Latent SVMs for Human Detection with a Locally Affine Deformation Field

Ľubor Ladický[1]
lubor@robots.ox.ac.uk
Philip H.S. Torr[2]
philiptorr@brookes.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] School of Technology
Oxford Brookes University
Oxford, UK

Abstract

Methods for human detection and localization typically use histograms of gradients (HOG) and work well for aligned data with low variance. For methods based on HOG despite the fact the higher resolution templates capture more details, their use does not lead to a better performance, because even a small variance in the data could cause the discriminative edges to fall into different neighbouring cells. To overcome these problems, Felzenszwalb *et al*. proposed a star-graph part based deformable model with a fixed number of rigid parts, which could capture these variations in the data leading to state-of-the-art results. Motivated by this work, we propose a latent deformable template model with a locally affine deformation field, which allows for more general and more natural deformations of the template while not over-fitting the data; and we also provide a novel inference method for this kind of problem. This deformation model gives us a way to measure the distances between training samples, and we show how this can be used to cluster the problem into several modes, corresponding to different types of objects, viewpoints or poses. Our method leads to a significant improvement over the state-of-the-art with small computational overhead.

1 Introduction

Human detection is typically formulated as a problem where the objective is to find all the people within an image and enclose each one of them by a tight bounding box. Dalal and Triggs [4] introduced the histograms of oriented gradients (HOG) feature for this problem over cells composing the bounding box, efficiently matching object shape with the learnt rigid template of edge directions. This method was originally applied to pedestrian detection, but it turned out to give good performance for a wide range of object classes with distinctive shape. Intuitively, a higher dimensional template should capture more small details and should lead to a better performance. However, even under small local deformations of the data it is impossible to align the data properly and the discriminative edges often fall into the neighbouring cell. To overcome this problem, Felzenszwalb *et al*. [8] proposed a star-graph part based model (later generalized to a multi-layer hierarchy [3]) allowing a predetermined number of rigid parts to change their relative location with respect to the centre of the object.

http://dx.doi.org/10.5244/C.26.10

Large intra-class variance was modelled by splitting training samples based on their aspect ratio and training a classifier for each set of training samples independently. This procedure works if the different aspect ratio corresponds to a different viewpoint, such as for example for a car. However, it is not very suitable for human detection, where different human poses often have the same aspect ratio and the method does not learn an independent model for each one of them.

Motivated by this work, we propose a new latent variable SVM allowing for any deformations of the template, expressed in terms of a deformation field. Rather than restrict ourselves to a star-graph model, we allow the template to deform according to a locally affine deformation field. We propose tractable optimisation for learning parameters of the model and for evaluation. Our deformation model can be seen as the generalisation of [18], which refines the template beyond translation and scaling with an additional transformation selected from a finite set of possible perturbations covering aspect ratio change and small in plane rotations. We show how the deformation field could be used to measure the similarity of the training samples and thus could be used to cluster the problem into several poses or viewpoints using more suitable measure than the aspect ratio. Such clustering has already been used for the pose estimation problem [12].

The only similar approach to ours, that tries to measure the similarities between objects by matching them using a deformation field, has been proposed for the detection rescoring [13] and the classification problem [5]. However, the optimisation of the unconstrained pairwise random field is computationally too heavy and can not be used for a sliding window detector.

2 Previous Work

First we describe the formulation of Dalal and Triggs [4]. The linear support vector machine (SVM) classifier response for a given image sub-window is based on histograms of oriented gradients (HOG) evaluated on a regular grid of $n = n_x \times n_y$ (in general overlapping) cells, where each cell is a rectangular region of a fixed size $S_x \times S_y$ centred at the point $c_{ij} = [x_{ij}, y_{ij}]$. Let $\mathbf{h}(c_{ij}) \in \mathbb{R}^m$ be the corresponding histogram of gradients with m directions over the cell, centred at the point c_{ij}, and $\mathbf{h}(\mathbf{c}) \in \mathbb{R}^{mn}$ concatenated histograms over all cells. The linear discriminant function takes the form :

$$H(\mathbf{c}) = \mathbf{w}^* \cdot \mathbf{h}(\mathbf{c}) + b^* = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{l=1}^{m} w^*_{ijl} h_l(c_{ij}) + b^*, \tag{1}$$

where $H(\mathbf{c}) > 0$ indicates a positive detection, negative otherwise. The weights $\mathbf{w}^*$ and bias b^* are trained by solving the optimization problem using M training samples with ground truth labels $z^k \in \{-1, 1\}$ as:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1, .., M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \left(\mathbf{w}.\mathbf{h}(\mathbf{c}^k) + b \right), \end{aligned} \tag{2}$$

where $\mathbf{h}(\mathbf{c}^k)$ are the concatenated histograms of k-th training sample and λ is the regularisation strength. Typically, data sets contain only positive bounding boxes and negative ones are randomly sampled. An improvement over this approach can be gained by iteratively bootstrapping the training set using the hard negatives obtained by the classifier from the previous iteration [4, 19].

The rigid formulation has been extended [8] to a more flexible one with a fixed number of rigid parts using Latent SVM, with the part locations as latent variables. The model is trained by alternating between estimation of the locations of parts given the weight vectors and estimating optimal weights and bias given the state of the latent variables.

3 Deformable Template with MRF Priors

In this work we propose a model that allows the deformation of an object using the deformation field $\mathbf{d} = [\mathbf{d^x}, \mathbf{d^y}]$ containing an optic flow like deformation parameters $[d^x_{ij}, d^y_{ij}]$ for each cell centre. This can be thought of a set of latent variables (as in the Latent CRF). However, the form of prior we shall choose will be much richer than in [8] and a generalisation in that we will allow for a more general deformations of the template.

Formally, the deformation can be defined using a deformation function $D^{\mathbf{d}}(\mathbf{c})$ transforming each cell centre relative to its size $S_x \times S_y$ as :

$$D^{d_{ij}}(c_{ij}) = D^{d_{ij}}([x_{ij}, y_{ij}]) = [x_{ij} + d^x_{ij} S_x, y_{ij} + d^y_{ij} S_y], \tag{3}$$

where d^x_{ij} and d^y_{ij} are the deformations relative to the size of the cell. We restrict deformations $d_{ij} = [d^x_{ij}, d^y_{ij}]$ to the interval $\mathcal{L} = (-d_{max}, d_{max}) \times (-d_{max}, d_{max})$ with discrete steps (0.5 cell used in experiments). The deformation field $\mathbf{d}$ is treated as a set of latent variables jointly estimated with the parameters (weights and bias) of the SVM classifier. To penalize improbable deformations the regularisation term $R(\mathbf{d})$ is introduced. The classifier for our deformable template then takes the form :

$$H(\mathbf{c}) = \max_{\mathbf{d}} \left(\mathbf{w}^* \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) + b^* - R(\mathbf{d}) \right), \tag{4}$$

where the regularisation term $R(\mathbf{d})$ for the deformation field $\mathbf{d}$ takes the form of the pairwise Markov Random Field (MRF) cost :

$$R(\mathbf{d}) = \theta_p \sum_{(ij,fg) \in \mathcal{E}} \psi_p(d_{ij} - d_{fg}), \tag{5}$$

where $\mathcal{E}$ is the set of pairs of neighbouring cells and ψ_p pairwise potential enforcing neighbouring patches to take similar deformation. This kind of regularisation is typically used for the optical flow problem. Experimentally the most successful pairwise potential for these problems takes the form of quadratic truncated cost $\psi_p(d_{ij} - d_{fg}) = \min(||d_{ij} - d_{fg}||^2, T)$, where T is the truncation parameter.

In the standard optical flow problem we are typically matching the same object between images, and this kind of regularisation is sufficient and not prone to over-fitting. However, in the detection problem we match two different things – an object and a template. A too low pairwise weight θ_p could lead to over-fitting making an object "fall apart". On the other hard a too high pairwise weight makes deformations impossible. Experimentally, the gap

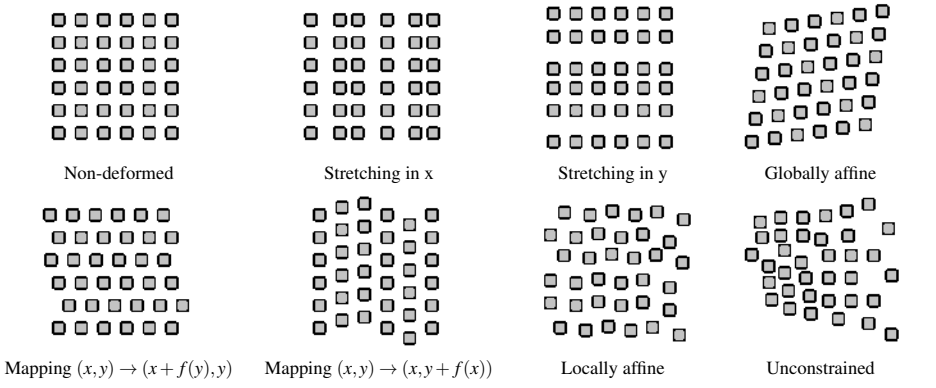


Figure 1: Expressive power of the locally affine deformation field. As shown in the figure, the locally affine constraints allow for stretching or mapping of the template in both axes, global affine transformation of the template or the combination of all of them resulting in the general locally affine transformation, in which any 2×2 neighbouring cells transform into a parallelogram. A typical example of a locally affine transformation in practice is the deformation field used in the Microsoft Windows logo, which can be composed of the mapping of the x axis with the global affine transformation. For comparison, the example of a general unconstrained transformation of the template is shown in the last image.

between these two cases is very small and often θ_p needs to be tuned per instance to get desired output.

One way to deal with the problem of matching of an object to an object in the database was proposed in [14] using the flow of the more discriminant dense SIFT [15]features. Such approach is computationally too heavy for the sliding window detection. The way we explore here, is to keep θ_p low, but enforce certain higher order structures in the deformation field. One such useful type of structure is to constrain the motion to be a *local affine* deformation field; this means each 2×2 neighbouring cells can deform into a parallelogram. More formally we will call a deformation field locally affine and write $\mathbf{d} \in \mathcal{A}$ if:

$$\forall ij \in \{1,..,n_x-1\} \times \{1,..,n_y-1\} : \; d_{i,j} + d_{i+1,j+1} = d_{i+1,j} + d_{i,j+1}. \tag{6}$$

where $d_{i,j} = d_{ij}$. Locally affine deformation fields allow for a large variety of stretching or bending while keeping its structure consistent, forcing the template not to fall apart. Examples of locally affine deformation of the regular template describing the expressive power of such transformation are shown in the Figure 1. As we show later, the locally affine constraints allow for a novel fast inference method.

The latent SVM optimisation problem for learning the weights $\mathbf{w}^*$ and the bias b^* becomes:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1,..,M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \max_{\mathbf{d} \in \mathcal{A}} \left(\mathbf{w} \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c}^k)) + b - R(\mathbf{d}) \right). \end{aligned} \tag{7}$$

4 Learning the Parameters of the Deformable Model

The optimisation problem (7) for the training stage is non-convex. However, we can follow the same approach as [8] and iteratively estimate the weight vector $\mathbf{w}$ with the bias b, and the deformation field $\mathbf{d}$ for each training sample.

First, the problem of finding the optimal weight vector $\mathbf{w}$ and bias b, given the deformation fields $\hat{\mathbf{d}}^k$ for each training example becomes:

$$\begin{aligned}(\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda_w ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1,..,M\}: \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \left(\mathbf{w} \cdot \mathbf{h}(D^{\hat{\mathbf{d}}^k}(\mathbf{c}^k)) + b - R(\hat{\mathbf{d}}^k)\right)\end{aligned} \tag{8}$$

and can be solved using any standard SVM algorithms [1, 11, 17]. The problem of finding the optimal deformation field $\mathbf{d}^*$ for each training example given current weights $\hat{\mathbf{w}}$ becomes:

$$\begin{aligned}\mathbf{d}^{k*} &= \arg\max_{\mathbf{d}^k \in \mathcal{A}} \left(\hat{\mathbf{w}} \cdot \mathbf{h}(D^{\mathbf{d}^k}(\mathbf{c}^k)) - R(\mathbf{d}^k)\right) \\ &= \arg\min_{\mathbf{d}^k \in \mathcal{A}} \sum_{(ij,fg)\in\mathcal{E}} \psi_p(d^k_{ij} - d^k_{fg}) - \hat{\mathbf{w}} \cdot \mathbf{h}(D^{\mathbf{d}^k}(\mathbf{c}^k)).\end{aligned} \tag{9}$$

The last term can be decomposed into the sum of the independent functions of deformations for each cell d^k_{ij} as:

$$\hat{\mathbf{w}} \cdot \mathbf{h}(D^{\mathbf{d}^k}(\mathbf{c}^k)) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{l=1}^{m} \hat{w}_{ijl} h_l(D^{d^k_{ij}}(c^k_{ij})). \tag{10}$$

By defining $\psi_u(d^k_{ij}) = -\sum_{l=1}^{m} \hat{w}_{ijl} h_l(D^{d^k_{ij}}(c^k_{ij}))$ the optimisation procedure to find the optimal deformation field becomes :

$$\mathbf{d}^{k*} = \arg\min_{\mathbf{d}^k \in \mathcal{A}} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \psi_u(d^k_{ij}) + \sum_{(ij,fg)\in\mathcal{E}} \psi_p(d^k_{ij} - d^k_{fg}), \tag{11}$$

which is the max-a-posteriori (MAP) estimation of the pairwise MRF problem with $|\mathcal{L}^x||\mathcal{L}^y|$ labels under the additional locally affine deformation field constraints. Without these constraints and with convex pairwise function $\psi_p(\cdot)$ the problem would be approximately solvable by estimating $\mathbf{d}_x$ and $\mathbf{d}_y$ iteratively using graph cut [2] with Ishikawa's graph construction [10]. However, in practice it would be computationally too heavy for a sliding window approach even with a linear pairwise cost as used in [5]. Instead, we propose an alternative computationally very efficient approach, which solves the problem under the additional locally affine deformation field constraints for any form of pairwise costs.

We start with the observation that the deformation of all cells in the first row and in the first column of the deformation field fully determines the deformation of any other cell. Intuitively we can fill the deformations inductively as $d_{i,j} = d_{i-1,j} + d_{i,j-1} - d_{i-1,j-1}$. Locally affine constraints can be satisfied for any deformations of the cells in the first row and in the first column. Thus, any locally affine deformation field can be reached by two moves – the

first in which we move each row j by a deformation $\Delta^r d_j = (\Delta^r d_j^x, \Delta^r d_j^y)$ and the second in which we move each column i by a deformation $\Delta^c d_i = (\Delta^c d_i^x, \Delta^c d_i^y)$. Trivially, both of these moves do not break the local affinity property and can lead to any deformation of the cells in the first row and in the first column and thus to any arbitrary locally affine deformation field. Formally, the transformation function for the row move is defined as:

$$T^{\Delta^r \mathbf{d}}(d_{ij}) = d_{ij} + \Delta^r d_j, \forall ij \in \{1,..,n_x\} \times \{1,..,n_y\}. \tag{12}$$

The optimisation problem to find the optimal row move is:

$$\Delta^r \mathbf{d}^* = \arg\min_{\Delta^r \mathbf{d}} \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} \psi_u(d_{ij}^k + \Delta^r d_j) + \sum_{(ij,fg)\in\mathcal{E}} \psi_p(d_{ij}^k - d_{fg}^k + \Delta^r d_j - \Delta^r d_g). \tag{13}$$

Pairwise costs between cells in the same row do not change. Thus, by defining $\psi_u^r(\Delta^r d_j) = \sum_{i=1}^{n_x} \psi_u(d_{ij}^k + \Delta^r d_j)$ and $\psi_p^r(\Delta^r d_j, \Delta^r d_g) = \sum_{(ij,fg)\in\mathcal{E}, j\neq g} \psi_p(d_{ij}^k - d_{fg}^k + \Delta^r d_j - \Delta^r d_g)$ the optimisation problem becomes:

$$\Delta^r \mathbf{d}^* = \arg\min_{\Delta^r \mathbf{d}} \sum_{j=1}^{n_y} \psi_u^r(\Delta^r d_j) + \sum_{(j,g)\in\mathcal{E}^r} \psi_p^r(\Delta^r d_j, \Delta^r d_g), \tag{14}$$

where $\mathcal{E}^r$ is the set of pairs of the neighbouring rows. For a 4- or an 8- neighbourhood of $\mathcal{E}$ this problem is a simple MRF problem over the chain of the n_y variables and can be solved very efficiently using dynamic programming. Equivalently, the transformation function for the column move is defined as:

$$T^{\Delta^c \mathbf{d}}(d_{ij}) = d_{ij} + \Delta^c d_i, \forall ij \in \{1,..,n_x\} \times \{1,..,n_y\}. \tag{15}$$

The optimisation problem to find the optimal column move is:

$$\Delta^c \mathbf{d}^* = \arg\min_{\Delta^c \mathbf{d}} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \psi_u(d_{ij}^k + \Delta^c d_i) + \sum_{(ij,fg)\in\mathcal{E}} \psi_p(d_{ij}^k - d_{fg}^k + \Delta^c d_i - \Delta^c d_f). \tag{16}$$

By defining $\psi_u^c(\Delta^c d_i) = \sum_{j=1}^{n_y} \psi_u(d_{ij}^k + \Delta^c d_i)$ and $\psi_p^c(\Delta^c d_i, \Delta^c d_f) = \sum_{(ij,fg)\in\mathcal{E}, i\neq f} \psi_p(d_{ij}^k - d_{fg}^k + \Delta^c d_i - \Delta^c d_f)$ the optimisation problem becomes:

$$\Delta^c \mathbf{d}^* = \arg\min_{\Delta^c \mathbf{d}} \sum_{i=1}^{n_x} \psi_u^c(\Delta^c d_i) + \sum_{(i,f)\in\mathcal{E}^c} \psi_p^c(\Delta^c d_i, \Delta^c d_f), \tag{17}$$

where $\mathcal{E}^c$ is the set of pairs of the neighbouring columns. Again, for a 4- or an 8- neighbourhood of $\mathcal{E}$ this problem can be solved using dynamic programming. The local optimum of the locally affine deformation field $\mathbf{d}$ is found by iterating between the row and column moves. Experimentally, the local minimum is typically found after 2 iterations. In case we want to speed up the estimation of the deformation field we can split each of the row and column moves into two moves for each of the coordinate $\mathbf{d}^x$ and $\mathbf{d}^y$. The row moves in $\mathbf{d}^y$ and the column moves in $\mathbf{d}^x$ correspond to the stretching or shrinking of the template, the row moves in $\mathbf{d}^x$ and the column moves in $\mathbf{d}^y$ to the bending on the template.

The optimisation problem during the test phase is equivalent to the estimation of the optimal deformation field in the training phase:

$$H(\mathbf{c}) = \max_{\mathbf{d}\in\mathcal{A}} \left(\mathbf{w}^* \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) + b^* - R(\mathbf{d}) \right), \tag{18}$$

and solved in the same manner.

5 Learning of Different Viewpoints or Poses

Typically, different viewpoints are modelled by splitting the positive samples based on the aspect ratio and trained independently for each aspect ratio [8]. However, this approach does not model different poses with similar bounding boxes independently. We can take an advantage of our deformation field model and cluster the problem into subproblems based on the similarity of training samples.

We start by defining the non-commutative similarity measure $S^*_{k\to p}$ between two instances k and p as the scalar product between the reference feature vector of the instance p and the feature vector of the deformed instance k penalized by the deformation field regularisation $R(\mathbf{d})$ as:

$$S^*_{k\to p} = \max_{\mathbf{d}\in\mathcal{A}} \left(\mathbf{h}(\mathbf{c}^p) \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c}^k)) - R(\mathbf{d}) \right). \tag{19}$$

If the foreground-background masks are given, then to align the samples we can either replace the histograms of gradients with the foreground/background ratio or by the multiplication of them. The $M \times M$ similarity matrix S is then filled with the values:

$$S_{kp} = \max(S^*_{k\to p}, S^*_{mirror(k)\to p}), \tag{20}$$

where $mirror(k)$ is the mirrored instance k. Because the feature vector is non-negative and $R(\mathbf{0}) = 0$, the values in the similarity matrix can be any arbitrary non-negative numbers. In case the instances k and p differ too much in the aspect ratio (by 25% in experiments), then the similarity measure $S_{kp} = 0$. We formulate the problem of clustering the instances into T clusters as a search for the reference subset $\mathcal{P}$ of training samples, such that the $\sum_{k=1}^{M} \max_{p\in\mathcal{P}} S_{kp}$ is maximised. We solve this problem similarly to the standard k-medoid approach, where we randomly pick the set of centres $\mathcal{P}$ and then iteratively find the most similar centre for each training sample and using brute force find the centre in each cluster so that it maximises the cost function. Because the number of training samples is typically not huge, this procedure is very fast. To avoid very bad local optima the first iteration is done using twice the number of centres and then the desired number of largest clusters is kept for the next iterations. This procedure gives us not only the centres of the clusters and the membership of each training sample, but also their initial latent deformation field and the latent variable deciding whether we mirror the sample or not.

6 Experiments

Most existing human detection data sets typically consist of fully visible pedestrians [4] with low variability of poses. In such cases it is optimal to use just one model, which successfully captures all common poses in the street scene scenario. Thus, we tested our method on the more challenging Buffy data set of [9], which consists of images with large variety of poses and truncations by the edge of the image, which makes it suitable for our clustering method. The buffy data set consists of 748 images from episodes $s5e2$ – $s5e6$, with episode $s5e3$ used for training, episode $s5e4$ for validation and episodes $s5e2$, $s5e5$ and $s5e6$ for testing. Human detection, using an upper body detector, is typically used there as the preprocessing step for the pose estimation [6, 9, 16].

The training samples are clustered into 10 models. Clustering of the training samples into multiple models together with their corresponding HOG template deformed by the latent deformation field is shown in the Figure 2. The different clusters typically correspond to not

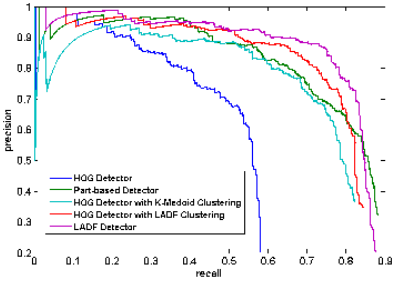
Figure 2: Clustering of the training samples into several models. Instances in the same row belong to the same cluster. Instances shown correspond to the largest four (out of ten) clusters. Overlaid on each training sample is the trained HOG template for the corresponding model deformed by the locally affine deformation field estimated as the set of latent variables. In the case of an optimal match with the mirrored template, the image is shown mirrored. As seen from the figure, the clusters typically correspond to not only different aspect ratios. but also to different poses. The first row is the cluster containing instances with a relatively large head compared to their shoulders (typically women with long hair). The second cluster contains fully visible humans standing upright. The third cluster contains humans holding their hips and the fourth one humans with large shoulders (typically men) comparing to the size of the head, slightly bent so the head is to the left side of the template.

only different aspect ratio but also to different poses. The HOG templates are trained using 4 bootstrapping iterations. During the evaluation a sliding window is initialised at every 3×3 cell. The resulting bounding boxes are obtained as a bounding box surrounding the cells in the corners of the template. Highly overlapping boxes are suppressed using standard non-maxima suppression. Several local optima should correspond to the different instances of object of interest. Qualitative results are shown in the Figure 3. The detection is considered as correct if the intersection vs union measure [7] is above 0.5. The performance is measured as the average precision(AP) – area under the precision vs recall curve [7]. We compared our deformable LADF detector (10 models) with the rigid HOG detector – root filter of [8] clustered using aspect ratio (3 models), the state-of-the-art part-based detector [8] clustered using aspect ratio (3 models), the rigid HOG detector with K-medoid clustering (10 models) and the rigid HOG detector with LADF training and clustering (10 models). The results of the first two detectors are obtained using the publicly available code of [8], the results of the other detectors are obtained using our own code. Our LADF detector significantly outperformed existing approaches. The results of the HOG detector using the models trained using LADF training suggest, that both the alignment and the clustering of the training samples play crucial role to achieve good performance. Quantitative comparison in terms of performance, training and test time is shown in the figure 4.

7 Conclusions

In this paper we propose a novel deformable human detector with a latent locally affine deformation field. We show how the classifier can be learnt and its deformation field efficiently estimated. We also show how the deformation field could be used to cluster the training

Figure 3: Typical results on the Buffy data set. Positive detections are overlaid with the learnt HOG template of the corresponding model, deformed by the deformation field. Large majority of the persons are correctly detected. Typical mistakes include missed detections of hard instances (A3, A4, C3, E4), false positive detections (A5, B2), detection with insufficient overlap with the ground truth (B3) or detection by a wrong model with a wrong truncation (A3).



Method	AP	Training time	Test time
HOG detector	47.65%	1h	1.4s
Part-based detector	72.38%	19.5h	6s
HOG detector with K-Medoid Clustering	69.10%	1.5h	4.7s
HOG detector with LADF Clustering	73.78%	2.5h	4.7s
LADF Detector	76.03%	2.5h	50s

Figure 4: Quantitative comparison of our deformable detector with the rigid HOG detector - root filter of [8] clustered using aspect ratio (3 models), the state-of-the-art part-based detector [8] clustered using aspect ratio (3 models), the rigid HOG detector with K-medoid clustering (10 models), the rigid HOG detector with LADF training and clustering (10 models) and the full LADF detector (10 models). On the left side we show the precision vs recall curves of all these 4 methods and on the right side a comparison in terms of average precision - AP, training and test time.

samples based on their pose or viewpoint. We tested the algorithm on the challenging Buffy data set and showed promising results. We assume, our random field formulation with the locally affine constraints could be used in the future for other computer vision tasks, where the warping of one object into another is desired, such as tracking or optical flow estimation [14].

Acknowledgements. We are grateful for financial support from ERC grant VisRec no. 228180.

References

[1] A. Bordes, L. Bottou, and P. Gallinari. SGD-QN: Careful quasi-newton stochastic gradient descent. *JMLR*, 2009.

[2] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *PAMI*, 2004.

[3] Y. Chen, L. Zhu, and A. L. Yuille. Active mask hierarchies for object detection. In *ECCV*, 2010.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *ICCV*, 2011.

[6] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2011.

[8] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[10] H. Ishikawa. Exact optimization for markov random fields with convex priors. *PAMI*, 2003.

[11] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[12] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[13] L. Ladicky. Global structured models towards scene understanding. *PhD thesis, Oxford Brookes University*, 2011.

[14] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008.

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[16] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[17] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, 2007.

[18] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial truncation. In *NIPS*, 2009.

[19] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2004.