A Statistical Field Model for Pedestrian Detection

Ying Wu and Ting Yu and Gang Hua
Department of Electrical & Computer Engineering
Northwestern University, Evanston, IL 60208
{yingwu,tingyu,ganghua}@ece.northwestern.edu

Abstract

This paper presents a new statistical model for detecting and tracking deformable objects such as pedestrians, where large shape variations induced by local shape deformation can not be well captured by global methods such as PCA. The proposed model employs a Boltzmann distribution to capture the prior of local deformation, and embeds it into a Markov network which can be learned from data. A mean field variational analysis of this model provides computationally efficient algorithms for computing the likelihood of image observations and facilitate fast model training. Based on that, effective detection and tracking algorithms for deformable objects are proposed and applied to pedestrian detection and tracking. The proposed method has several advantages. Firstly, it captures local deformation well and thus is robust to occlusions and clutter. In addition, it is computationally tractable. Moreover, it divides deformation into local deformation and global deformation, then conquers them by combining bottom-up and top-down methodologies. Extensive experiments demonstrate the effectiveness of the proposed model for deformable objects.

1 Introduction

The research of human detection and tracking has received more and more attentions in recent years, due to the contemporary emerging applications such as perceptual interfaces, ubiquitous computing and smart video surveillance. Different applications are concerned about different image resolutions of the subjects, thus incur different techniques. For example, in perceptual interfaces, accurate motions of body parts should be determined for commanding and interaction, thus human articulation is of great interests. In smart video surveillance, since the human is typically located at small regions in the images, we generally treat a human as an entity, i.e., as a deformable object. Due to the tremendous variance in the visual appearance of human induced by factors like clothing, deformation and lighting, detecting human seems to be more challenging than detecting faces. In this paper, we investigate human detection and tracking from the perspective of deformable shapes.

The human shape is more or less unique in the real world, and thus provides a powerful clue to be distinguished from other objects. For example, it is easy for us to visually recognize human-like silhouettes or contours. However, a real challenge for computers is to handle the large variations embedded in the shape deformations induced by various factors, including rigid motion such as rotation, global deformation such as scaling and shearing, and local deformation caused by body postures, view changes and clothing styles. Setting aside the deformation caused by rigid motion and global deformation, the local deformation of human shapes is quite complicated and is of very high degrees of freedom. Thus, any successful approach to human detection and tracking must have effective representations and prior models to accommodate the large variations in the local shape deformation. Some examples includes the active shape model [2], exemplar hierarchy model [4], the metric mixture model [13], and generative models [7].

Considering the complexity and the degrees of freedom of the local deformation in human shapes, we propose a new statistical model based on a mean field Markov network to capture complicated priors of the local shape deformation. This model is different from the existing methods as described in Section 2. The proposed model is a graphical model. The hidden layer representing the hidden scene (i.e., the deformable contour) is a Markov network, in which each scene node is associated with an observation model describing the conditional likelihood of image observations of this scene node. The structure of the proposed model is similar to that in [3], but the difference between them is that our method employs the Boltzmann prior for the hidden layer and enables rigorous and elegant analytical results by performing variational analysis, through which the image likelihood can be estimated for detection and model parameters can be learnt from training data. In addition, another theoretical benefit is that the image likelihood estimates are lower bounded. Although the complexity of the model structure prevents tractable exact analysis, we obtain a computationally efficient mean field approximation for the model through the probabilistic variational methods advocated by Jordan et.al [8, 6]. Since local shape deformations are embedded in such a statistical model, it enables effective and efficient detection and tracking algorithms for deformable objects such as pedestrians.

The proposed method enjoys a number of advantages. Firstly, since the model employs a network rather than a vector to describe a shape, it can sufficiently capture the local constraints in the deformation by the local network structure and enable accurate modeling of the local deformation priors. Secondly, since the model captures shape deformation and performs image measurements in a distributed fashion, it is more robust against occlusion than the global approaches (such as PCA) in which image measurements have to be performed in a centralized fashion, i.e., conditioned on all the deformable parameters. Thirdly, having an observation layer effectively addresses the modeling of observation noise and thus it will be more robust against cluttered backgrounds. Fourthly, the mean field approximation provides a computationally efficient way to compute the likelihood of image observations, to infer the hidden states of the model, and to facilitate fast learning. Last but not least, it facilitates the integration of the top-down and bottom-up approaches in tracking deformable objects, where the top-down approaches involve evaluating a large number of hypotheses, and the bottom-up approaches needs large efforts in grouping and detection. Given the large number of DoFs in a deformable object such as a pedestrian, either approach would not be satisfactory, because the number of hypotheses would be tremendous and grouping a deformable object is difficult. The proposed tracking method is able to balance these two methodologies and combines the advantages of both: the global deformation is handled in a top-down fashion by particle filtering, while the local deformation is coped by an bottom-up approach by directly evaluating the likelihood of image observations.

2 Related Work

The research of deformable shapes has a long history, and different approaches have been investigated. For all these methods, three important issues should be addressed, i.e., the shape representation $\mathbf{X}$, the shape prior $p(\mathbf{X})$ and the conditional likelihood of image observation $p(\mathbf{Z}|\mathbf{X})$.

Different shape representations can be categorized into either parametric or non-parametric models. Examples of parametric representations includes Fourier descriptors, B-splines [1, 9], the deformable template [15], etc, where shape deformation is controlled by the shape parameters and smoothness constraints. A typical non-parametric representation is the point distribution model [2] where a shape is described by an ordered and labelled set of landmark points, and the shape deforms when the points change. Although it provides great flexibility, registration of landmark points is not a trivial task. An even radical approach is to use a 2D mask [7, 4, 13], where the shape deforms when multiplying by a sparse permutation matrix [7], or selecting different exemplars [4, 13]. In all these representations, a deformable shape is mapped to a point in a vector space (i.e., the shape space), although the dimensionality varies for different approaches. The proposed approach adopts a 2D representation to ease the task of shape alignment.

Obviously, in reality, a shape can not be allowed for arbitrary deformation, thus we should characterize the allowable shape space by the deformation prior model $p(\mathbf{X})$. An idea is to reduce the dimensionality of $\mathbf{X}$ and model the variance of deformation by a multivariate Gaussian distribution in a lower-dimensional subspace. This is the spirit of principal component analysis (PCA), and has been widely adopted for learning deformation priors [2, 1]. Since PCA identifies a linear subspace and catches linear correlations, it is powerful to capture and decorrelate global deformation, but insufficient for local deformation. Thus, it motivated methods of using mixture distributions [7] or exemplar databases [4, 13]. Although mixture distributions can represent arbitrarily complicated densities in theory, it becomes unrealistic when the number of mixtures increase tremendously. An inhomogeneous Gibbs model was proposed to alleviate this problem for face deformation [10]. Our proposed approach stands out from the above by characterizing $p(\mathbf{X})$ as a Boltzmann distribution and embedding the prior into a Markov network, where the mean field variational analysis is employed for analysis. (details in Section 3 and 4).

Different approaches have been investigated to *fit* a shape model to image observations. This can be done through minimizing an energy function [9], or based on the Bayesian framework where it is important to characterize the conditional likelihood of image observation $p(\mathbf{Z}|\mathbf{X})$. Analytical forms can be obtained by assuming the independence among a set of discrete points on shape contours [1, 11]. To bypass the independence assumption, the conditional likelihood can be modelled as a metric exponential density obtained from the Chamfer distance based on exemplars [13]. When separating global motion from deformation, the likelihood conditioned on only global motion can be obtained by the mixture (integral) of all exemplar components in the metric mixture model [13]. The proposed approach also provides tractable ways to calculate the likelihood only conditioned on global motion, but the differences from [13] are: (a) in our model $p(\mathbf{Z}|\mathbf{X})$ factorizes by independent components, and (b) $p(\mathbf{Z})$ is an integral over almost infinite number of $\mathbf{X}$ instead of a finite set of exemplars, and our method obtains a lower bound of $p(\mathbf{Z})$ through mean field approximation.

3 The Representation

In general, global deformation has less degrees of freedom than local deformation. Thus, global approaches such as PCA are suitable for capturing global deformation by finding a set of deformation basis. Since these approaches are not suitable for local deformation, local methods which

have large DoFs should be employed for representing local deformation. Here, we use a two-layer graphical model as the representation as in Figure 1. The model consists of a

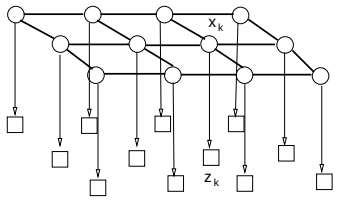

Figure 1: Markov network for deformable objects.

mixture of undirected and directed links. The hidden layer is an undirected graph $G_x = \{V, E\}$, where each vertex (or node) represents the hidden scene x_k to be inferred. x_k takes binary values, i.e., $x_k \in \{0, 1\}$, where $x_k = 1$ means that site k is on the object contour. Each hidden node is connected to its neighborhood nodes $\mathcal{N}(k)$.

The prior of the scene (i.e., the deformation) is described by the joint probability of all hidden nodes, i.e., $\mathbf{X} = \{x_1, \ldots, x_n\}$. We assume $p(\mathbf{X})$ to be a Gibbs distribution, and thus can be factorized as:

$$p(\mathbf{X}) = \frac{1}{Z_c} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j) \prod_{i \in V} \psi_i(x_i) \tag{1}$$

where ψ_i and ψ_{ij} are the potential functions associated with site $i \in V$ and the link $(i, j) \in E$, and Z_c is a normalization term or the partition function.

To model the prior of local shape deformation, we work with a special case where $x_i \in \{0, 1\}$, thus $p(\mathbf{X})$ becomes a Boltzmann distribution, i.e.,

$$p(\mathbf{X}) = \frac{1}{Z_c} \prod_{(i,j)\in E} e^{\alpha_{ij} x_i x_j} \prod_{i \in V} e^{\beta_i x_i} \tag{2}$$

where $\{\alpha_{ij}, \beta_i : \forall (i, j) \in E, i \in V\}$, are parameters which can be learnt from training data (see Section 5).

In addition, each hidden node x_k is associated with an observation node z_k representing the image observation produced by x_k, which is characterized by the conditional probability $p(z_k|x_k)$. The observation of the scene is the collection of the image observations on each scene site, i.e., $\mathbf{Z} = \{z_1, \ldots, z_n\}$. We have:

$$p(\mathbf{Z}|\mathbf{X}) = \prod_{k=1}^{n} p_k(z_k|x_k). \tag{3}$$

Thus, the model in Figure 1 is fully characterized by $\{\alpha_{ij}, \beta_i, p_i\}$, where $p_i = p_i(z_i|x_i)$, and we denote the model by $\lambda = \{\alpha_{ij}, \beta_i, p_i\}$.

This model is suitable for local deformation, because (a) it models the constraints among neighbor sites rather than treating them independently; (b) the Boltzmann distribution can capture complex distributions which can not be represented by Gaussian or mixture of Gaussian; and (c) the observation model provides clues for Bayesian inference from image data.

A core issue in detection and tracking is the calculation of the image likelihood $p(\mathbf{Z}|\lambda)$. However, this is not a trivial problem though, since it involves the integral of all possible configurations of $\mathbf{X}$, i.e.,

$$p(\mathbf{Z}|\lambda) = \int_{\mathbf{X}\in\mathcal{X}} \prod_{i=1}^{n} p_i(z_i|x_i) p(\mathbf{X}) d\mathbf{X}. \tag{4}$$

The key to solve this problem is to design an effective inference algorithm to estimate the posterior $p(\mathbf{X}|\mathbf{Z}, \lambda)$ and its marginals $p(x_i|\mathbf{Z}, \lambda)$. Without causing any confusion, we usually denote $p(\cdot|\lambda)$ by $p(\cdot)$ for short.

4 Mean Field Approximation

Given the high dimensionality in the graphical model in Section 3, solving $p(\mathbf{Z}|\lambda)$ and $p(\mathbf{X}|\mathbf{Z}, \lambda)$ involves computationally intensive multi-dimensional integral over $p(\mathbf{X}, \mathbf{Z}|\lambda)$. Although the Markovian property of the structure of $p(\mathbf{X}|\lambda)$ simplifies the problem, the exact analysis for such a model is still prohibitive due to the loops in the graphical model.

Thus, approximated but computationally efficient analysis methods are of special interests. *Variational approximation* is one of these methods. The core idea is to employ an analytical and simple variational distribution $Q(\mathbf{X})$ to approximate the posterior probability $p(\mathbf{X}|\mathbf{Z})$, such that the Kullback-Leibler (KL) divergence of these two distributions is minimized.

To see this clearly, we follow Jaakkola & Jordan [6] to formulate an optimization problem to solve $p(\mathbf{Z})$ and $p(\mathbf{X}|\mathbf{Z})$ simultaneously. The objective function is:

$$\begin{aligned} J(Q) &= \log p(\mathbf{Z}) - KL(Q(\mathbf{X})||p(\mathbf{X}|\mathbf{Z})) \\ &= H(Q) + E_Q[\log p(\mathbf{X}, \mathbf{Z})] \end{aligned} \tag{5}$$

where $H(Q)$ is the entropy of $Q(\mathbf{X})$ and $E_Q[\cdot]$ is the expectation w.r.t. $Q(\mathbf{X})$. It is easy to see $J(Q)$ is a lower bound of $\log p(\mathbf{Z})$. By maximizing the lower bound $J(Q)$ w.r.t. Q, we can obtain an optimal approximation of $p(\mathbf{X}|\mathbf{Z})$ by Q^*, and a closest $\log p(\mathbf{Z})$ by $J(Q^*)$.

Choosing the variational distributions $Q(\mathbf{X})$ could be arbitrary, but an appropriate $Q(\mathbf{X})$ would make difference on analyzing. Although substantial creativity can be required to find the appropriate forms for $Q(\mathbf{X})$ [8], we adopt a fully factorized form for simplicity:

$$Q(\mathbf{X}) = \prod_{i}^{n} Q_i(x_i) \tag{6}$$

where $Q_i(x_i)$ is an independent distribution of the hidden node x_i. Then, $H(Q) = \sum_i H(Q_i)$.

Based on the factorization of $Q(\mathbf{X})$, it can be shown that the best variational density is made of a set of interrelated Gibbs distributions:

$$Q_i(x_i) = \frac{1}{Z_i} e^{E_Q[\log p(\mathbf{X},\mathbf{Z})|x_i]}, \quad i = \{1, \ldots, M\} \quad (7)$$

where Z_i is a constant and $E_Q[\log p(\mathbf{X}, \mathbf{Z})|x_i]$ is the conditional expectation given x_i.

Eq. 7 gives a general solution. Moreover, it is easy to show the factorization of $p(\mathbf{X})$ in Eq. 2 enables further simplification, i.e.,

$$Q_i(x_i) \longleftarrow \frac{1}{Z_i'} p_i(z_i|x_i)\psi_i(x_i)M_i(x_i), \quad \text{where}$$

$$M_i(x_i) = \exp\{\sum_{k\in\mathcal{N}(i)} \int_{x_k} Q_k(x_k) \log \psi_{ik}(x_i, x_k)\}, \quad (8)$$

where Z_i' is a constant, and $\mathcal{N}(i)$ is the neighborhood of the site i. The iterative updating of $Q_i(x_i)$ will monotonically increase $J(Q)$ and eventually reach an equilibrium. These updating equations are called *mean field equations*. From Eq. 8, the variational belief of a hidden node x_i is determined by three factors: the local conditional likelihood $p_i(z_i|x_i)$, the local prior $\psi_i(x_i)$, and the neighborhood prior from the constraints of the neighborhood nodes $x_{\mathcal{N}(i)}$.

Thus, we can treat the term $p_i(z_i|x_i)\psi_i(x_i)$ as the local belief of x_i, and treat the term $M_i(x_i)$ as the "message" propagated through the nearby nodes of x_i. This method is actually different from belief propagation [3], due to its use of variational analysis and to the different contents in the "messages". In our method, the computation of $M_i(x_i)$ is easier than belief propagation, due to the factorization in the variational distribution. In addition, we can clearly see from this equation that the computation is significantly reduced by avoiding multi-dimensional integral, since Eq. 8 involves only one dimensional integral.

For deformable shapes, considering $x_i \in \{0, 1\}$, we use:

$$Q(\mathbf{X}) = \prod_i^n \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}, \quad (9)$$

where $\{\mu_i\}$ are variational parameters to be optimized. Under this variational distribution, the mean field equations Eq. 8 can be simplified as:

$$\mu_i = \frac{p_i(z_i|x_i=1)m_i}{p_i(z_i|x_i=0) + p_i(z_i|x_i=1)m_i},$$

$$\text{where} \quad m_i = \exp\{\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mu_j + \beta_i\} \quad (10)$$

Similar results have also been obtained by Jordan et al. [8], Peterson and Anderson [12]. Then we have:

$$\begin{aligned} J(Q) &= \sum_i H(Q_i) + \sum_{(i,j)\in E} \alpha_{ij}\mu_i\mu_j + \sum_{k\in V} \mu_k\beta_k \\ &+ \sum_{k\in V} (1-\mu_k) \log p_k(z_k|x_k=0) \\ &+ \sum_{k\in V} \mu_k \log p_k(z_k|x_k=1) - \log Z_c \quad (11) \end{aligned}$$

We admit that $J(Q)$ can not be fully computed, due to the complexity of calculating $\log Z_c$. But the computation of $\widetilde{J}(Q) = J(Q) + \log Z_c$ is computationally tractable in practice. Fortunately, it is not necessary to calculate $\log Z_c$, because once we find an optimal mean field distribution of Q^*, we readily have:

$$p(\mathbf{Z}) \propto e^{\widetilde{J}(Q^*)},$$

which is enough for other related computing (such as detection and tracking in Section 6).

5 Learning

This section discusses the problem of learning model parameters $\lambda = \{\alpha_{ij}, \beta_i, p_i\}$ from data. The training of $\{\alpha_{ij}, \beta_i\}$ and $\{p_i\}$ can be separated. The initial model is constructed by the following way:

1. collecting a set of labelled (annotated) training examples, $\mathcal{L} = \{\mathbf{X}^k, \mathbf{Z}^k, k = 1, \ldots, K_1\}$. For the application of deformable shapes, $\mathbf{x}_i \in \{0, 1\}$, and $\mathbf{z}_i$ is the average edge direction over a small image patch associated with $\mathbf{x}_i$ in our applications. We quantize $\mathbf{z}_i$, and use histogram to model its distribution. If the target is very small, z_i simply takes binary value to indicate if it is a detected edge point or not.

2. learning $p_i(z_i|x_i)$ for each x_i. Due to the factorization of $p(\mathbf{Z}|\mathbf{X})$, i.e., Eq. 3, each individual $p_i(z_i|x_i)$ can be learned independently. Each $p_i(z_i|x_i)$ is represented by a histogram in our experiments.

3. learning $\{\alpha_{ij}, \beta_i\}$ by the following steps:

   3.a calculating sufficient statistics $S_{ij} = E_p[x_i x_j]$ and $S_i = E_p[x_i]$ from the supervised training data $\{\mathbf{X}^k\}$;

   3.b initialize a model $\lambda_b^0 = \{\alpha_{ij}^0, \beta_i^0\}$;

   3.c collecting synthesized samples of $\{\mathbf{X}_g^k\}$ by Gibbs sampling of $p(\mathbf{X}|\lambda_b)$;

   3.d calculating sufficient statistics $G_{ij} = E_{\lambda_b}[x_i x_j]$ and $G_i = E_{\lambda_b}[x_i]$ from the synthesized data;

3.e adjusting the parameters by:

$$\Delta\alpha_{ij} \propto (G_{ij} - S_{ij}) \quad (12)$$
$$\Delta\beta_i \propto (G_i - S_i) \quad (13)$$

3.f go to step 3.c;

In our experiments, we select:

$$\alpha_{ij}^0 = \log\frac{S_{ij}}{1-S_{ij}}, \quad \text{and} \quad \beta_i^0 = \log\frac{S_i}{1-S_i}$$

as the initialization. We observed the convergence in less than 50 iterations.

Once an initial model is trained, then we finely tune the model by using a large set of unlabelled training examples $\mathcal{U} = \{\mathbf{Z}^k, k = 1, \ldots, K_2\}$ which are cheaply available. The process is an EM iteration:

- **E-step**: $\forall \mathbf{Z}^k \in \mathcal{U}$, infer the posterior $p(x_i^k|\mathbf{Z}^k, \lambda^t)$ based on variational mean field approximation in Eq. 8, i.e., we obtain the set of variational parameters $\{\{\mu_i\}^k\}$, where $k = 1, \ldots, K_2$.
- **M-step**: estimate the model parameters $\lambda^{t+1} = \{\alpha_{ij}^{t+1}, \beta_i^{t+1}, p_i^{t+1}\}$, given a fixed $\{\{\mu_i\}^k\}$ by a stochastic gradient descent:

$$\Delta\alpha_{ij} \propto \frac{\partial J(Q)}{\partial\alpha_{ij}} \approx \mu_i\mu_j - E_Q[x_ix_j] \quad (14)$$
$$\Delta\beta_i \propto \frac{\partial J(Q)}{\partial\beta_i} \approx \mu_i - E_Q[x_i] \quad (15)$$

where $E_Q[x_ix_j]$ and $E_Q[x_i]$ are sufficient statistics calculated w.r.t. the variational distributions. And the method of estimating p_i is the same as the step 2 in the above supervised training.

6 Pedestrian Detection and Tracking

6.1 Detection

Pedestrian detection involves two mean field models: λ_0 represents the negative hypothesis, i.e., no pedestrian presence, and λ_1 for the positive hypothesis, i.e., pedestrian presence. The detection algorithm scans the shape space $\mathcal{Y}$ which accommodates different locations $\mathbf{u}$, orientations θ, and scales s, i.e., $\mathbf{y} = \{\mathbf{u}, \theta, s\} \in \mathcal{Y}$. In our experiment, we scan all locations and 5 scales.

For each $\mathbf{y}$, we collect the edge map of the associated image patch and treat it as the observation $\mathbf{Z} = \mathbf{I}(\mathbf{y})$. We can perform likelihood ratio detection for each given $\mathbf{y}$:

$$\log p(\mathbf{Z}|\mathbf{y}, \lambda_1) - \log p(\mathbf{Z}|\mathbf{y}, \lambda_0) > \tau_o \geq 0. \quad (16)$$

Since it is unrealistic to calculate $p(\mathbf{Z}|\mathbf{y}, \lambda)$ (in Eq. 4), the variational analysis in Section 4 nicely provides a mean field solution as an approximation, i.e.,

$$\log p(\mathbf{Z}|\mathbf{y}, \lambda) \approx J(Q^*(\mathbf{X}|\mathbf{y}, \lambda)),$$

where $Q^*(\mathbf{X}|\mathbf{y}, \lambda)$ is the optimal mean field approximation of the posterior $p(\mathbf{X}|\mathbf{Z}, \mathbf{y}, \lambda)$. Thus, the detection rule for each given $\mathbf{y}$ becomes:

$$\widetilde{J}(Q^*(\mathbf{X}|\mathbf{y}, \lambda_1)) - \widetilde{J}(Q^*(\mathbf{X}|\mathbf{y}, \lambda_0)) > \tau, \quad (17)$$

where $\widetilde{J}(Q^*(\mathbf{X}|\mathbf{y}, \lambda_k)), k = \{0, 1\}$ can be obtained according to Eq. 5 once the mean field iteration stops at $Q^*(\mathbf{X}|\mathbf{y}, \lambda_k)$ according to Eq. 8. There are two factors affecting the threshold τ: (a) $J(Q^*|\lambda_k)$ only provides an optimal lower bound of $\log p(\mathbf{Z}|\lambda_k)$, and (b) we generally only calculate $J(Q^*|\lambda_k)$ up to a constant difference $\log Z_c^k$ (see Eq. 11). Thus, we do not simply set $\tau = 0$, but train this threshold from supervised examples to reduce the rate of false alarm and miss detection.

6.2 Tracking

Different from detection, only the pedestrian model λ_1 is involved in tracking, where the task is to estimate the posterior density of $p(\mathbf{y}_t|\underline{\mathbf{I}}_t, \lambda_1)$, where $\mathbf{y_t} = \{\mathbf{u}_t, \theta_t, s_t\}$ is the same as in the detection problem, and $\underline{\mathbf{I}}_t = \{\mathbf{I}_1, \ldots, \mathbf{I}_t\}$. According to Bayesian rule, we have:

$$p(\mathbf{y}_t|\underline{\mathbf{I}}_t, \lambda_1) \propto p(\mathbf{I}_t|\mathbf{y}_t, \lambda_1) \int_{y_{t-1}} p(\mathbf{y}_t|\mathbf{y}_{t-1}) p(\mathbf{y}_{t-1}|\underline{\mathbf{I}}_{t-1}, \lambda_1). \quad (18)$$

The dynamic process can be represented as a dynamic Bayesian network in Figure 2. Clearly, the hidden factor

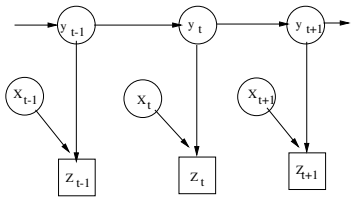

Figure 2: The graphical model representing the dynamic process.

$\mathbf{X}_t$ of local deformation has been integrated out in the observation process, which is powerful for tracking since it leaves less motion parameters to be estimated. It is clear that the visual dynamics is governed by the observation model $p(\mathbf{I}_t|\mathbf{y}_t, \lambda_1)$ and the motion model $(\mathbf{y}_t|\mathbf{y}_{t-1})$ such as a const acceleration model. Since we have:

$$p(\mathbf{I}_t|\mathbf{y}_t, \lambda_1) = p(\mathbf{Z}(\mathbf{y}_t)|\lambda_1) \propto e^{\widetilde{J}(Q^*(\mathbf{X}_t|\mathbf{y}, \lambda_1))},$$

the local deformation has been absorbed in the calculation of data likelihood which is based on the mean field inference. The tracking algorithm is easily implemented using particle filtering [1, 5], where each particle represents a sample of $\mathbf{y}_t$. Detailed results will be reported in Section 7.

7 Experiments

7.1 Training and Model Validation

We trained two models, one for the human λ_1 and the other for the background λ_0. To train λ_1, the training data of various people were collected and their contours were extracted. The we resized and aligned all the contours by compensating the global motions. Using the extracted contours and the corresponding image observations, we obtained a set annotated of $3,000$ training data. All training images are aligned to the center of mass. Some examples are shown in Figure 3(a). Training λ_0 is easier than λ_1, since the alignment step is not needed, and a set of $10,000$ training data were collected randomly from the training sequences to train λ_0. Some of them are shown in Figure 3(b).

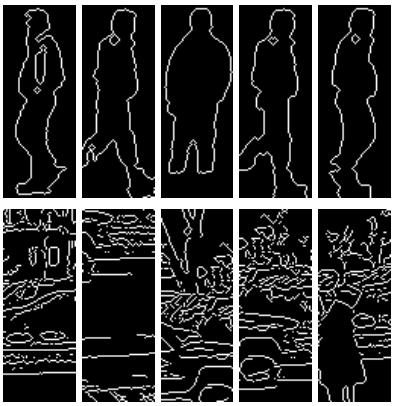

Figure 3: The upper row are examples of training data for human λ_1, and the bottom row for nonhuman λ_0.

It is important to know if the trained Boltzmann model really captures the distribution of $p(\mathbf{X})$. Although there is no quantitative means to validate that, a plausible way for a rough validation is to sample the prior Boltzmann distribution $p(\mathbf{X})$ and then perform a subjective evaluation. To synthesis an image, we first draw a sample of $\mathbf{X} = \{x_1, \ldots, x_n\}$ by Gibbs sampling from $p(\mathbf{X})$ in Eq. 2, then for each x_i, a sample of z_i is drawn from $p_i(z_i|x_i)$. Putting together z_i produces a synthesis image. Through our subjective evaluations, the trained models were able to synthesize reasonably good data. Some synthesized data based on λ_1 and λ_0 are shown in Figure 4.

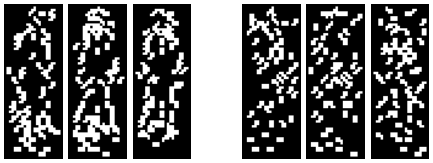

Figure 4: Examples of synthesized data. Left ones are sample from λ_1 and right ones from λ_0.

7.2 Pedestrian Detection

We performed extensive experiments and quantitative evaluation of the proposed approach to pedestrian detection, and we are especially interested in the investigation of the capacity of the mean field model of capturing the tremendous shape variations and its performance and robustness to partial occlusion.

7.2.1 Performance Evaluation

To provide quantitative evaluation of the proposed approach, we constructed a testing database which contains 1000 images collected from various occasions. We manually annotated the ground truth detection for each image. The ROC curve is shown in figure 5, which shows that at 80% detection rate, the detector has a false positive rate of about 1/200,000 which corresponds to about one false alarm per frame for 320×240 images. This is comparable to the most recent method reported in [14].

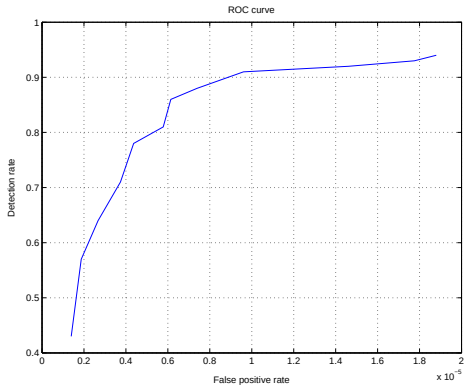

Figure 5: ROC curve of the proposed detector.

The mean field model is capable of capturing the local deformation caused by the view changes of the human. In our test data, there are a large volume of images where the pedestrians present various profiles. Some of the detection examples are shown in Figure 6.

In addition, the model is also able to detect the target from noisy environments. Some of the results are shown in Figure 7. The robustness comes from the observation models of λ_1 and λ_0. We did observe the case where in a region the edge map is pervasive and it is impossible to tell where the person is from the edge map.

7.2.2 Evaluation on Partial Occlusion

More interestingly, the proposed algorithm works even when the target is partially occluded. Samples of our results on detection under occlusion are shown in Figure 8. This feature is unique, since the robustness to partial occlusion is an intrinsic benefit of the proposed distributed shape

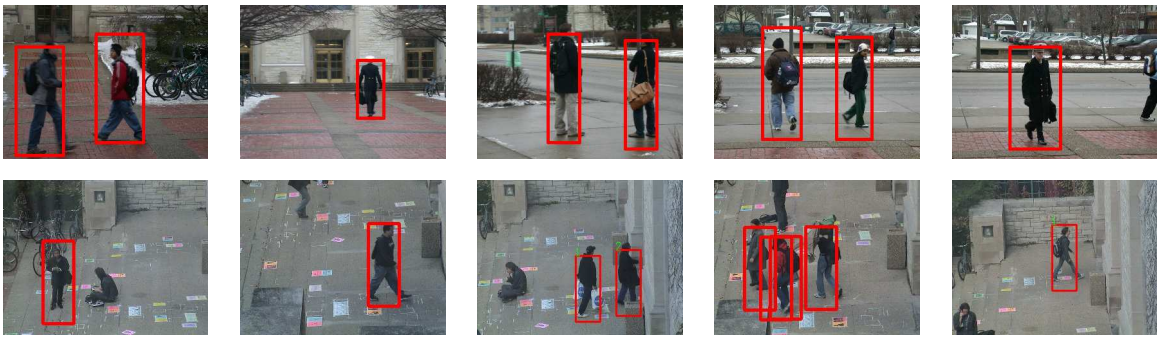

Figure 6: Pedestrian detection under various views.

Figure 7: Pedestrian detection in various backgrounds and noisy environments.

model. This is true because we did deliberately not include the occlusion cases in training. On the other hand, centralized shape models such as PCA can not cope with this problem since it is infeasible to include all possible occlusion situations in training.

To have a quantitative study on the robustness of our method, we created another testing database which consists of 3 subsets, each of which contains 100 images under a certain rough percentage of occlusion (less than 20%, between 20% and 40%, and over 40%, respectively). The ROC curves for these occlusion cases were obtained as shown in figure 9. These ROC curves show that the performance

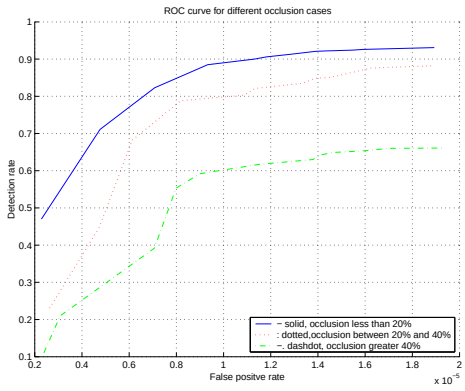

Figure 9: ROC curves for the three occlusion subsets.

of the proposed method does not degrade much when the percentage of occlusion is under 40%, since 80% detection rate can be achieved with comparable false positive rate as the case without occlusion. But when the occlusion is over 40%, the detection rate drops a lot. Although such quantitative measures are rough, they do verify the robustness of the proposed approach to partial occlusion.

7.3 Pedestrian Tracking

Tracking deformable objects is a challenging problem, especially when the camera is not fixed and the target presents large shape deformation, as in the demonstration of this section. Since the mean field approximation also gives the data likelihood (given a global motion) by integrating our all possible local deformation, this is powerful and ideal for tracking deformable targets, as described in Section 6.2. We did extensive experiments and verified this idea. In our experiments, a particle filter was applied to track the targets, i.e., the global motion $\mathbf{y} = \{\mathbf{u}, \theta, s\}$ is estimated. 400 particles were used. Some sample frames are shown below in Figure 10. In this sequence, camera is not fixed, and the pedestrian walks and rotates, and scale changes are also included.

8 Discussion and Conclusions

Characterizing priors for shape deformation is critical for analyzing deformable objects. Global approaches such as PCA prove to be effective to capture global deformation and reveal global correlations. However, global approaches are not suitable for representing local deformation, which is important for many real world applications, such as pedestrian detection and tracking. In this paper, we described a local approach to model local deformation based on a Markov network, where a Boltzmann distribution was employed to capture the complicated prior for local deformation, and

Figure 8: Detection under occlusion.

Figure 10: Pedestrian tracking based on the mean field Boltzmann model.

a variational mean field approximation was presented for computationally efficient inference, likelihood calculation and model training. Based on this model, the detection and tracking problems were also investigated. The success of applying the proposed method to pedestrian detection and tracking showed its effectiveness and applicability.

Aligning training data in the proposed approach is easier than the approach labelling landmark data in [2], but it leaves a problem: how sensitive is the trained model to the alignment errors? In our future work, we will investigate the capacity of the Markov network with Boltzmann prior, i.e., to what extent the model can capture deformations. In addition, better image observation models will be studied to reduce the rate of false alarm.

Acknowledgments

This work was supported in part by National Science Foundation (NSF) Grant IIS-0308222, IIS-0347877, Northwestern startup funds and the Murphy Fellowships.

References

[1] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, London, 1998.

[2] T. F. Cootes, C. J. Taylor, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61:38–59, Jan. 1995.

[3] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *Int'l Journal of Computer Vision*, 40:25–47, 2000.

[4] D. M. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *IEEE Int'l Conf. on Computer Vision*, pages 87–93, Corfu, Greece, Sept. 1999.

[5] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. of European Conf. on Computer Vision*, pages 343–356, Cambridge, UK, 1996.

[6] T. S. Jaakkola. Tutorial on variational approximation methods. MIT AI Lab TR, 2000.

[7] N. Jojic, N. Petrovic, B. Frey, and T. S. Huang. Transformed hidden Markov models: Estimating mixture models and inferring spatial transformations in video sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, June 2000.

[8] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2000.

[9] M. Kass, A. Witkin, and D. Terzopoulos. Snake: Active contour models. In *Int'l Conf. on Computer Vision*, pages 259–268, 1987.

[10] C. Liu, S. C. Zhu, and H.-Y. Shum. Learning inhomogeneous gibbs model of faces by minimax entropy. In *IEEE Int'l Conf. on Computer Vision*, Vancouver,Canada, July 2001.

[11] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 572–578, Greece, 1999.

[12] C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, pages 995–1019, 1987.

[13] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[14] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 734–741, Nice, France, Oct. 2003.

[15] A. Yuille. Deformable templates for face recognition. *J. of Cognitive Neuroscience*, 3, 1991.