A Field Model for Human Detection and Tracking

Ying Wu, *Member*, *IEEE*, and Ting Yu, *Student Member*, *IEEE*

Abstract—The large shape variability and partial occlusions challenge most object detection and tracking methods for nonrigid targets such as pedestrians. This paper presents a new approach based on a two-layer statistical field model that characterizes the prior of the complex shape variations as a Boltzmann distribution and embeds this prior and the complex image likelihood into a Markov field. A probabilistic variational analysis of this model reveals a set of fixed-point equations characterizing the equilibrium of the field. It leads to computationally efficient methods for calculating the image likelihood and for training the model. Based on that, effective algorithms for detecting nonrigid objects are developed. This new approach has several advantages. First, it is intrinsically suitable for capturing local nonrigidity. In addition, due to the distributed likelihood, this approach is robust to partial occlusions. Moreover, the two-layer structure provides large flexibility of modeling the image observations, which makes the new method robust to clutters. Extensive experiments demonstrate its effectiveness.

Index Terms—Object detection, shape, Markov random fields, image models, machine learning, statistical computing, probabilistic algorithms.

1 Introduction

The research into human detection and tracking has received more and more attention in recent years, due to the drive from many emerging applications, such as perceptual interfaces, ubiquitous computing, and smart video surveillance [4], [10], [26], [27]. Different applications are concerned with different image resolutions of the subjects, thus requiring different techniques. For example, in perceptual interfaces, the motions of the human body parts need to be determined for action recognition; thus, these applications require fairly high resolution for analyzing the articulated motion of the body parts. In contrast, in many video surveillance applications, since the human typically is associated with small image regions, the human needs to be treated as a nonrigid entity for detection and tracking, while the detailed motion of the body parts is no longer the major focus here. This paper addresses the detection and tracking problem in the latter context.

A critical issue in object detection and tracking is to calculate the likelihood $p(\mathbf{Z}|\mathbf{y})$ of the image measurements $\mathbf{Z}$ given the rigid motion parameters $\mathbf{y}$ (such as the location, the orientation, and the scale) of the target. If the target presents apparent visual invariants (or features), calculating the image likelihood is straightforward. For example, despite the uncertainty in the visual appearances, frontal faces have a similar image pattern that allows the use of the Harr features for face detection [35]. On the other hand, if apparent invariants are not available, the image likelihood $p(\mathbf{Z}|\mathbf{y})$ needs to be broken into a set of conditionals $p(\mathbf{Z}|\mathbf{y}, \mathbf{X})$ and integrate them, where $\mathbf{X}$, for example, can be the nonrigid motion of the target. If $\mathbf{X}$ is complicated, calculating such an integration can be very difficult, leading to the nontrivial nature of detection and tracking in this scenario. Unfortunately, this is the case when treating the human as a nonrigid entity.

Although the research of object detection has greatly moved forward with the success of face detection, these face detection methods may not apply to human detection. The visual appearances of the human present tremendous variability while lacking apparent visual invariants, as the diversified clothing and the body articulation may significantly change the image of a person. In addition, handling partial occlusion is more concerned in detecting humans because in practice the detector should be robust to the missing body parts, but this challenges most existing methods. Therefore, it is desirable to investigate new human detection methods that cope with the large uncertainty of the visual appearances and are robust to partial occlusions. This paper addresses these two difficulties.

Because the human-like shapes are more or less unique in the real world, they may provide a powerful clue for human detection and tracking [8], [11], [34], [38]. The difficulty of analyzing human shapes lies in the fact that the local shape nonrigidity has a large number of degrees of freedom thus having very complicated uncertainties, which makes rule-based methods unsuitable. Thus, learning-based methods are generally adopted for learning the shape distributions from training data. If the uncertainty is simple, parametric methods such as Gaussian models or Gaussian mixture models can do a good job, e.g., in face shape [5]. Unfortunately, because of the local nonrigidity, the uncertainty in human shapes is too complex to be sufficiently modeled by reasonable Gaussian mixture models. On the other hand, nonparametric methods, e.g., by using exemplars [11], can be quite flexible. However, a huge set of exemplars is needed to represent the concept of the human-like shapes in order to accommodate the possible variations. As a trade-off, because the Gibbs distribution is flexible to capture a large variety of

• *The authors are with the Department of Electrical and Computer Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208. E-mail: {yingwu, tingyu}@ece.northwestern.edu.*

Manuscript received 27 July 2004; revised 16 June 2005; accepted 19 Sept. 2005; published online 13 Mar. 2006.
Recommended for acceptance by M. Srinivasan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0386-0704.

densities, it can naturally be employed for this task. This idea has been exploited for modeling the face deformations [20], where the face is represented as a random vector and an inhomogeneous Gibbs distribution can be learned from training data. Although this model is complicated and needs quite complex training, its excellent performance suggests that the Gibbs distribution is useful for characterizing the large and complex variability of the human-like shapes.

Unfortunately, it is difficult for the above learning-based methods to handle partial occlusions because it is not practical, if not impossible, to have the training data that cover all possible situations of partial occlusions. Actually, this difficulty lies in the fact that these methods represent a pattern as a centralized random feature vector. Thus, the missing elements due to occlusion will greatly change the feature vector, thus affecting the classification dramatically. Different from such vectorized methods, the component-based methods [22], [29] divide the entity into parts and take advantage of the structures or correlations of the parts. Their excellent performances on detecting partial occluded targets suggest the needs beyond the vectorized models.

The contribution of this paper is a new nonvectorized method based on a two-layer field model for detecting and tracking complex targets such as the human. This new method stands out because of its robustness to partial occlusions, which is difficult for the vectorized methods.

Different from most existing methods, this new approach embeds the complicated nonrigid shape prior into a statistical field and distributes the complex image likelihood to the local sites of the field. This new model has two layers. The hidden layer is a hidden Markov field that captures the shape prior. Every node of this Markov field is associated with an observation node describing the conditional likelihood of image observations of this hidden node, thus constituting the observation layer of this field model. The proposed method models the prior of the nonrigid human shapes as a Boltzmann distribution. Although it is a special Gibbs distribution, our method is different from [20] because the proposed method does not characterize the Boltzmann distribution directly in a vector space, but distribute it into the Markov field. This treatment results in quite simple inference and learning algorithms in both theory and practice. Although the structure of this field model is similar to that in [9], the difference is apparent, as our method employs probabilistic variational analysis that leads to rigorous and elegant analytical approximations [15], [17]. Another theoretical benefit is that the image likelihood estimates are lower bounded. This new approach enables effective and efficient detection and tracking algorithms for many nonrigid objects such as pedestrians.

This new approach has a number of advantages over many existing methods. First, since this model employs a field rather than a vector to describe a shape, it can sufficiently capture the local variability of the shape by the local network structure, thus enabling accurate modeling of the complex shape prior. Second, since the model captures shape variability and performs image measurements in a distributed fashion, it is more robust against occlusion than the vector-based global approaches (such as PCA) in which image measurements have to be performed in a centralized fashion, i.e., conditioned on all shape parameters. In addition, having an observation layer leads to more flexibility and robustness for handling cluttered backgrounds. Third, the variational 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 integrates the top-down and bottom-up methodologies for tracking nonrigid objects. The top-down approaches involve evaluating a large number of hypotheses, and the bottom-up approaches require large efforts in grouping and detection. Given the huge variability in the nonrigid human shapes, neither approach would be satisfactory because the number of hypotheses would be tremendously large and grouping a nonrigid object is very difficult. The proposed tracking method is able to balance these two methodologies and to combine the advantages of both: The global variability is handled in a top-down fashion by particle filtering [2], [14], while the local nonrigidity is coped with by a bottom-up approach by directly evaluating the likelihood of image measurements.

The paper is organized as follows: After a brief description of the related work in Section 2, this paper presents the two-layer field model in Section 3. The probabilistic variational analysis of this model is given in Section 4, and the learning algorithm is presented in Section 5. Section 6 describes our methods for pedestrian detection and tracking, and our extensive results are reported in Section 7. The paper concludes in Section 8.

2 Related Work

In the past few decades, many methods have been proposed for object detection, mainly for the human face and cars. They are based on different classification schemes. Neural network have been employed for detecting faces [31] by classifying the candidate image patches into face or nonface classes. A learned histogram model for wavelet coefficients can be used for face/car detection [32] because the histograms approximate the distributions of object features for discrimination. In addition, combined with the Harr features, the AdaBoost classifier has been very successful for frontal face detection [35] and has been extended for pedestrian detection [36] with the help of motion information. Support vector machines are also widely used in the detection tasks [7], [23], [25].

But, most existing approaches seem not to be suitable for detecting targets with large shape variations, such as the pedestrian. For example, methods based on raw pixel features, such as [24], [31], cannot handle large variability in the appearance of pedestrians. It turns out that the shape features of these deformable targets need to be used. It is suggested that both local and global cues should be combined for such a challenging task [19].

The research on nonrigid shape analysis has a long history, and various approaches have been proposed and investigated. For all these methods, three important common issues should be addressed, i.e., the shape representation $\mathbf{X}$, the shape prior $p(\mathbf{X})$, and the conditional

likelihood of image observation $p(\mathbf{Z}|\mathbf{X})$. (Here, the rigid motion $\mathbf{y}$ is dropped for clarity.)

Different shape representations can be categorized into either parametric or nonparametric models. Examples of parametric representations includes Fourier descriptors, B splines [2], [18], the deformable template [39], etc., where shape deformation is controlled by the shape parameters and smoothness constraints. A typical nonparametric representation is the point distribution model [5] where a shape is described by an ordered and labeled 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 more radical approach is to use a 2D mask [11], [16], [34], where the shape deforms when multiplied by a sparse permutation matrix [16], or by selecting different exemplars [11], [34]. 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. These vectorized models are global since the image observations are conditioned on all the shape parameters. Thus, these global methods are generally not likely to be able to cope with partial occlusions, unless the training data have incorporated all possible occlusion situations. In this paper, rather than using a global representation, a field representation is proposed, with which the complex variability of the nonrigid shape and the occlusion difficulties can be handled easily.

Obviously, in reality, a shape cannot be allowed to deform arbitrarily; thus, the allowable shape space needs to be characterized by having a shape deformation prior model $p(\mathbf{X})$. A possible approach is to reduce the correlations among different shape parameters and model the variance of deformation by a multivariate Gaussian distribution in a lower-dimensional subspace. This is the spirit of the principal component analysis (PCA) and has been widely adopted for learning deformation priors [2], [5]. Since PCA identifies a linear subspace and catches linear correlations, it is powerful to capture and decorrelate certain global deformations, but insufficient for the local nonrigidity. Thus, it motivated the methods that use mixture distributions [16] or exemplar databases [11], [34]. Although mixture distributions can represent arbitrarily complicated densities in theory, it becomes unrealistic when the number of mixtures increases tremendously. To alleviate this difficulty, an inhomogeneous Gibbs model has been proposed and applied successfully to face deformation [20], although the model needs expensive training. The approach proposed in this paper also models the shape deformation prior, but instead of modeling the prior in a global fashion, our approach is based on the field representation and the prior, i.e., a Boltzmann distribution (a specific Gibbs distribution), is distributed into a Markov field, and a variational analysis is employed for analytical results (details in Sections 3 and 4). The new approach stands out from the existing methods by this new representation.

Different approaches have been investigated to *fit* a shape model to image observations. This can be done through minimizing an energy function [18], or based on the Bayesian framework where it is important to characterize the conditional likelihood of image observation $p(\mathbf{Z}|\mathbf{X})$.

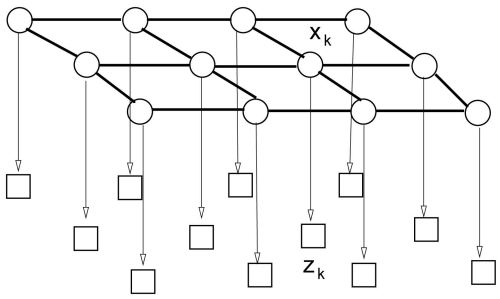

Fig. 1. A two-layer field representation for nonrigid objects.

Analytical forms can be obtained by assuming the independence among a set of discrete points on shape contours [2], [21]. To bypass the independence assumption which may be invalid in reality, the conditional likelihood can be modeled as a metric exponential density obtained from the Chamfer distance based on exemplars [34]. When separating global motion from local nonrigidity, the likelihood conditioned on only global motion can be obtained by the mixture (integral) of all exemplar components in the metric mixture model [34]. The proposed approach in this paper also provides tractable ways to calculate the likelihood only conditioned on global motion, but the differences from [34] are: 1) In our model, $p(\mathbf{Z}|\mathbf{X})$ factorizes by independent components and 2) $p(\mathbf{Z})$ 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 $p(\mathbf{Z})$.

There have been many excellent works on nonrigid shape matching [1], [3], [6], [30], [33]. These methods are more concerned with the matching of extracted shapes for shape registration, where the nonrigid motion needs to be explicitly estimated, while this paper is more concerned with integrating out the variability of the nonrigid motion. In addition, since this paper is based on field model, it is also related to Markov random fields (MRF) that have been widely used for image restoration [13], texture analysis [40], surface reconstruction [12], etc. This paper extends MRF to a two-layer model that consists of a random field and an image observation layer. It is more like the Markov network [9], [37] and the detailed differences will be presented in later sections. More importantly, this paper deals with the nonrigid target detection and tracking problem that has not been addressed by the above methods.

3 The Field Representation

Global methods such as PCA are suitable for capturing the global deformation with a set of uncorrelated deformation bases. But, they tend to ignore the detailed local variations induced by the nonrigidity and they are generally vulnerable to partial occlusion. Therefore, it is desirable to have a model that can handle the large number of degrees of freedom of the local nonrigidity and is robust to occlusion. In this paper, a two-layer field model is proposed as the representation. This is not a vectorized and centralized model but a field and distributed model, as shown in Fig. 1.

This field model consists of two layers. The hidden layer is a hidden random field that represents the shape, modeled as an undirected graph $G_x = \{V, E\}$, where each vertex (or node, or site) represents the hidden shape scene x_k to be

inferred. In this model, x_k takes binary values, i.e., $x_k \in \{0,1\}$, where $x_k = 1$ means that node k is on the object's contour. Each hidden node is connected to its neighborhood nodes $\mathcal{N}(k)$, thus forming a field.

This hidden random field captures the priors of the shape and needs to be inferred from image observations. The feasible shape changes are described by the joint probability of all hidden nodes, i.e., $\mathbf{X} = \{x_1, \ldots, x_n\}$. By assuming $p(\mathbf{X})$ to be a Gibbs distribution, it can be embedded in the random field and be equivalently 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. Specifically, because x_i is binary in our setting for modeling the shape, $p(\mathbf{X})$ 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 learned from training data (see Section 5).

The other layer is the observation layer, through which the shape is associated with its image measurements. As shown in Fig. 1, 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 shape is the collection of the image observations for all sites, i.e., $\mathbf{Z}(\mathbf{y}) = \{z_1, \ldots, z_n\}$, where $\mathbf{y}$ is the global motion. This is a distributed likelihood model. Without causing confusion, $\mathbf{Z}(\mathbf{y})$ is denoted as $\mathbf{Z}$ for short in later sections. We have:

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

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

This two-layer field model is suitable for describing local nonrigidity and is robust to occlusion because of the following reasons: 1) Since the neighborhood sites of the shape are generally correlated, this model captures the correlations and constraints among neighboring sites rather than simply treating them independently, thus resulting in more accurate modeling. 2) The Boltzmann distribution can capture complex distributions which cannot be represented by Gaussian or mixture of Gaussian, thus providing more powerful priors. 3) Because the observation model $p(\mathbf{Z}|\mathbf{X})$ is distributed over the field (i.e., the shape), wrong estimates on some part of the field may not ruin the other parts of the shape, thus leading to the robustness to partial occlusion.

Within this model, two key problems need to be solved:

1. *Calculating the likelihood* $p(\mathbf{Z}|\lambda)$. This is not a trivial problem, 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 that estimates the posterior $p(\mathbf{X}|\mathbf{Z}, \lambda)$ and its marginals $p(x_i|\mathbf{Z}, \lambda)$.
2. *Learning model parameters* λ. These parameters need to be estimated from training data. Without causing any confusion, we usually denote $p(\cdot|\lambda)$ by $p(\cdot)$ for short.

The learning problem is closely related to the likelihood problem because the solution to the learning problem relies on the inference of the model (i.e., the estimation of $p(\mathbf{X}|\mathbf{Z}, \lambda)$). Therefore, this paper presents an analytical approximation to the likelihood in Section 4, and the solution to the learning problem in Section 5.

4 Variational Inference

The field model introduced in Section 3 is a high-dimensional stochastic system because it consists of a large number of random variables (or nodes). Thus, solving the observation likelihood $p(\mathbf{Z}|\lambda)$ and the posterior $p(\mathbf{X}|\mathbf{Z}, \lambda)$ involves computationally intensive multidimensional 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 loopy structures of this field model.

Thus, approximated but computationally efficient analysis methods are of special interest. *Probabilistic variational approximation* is one of these methods. Here, the general approach of the variational analysis for the field model is given in Section 4.1, and the deduced Boltzmann field for nonrigid shapes is presented in Section 4.2.

4.1 Probabilistic Variational Analysis

The core idea of probabilistic variational analysis is to find an analytical and simple variational distribution $Q(\mathbf{X})$ from a variational family to approximate the complicated 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 [15] and formulate an optimization problem to solve $p(\mathbf{Z})$ and $p(\mathbf{X}|\mathbf{Z})$ simultaneously. An objective function can be written as:

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

where $H(Q)$ is the entropy of $Q(\mathbf{X})$ and $E_Q[\cdot]$ denotes the expectation with regard to $Q(\mathbf{X})$. It is easy to see that $\log p(\mathbf{Z})$ is lower bounded by $J(Q)$ since the KL divergence is nonnegative. By maximizing the lower bound $J(Q)$ with regard to Q, an optimal approximation of $p(\mathbf{X}|\mathbf{Z})$ can be obtained by Q^*, and a closest value of $\log p(\mathbf{Z})$ by $J(Q^*)$.

The spirit of this variational approach is to find the best approximation of $p(\mathbf{X}|\mathbf{Z})$ within a given variational family $Q(\mathbf{X})$. When such a variational family has good analytical properties, such as having independent components, or

sparse correlations, or factorized forms, analytical approximation can generally be expected. Although the selection of the variational family $Q(\mathbf{X})$ can be arbitrary, an appropriate $Q(X)$ will make a big difference on analysis. Here, a fully factorized form is adopted:

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

where $Q_i(x_i)$ is an independent distribution of the hidden node x_i. Then, the entropy of the variational distribution can be written as:

$$H(Q) = \sum_i H(Q_i).$$

Such a fully factorized variation leads to the mean field approximation. To see this clearly, we minimize the KL divergence with respect to $Q(\mathbf{X})$. It can be easily shown (see the Appendix) that the optimal approximation is made of a set of interrelated Gibbs distributions:

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

where Z_i is a normalization constant, and $E_Q[\log p(\mathbf{X}, \mathbf{Z})|x_i]$ is the conditional expectation given x_i. The set of equations in (7) are fixed point equations. The iterative updating of $Q_i(x_i)$ will monotonically increase $J(Q)$ and eventually reach an equilibrium. These equations can be called as *mean field equations*.

Equation (7) gives a general solution with a very general form of $Q(\mathbf{X})$. Furthermore, when taking advantage of the special factorization property of $p(\mathbf{X})$ in (2) and $p(\mathbf{Z}|\mathbf{X})$ in (3), a further simplification can be easily obtained. Given the structure of this field model, it is easy to shown that:

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

where

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

where Z_i' is a normalization constant, and $\mathcal{N}(i)$ is the neighborhood of the site i. The iterative updating of $Q_i(x_i)$ based on these mean field equations will monotonically increase $J(Q)$ as well and eventually reach an equilibrium. From (8), it is interesting to notice that 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 $M_i(x_i)$ from the constraints of the neighborhood nodes $x_{\mathcal{N}(i)}$. This can be treated as a generalized Bayesian rule for the field model.

Thus, the term $p_i(z_i|x_i)\psi_i(x_i)$ can be treated 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 the belief propagation algorithm [9], 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 because of the factorization in the variational distribution. In addition, it is clear from this equation that the computation is significantly reduced by avoiding multidimensional integrals, noticing (8) involves only one-dimensional integral.

4.2 Boltzmann Field

The derivation described above was only based on the factorization properties of $Q(\mathbf{X})$, $p(\mathbf{X})$, and $p(\mathbf{Z}|\mathbf{X})$. Thus, the result is quite general. When using the field model for nonrigid shapes, since x_i are binary random variables (i.e., $x_i \in \{0, 1\}$), a Boltzmann distribution for $p(\mathbf{X})$ can be employed, as introduced in Section 3 and (2). Since x_i is binary, we can choose a specific variational distribution here:

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

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

$$\mu_i = \frac{p_i(z_i|x_i=1)m_i}{p_i(z_i|x_i=0) + p_i(z_i|x_i=1)m_i}, \tag{10}$$

where

$$m_i = \exp\left\{ \sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mu_j + \beta_i \right\}.$$

It can be called a *Boltzmann field*. This set of mean field equations in (10) are much simpler than (8) since they only involve a finite set of variational parameters, rather than a set of Gibbs distributions. As a result, the computation is quite straightforward. Similar results have also been obtained by Jordan et al. [17] and Peterson and Anderson [28].

Then, based on this particular variational setting and the result above, (5) becomes:

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

We admit that $J(Q)$ cannot be fully computed, because of the complexity of calculating $\log Z_c$. Instead, it is simple to compute $\widetilde{J}(Q) = J(Q) + \log Z_c$ in practice. Fortunately, it is not necessary to calculate $\log Z_c$, because once an optimal mean field Boltzmann approximation (denoted by Q^*) can be found based on (9) and (10), we readily have:

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

which is enough for our application of 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 training data. The solution of this model-learning problem is in the expectation-maximization (EM) framework, where the core of the

expectation step is the inference of the hidden field described in Section 4. In our method, the training of $\{\alpha_{ij}, \beta_i\}$ and $\{p_i\}$ can be separated. Considering the difficulty of collecting the training data with the known hidden variables (i.e., the annotated training data), this paper proposes a method of using both annotated and un-annotated training data in a semisupervised fashion. The proposed learning method is based on the Gibbs sampling technique and the Expectation-Maximization iterations.

The initial model is constructed by the following way:

1. Collecting a set of annotated training examples, $\mathcal{L} = \{\mathbf{X}^k, \mathbf{Z}^k\}_{k=1}^{K_1}$, where $\mathbf{X}^k$ and $\mathbf{Z}^k$ denote the kth annotated training sample. For each sample, the ith hidden node of the field model takes binary values $x_i \in \{0, 1\}$, and the observation of this hidden node z_i is the average edge direction over a small image patch associated with x_i in our nonrigid shape applications. z_i is quantized and its distribution is modeled as a histogram. If the target shape 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., (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:
   a. calculating sufficient statistics $S_{ij} = E_p[x_i x_j]$ and $S_i = E_p[x_i]$ from the annotated training data $\{\mathbf{X}^k\}_{k=1}^{K_1}$;
   b. initialize a model $\lambda_b^0 = \{\alpha_{ij}^0, \beta_i^0\}$;
   c. producing synthesized samples of $\{\mathbf{X}_g^k\}_{k=1}^{N}$ by Gibbs sampling of $p(\mathbf{X}|\lambda_b)$;
   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 $\{\mathbf{X}_g^k\}_{k=1}^{N}$;
   e. adjusting the parameters by:

$$\Delta\alpha_{ij} \propto (G_{ij} - S_{ij}), \tag{12}$$

$$\Delta\beta_i \propto (G_i - S_i); \tag{13}$$

   f. go to Step 3c.

In our experiments, we select

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

as the initialization. This can be explained by noticing the fact that both sufficient statistics $S_{ij} = E_p[x_i x_j]$ and $S_i = E_p[x_i]$ are in $[0, 1]$. We choose logistic functions: $S_{ij} = \frac{1}{1+e^{-\alpha_{ij}}}$ and $S_i = \frac{1}{1+e^{-\beta_i}}$. This leads to the above initial guess for training α_{ij} and β_i. In all of our experiments, we observed the convergence in less than 50 iterations.

Once the model is initialized, the model can be finely tuned by using a large set of un-annotated training examples $\mathcal{U} = \{\mathbf{Z}^k\}_{k=1}^{K_2}$ 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 (8), i.e., the set of variational parameters $\{\{\mu_i\}^k\}_{k=1}^{K_2}$ is obtained.
- **M-step**: estimate the model parameters $\lambda^{t+1} = \{\alpha_{ij}^{t+1}, \beta_i^{t+1}, p_i^{t+1}\}$, given a fixed $\{\{\mu_i\}^k\}_{k=1}^{K_2}$ by a stochastic gradient descent:

$$\Delta\alpha_{ij} \propto \frac{\partial J(Q)}{\partial \alpha_{ij}} \approx \mu_i\mu_j - E_Q[x_i x_j], \tag{14}$$

$$\Delta\beta_i \propto \frac{\partial J(Q)}{\partial \beta_i} \approx \mu_i - E_Q[x_i], \tag{15}$$

  where $E_Q[x_i x_j]$ and $E_Q[x_i]$ are sufficient statistics calculated with regard to the variational distributions. The method of estimating p_i is the same as in Step 2 in the above supervised training.

6 Pedestrian Detection and Tracking

A suitable representation for the nonrigidity of a pedestrian is critical for detection and tracking. In this section, we approach these tasks using the proposed field model.

6.1 Pedestrian Detection

Pedestrian detection involves two mean field models: λ_0 corresponds to the negative hypothesis H_0, i.e., no pedestrian presence, and λ_1 to the positive hypothesis H_1, i.e., pedestrian presence. The detection algorithm scans different extrinsic shape poses, including locations $\mathbf{u}$, orientations θ, and scales s, denoted by $\mathbf{y} = \{\mathbf{u}, \theta, s\}$. For different scales, we keep the dimension and the number of hidden nodes of the field model the same, but use different sizes of image patches for the observation nodes. In our experiment, we scan all image locations and over five scales.

For each extrinsic shape pose $\mathbf{y}$, we collect the edge map of the corresponding image patch and treat it as the image observation $\mathbf{Z} = \mathbf{I}(\mathbf{y})$ of the hidden Markov field. Likelihood ratio detection is performed on each given $\mathbf{y}$ to determine the pedestrian presence on this particular $\mathbf{y}$:

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

Since it is unrealistic to calculate $p(\mathbf{Z}|\mathbf{y}, \lambda)$ (in (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:

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

where $\widetilde{J}(Q^*(\mathbf{X}|\mathbf{y}, \lambda_k)), k = \{0, 1\}$ can be obtained according to (5) once the mean field iteration converges at $Q^*(\mathbf{X}|\mathbf{y}, \lambda_k)$ according to (8).

There are two factors affecting the threshold τ: 1) $J(Q^*|\lambda_k)$ only provides an optimal lower bound of $\log p(\mathbf{Z}|\lambda_k)$ and 2) we generally only calculate $J(Q^*|\lambda_k)$ up to a constant difference, i.e., $\log Z_c^k$ (see (11)). Thus, we do not simply set $\tau = 0$, but train this threshold from supervised examples to reduce the rate of false alarm and missed detection.

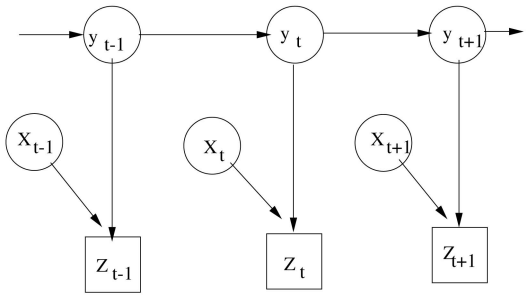

Fig. 2. The dynamic process for tracking a nonrigid target.

6.2 Pedestrian 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:

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

The dynamic process can be represented as a dynamic Bayesian network in Fig. 2.

Clearly, the hidden factor $\mathbf{X}_t$ of local nonrigidity has been integrated out in the observation process. This is powerful for tracking since it leaves no extra motion parameters to be estimated.

It is clear that the visual dynamics is governed by the dynamics model $p(\mathbf{y}_t|\mathbf{y}_{t-1})$ and the observation model $p(\mathbf{I}_t|\mathbf{y}_t, \lambda_1)$. 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 nonrigidity has been absorbed in the calculation of data likelihood which is based on the mean field inference. In our experiments, the dynamics model is characterized as a constant acceleration model and the parameters are learned from an annotated training sequence. Once the MAP solution

$$\mathbf{y}_t^* = \arg\max p(\mathbf{y}_t|\underline{\mathbf{I}}_t, \lambda_1)$$

is obtained, the local nonrigidity can be revealed by

$$p(\mathbf{X}_t|\mathbf{I}_t, \mathbf{y}_t^*, \lambda_1) \approx Q^*(\mathbf{X}_t|\mathbf{y}_t^*, \lambda_1).$$

Because the image likelihood $p(\mathbf{I}_t|\mathbf{y}_t, \lambda_1)$ can be calculated, the tracking algorithm can be easily implemented using particle filtering [2], [14], where each particle represents a sample of $\mathbf{y}_t$. Detailed results will be reported in Section 7.

7 Experiments

In order to validate the proposed approach and demonstrate the applicability of this field model, we performed experiments on pedestrian detection and tracking and compared the proposed detection method with the AdaBoost detector.

7.1 Training and Model Validation

Two models were trained, one for the human λ_1 and the other for the background λ_0. In our experiments, the size of the field was set to 12×6 and each of the node covers an image patch, whose size depends on scale. The experiments used 16×16 patches for the finest scale, and coarser scales correspond to smaller patches. Five scales were used, where the coarsest scale takes 5×5 patches. Neighborhood image patches overlap.

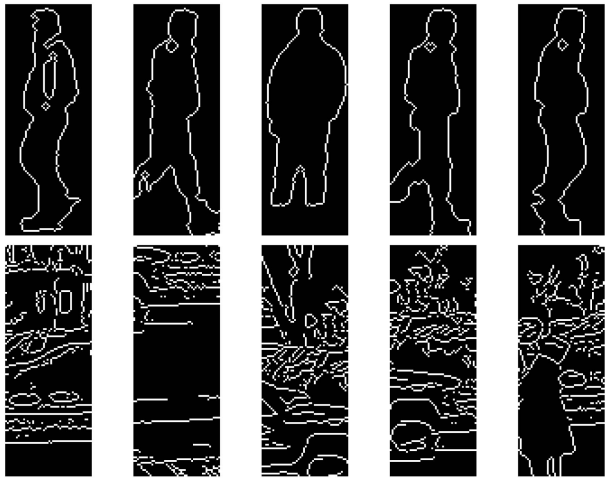

Fig. 3. The upper row are examples of annotated training data for human λ_1 and the bottom row for nonhuman λ_0.

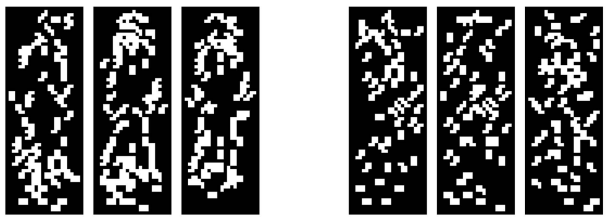

Fig. 4. Examples of synthesized data. The left ones are samples from λ_1 and the right ones from λ_0.

To train λ_1, the training data of various people were collected and their contours were extracted. Then, we resized and aligned all the contours by compensating for their extrinsic poses. Using the extracted contours and the corresponding image observations, we obtained a set of 3,000 annotated training data. All training images are aligned to the center of mass. Some examples are shown in upper row of Fig. 3. 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 the bottom row of Fig. 3. In addition to these annotated training data, 30,000 unannotated data to tune the model were also used, based on the method described in Section 5.

It is important to know if the trained Boltzmann field model really captures the true shape prior $p(\mathbf{X})$. Although there is no quantitative means to validate that, a plausible way for a rough validation is to sample the learned prior Boltzmann distribution $p(\mathbf{X})$ and the learned image likelihood distribution $p(\mathbf{Z}|X)$, and then perform a subjective evaluation. To synthesize an image, a sample of $\mathbf{X} = \{x_1, \ldots, x_n\}$ is first drawn by Gibbs sampling from $p(\mathbf{X})$ in (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 synthesized 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 Fig. 4.

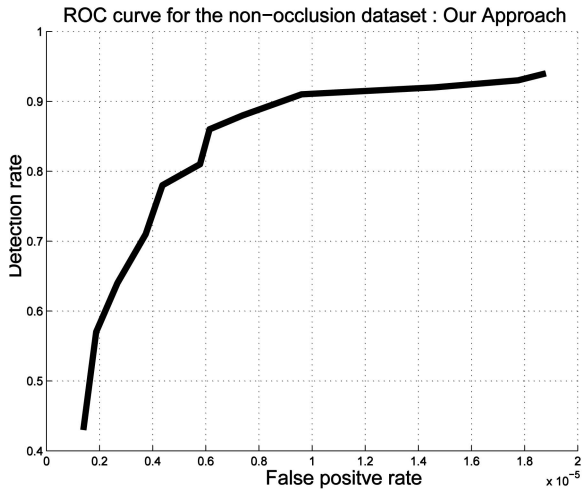

Fig. 5. ROC curve of the proposed pedestrian detector.

7.2 Pedestrian Detection

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

7.2.1 Performance Evaluation

To provide quantitative evaluation of the proposed approach, we constructed a testing database which contains 1,000 images collected from various occasions. We manually annotated the ground truth detection for each image. The ROC curve is shown in Fig. 5. This curve shows that at 80 percent 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 [36].

Our extensive experiments show that the proposed field model is capable of capturing the nonrigidity caused by the view changes of the pedestrian. In our test data, there is a large volume of images where the pedestrians present various profiles. Some of the detection examples are shown in Fig. 6. The algorithm can also easily detect multiple targets. In the bottom right image, a false alarm was observed. In these results, the algorithm did not detect the sitting persons and the cyclist. This is reasonable, because the upper body of these examples are largely inclined and our training set did not contain such cases.

In addition, this field model is also able to detect the target from various environments. Some of the results are shown in Fig. 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 from the edge map where the person is.

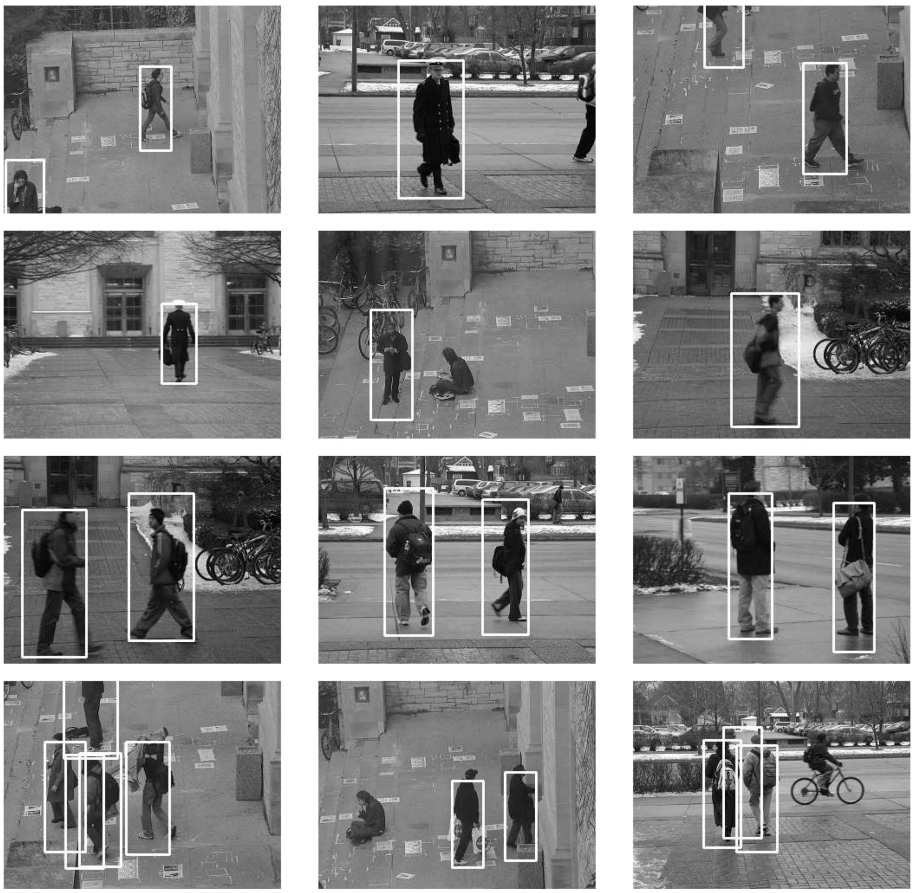

Fig. 6. Pedestrian detection under various views.

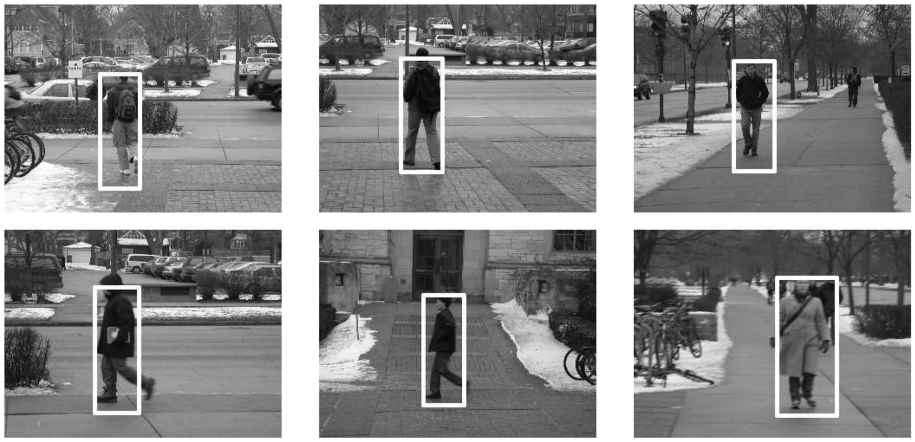

Fig. 7. Pedestrian detection in various environments.

Besides detection, a question of great interest is to reveal the value of hidden field. For each detected region, when displaying the corresponding mean field $\{\mu_i\}$, a clear pedestrian contour can be seen. Some examples are shown in Fig. 8.

In addition, the proposed detection algorithm is efficient. Currently, our unoptimized C++ implementation runs at about two frames/second on a Pentium IV 2GHz PC for 320×240 images. We believe there is much room for improving the implementation. Beyond most existing methods, the proposed field model enables parallel computing, since the mean field updating on the set of sites is intrinsically parallel. In addition, when building real systems, the proposed detection method can be easily combined with background subtraction, motion detection, or other remedies to further reduce the false alarm rate while not decreasing the detection rate. (We did not perform these postprocessing in our experiments, in order to provide a true ROC of the new model.)

7.2.2 Evaluation on Partial Occlusion

More interestingly, the proposed field model works well even when the target is partially occluded. Sample results on the detection under occlusion are shown in Fig. 9. This feature is unique, since the robustness to partial occlusion is an intrinsic benefit of the proposed field model. This is truly owing to the properties of the field model because it is not necessary for the proposed method to deliberately include the occlusion cases in training data. On the contrary, vectorized shape models, such as active shape models [5], cannot cope with this problem since it is generally infeasible to include all possible occlusion situations in training.

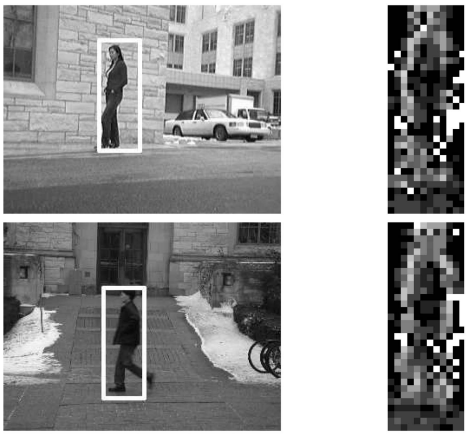

Fig. 8. The mean field inference of the hidden Markov field. The right column shows the estimated mean field $\{\mu_i\}$ of the detected regions on the left column.

To perform a quantitative study on the robustness of our method, another test database was created, consisting of three subsets, each of which contains 100 images under a certain rough percentage of occlusion (less than 20 percent, between 20 percent and 40 percent, and over 40 percent, respectively). The ROC curves for these occlusion cases were obtained and are shown in Fig. 10.

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

7.2.3 Comparison with the AdaBoost Detector

We compare the performance of the proposed method with the AdaBoost detector [35], which is by far one of the best for face detection and is widely used for various object detection tasks. To have a comprehensive comparison, six different data sets were used:

- Data set A is a set of 1,000 images including both nonocclusion and occlusion cases.
- Data set B is a set of 1,000 images of nonocclusion cases only.
- Data set C is a set of 300 images of various occlusion cases.
- Data set D is a set of 100 images, each of which presents over 40 percent occlusion.
- Data set E is a set of 100 images, each of which presents 20 percent to 40 percent occlusion.

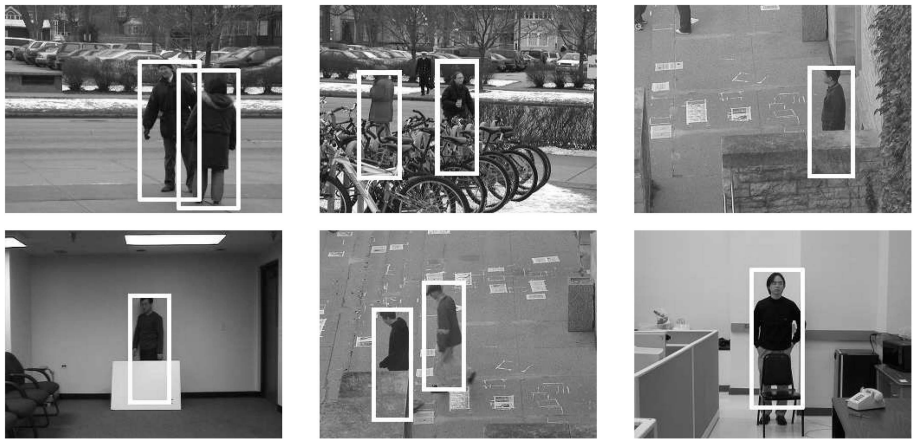

Fig. 9. Sample results of pedestrian detection under partial occlusions.

- Data set F is a set of 100 images, each of which presents less than 20 percent occlusion.

The ROCs on these data sets are shown in Fig. 11 and Fig. 12.

Fig. 11a shows the two ROCs on data set A that contains a mixture of nonocclusion and occlusion cases, and Fig. 11b on data set B of all nonocclusion cases. These ROCs show that our method has overall 5-10 percent higher detection rate than the AdaBoost detector. With high false alarm rates, both methods have high detection rates (over 90 percent).

Fig. 12a shows the ROCs on data set C, and gives the comparison of the two methods on general occlusions. It is apparent from this figure that our method significantly outperforms the AdaBoost detector. With high false alarm rates, our method can obtain over 80 percent detection while Adaboost is merely 70 percent. Figs. 12b, 12c, and 12d show the ROCs on various degrees of occlusions. If the target present with over 40 percent occlusion, AdaBoost detector hardly works, while our method has over 60 percent detection. When the target has a moderate occlusion (between 20 and 40 percent), our method also significantly outperforms AdaBoost. When the occlusion is less than 20 percent, the two methods are comparable, but our method is slightly better.

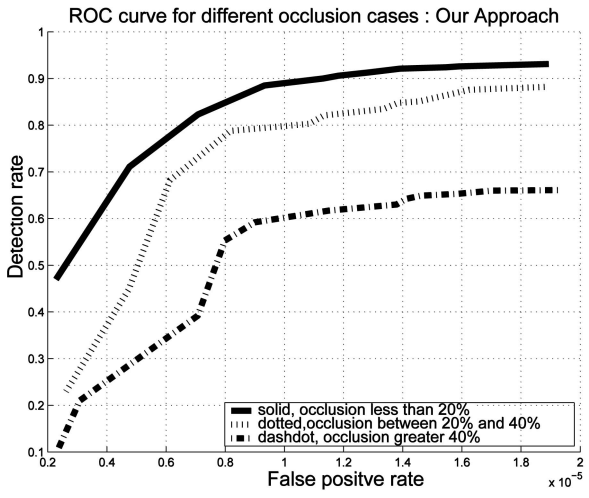

Fig. 10. ROC curves on the three testing subsets under different occlusion percentages.

7.3 Pedestrian Tracking

Tracking nonrigid objects is a challenging problem, especially when the camera is not fixed and the target presents large shape variances, as in the demonstration of this section. Since the mean field approximation also gives the data likelihood (given a global motion) by integrating out all possible local nonrigidity, this is powerful and ideal for tracking nonrigid 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., to estimate the extrinsic pose parameters $\mathbf{y} = \{\mathbf{u}, \theta, s\}$ through the video. Four hundred particles were used in the experiments.

Some sample frames are shown below in Fig. 13. Actually, this is a difficult sequence for many tracking schemes. One difficulty is that the camera is not static and tracking methods based on background subtraction cannot be applied. In addition, when the pedestrian walks and rotates, the visual appearances change dramatically and present nonstationary characteristics, which is a very difficult problem for visual tracking in general. This example shows the effectiveness of the proposed field model. Our method can handle such a difficult scenario because the image likelihoods have integrated all the shape deformations. The C++ implementation of the proposed tracking algorithm runs at over 15 frames/second on a Pentium IV 2GHz PC.

8 Discussion and Conclusions

Characterizing priors of nonrigid shapes is critical for analyzing nonrigid objects. Global or vectorized approaches such as PCA prove to be effective in capturing global deformation by reducing global correlations. However, these vectorized models are neither suitable for handling local nonrigidity nor robust to partial occlusion, which are both important for many real-world applications such as pedestrian detection and tracking. This paper proposed a new statistical method to capture the local nonrigidity based on a two-layer field model, where a Boltzmann distribution was

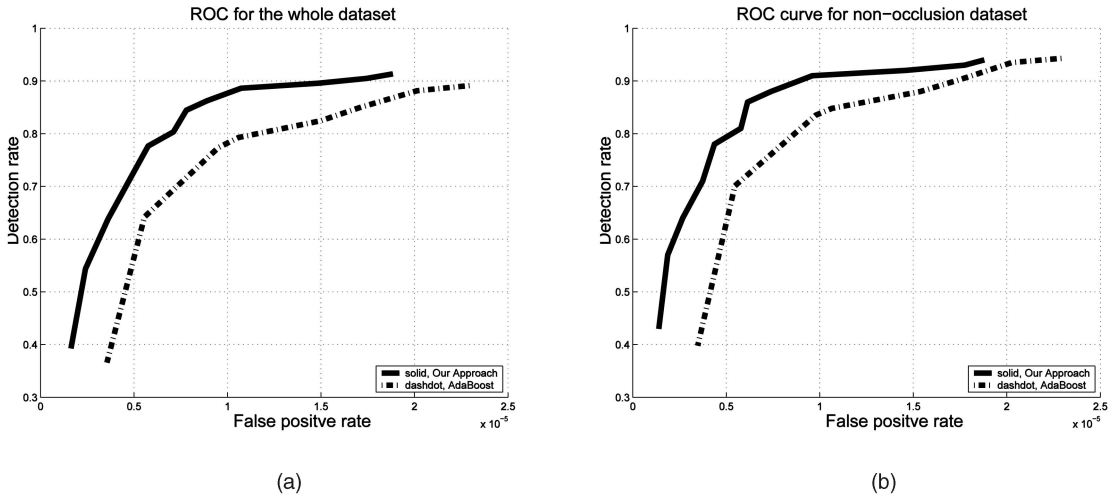

Fig. 11. ROC curves on Data set A and B. (a) ROC on Data set A. (b) ROC on Data set B.

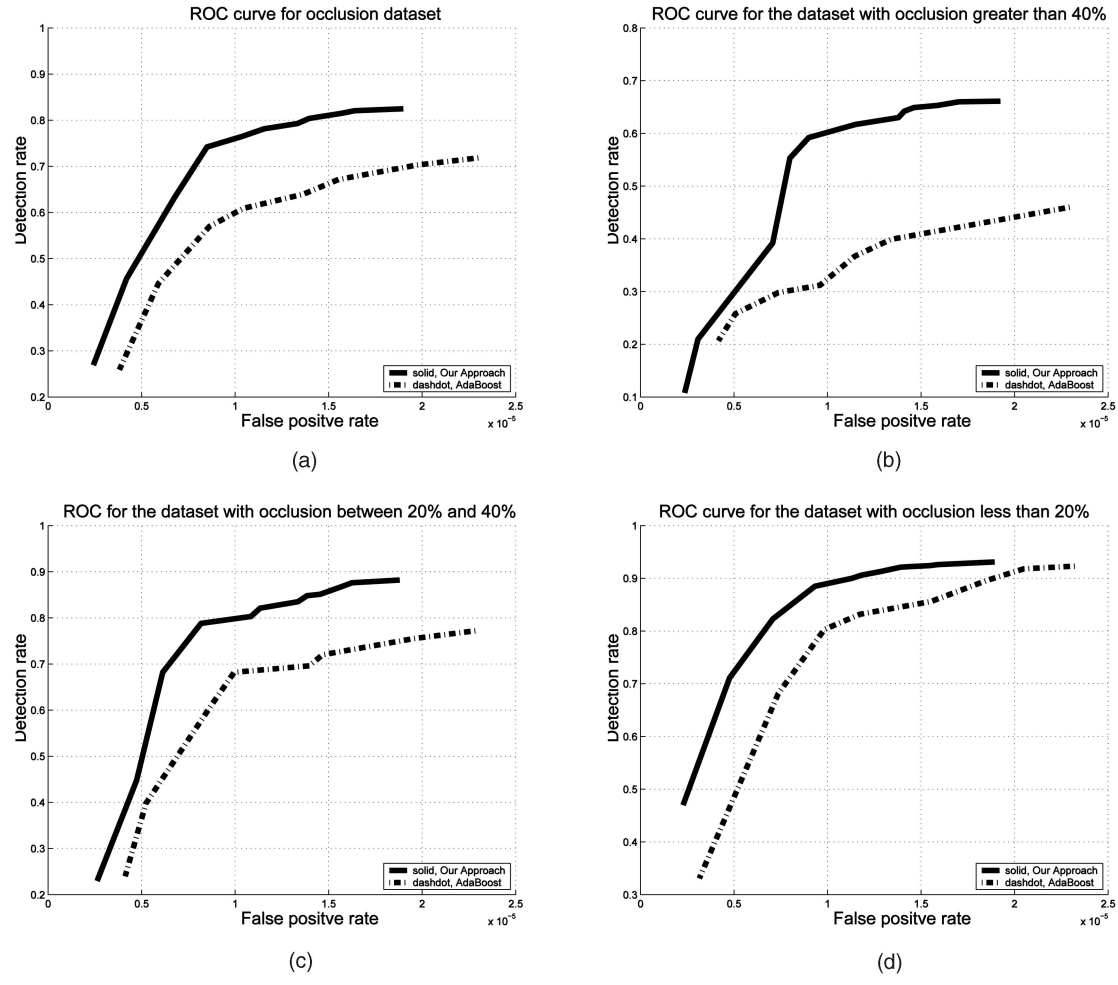

Fig. 12. ROC curves on Data set C, D, E, and F. (a) ROC on Data set C. (b) ROC on Data set D. (c) ROC on Data set E. (d) ROC on Data set F.

employed to characterize the complicated prior for local variability and a variational mean field approximation was presented for computationally efficient inference, likelihood calculation and model training. Due to the distributed likelihood model, this new field method is robust to occlusion. Based on the framework of this field model, the detection and tracking problems were also investigated and results presented. The success of applying the proposed method to pedestrian detection and tracking showed its effectiveness and general applicability.

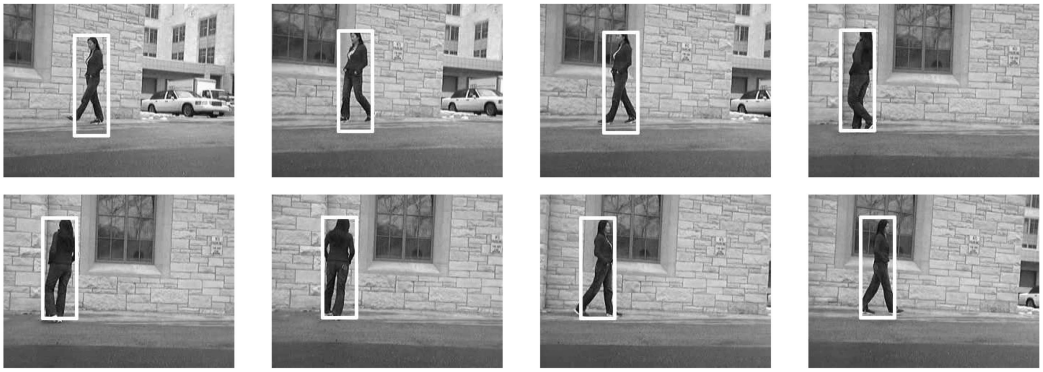

Fig. 13. Tracking a nonrigid target based on the mean field Boltzmann model.

Aligning training data in the proposed approach is easier than labeling landmark data in the active shape model [5], but it does leave a problem: How sensitive is the trained model to the alignment errors? We leave this for further studies. In addition, in our future work, we plan to investigate the capacity of the proposed Boltzmann field model, i.e., to what extent the model can capture local nonrigidity. Moreover, better image observation models will be studied to reduce the false alarm rate.

APPENDIX

This appendix gives the derivation of the mean field approximation of (7). Based on (5) and (6), we have:

$$J(Q_i) = H(Q_i) + \sum_{k \neq i} H(Q_k) + \int_{x_i} Q_i E_Q[\log p(\mathbf{X}, \mathbf{Z}) | x_i].$$

Since Q_i is a distribution, we can construct a Lagrangian:

$$L(Q_i) = J(Q_i) + \Lambda\left(\int_{x_i} Q_i - 1\right).$$

Then, the derivative of $L(Q_i)$ with regard to Q_i gives:

$$\frac{\partial L(Q_i)}{\partial Q_i} = -\log Q_i - 1 - E_Q[\log p(\mathbf{X}, \mathbf{Z}) | x_i] + \Lambda.$$

Once we set the derivative to zero, we obtain:

$$Q_i = e^{-1+\Lambda+E_Q[\log p(\mathbf{X},\mathbf{Z})|x_i]} = \frac{1}{Z_i} e^{E_Q[\log p(\mathbf{X},\mathbf{Z})|x_i]}.$$

ACKNOWLEDGMENTS

This work was supported in part by US National Science Foundation (NSF) Grant IIS-0308222, IIS-0347877 (CAREER), Northwestern startup funds, and the Murphy Fellowships. The authors also greatly thank the reviewers for the constructive comments and suggestions.

REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, pp. 509-522, 2002.
[2] A. Blake and M. Isard, *Active Contours.* Springer-Verlag, 1998.
[3] H. Chui and A. Rangarajan, "A New Algorithm for Nonrigid Point Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 44-51, June 2000.
[4] R. Collins, A. Lipton, and T. Kanade, "Special Issue on Video Surveillance and Monitoring," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 745-746, 2000.
[5] T.F. Cootes, C.J. Taylor, and J. Graham, "Active Shape Models —Their Training and Application," *Computer Vision and Image Understanding,* vol. 61, pp. 38-59, Jan. 1995.
[6] J. Coughlan and S. Ferreira, "Finding Deformable Shapes Using Loopy Belief Propagation," *Proc. European Conf. Computer Vision,* vol. 3, pp. 453-468, 2002.
[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 886-893, June 2005.
[8] L. Davis, I. Haritaouglu, and D. Harwood, "Ghost: A Human Body Part Labeling System Using Silhouettes," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 77-82, 1998.
[9] W. Freeman, E. Pasztor, and O. Carmichael, "Learning Low-Level Vision," *Int'l J. Computer Vision,* vol. 40, pp. 25-47, 2000.
[10] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding,* vol. 73, pp. 82-98, Jan. 1999.
[11] D.M. Gavrila and V. Philomin, "Real-Time Object Detection for 'Smart' Vehicles," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 87-93, Sept. 1999.
[12] D. Geiger and F. Girosi, "Parallel and Determinstic Algorithms from MRFs: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, pp. 401-412, 1991.
[13] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.
[14] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," *Proc. European Conf. Computer Vision,* pp. 343-356, 1996.
[15] T.S. Jaakkola, "Tutorial on Variational Approximation Methods," technical report, MIT Artificial Intelligence Lab., 2000.
[16] N. Jojic, N. Petrovic, B. Frey, and T.S. Huang, "Transformed Hidden Markov Models: Estimating Mixture Models and Inferring Spatial Transformations in Video Sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 26-33, June 2000.
[17] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning,* vol. 37, pp. 183-233, 2000.
[18] M. Kass, A. Witkin, and D. Terzopoulos, "Snake: Active Contour Models," *Proc. Int'l Conf. Computer Vision,* pp. 259-268, 1987.
[19] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 878-885, June 2005.

[20] C. Liu, S.C. Zhu, and H.-Y. Shum, "Learning Inhomogeneous Gibbs Model of Faces by Minimax Entropy," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 281-287, July 2001.
[21] J. MacCormick and A. Blake, "A Probabilistic Exclusion Principle for Tracking Multiple Objects," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 572-578, 1999.
[22] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images by Components," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 4, pp. 349-361, Apr. 2001.
[23] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian Detection Using Wavelet Templates," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 193-199, 1997.
[24] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 130-136, 1997.
[25] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," *Int'l J. Computer Vision,* vol. 38, pp. 15-33, 2000.
[26] V. Pavlović, R. Sharma, and T.S. Huang, "Visual Interpretation of Hand Gestures for Human Computer Interaction: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 677-695, July 1997.
[27] A. Pentland, "Looking at People: Sensing for Ubiquitous and Wearable Computing," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 107-119, Jan. 2000.
[28] C. Peterson and J. anderson, "A Mean Field Theory Learning Algorithm for Neural Networks," *Complex Systems,* pp. 995-1019, 1987.
[29] D. Ramanan and D. forsyth, "Finding and Tracking People from the Bottom Up," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 467-474, June 2003.
[30] A. Rangarajan, J. Coughlan, and A. Yuille, "A Bayesian Network Framework for Relational Shape Matching," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 671-678, Oct. 2003.
[31] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 23-38, Jan. 1998.
[32] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 746-751, 2000.
[33] S. Sclaroff and A. Pentland, "Modal Matching for Correspondence and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, pp. 545-561, 1995.
[34] K. Toyama and A. Blake, "Probabilistic Tracking in a Metric Space," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 50-57, July 2001.
[35] P. Viola and M. Jones, "Rapid Object Detection Using A Boosted Cascade of Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, Dec. 2001.
[36] P. Viola, M. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 734-741, Oct. 2003.
[37] Y. Weiss, "Correctness of Local Probability Propagation in Graphical Models with Loops," *Neural Computation,* vol. 12, pp. 1-41, 2000.
[38] Y. Wu, T. Yu, and G. Hua, "A Statistical Field Model for Pedestrian Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 1023-1030, June 2005.
[39] A. Yuille, "Deformable Templates for Face Recognition," *J. Cognitive Neuroscience,* vol. 3, pp. 59-70, 1991.
[40] S.C. Zhu, Y.N. Wu, and D.B. Mumford, "FRAME: Filters, Random Field and Maximum Entropy—Towards a Unified Theory for Texture Modeling," *Int'l J. Computer Vision,* vol. 27, pp. 1-20, 1998.

Ying Wu received the BS degree from Huazhong University of Science and Technology, Wuhan, China, in 1994, the MS degree from Tsinghua University, Beijing, China, in 1997, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001. From 1997 to 2001, he was a research assistant at the Beckman Institute for Advanced Science and Technology at UIUC. During the summers of 1999 and 2000, he was a research intern with Microsoft Research, Redmond, Washington. Since 2001, he has been an assistant professor in the Department of Electrical and Computer Engineering at Northwestern University, Evanston, Illinois. His current research interests include computer vision, computer graphics, machine learning, multimedia, and human-computer interaction. He serves as an associate editor of the SPIE *Journal of Electronic Imaging*. He received the Robert T. Chien Award at UIUC in 2001, and is a recipient of the US National Science Foundation CAREER award. He is a member of the IEEE and the IEEE Computer Society.

Ting Yu received the BS and MS degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2000 and 2002. He is currently a PhD candidate in the Department of Electrical and Computer Engineering at Northwestern University, Evanston, Illinois. During the summers of 2004 and 2005, he was a research intern with the NEC Labs America, Cupertino, California, and Microsoft Research, Redmond, Washington, respectively. His research interests include computer vision, image/video processing and analysis, statistical learning, and data mining. He received the Walter P. Murphy Fellowship at Northwestern in 2002, and the Motorola Graduate Scholarship and Excellent Student Scholarships at Tsinghua in 2001, 1999, and 1997. He is a student member of the IEEE.