Detecting Humans under Partial Occlusion using Markov Logic Networks

Raghuraman Gopalan
Dept. of ECE
University of Maryland
College Park, MD 20742 USA
raghuram@umiacs.umd.edu

William Schwartz
Dept. of Computer Science
University of Maryland
College Park, MD 20742 USA
schwartz@cs.umd.edu

ABSTRACT

Identifying humans under partial occlusion is a challenging problem in unconstrained scene understanding. In contrast to many existing works that model human appearance *in isolation*, we address this problem by studying the semantic context between human face and other body parts using Markov logic networks. By learning a set of probabilistic first-order logic rules that capture interactions between body parts under varying degrees of occlusion, and the relationship they share with the neighboring spatial windows, we obtain a graphical model representation of these instances to facilitate inference. We illustrate the efficacy of our method through experiments on standard human detection datasets, and an internally collected dataset with several occluding humans.

Categories and Subject Descriptors

I.4.8 [**Computing methodologies**]: Image processing and computer vision—*Scene analysis, Object recognition*

General Terms

Algorithm

Keywords

Human detection, Occlusion, Context, Learning

1. INTRODUCTION

Detecting humans from still images is an extensively studied problem in computer vision. It is a challenging problem due to the presence of scene-induced variations like pose, lighting, inter- and intra- person occlusions, among others. Occlusions, in particular, pose a significant challenge since there is no analytical model for the nature of variations it could inflict on the person's appearance.

The vast majority of the literature on human detection can be classified into the following two categories: holistic window-based and part-based. In the first category, features extracted from sliding windows are used holistically to perform classification. Some examples include the work of Dalal and Triggs [1] that used histogram of oriented gradients (HOG) features, and Tuzel et al [14] using covariance matrices. Part-based approaches, on the other hand, view a human as a combination of different sub-parts and model their interaction to perform classification. For instance, Wu and Nevatia [16] used edgelet features in a boosting framework to learn nested cascade detectors for each part. Mikolajczyk et al [5] used probabilistic combination to fuse information across body parts.

However, there are only very few works that explicitly account for the effect of occlusion in detecting humans. For instance, Shet et al [12] used a bilattice based logical reasoning framework to detect occluding humans by considering the responses of different low level parts-based detectors as logical facts. Schwartz et al [11] combined the outputs of detectors corresponding to different human parts, and the face using first order logic to model cases that study interactions between them.

Contributions: However, logic (by itself) can not accommodate for the probabilistic nature of the real world, and hence a more formal approach that accommodates the uncertainties of the visual scene is needed. Further, by focusing on different aspects of human in isolation (within a single window), the existing works do not account for the information conveyed by the surrounding scene. Since human vision perceives the real world by associating a set of contextual constraints prevalent in nature [13], we propose

- Modeling contextual information between different human parts, and the relation they share with the surrounding spatial windows.
- Learning the source of context using the framework of Markov logic networks [8], which integrates first-order logic with probabilistic reasoning to perform robust inference.

Outline of the paper: We first motivate our approach in Section 2 and discuss the details of the detectors, and the sources of context. In Section 3 we introduce the basics of Markov logic networks, including the construction of the network and performing inference. We also highlight our adaptation of this framework in terms of detecting humans under occlusion. We then validate our method in Section 4 on datasets, with and without occlusion, and provide comparisons with the existing methods. We also analyze the advantage of Markov logic when compared with other ways

PerMIS'10, September 28-30, 2010 Baltimore, MD USA.

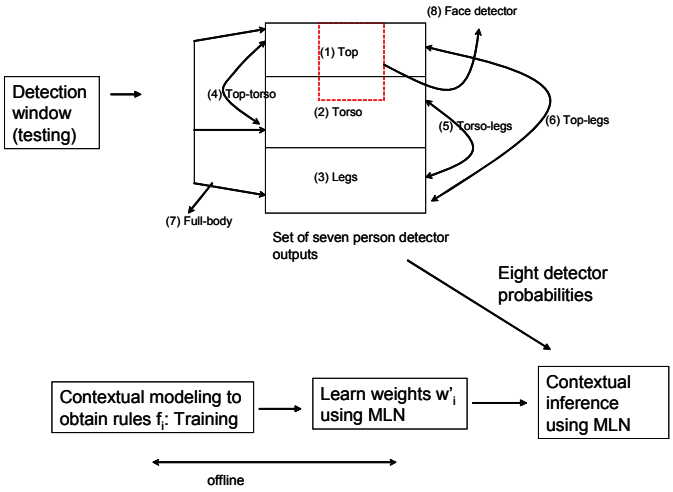

Figure 1: An overview of the proposed approach

of integrating information from detectors. We then discuss the merits and demerits of our approach, and conclude the paper in Section 5.

2. AN OVERVIEW OF OUR APPROACH

Our method to detect humans under partial occlusion stems from the notion of parts. We specifically focus on modeling the contextual interactions of these parts in performing inference. Given a set of N detectors corresponding to different body parts, and a set of detection windows $\{d_i\}_{i=1}^{M}$ in a close spatial neighborhood, we propose the following:

- Learn the relation shared between the N detectors under varying degree of occlusion - **Intra-window context**
- Learn the relation of a window with the surrounding windows under visual uncertainties - **Inter-window context**
- Formulate the contextual information with a set of logic rules, and perform probabilistic inference within the framework of Markov logic networks.

We discuss more details in the following sections.

2.1 Design of detectors

We divide the human into the following $N = 8$ parts, full body, top, torso, legs, top-torso, top-legs, torso-legs, and face (trained on a higher resolution than other human parts). An illustration is given in figure 1. We then learn individual detectors for these parts using training data with and without occlusion. The detectors are based on Partial least squares (PLS) [15], and we use the following set of features [10] to learn the appearance variations of these parts: co-occurrence matrices [14] to extract texture features, edge information using HOG [1], and color frequency [10].

Let $\mathbb{X} \in R^m$ denote an m-dimensional space of feature vectors and similarly let $\mathbb{Y} \in R$ be a 1-dimensional space representing the class labels. Let the number of samples (training patches) be n. PLS decomposes the zero-mean matrix $\mathbf{X}$ $(n \times m)$ and zero-mean vector $\mathbf{y}$ $(n \times 1)$ into

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1)$$

$$\mathbf{y} = \mathbf{U}\mathbf{q}^T + \mathbf{f} \quad (2)$$

where $\mathbf{T}$ and $\mathbf{U}$ are $(n \times p)$ matrices containing p extracted latent vectors, the $(m \times p)$ matrix $\mathbf{P}$ and the $(1 \times p)$ vector $\mathbf{q}$ represent the loadings and the $(n \times m)$ matrix $\mathbf{E}$ and the $(n \times 1)$ vector $\mathbf{f}$ are the residuals. The PLS method, using the nonlinear iterative partial least squares (NIPALS) algorithm [15], constructs a set of weight vectors (or projection vectors) $W = \{w_1, w_2, ..., w_p\}$ such that

$$[cov(t_i, u_i)]^2 = \max_{|\mathbf{w}_i|=1} [cov(\mathbf{X}\mathbf{w}_i, \mathbf{y})]2 \quad (3)$$

where $\mathbf{t}_i$ is the i^{th} column of matrix $\mathbf{T}$,$\mathbf{u}_i$ the i^{th} column of matrix $\mathbf{U}$ and $cov(t_i, u_i)$ is the sample covariance between latent vectors $\mathbf{t}_i$ and $\mathbf{u}_i$. After the extraction of the latent vectors $\mathbf{t}_i$ and $\mathbf{u}_i$, the matrix $\mathbf{X}$ and vector $\mathbf{y}$ are deflated by subtracting their rank-one approximations based on $\mathbf{t}_i$ and $\mathbf{u}_i$. This process is repeated until the desired number of latent vectors had been extracted. The dimensionality reduction is performed by projecting the feature vector $\mathbf{v}_i$, extracted from a i^{th} detection window, onto the weight vectors $W = \{w_1, w_2, .., w_p\}$, obtaining the latent vector $\mathbf{z}_i$ $(1 \times p)$ as a result. This vector is then used in classification to obtain the detection probability $p_{w,N}$ of a particular window, w belonging to the N^{th} part.

2.2 The sources of context

The set of $p_{d_i,N}$ for all N detectors across windows $d_i, i = 1$ *to* M are our primary information using which we learn the contextual information. The choice of M depend on the application of interest, and we chose $M = 10$ windows equally spaced around a center window d_c in an image at which the decision about a human is to be taken. This process will be repeated across all possible locations (d_c) in an image, in a sliding window fashion.

This type of context, being built from the intermediate detector probabilities, is generally referred as the *semantic context* [2]. We now learn contextual rules using $p_{d_i,N}$ as follows,

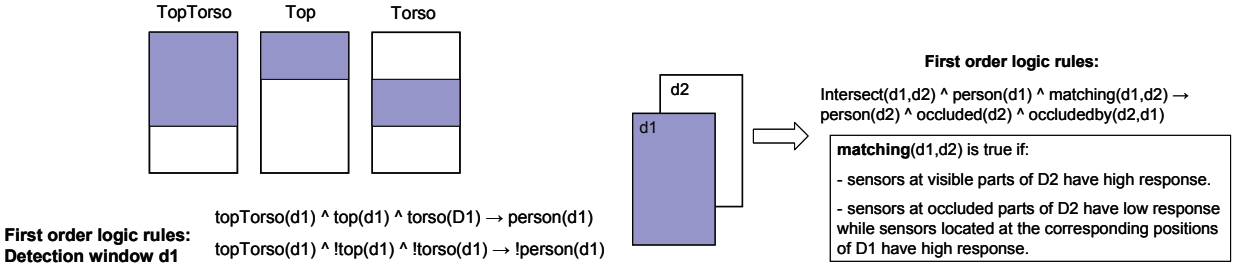

Figure 2: Examples of Contextual modeling. (a) Intra-window context. (b) Inter-window context

- **Intra-window context:** This is done for a single d_c across all N. The main focus behind these rules is to see how different parts of humans supplement information under varying degrees of visual uncertainties such as, occlusions, pose among others. The general terminology for these rules can be written as,

$$p_{d_c,1} \circledast p_{d_c,2} \circledast ... \circledast p_{d_c,N} = Out_{d_c} \quad (4)$$

  where $\circledast$ denotes to function operation between the probabilities, which can be either the union $\vee$, intersection $\wedge$, or inversion of probabilities $\neg$. Out refers to the decision pertaining to a window d_c, which can be any semantic state such as, human present, human absent, human present with occlusion, which part is occluded among others.

- **Inter-window context:** We now analyze how $p_{d_i,[1..N]}$ interact with $p_{d_j,[1...N]}, \forall j = 1\ to\ M, j \neq i$. This is motivated by the fact the an object does not occur in isolation, but share some properties with the surrounding scene. Rules to encode this information take the following form. For a center detection window d_c,

$$p_{d_c,j_1} \circledast p_{d_i,j_2} \circledast Out_{d_c} \circledast Relation(d_c, d_i) = Out^{\star}_{d_c} \quad (5)$$

  where $Out^{\star}_{d_c}$ refers to semantic decision like human present or absent, which window is occluding others etc., and *Relation* refers to high level analysis of two detection windows in that, how their holistic probabilities are matching, are they intersecting, among others. An illustration is given in figure 2.

Let us now refer all possible rules (or formulas) in the form of (4,5) as $\mathbb{F} = \{f_i\}_{i=1}^{N_f}$. With these set of rules and training samples belonging to positive (human) and negative (non-human) class illustrating occlusions, we use Markov logic networks [8] to perform inference.

3. INFERENCE USING MARKOV LOGIC NETWORKS

Markov logic networks (MLN) is a first-order knowledge base (KB) with a weight w_i' attached to each formula f_i. Together with a set of constants representing objects in the domain (in our case, detection windows d_i's), it specifies a ground Markov network containing one feature for each possible grounding of a first-order formula in the KB, with the corresponding weight. Inference in MLNs is performed by Markov Chain Monte Carlo (MCMC) over the minimal subset of the ground network required for answering the query. Weights are efficiently learned from relational databases by iteratively optimizing a pseudo-likelihood measure. Optionally, additional clauses are learned using inductive logic programming techniques.

A first-order KB can be seen as a set of hard constraints on the set of possible worlds: if a world violates even one formula, it has zero probability. The basic idea in MLNs is to soften these constraints: when a world violates one formula in the KB it is less probable, but not impossible. The fewer formulas a world violates, the more probable it is. Each formula has an associated weight that reflects how strong a constraint it is: the higher the weight, the greater the difference in log probability between a world that satisfies the formula and one that does not, other things being equal.

3.1 Constructing the network

With this idea, an undirected network, called a Markov Network, is constructed to predict the outcomes (or events like a presence of human, the type of occlusion) such that,

- Each of its nodes correspond to a ground atom (an event) e_k.
- If a subset of ground atoms $e_{\{i\}} = \{e_k\}$ are related to each other by a formula f_i, then a clique C_i over these variables is added to the network. C_i is associated with a weight w_i' and a feature fe_i is defined as follows

$$fe_i(e_{\{i\}}) = \begin{cases} 1,\ if\ f_i(e_{\{i\}})\ is\ true \\ 0, otherwise \end{cases} \quad (6)$$

Thus first-order logic formulae in our knowledge base serve as templates to construct the Markov Network. This network models the joint distribution of the set of all ground atoms, E, each of which is a binary variable. It provides a means for performing probabilistic inference.

$$P_E = P(Event = E) = \frac{1}{Z} \exp(\sum_i w_i' fe_i(e_{\{i\}})) \quad (7)$$

$$P_E = \frac{1}{Z} \exp(\sum_i w_i' f_i) \quad (8)$$

where Z is the normalizing factor.

3.2 Inference

Based on the constructed Markov network, the marginal distribution of any event E given some evidence (observations) can be computed using probabilistic inference. Since the structure of the network may be very complex (e.g. containing undirected cycles), exact inference is often intractable. MCMC sampling is a good choice for approximate

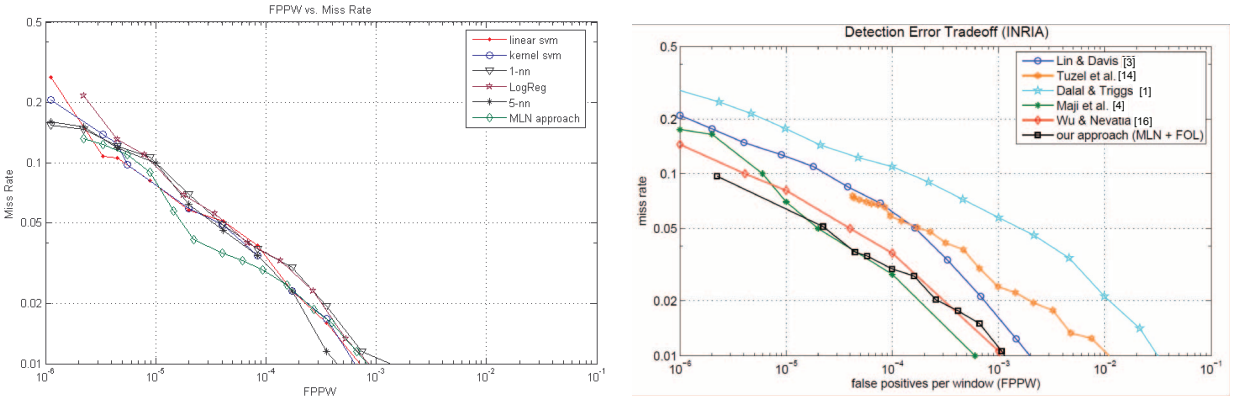

Figure 3: Experimental validation. (a) Comparing different combination strategies on INRIA [1] dataset. (b) Detection performance in INRIA dataset.

reasoning. In MLN, the probability that a ground atom E_i is equal to e_i given its Markov blanket (neighbors) B_i is

$$P(E_i = e_i | Bi = bi) = \frac{\exp(\sum_{fe_j \in f_i} w'_j fe_j(E_i = e_i, B_i = b_i))}{Z_1} \quad (9)$$

where $Z_1 = \sum_{k1 \in \{0,1\}} \exp(\sum_{fe_j \in f_i} w'_j fe_j(E_i = k1, B_i = b_i))$.

where f_i is the set of all cliques that contain E_i and fe_j is computed as in (6). Basic MCMC (Gibbs sampling) is known to have difficulty dealing with deterministic relations, which are unavoidable in our case. It has been observed that using simulated tempering [8] gives better performance than the basic Gibbs sampling. Simulated tempering is a MC method that is closely related to simulated annealing. However, instead of using some fixed cooling schedule, a random walk is also performed in the temperature space whose structure is predetermined and discrete. These moves aim at making the sampling better at jumping out of local minima.

Hence, with the set of contextual rules $\mathbb{F}$ as the input to MLN, we learn the weights w'_i and perform inference over detection windows d_i's using the open source Alchemy system [8].

4. EXPERIMENTS

We evaluated our method on datasets containing humans with and without occlusion. For the first task we used the widely followed benchmark of INRIA pedestrian detection images [1], with the main goal of comparing the performance of our method with other standard approaches. Then to study performance under occlusion, we used images from an internally collected dataset.

4.1 Training stage

We estimate the parameters of our system using a 10-fold cross-validation procedure on the training dataset provided by INRIA Person Dataset. This dataset has 2416 positive samples of size 64×128 pixels and images containing no humans, used to obtain negative exemplars. We sample this set to obtain our validation set containing 2000 positive samples and 10000 negative samples. This was used to estimate the detection thresholds of seven human part detectors. For the face detector, high resolution images of 240×320 pixels from the CMU-MIT dataset [9] were used (since faces generally require higher resolution for detection when compared to humans).

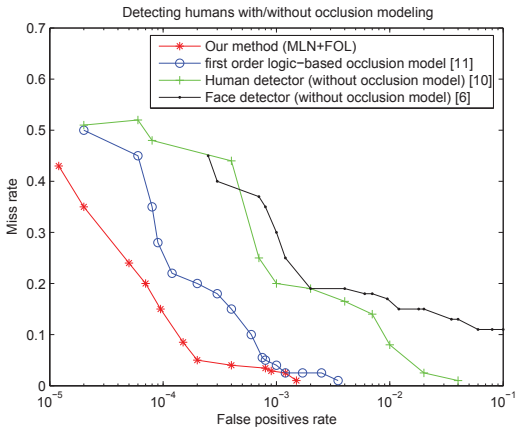

Figure 4: Detecting humans under occlusion on an internally collected dataset

4.2 Testing

We first evaluate whether the integration of the 8 detectors using MLN is better than other standard combination method. The INRIA testing dataset was used for this purpose, and the following methods were compared with: linear and kernel SVM [7], logistic regression and k-nearest neighbor. We provide the results in figure 3(a) where we see that MLN offers better performance. We then compared our approach with other standard methods ([1, 16, 3, 14, 4]) for human detection. As noted before, this dataset has very few occluded images. The results are given in figure 3(b), where we see that in regions with stringent requirements on false positives, our method has very low miss rates.

We then used our internally collected dataset with lots of occluding humans to test the proof of concept. We had nearly 70 images containing nearly 300 humans. We compare our method with methods for detecting humans [10] and faces [6] without occlusion handling, and a first-order logic based method to model occlusions [11]. The results are

Figure 5: Sample detection results on images from an internally collected dataset. Semantic decisions on occluding humans are a part of the output.

given in figure 4, and some examples of detection in figure 5 which, in addition to providing the location information of humans, gives semantic knowledge of occluding persons. These results exemplify the efficacy of our approach.

5. CONCLUDING REMARKS

We proposed a method to model contextual information of humans, and perform inference using Markov logic networks to handle partial occlusions. Through this, we illustrated the importance of context, and the use of probabilistic interpretation of first-order logic to perform robust inference under visual uncertainties. A study of more formal model for occlusion and other sources of context, and incorporating them in this framework is an interesting future work to improve robustness to noise.

6. ACKNOWLEDGEMENTS

This work was supported by a MURI Grant N00014-08-1-0638 from the Office of Naval Research.

7. REFERENCES

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[2] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278, 2009.

[3] Z. Lin and L. S. Davis. A pose-invariant descriptor for human detection and segmentation. In *European Conference on Computer Vision*, pages 423–436, 2008.

[4] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, 2008.

[5] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *European Conference on Computer Vision*, pages 69–82, 2004.

[6] H. Moon, R. Chellappa, and A. Rosenfeld. Optimal edge-based shape detection. *IEEE Transactions on Image Processing*, 11(11):1209–1227, 2002.

[7] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–137. Published by the IEEE Computer Society, 1997.

[8] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.

[9] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 20, pages 23–38, 1998.

[10] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *Proceedings of the International Conference on Computer Vision*, pages 24–31, 2009.

[11] W. R. Schwartz, R. Gopalan, R. Chellappa, and L. S. Davis. Robust human detection under occlusion by integrating face and person detectors. In *International Conference on Biometrics*, pages 970–979, 2009.

[12] V. Shet, J. Neumann, V. Ramesh, and L. Davis. Bilattice-based logical reasoning for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, 2007.

[13] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[14] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, 2007.

[15] H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, 6:581–591, 1985.

[16] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *IEEE International Conference on Computer Vision*, pages 90–97, 2005.