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

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 [1] 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*. [2] proposed a star-graph part based model allowing a predetermined number of rigid parts to change their relative location with respect to the centre of the object. 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.

The classifier for our deformable template then takes the form :

$$H(\mathbf{c}) = \max_{\mathbf{d}\in\mathcal{A}} \left(\mathbf{w}^* \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) + b^* - R(\mathbf{d}) \right), \qquad (1)$$

where $\mathbf{c}$ is the set of cells, $\mathbf{h}(D^{\mathbf{d}}(\mathbf{c}))$ are the histograms of oriented gradients on the template deformed by the deformation field $\mathbf{d}$, $R(\mathbf{d})$ is the regularisation cost taking the form of the pairwise Markov Random Field (MRF) and $\mathcal{A}$ is the set of locally affine deformation fields 1, in which any 2×2 neighbouring cells transform into a parallelogram.

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

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

where M is the set of training samples and $z^k \in \{-1, 1\}$ is the label of the k-th training sample. This problem is non-convex, however, we can follow the same approach as [2] and iteratively estimate the weight vector $\mathbf{w}$ with the bias b, and the deformation field $\mathbf{d}$ for each training sample.

The problem of finding the optimal weight vector $\mathbf{w}$ and bias b given estimated deformation fields for each training sample is a standard SVM problem and can be solved with any standard SVM algorithm. The problem of finding the optimal deformation field given weight vector is the max-a-posteriori (MAP) estimation of the pairwise MRF problem under the additional locally affine deformation field constraints. 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. 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 $\Delta^r d_j = (\Delta^r d^x_j, \Delta^r d^y_j)$ and the second in which we move each column i by a deformation $\Delta^c d_i = (\Delta^c d^x_i, \Delta^c d^y_i)$. 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. Both of these subproblems can be solved exactly using dynamic programming.

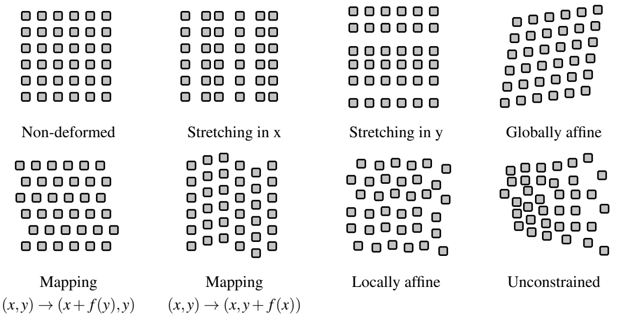

Figure 1: Expressive power of the locally affine deformation field. The locally affine constraints allow for stretching or mapping of the template in both axes, global affine transformation of the template or the combination of all of them resulting in the general locally affine transformation, in which any 2×2 neighbouring cells transform into a parallelogram.

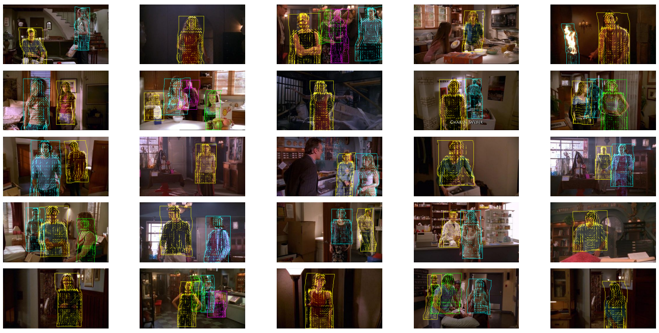

Figure 2: Typical results on the Buffy data set. Positive detections are overlaid with the learnt HOG template of the corresponding model, deformed by the deformation field.

Typically, different viewpoints are modelled by splitting the positive samples based on the aspect ratio and trained independently for each aspect ratio. However, this approach could not model different poses with similar bounding boxes independently and the star-graph model (or alternatively our locally affine deformation field) could not capture this kind of deformations. We can take an advantage of our deformation field model and cluster the problem into subproblems based on the similarity of training samples, defined as their scalar product, regularised by the MRF cost of the deformation field which transforms one training sample to another.

We tested our method on the more challenging Buffy data set of [3], 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. Our method significantly outperformed other state-of-the-art approaches [1, 2]. We assume, our locally affine deformation field formulation could be used in the future for other computer vision tasks, such as tracking or optical flow estimation.

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

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

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