Training Deformable Object Models for Human Detection Based on Alignment and Clustering

Benjamin Drayer and Thomas Brox

Department of Computer Science,
Centre of Biological Signalling Studies (BIOSS),
University of Freiburg, Germany
{drayer,brox}@cs.uni-freiburg.de

Abstract. We propose a clustering method that considers non-rigid alignment of samples. The motivation for such a clustering is training of object detectors that consist of multiple mixture components. In particular, we consider the deformable part model (DPM) of Felzenszwalb et al., where each mixture component includes a learned deformation model. We show that alignment based clustering distributes the data better to the mixture components of the DPM than previous methods. Moreover, the alignment helps the non-convex optimization of the DPM find a consistent placement of its parts and, thus, learn more accurate part filters.

1 Introduction

Much variability among images of persons is due to viewpoint and deformation/articulation. This variation makes it hard to pick discriminative features. Exemplar based classifiers are less affected by variation, as they learn the samples by heart, but they do not generalize well. The deformable part model (DPM) [9] has introduced two complementary concepts to deal with variation: mixture components address very different viewpoints and deformable parts can handle smaller viewpoint changes and articulation. Mixture components and a hierarchical part structure are also used in other recognition models, such as convolutional neural nets [16]. With their deep hierarchy of mixture components, they implement these concepts even more rigorously. However, the idea to also model typical deformations by including deformation costs is unique to the deformable part model.

As viewpoint labels and the part placement are not given in typical training sets, these need to be inferred in conjunction with the classifier training. This makes the training procedure a rather tough non-convex optimization problem, where a good initialization is crucial. Felzenszwalb et al. [9] address this by using the bounding box aspect ratio to initialize mixture components. Others, building on [9], suggested clustering the HOG descriptors [1, 5, 12]. In all these cases, parts are initialized at high energy positions of the root filters' positive weights. Initially, they do not account for deformation/movement within the

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 406–420, 2014.

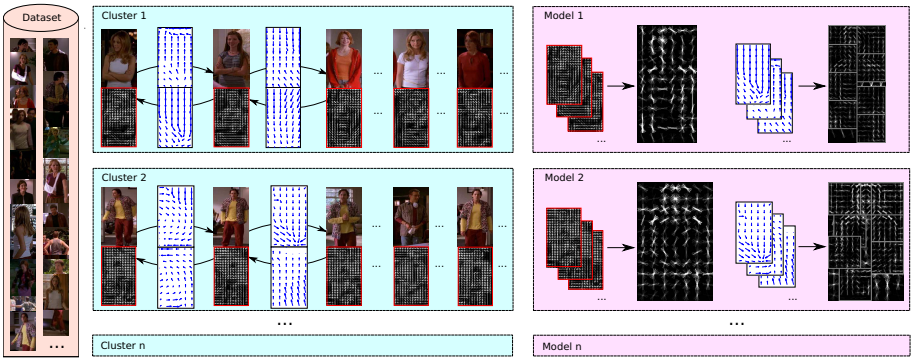

Fig. 1. On a training dataset (left), we compute alignment vector fields between all pairs of samples. This allows us to cluster the samples based on alignment-normalized similarities (middle). These clusters serve as mixture components for the deformable part model (right). The non-aligned samples in a cluster serve to train the root filter. Thanks to the alignment, we know a good initial placement of parts, which leads to more detailed part filters.

training samples, which leads to very blurred part filters in the first training iterations.

In this paper, we suggest running an alignment procedure on the training samples. This has two positive effects: (1) We enhance the clusters by enforcing that all samples in a cluster are similar up to a "regular" deformation, which directly results in stronger mixture components. (2) The initial part placement is improved.

In particular, we allow for deformations that can be well represented by the deformable model. We use distances for clustering that reflect this space of deformations better than typical distances defined on non-aligned samples. As a consequence, we obtain mixture components that generalize better over deformations while the classifier can learn more detailed structures that are specific to the respective component.

Additionally, the alignment allows us to initialize the part placement, since it tells us where the part should be placed in each training sample. Hence, the optimization of the DPM can train quite distinct part filters already in the first iteration. As a consequence, the final part filters can capture more detailed structures. An overview of our method is illustrated in Figure 1.

2 Related Work

Appearance based clustering of training data in the context of the deformable part model by Felzenszwalb et al. [9] was proposed by Gu et al. [12]. The bounding box aspect ratio is supplemented by a distance on the HOG descriptors as a criterion to define the mixture components. A pure appearance based clustering was proposed by Divalla et al. [5]. Clustering was also used in conjunction with

simple template based models, e.g., in Aghazadeh et al. [1] and Hariharan el al. [13]. An extreme variant, where each training sample defines its own mixture component, was proposed in [18]. Based on that, Dong et al. [6] use a combination of appearance and shape, where shape is obtained from the respective Exemplar-SVM.

Another line of work defines mixture components with the help of additional supervision. This has been proposed in Zhu et al. [22, 23], where landmark annotation, respectively human clustering, is used. Two other approaches that fall into this category are the poselets of Bourdev et al. [3, 11] using keypoint annotation and Azizpour et al. [2], where object parts are annotated in the training images.

Alignment of training examples was also considered in the work of Gu et al. [11]. For each manually selected cluster representative, its 32 nearest neighbors are added to form a cluster. Both, their and our approach, align the examples with respect to the cluster representative. Different from their approach, we select the representatives automatically and apply an unsupervised, non-rigid alignment, whereas Gu et al. employ a transformation matrix optimizing the Procrustes distance between the keypoints.

In terms of the overall framework, the work of Ladicky et al. [17] is most related to ours, as it is the only one that combines the definition of mixture components with an unsupervised alignment procedure. Ladicky et al. rely on a locally affine model, which allows efficient optimization of the alignment variables in the structured SVM. In contrast, we have a more general non-rigid deformation model for clustering and provide a strong initialization for the star-model of Felzenszwalb et al. [9].

3 Alignment

Although HOG [4] is robust to some local deformation, clustering in HOG space generally cannot deal with larger image transformations and deformations. For example, already small rotation of the object in an image makes clustering fail. To address this problem, we use a distance we proposed in [7] that normalizes out all spatial transformations including non-rigid deformations and considers the similarity of the aligned HOG features as well as the deformation energy. In the scope of detection, we benefit from the alignment twice: The clusters improve and within a cluster we obtain correspondences for the various object parts.

For each pair of examples we aim for the optimum deformation field that aligns one example to the other. The cost function consists of the data term E_D, that aims for maximum feature overlap and a pairwise regularization term E_P, that penalizes strong deformations:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + E_P(\mathbf{u}) \tag{1}$$

This cost function is minimized with respect to the deformation field $\mathbf{u}$. The globally normalized version of the HOG features $F/\left\|F\right\|_2$ allows us to compare the alignment energies of different image pairs.

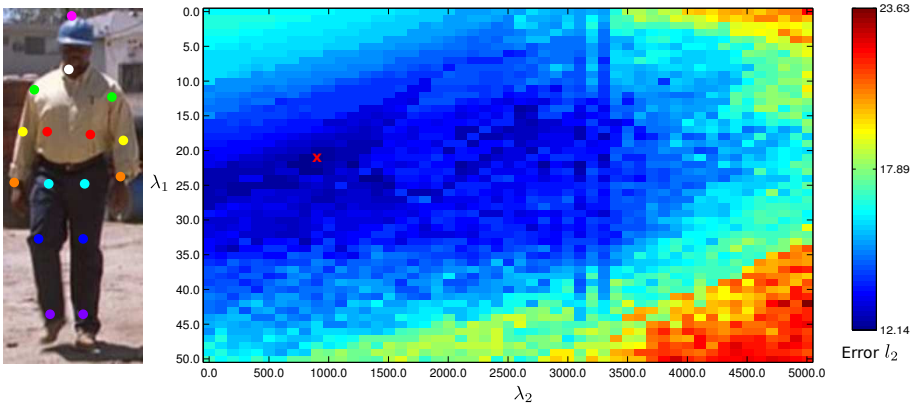

Fig. 2. Optimization of the weight parameters. **Left:** Exemplar of the given keypoint annotation. **Right:** l_2 distance between the corresponding keypoints after alignment with parameters λ_1 and λ_2, averaged over the image pairs. The optimum of 12.14 is achieved with $\lambda_1 = 21$ and $\lambda_2 = 900$ (marked in red). The l_2 distance before alignment is 16.29.

3.1 Data Term

We use a weighted combination of the l_1-norm and the dot product, as we introduced in [7]. The intuition behind this choice is to benefit from the robustness of the l_1-norm and the capability of the dot-product to match features of different magnitude. A grid search on the weighting parameters λ_1 and λ_2 shows that the best results are achieved by a combination of both distances; see Figure 2. One may expect this result, because the l_1-norm has problems matching features of different magnitude, whereas the dot product tends to many-to-one correspondences. The data term reads:

$$E_D(\mathbf{u}) = \sum_{\mathbf{x}} \lambda_1 \left| F_2(\mathbf{x}+\mathbf{u}(\mathbf{x})) - F_1(\mathbf{x}) \right|_1 - \lambda_2 \langle F_2(\mathbf{x}+\mathbf{u}(\mathbf{x})), F_1(\mathbf{x}) \rangle, \quad (2)$$

where $\mathbf{x}$ are the coordinates of the grid points and $F_1(\cdot)$, $F_2(\cdot)$ denote the feature representation of the images being aligned by $\mathbf{u}$. The influence of the deformation cost E_P is implicitly handled by the weighting parameters λ_1 and λ_2.

We select these parameters automatically using the keypoint annotation provided in the Buffy training set, which coincide with the human joints; see Figure 2. The usage of keypoints in our work is restricted to the optimization of the weighting parameters. For the non-rigid alignment, only HOG features are used.

In order to select the parameters, we take n pairs of images $(I_1, J_1), ..., (I_n, J_n)$ and the corresponding keypoint pairs $(p_1, q_1), ..., (p_n, q_n)$. With $u_{\lambda_1\lambda_2}^{(i,j)}$ we denote the alignment between image pair (I_i, J_i) under the parameters λ_1, λ_2, where the optimal parameters correspond to the alignment that minimizes the l_2-distance

between corresponding keypoints:

$$\underset{\lambda_1,\lambda_2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\| q_i - u^{(i,j)}_{\lambda_1\lambda_2}(p_i) \right\|_2 . \tag{3}$$

For a set of $n = 20$ pairs, we have an average l_2-distance of 16.29 without alignment. The best result is achieved with $\lambda_1 = 21, \lambda_2 = 900$, as shown in Figure 2, yielding a distance of 12.14.

3.2 Deformation Cost

The deformation cost is defined as the total variation of the deformation field $\mathbf{u}$:

$$E_P(\mathbf{u}) = \sum_{\mathbf{x},\mathbf{y}\in\mathcal{N}(\mathbf{x})} |\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{y})|_1 , \tag{4}$$

where $\mathcal{N}(\mathbf{x})$ denotes the neighborhood of $\mathbf{x}$. In our experiments, we use a 4-connected neighborhood. The total variation regularization prefers piecewise constant deformation fields and allows for discontinuities in the deformation field. This is necessary for handling the typical challenges of the dataset, e.g. raising arms, change in viewpoint and occlusion of body parts.

The resulting optimization problem can be solved efficiently with the Fast Primal-Dual solver of [14, 15]. On average, the alignment of a pair takes 0.11 seconds. The result is an approximation, but in practice it is very close to the global optimum [1].

4 Clustering

4.1 Pairwise Distances and Spectral Clustering

We directly use the energy E in (1) to define pairwise distances for clustering, which includes both the matching cost and the deformation cost. Figure 3 illustrates that the nearest neighbors more often contain the same instance or other similar instances if alignment is taken into account.

Based on the pairwise distances we apply spectral clustering [19, 21], by constructing the affinity matrix A in the following way:

$$A(i,j) = e^{-\frac{E(i,j)}{2\sigma^2}}, \tag{5}$$

with $\sigma = 0.7$. The alignment procedure and the derived distances are most informative for small distances. In case of large distances, a good alignment cannot be found, which indicates that the samples do not match. Therefore, we only keep the affinities of the 20 nearest neighbors of each sample and set all other affinities to zero. This ensures sufficient connectivity and keeps the graph of the dataset from splitting into tiny clusters, while at the same time a strong

[1] This can be read from the lower and upper bounds computed during optimization.

Fig. 3. Nearest neighbor queries showing query image **Q** and its 10 nearest neighbors. The 1st and 3rd row show the query result using HOG features without alignment. The 2nd and 4th row use alignment energy E as distance and show the nearest neighbors warped to the query image. The alignment based distance returns more similar instances, demonstrating that it is more invariant to deformations than simple HOG based distances.

preference is given to the most similar pairs. Since the 20 nearest neighbors are not the same for a pair of samples and the alignment is not enforced to be a diffeomorphism, the affinities are not symmetric. Thus, symmetry of the final affinity matrix is enforced by using $A + A^t$.

4.2 Clustering Performance

In order to evaluate the clustering performance and to justify some design choices quantitatively, we add some additional annotation to the training data of the Buffy dataset. Manually specifying unique ground truth clusters is impossible. There are many cases in which even humans do not agree on whether an object belongs to one or another cluster and setting the right number of clusters is even more difficult. Hence, rather than specifying clusters, we label pairs of examples by assigning them to one out of three categories. The first category comprises the pairs that are clearly similar and should end up in one cluster. The second category contains all pairs that are clearly different and should end up in different clusters. The last category is 'unknown' that takes all pairs, for which an assignment to one of the two other categories is difficult to make.

Based on this annotation, we can compute true and false positives, as well as true and false negatives. In clustering, the Jaccard index, the Rand index and

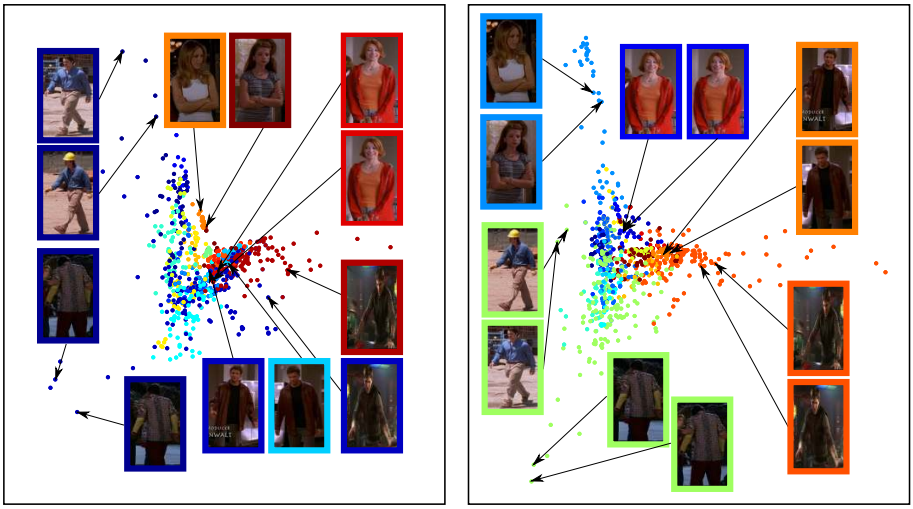

Fig. 4. Qualitative comparison between clustering in HOG space (no alignment) **(left)** and alignment based clustering **(right)**. For the visualization we used a 2D embedding (obtained with multidimensional scaling [20]) of the corresponding affinity matrices. Different colors indicate different clusters. Without alignment, some instances that should belong to the same cluster, e.g., the two cross-armed women, the men at the table, and the upper frontal bodies, end up in different clusters. With alignment similar examples are mapped closer together, e.g., the men walking to the left and the men from behind.

F-score are frequently used quality measures obtained from these values. However, one has to decide on the number of clusters using these measures, which is hard. In our scenario, it is more appropriate to compute precision and recall for the clustering. This way we do not need to fix the number of clusters and the average precision (AP) serves as measure for the overall quality.

As the number of pairs is quadratic in the number of examples, the effort to annotate all pairs is too high. However, we can resort to random sampling and the effect of large numbers. Labeling m pairs of images resembles the true distribution with a maximum deviation of

$$\varepsilon = z_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{m}}. \tag{6}$$

Here, $z_{(1-\frac{\alpha}{2})}$ is the z-quantile of the normal distribution and s denotes the standard deviation. Due to the underlying Bernoulli distribution with $s^2 = p(1-p)$, the upper bound of the standard deviation is 0.5. We manually labeled 4000 pairs of images and considering a confidence of 95% we obtain a maximum deviation of $\varepsilon \leq 1.6\%$. This means the computed AP for the clustering is the center of a confidence interval with length ≤ 3.2.

We evaluate the affinities of Equation 5 by varying the exponent corresponding to the distance. Namely, the weighted combination of l_1-norm and dot product,

Table 1. Average precision on the clustering task with and without alignment. Alignment improves the clustering AP by 4%. The deformation cost E_P alone is not sufficient for clustering.

E_D, $\mathbf{u} = 0$ (without alignment)	43.62
E_D	47.88
E_P (deformation cost only)	19.72
$E = E_D + E_P$	**48.04**

Table 2. Comparison of the DPM clustering and ours for a fixed number of clusters. The left column shows the F-measure of the initial clusters and the right column after reassignment by the mixture model. We evaluated the DPM with 3 and 5 aspect ratios, corresponding to 6 and 10 clusters. Our alignment based clustering yields stronger clusters than DPM.

	At initialization	After DPM training
DPM $K = 6$	0.3831	0.5012
DPM $K = 10$	0.4664	0.5013
$E = E_D + E_P$, $K = 10$	0.5251	**0.5308**

the energy of the data term E_D, the smoothness term E_P and the total energy of the alignment E. We achieve an improvement of more than 4% AP using the energy of the alignment instead of the unaligned features, see Table 1. Despite having an uncertainty of 1.6% this improvement is statistically significant.

We also evaluate how well the internal optimization of the mixture components of the DPM perform on the clustering task. The DPM first clusters the samples based on the bounding box aspect ratio and then splits each group into a so-called left- and right-facing cluster based on appearance. We evaluate 6 and 10 clusters (corresponding to 3 and 5 aspect ratios) and report the performance of the initial clusters and the final ones (after running the full DPM training). Since in this experiment the number of clusters is fixed and we only know a single precision-recall point, we compare performance based on the F-measure. As one may expect, Table 2 shows that optimization of the mixture components by the DPM improves the clustering performance compared to the initial cluster assignment. Therefore, the proposed clustering using non-rigid alignment E yields better clusters than the DPM.

A qualitative comparison of the alignment based clustering is given in Figure 4. The alignment based distance provides cleaner clusters than the purely appearance based distance. This is because the non-rigid alignment better deals with deformation in the data and hence provides more meaningful matches (see also Figure 3).

5 Training the Deformable Part Model

Given n training examples $(x_1, y_1), ..., (x_n, y_n)$ with $y_i \in \{-1, 1\}$, the DPM [9] tries to infer the assignment $z_i \in \{1, ..., K\}$ of a sample i to one of K mixture

components, the hyperplane parameters w_{kj}, b_{kj} of a linear SVM for each mixture component k and part $j \in \{1, ...M\}$, and the deformation parameters θ_{kj} in a joint optimization process. The objective is highly non-convex. To initialize the model, [9] first train a model with just the root filters and no parts. The root filters then serve as an initialization for the parts.

5.1 Mixture Components

Without parts, the training objective becomes that of a mixture of linear SVMs:

$$\begin{aligned} \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{k=1}^{K} \|w_{k0}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i, \\ y_i \cdot s_i^{z_i} \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0, \\ z_i = \underset{k}{\operatorname{argmin}}\, s_i^k, \\ s_i^k = w_{k0}^\top F(x_i) + b_{k0} \end{aligned} \tag{7}$$

with regularization parameter C and slack variables ϵ_i. $F(x_i)$ denotes the feature representation (HOG) of sample i. The optimum parameters are estimated by alternating optimization of SVM parameters w_{k0}, b_{k0} and the latent assignment variables z_i. In fact, this strongly resembles clustering in an expectation-maximization style. So even without parts there is a strong dependency on the initialization.

In [9], the initial assignments z_i are based on the bounding box aspect ratio. Moreover, each cluster is split into left- and right-facing instances based on similarity of the HOG descriptors. We replace these initial mixture components simply by the clusters from the previous section.

5.2 Part Filters

[9] derives initial part filters from the root filters. The root filters are upsampled to the resolution of the part filters and parts are placed such that most of the positive weights of the root filters are covered. Clearly, no deformation is considered for the initial placement of the parts. This is left to the overall optimization over all model parameters.

We propose to use the non-rigid alignment from Section 3 to initialize the relative positions of the parts. It is easy to see that this can be reduced to the initialization procedure in [9] but using the clusters from Section 4 and warping all samples within a cluster k to one representative sample r_k of that cluster. This is possible, since for each pair of samples the respective deformation field $\mathbf{u}$ is available. The representatives r_k are selected as the samples with the lowest intra-cluster distance.

We train auxiliary root filters as in Section 5.1 but using the warped samples and high resolution HOG features (the same resolution as the part filters). We call these root filters *accurate*. Their only purpose is to obtain improved initial part filters, which again can be set by covering most of the positive weights.

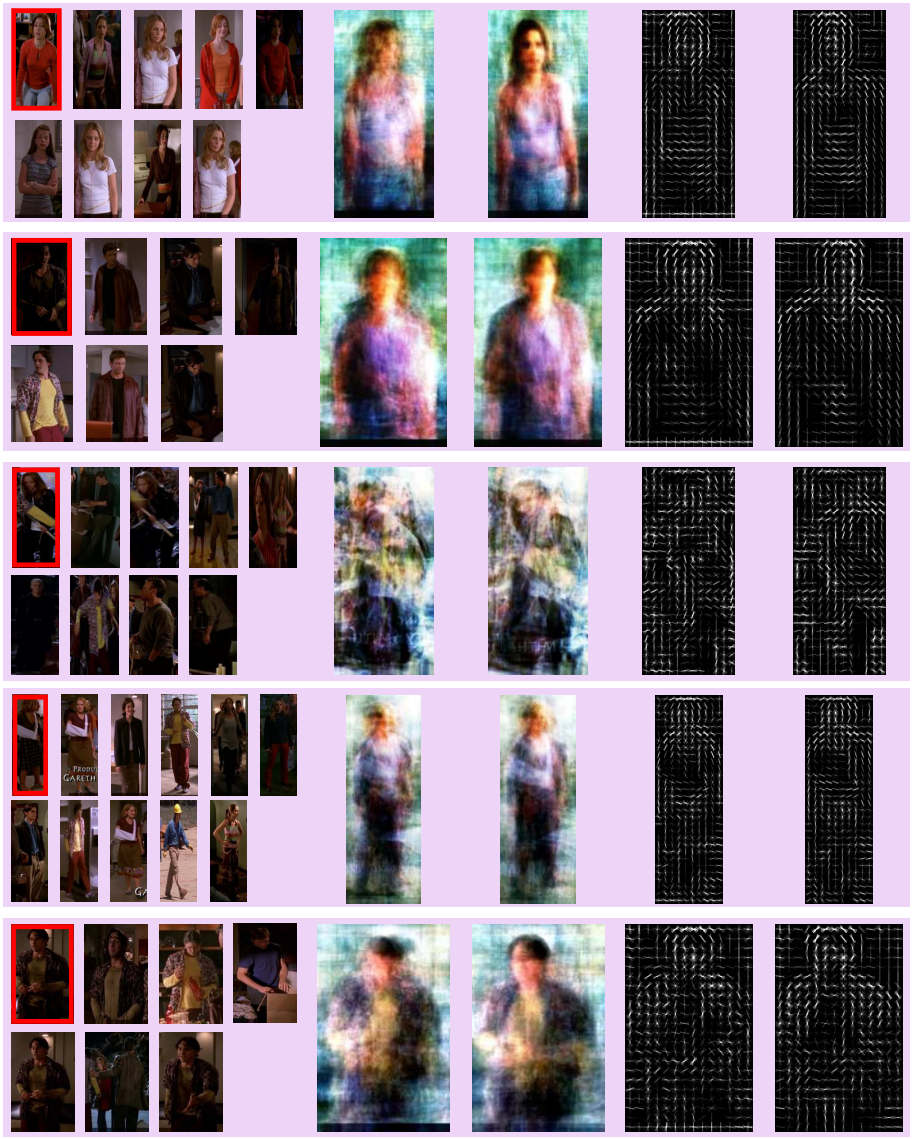

Fig. 5. Part initialization with accurate auxiliary root filters shown for 5 out of 10 mixture components. **Left:** Samples from the clusters, red marks the representative sample r_k. **Middle:** Overlay of all samples with (right) and without (left) alignment (images are histogram-normalized for better visualization). **Right:** Trained auxiliary root filter with (right) and without (left) alignment. Especially for the two top-most and the bottom cluster alignment leads to more detailed filters. In the head and shoulder region this is visible particularly well.

Table 3. Average precision (AP) on detection for various approaches on the Buffy dataset. Compared to the DPM [9] (with 3, 6 and 10 mixture components) we gain about 2% AP by clustering (DPM+c). Improving also the part filters (DPM+a) leads to another increase of 3% AP.

HOG $K = 3$	DPM $K = 3$	HOG $K = 6$	DPM $K = 6$	DPM $K = 10$	HOG clustering	Ladicky [17]	DPM+c	DPM+a
50.28	72.91	73.4	79.39	78.04	80.95	76.03	81.56	**84.57**

As shown in Figure 5, the accurate root filters are visibly more detailed since the warped samples agree on the same location of the most important structures of that filter. These details transfer to the initial part filters. We note that the higher accuracy of the root filters is due to the alignment procedure. Running the standard procedure just at a higher resolution has hardly any effect, since the local variation of the samples makes it impossible to learn consistent filters. The average sample image in Figure 5 clearly shows less detailed structures before warping than after warping.

After initialization, we combine the resulting part filters with the standard root filters (trained without aligning the samples to a reference sample). We denote this approach as DPM+a. The same approach with the standard part filter initialization is denoted as DPM+c. Both approaches run the final joint optimization over all parameters as in [9].

6 Experimental Evaluation

6.1 Dataset and Evaluation Method

For evaluation we use the Buffy dataset from [10] and PASCAL VOC 2007 [8]. The first was used in the closely related LADF detector by Ladicky et al. [17]. The dataset contains strong variation in illumination and truncated and occluded persons. It is composed of scenes from episodes $2 - 6$ of the fifth season of the TV-series "Buffy the Vampire Slayer". Since the dataset is based on videos and there are multiple samples of the same instance at different poses in the training set, it supports transitions between samples, whereas samples in datasets like PASCAL VOC are much harder to link. As in [17], we use $s5e2$, $s5e5$, $s5e6$ for testing and $s5e3$, $s5e4$ for training and validation. The training and test sets contain 276 and 472 images, respectively.

As usual, a detection is counted as positive if the intersection over union ratio with respect to ground truth bounding box is greater than 0.5. We report the average precision (AP) over the test set.

6.2 Results

Table 3 compares the AP of alignment based clustering (DPM+c) and initial part placement (DPM+a) against the classical DPM (with and without parts)

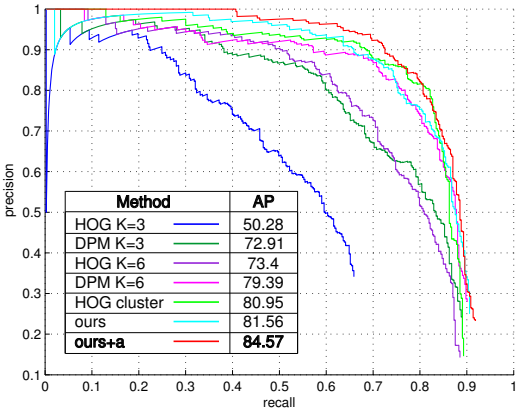

Fig. 6. Precision-recall curves of our method (red) on the Buffy dataset. Baselines are DPM without (dark blue and dark green) and with left-right-splitting (violet and magenta). We clearly see that appearance based clustering improves the detection results and that the alignment improves the results twice: first by optimizing the clustering, which results in a better detection performance and second, by the better part placement, which gives the major improvement.

and the LADF detector by Ladicky et al. [17][2]. The DPM model uses 3 aspect ratios, with and without left-right-splitting, resulting in $K = 6$ and $K = 3$ mixture components. For our approach, we used $K = 10$ components, which corresponds to the K components of the LADF detector. Figure 6 shows the precision-recall curves of methods for which code was publicly available. Typical detection results of our method are shown in Figure 7.

In the previous sections, we discussed the importance of good clustering as initialization for the mixture model learning. The alignment based clustering as a starting point for the mixture components (DPM+c) has indeed a positive effect on detection with the DPM. It also slightly improves over clustering without alignment indicating that the improvement is not only due to the larger number of mixture components[3]. The main boost in performance is due to the better part initialization, included in DPM+a, which is due to the accurate root filters trained on warped samples. In total, we get an improvement of more than 5% AP compared to the DPM.

[2] The reviewing process revealed some inconsistencies between the dataset used in [17] and the one that is available for download. For the public dataset a few corrections in the annotation have been made, which is why the direct comparison to [17] should be taken with a grain of salt. The comparison to DPM is fair and the DPM results reported in [17] can be almost exactly reproduced when using DPM version 3 or the newer version with left-right-splitting turned off.

[3] More mixture components are not necessarily advantageous because a larger number of components leads to a smaller number of samples to train each component.

Table 4. Average precision (AP) on detection for the DPM ($K = 6$) baseline, DPM+c and DPM+a on the PASCAL VOC 2007 test set. Classes with less variation or a denser sampling e.g. aeroplane improve with the alignment, whereas classes with larger variability, such as cat, are hard to align and performance drops.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
DPM [9]	28.9	**59.5**	**10.0**	15.2	**25.5**	49.6	**57.9**	**19.3**	**22.4**	25.2
DPM+c	29.7	58.2	9.7	16.3	22.9	**50.3**	52	14.8	18.9	27.9
DPM+a	**33.2**	57.4	9.7	**16.9**	25.0	48.6	52.3	13.3	20.2	**30.3**
	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
DPM [9]	23.3	**11.1**	56.8	48.7	**41.9**	12.2	17.8	**33.6**	45.1	**41.6**
DPM+c	24.9	10.3	57.2	48.7	36.8	**12.9**	17	24.1	45.8	40.9
DPM+a	**26.6**	6.5	**60.1**	**49.1**	38.4	9.8	**18.7**	29.7	**47.3**	39.8

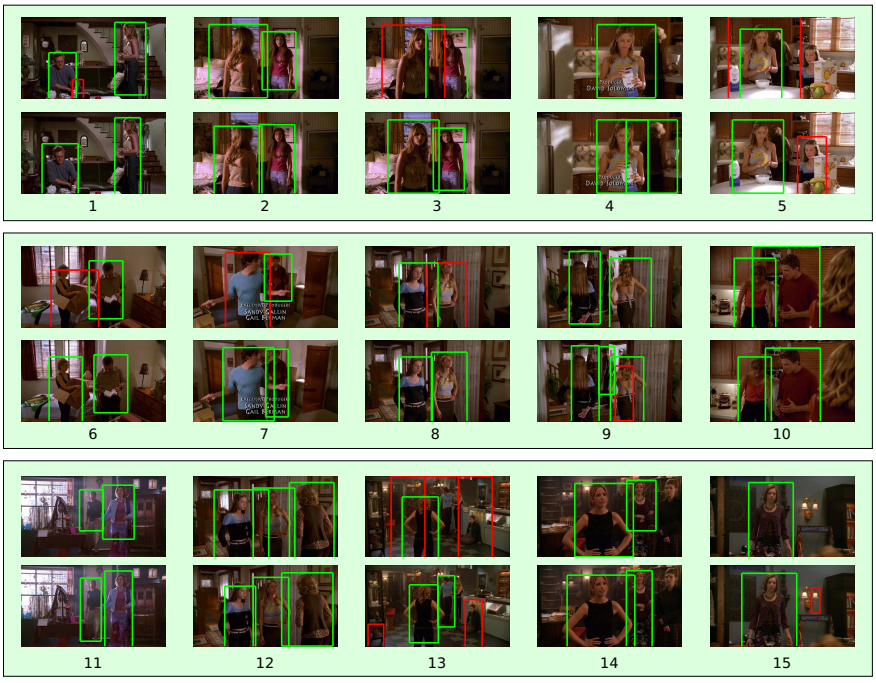

Fig. 7. Qualitative comparison between the detections of our model DPM+a(bottom rows) and the DPM (top rows). Green boxes indicate correct detections with an intersection over union ratio > 0.5; red boxes indicate false detections. With DPM+a, we capture a wider range of variation that does not only manifest in more detected people (ex 4, 7, 8, 12), but also in more precise bounding boxes (ex 6, 10, 11, 14). Both methods fail in case of highly occluded or truncated people, as in ex 5 and 10. Instances with few or no training data, as the sitting person in ex. 13, cannot be handled. These would require additional mixture components with corresponding additional training data.

The results of the object detection on PASCAL VOC 2007 are shown in Table 4. Both approaches DPM+c and DPM+a heavily depend on the alignment. If the specific class is too diverse in pose and appearance, such as for cat, dog, plant and person, alignment is too hard and even deteriorates the clustering and the DPM training. On the other hand, subcategories with a strong connectivity among the samples, such as for aeroplane, cow and horse, the DPM benefits from both clustering and alignment. We believe that the alignment is most beneficial when training samples from video showing the same instance in different poses are involved. Apart from the Buffy dataset, few data of that sort is yet available.

7 Conclusions

We have presented a new clustering method based on a pairwise non-rigid alignment. In the experiments, we have shown that such a strategy is most reasonable for datasets that allow for clear correspondences within subcategories such as in videos. Detailed analysis on the Buffy dataset has shown that the alignment improves the performance directly in terms of clustering AP

and indirectly by obtaining better detection results. Furthermore, we have demonstrated that alignment can help initialize the part placement in the deformable part model.

It is worth noting that we finally optimize the same energy model as the DPM. Apart from increasing the number of mixture components from $K = 6$ to $K = 10$ the model has not changed. The improvement is only due to a better initialization. This indicates that complex detection approaches which optimize many mutually dependent parameters can benefit from stronger optimization methods and sophisticated initialization strategies.

Acknowledgements. This study was supported by the Excellence Initiative of the German Federal and State Governments (EXC 294) and by the ERC Starting Grant VIDEOLEARN.

References

1. Aghazadeh, O., Azizpour, H., Sullivan, J., Carlsson, S.: Mixture component identification and learning for visual recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 115–128. Springer, Heidelberg (2012)
2. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 836–849. Springer, Heidelberg (2012)
3. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 168–181. Springer, Heidelberg (2010)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893. IEEE Computer Society (2005)

5. Divvala, S.K., Efros, A.A., Hebert, M.: How important are "Deformable parts" in the deformable parts model? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part III. LNCS, vol. 7585, pp. 31–40. Springer, Heidelberg (2012)
6. Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., Yan, S.: Subcategory-aware object classification. In: CVPR, pp. 827–834. IEEE (2013)
7. Drayer, B., Brox, T.: Distances based on non-rigid alignment for comparison of different object instances. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 215–224. Springer, Heidelberg (2013)
8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge (VOC2007) Results (2007), `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
10. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
11. Gu, C., Arbeláez, P., Lin, Y., Yu, K., Malik, J.: Multi-component models for object detection. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 445–458. Springer, Heidelberg (2012)
12. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 408–421. Springer, Heidelberg (2010)
13. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 459–472. Springer, Heidelberg (2012)
14. Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(8), 1436–1453 (Aug 2007)
15. Komodakis, N., Tziritas, G., Paragios, N.: Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. Computer Vision and Image Understanding 112(1), 14–29 (2008)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1106–1114 (2012)
17. Ladicky, L., Torr, P.H.S., Zisserman, A.: Latent svms for human detection with a locally affine deformation field. In: BMVC. BMVA Press (2012)
18. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV, pp. 89–96. IEEE Computer Society (2011)
19. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856. MIT Press (2001)
20. Seber, G.: Multivariate observations. Wiley (1984)
21. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 888–905 (1997)
22. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR, pp. 2879–2886. IEEE (2012)
23. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C.C.: Do we need more training data or better models for object detection? In: BMVC. BMVA Press (2012)