative displacement $q_{ts} = p_{ts} - p_{st}$, relative orientation $\theta_{ts} = \theta_t - \theta_s$, and scale difference $d_{ts} = d_t - d_s$ via the transformation

$$T_{st}(x_s, x_t) = (z_s, z_t, \sin(\theta_{ts}), \cos(\theta_{ts}), q_{ts}, d_{ts})^T. \tag{16}$$

Our truncated Gaussian pairwise potential is $\psi_{st}(x_s, x_t) \propto$

$$N(T_{st}(x_s, x_t) \mid \mu_{st}, \Sigma_{st}) \mathbb{I}_{\mathcal{A}}(d_s, \theta_s) \mathbb{I}_{\mathcal{A}}(d_t, \theta_t), \tag{17}$$

where the indicator function $\mathbb{I}_{\mathcal{A}}(\cdot)$ enforces valid angular components and non-negativity of the scale parameters by the constraint set $\mathcal{A} = \{d, \theta \mid d > 0,\ \sin^2\theta + \cos^2\theta = 1\}$.

The likelihood of pose x_s is obtained via contour points $c_s = B_s z_s + m_s$ defined in object-centered coordinates.

The mean m_s and transformation matrix B_s are learned via a PCA analysis of part-specific training data. A rotation matrix $R(\theta_s)$, scaling d_s, and translation $t(o_s)$ are then applied to draw the contour points in the image:

$$i_s(x_s) = d_s R(\theta_s) c_s + t(o_s), \quad c_s = B_s z_s + m_s. \quad (18)$$

Image likelihoods $\psi_s(x_s)$ for our synthetic-data experiments are determined from the distance of contours $i_s(x_s)$ to the closest observed edge, as shown in Figure 3. For real images, two complementary potentials capture information about boundary contours and skin color:

$$\psi_s(x_s) = \psi_s^{\text{contour}}(x_s)\psi_s^{\text{skin}}(x_s). \quad (19)$$

The contour likelihood is based on an SVM classifier trained on *histogram of oriented gradients* (HOG, Dalal & Triggs (2005)) features $h_s(i_s(x_s))$. SVM scores $f_s(h_s(i_s(x_s)))$ are mapped to calibrated probabilities via logistic regression (Platt, 1999), using a weight a_s and bias b_s learned from validation data:

$$\psi_s^{\text{contour}}(i_s(x_s)) = \frac{1}{1 + \exp(a_s f_s(h_s(i_s(x_s))) + b_s)}. \quad (20)$$

The skin color likelihood $\psi_s^{\text{skin}}(i_s(x_s))$ captures the tendency of lower arms to be unclothed, and is derived from a histogram model of skin appearance (Zuffi et al., 2012).

4.2. Synthetic Images

In this section we compare D-PMP with baseline methods on a set of synthetic images sampled from the DS model. Two experiments are conducted: in the first we assess each method's ability to retrieve the global MAP configuration, and in the second we evaluate how well each method retains multiple significant hypotheses in the posterior. For all methods we use 200 particles and run for 300 iterations.

Global MAP Estimate For this experiment we use a hand-constructed image containing four silhouettes arranged to spell "ICML" (Figure 3). The smooth distance likelihood produces significant modes of comparable size for each of the four figures, and their relative posterior probability is largely driven by the prior density. The third figure from the left (the letter "M") turns out to correspond to the global MAP since it is near the prior mean.

We assume a broadly sampled initial set of particles (Figure 3) and measure average error in body joints, across 10 runs, between the MAP estimate of all methods and the true MAP. Figure 4 shows a box plot of average body joint errors for each method. Other particle methods typically fail to discover the true MAP estimate, resulting in larger joint error compared to D-PMP, which locates the global MAP estimate in all but a single run. While both D-PMP and T-PMP typically produce high-probability configurations (Figure 4), the latter is sensitive to local optima, concentrating all particles on a single configuration (Figure 5). We also consider a hybrid method, D/T-PMP, in which D-PMP is run for the first 200 iterations and T-PMP for the final 100. This approach produces refined estimates which are near the global MAP in all runs, and have higher probability (better alignment) than either D-PMP or T-PMP.

Preserving Multiple Hypotheses In this experiment we sample, from the DS model prior, 9 *puppets* arranged in a 3×3 grid. A series of 6 images are generated, varying the relative distance between the puppets, and we measure the ability of each method to preserve hypotheses about significant modes as occlusion is increased (Figure 6). We use an oracle to select the torso particle closest to each ground-truth figure, and a Viterbi-style backward pass generates the modes consistent with each torso hypothesis. Figure 4 shows a line plot of median joint error versus puppet distance. Nine lines are plotted for each method, each line corresponding to a puppet. D-PMP maintains significantly better mode estimates compared to other methods.

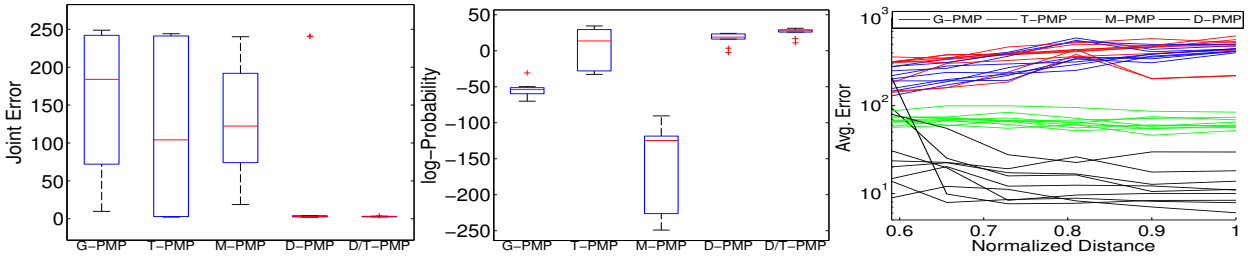
Figure 6 shows the final particle locations for one example run of each method. We observe sensitivity to local optima in T-PMP and G-PMP, which generally capture only one mode. M-PMP scatters particles widely, but does a poor job of concentrating particles on modes of interest.

4.3. Real Images

We demonstrate the robustness of our proposed algorithm on the *Buffy the Vampire Slayer* dataset (Ferrari et al., 2008), a widely used benchmark for evaluating pose estimation methods based on part-based models. The dataset consists of a standard partition of 276 test images and about 500 training images. We use a recent set of *stickmen* annotations for all figures in the dataset (Ladický et al., 2013). Images are partitioned into single- and multi-person groups, and results are reported on each set separately using different evaluation criteria. Detailed results for all images are provided in the supplemental material.

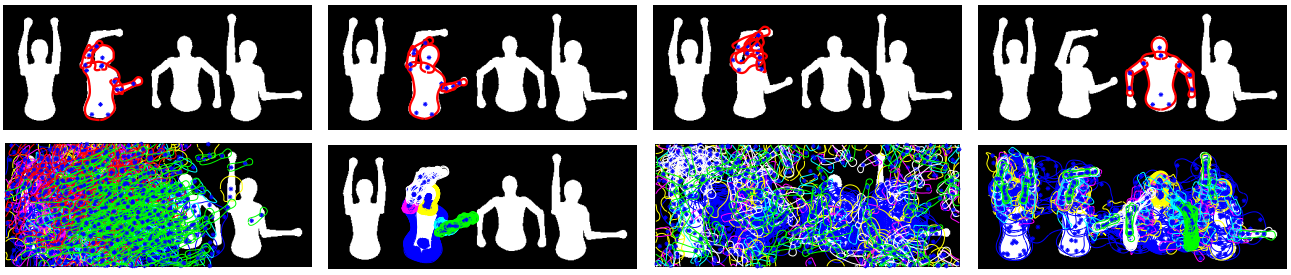
Inference and learning details We initialize with 100 particles for each body part sampled around candidate hypotheses generated from the *flexible mixture of parts* (FMP) pose estimation method (Yang & Ramanan, 2013). We prune FMP candidates with scale below a value of 0.5, and apply non-maximal suppression with overlap threshold 0.8. We run D-PMP, and our baseline particle methods, for 100 iterations per image. We also compare to the N-best maximal decoders computed on the raw FMP detections (Park & Ramanan, 2011), which uses a similarity metric to produce a diverse set of solutions, and has been shown to be more accurate than non-maximal suppression.

Detecting a single person For single-person images we use the standard evaluation criteria for this dataset, the *percentage of correctly estimated parts* (PCP), which is a detection metric based on the annotated stick representation. For a ground-truth part segment with endpoints g_1 and g_2, a predicted part segment with endpoints p_1 and



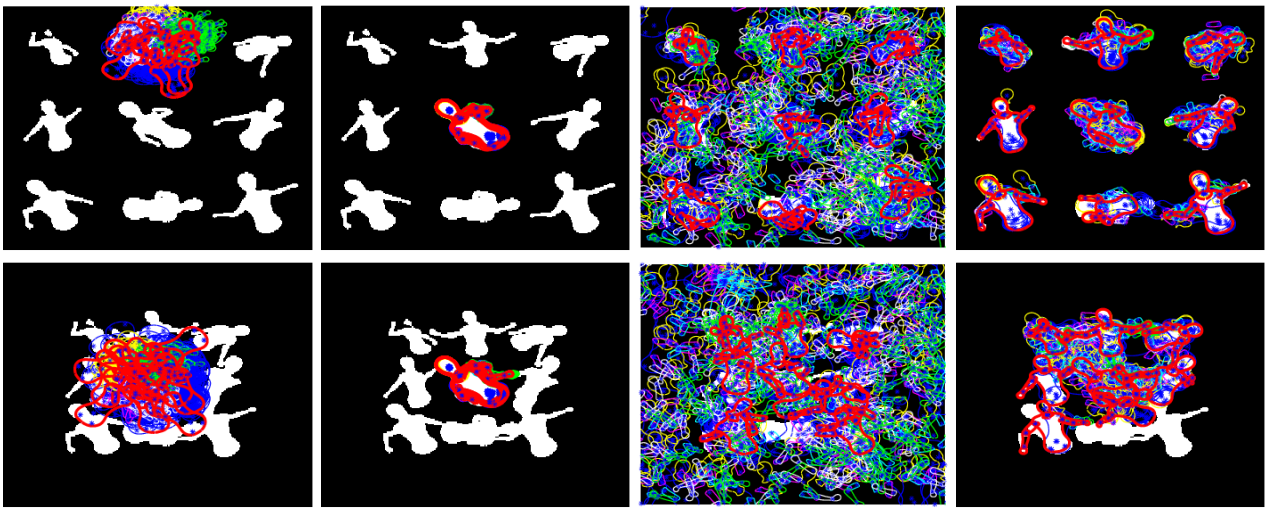
Joint Error | Log-Probability | Joint Error vs. Dist.

Figure 4. **Synthetic image experiments** *Left:* Box plots for 10 trials of the "ICML" experiment, where the joint error equals the L_2 distance from the true MAP pose, averaged over all joints. *Center:* Log-probability of the most likely configuration identified by each method. *Right:* Median joint error in the *distance experiment* of Figure 6, plotted versus the distance separating the 9 poses.



G-PMP | T-PMP | M-PMP | D-PMP

Figure 5. **Typical pose estimation results** We show the final MAP estimate (top) and 200 particles per part (bottom) for each method.



G-PMP | T-PMP | M-PMP | D-PMP

Figure 6. **Preserving multiple modes** Figures do not overlap at the furthest spacing (top), but extremities overlap at the closest spacing (bottom). Each method is run for 300 iterations from 30 random 200-particle initializations. The top 9 modes (red) are obtained by selecting the closest torso particle to each ground truth puppet, and from this a Viterbi backward pass generates the remaining limbs.

p_2 is *detected* if the average distance between endpoints is less than one-half the length of the ground truth segment: $\frac{1}{2}(\|g_1 - p_1\| + \|g_2 - p_2\|) \leq \frac{1}{2}\|g_1 - g_2\|$. The PCP score is the fraction of the full set of parts which are detected.[1]

[1]Ferrari et al. (2008) compute PCP relative to the number of images in the dataset which contain a detection, creating irregularities when varying the number of hypotheses. We instead normalize by the fixed number of images in the dataset.

Pose hypotheses are sorted according to their max-marginal value (or FMP score), and we report total PCP versus the number of hypothesized poses in Figure 8. We report scores averaging over all body parts, and separately for only the left and right lower arms, as these parts are the most difficult to detect accurately. While scores for the arms are uniformly lower as compared to total PCP, the trend is similar: given an identical model, D-PMP is sub-

Single Person **Multiple People**

Figure 7. **Preserving multiple hypotheses** *Left:* Single person images showing a MAP estimate (red) with poor arm placement. The second and third ranked solutions preserved by D-PMP, by max-marginal values, are shown for upper (magenta-cyan) and lower arms (white-green); they offer much greater accuracy. *Right:* The full set of particles at the final iteration of D-PMP shows multiple hypotheses retained about multiple people (top). For each person, we also plot the best pose in the set of retained hypotheses (bottom, red).

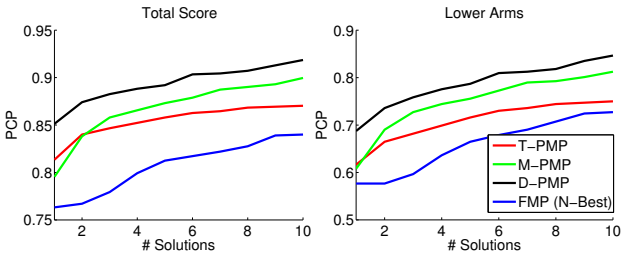


Figure 8. **Detection results for single person images** Average PCP score versus the number of hypotheses for all parts (left), and for only the lower arms (right). D-PMP shows the highest detection accuracy overall, with its diverse selection providing increasing gains as more hypotheses are considered.

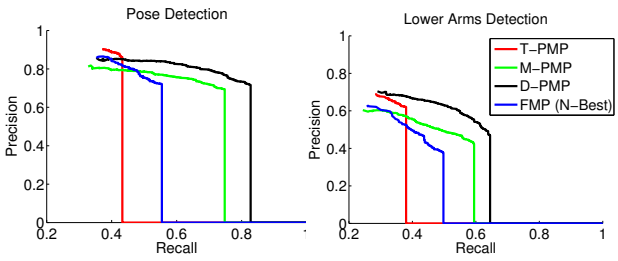


Figure 9. **Detection results for multi-person images** Precision-recall curves for body detections (left) and lower arm detections (right), determined via a PCP threshold of 0.5. A body is detected if either the torso or head is detected. D-PMP maintains fairly good precision for much higher levels of recall.

stantially more accurate than conventional particle max-product algorithms. We offer qualitative examples of how D-PMP preserves alternative (upper and lower arm) hypotheses in Figure 7.

Detecting multiple people For multiple people we report precision-recall, a standard metric for multi-class object detection. Figure 9 shows precision-recall for each method, where a body is considered detected if the torso or head PCP score is 1. We evaluate the challenging lower arm detection problem separately. The first point on each curve reports precision and recall based on the top-scoring pose in each image, and the curves are traced out by considering the top two, three, etc. hypotheses in each image.[2] D-PMP again outperforms all other methods, both for body detection as well as for lower arm detection. Figure 7 offers qualitative examples of D-PMP's ability to preserve hypotheses about multiple people in an image. Without an explicit model of multiple people, we are able to infer their existence by finding multiple diverse posterior modes.

[2]This differs slightly from the approach in the PASCAL VOC challenges, which consider the single top-scoring pose over *all* images, resulting in a curve starting at the top left corner.

5. Discussion

The diverse particle max-product (D-PMP) provides a general-purpose MAP inference algorithm for high-dimensional, continuous graphical models. While most existing methods are sensitive to initialization and prone to poor local optima, D-PMP's ability to preserve multiple local modes allows it to better reason globally about competing hypotheses. On a challenging pose estimation task, we show that D-PMP is robust to initialization, and we obtain accurate pose estimates for images depicting multiple people even without an explicit multi-person model. We believe the stability and robustness of D-PMP will prove similarly useful in many other application domains.

Acknowledgments We thank Rajkumar Kothapa, whose earlier insights about particle max-product algorithms (Kothapa et al., 2011) motivated this work. This research supported in part by ONR Award No. N00014-13-1-0644.

References

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to MCMC for machine learning. *JMLR*, 50(1-2):5–43, 2003.

Batra, D., Yadollahpour, P., Guzman-Rivera, A., and Shakhnarovich, G. Diverse M-best solutions in Markov random fields. In *ECCV*, pp. 1–16. Springer, 2012.

Besse, F., Rother, C., Fitzgibbon, A., and Kautz, J. PMBP: Patchmatch belief propagation for correspondence field estimation. In *BMVC*, 2012.

Cappé, O., Godsill, S. J., and Moulines, E. An overview of existing methods and recent advances in sequential Monte Carlo. *Proc. IEEE*, 95(5):899–924, 2007.

Charniak, E. and Johnson, M. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL*, pp. 173–180, 2005.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR*, pp. 886–893, 2005.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. Progressive search space reduction for human pose estimation. In *CVPR*, pp. 1–8. IEEE, 2008.

Fromer, M. and Globerson, A. An LP view of the M-best MAP problem. In *NIPS 22*, pp. 567–575, 2009.

Fromer, M. and Yanover, C. Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space. *Proteins: Structure, Function, and Bioinformatics*, 75(3):682–705, 2009.

Geman, S. and Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE PAMI*, 6(6):721–741, November 1984.

Ihler, A. and McAllester, D. Particle belief propagation. In *AISTATS*, pp. 256–263, 2009.

Ihler, A. T., Sudderth, E. B., Freeman, W. T., and Willsky, A. S. Efficient multiscale sampling from products of Gaussian mixtures. In *NIPS 16*. MIT Press, 2004.

Isard, M. PAMPAS: Real-valued graphical models for computer vision. In *CVPR*, volume 1, pp. 613–620, 2003.

Koller, D., Lerner, U., and Angelov, D. A general algorithm for approximate inference and its application to hybrid Bayes nets. In *UAI*, pp. 324–333, 1999.

Kothapa, R., Pacheco, J., and Sudderth, E. Max-product particle belief propagation. Master's project report, Brown University Dept. of Computer Science, 2011.

Ladický, L., Torr, P. H. S., and Zisserman, A. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, pp. 3578–3585. IEEE, 2013.

Meltzer, T., Yanover, C., and Weiss, Y. Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *ICCV*, volume 1, pp. 428–435. IEEE, 2005.

Nilsson, D. An efficient algorithm for finding the M most probable configurations in probabilistic expert systems. *Statistics and Computing*, 8(2):159–173, 1998.

Park, D. and Ramanan, D. N-best maximal decoders for part models. In *ICCV*, pp. 2627–2634. IEEE, 2011.

Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Sudderth, E. B., Ihler, A. T., Freeman, W. T., and Willsky, A. S. Nonparametric belief propagation. In *CVPR*, pp. 605–612, 2003.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *PAMI*, 30(6):1068–1080, 2008.

Trinh, H. and McAllester, D. Unsupervised learning of stereo vision with monocular cues. In *BMVC*, 2009.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Yadollahpour, P., Batra, D., and Shakhnarovich, G. Discriminative re-ranking of diverse segmentations. In *CVPR*, pp. 1923–1930. IEEE, 2013.

Yang, Y. and Ramanan, D. Articulated human detection with flexible mixtures-of-parts. *IEEE PAMI*, 35(12): 2878–2890, December 2013.

Yanover, C. and Weiss, Y. Finding the M most probable configurations using loopy belief propagation. In *NIPS 16*, pp. 289–296, 2003.

Zuffi, S., Freifeld, O., and Black, M. From pictorial structures to deformable structures. In *CVPR*, pp. 3546–3553. IEEE, 2012.