Deconstructing the Deformable Parts Model: Do More with Less

Brigit Schroeder[1], Baochen Sun[1], Kate Saenko[1], Karim Ali[1,2]

[1]University of Massachusetts, Lowell [2]University of California, Berkeley

Abstract. The deformable parts model (DPM) [1] is a successful detection model that continues to achieve state-of-the-art results in object detection. While the model's success is attributed to the deformable parts, the proposed system has many other design choices such as the use of multi-resolution HOG features, the sampling of parts from high energy areas, the use of a mixture model, etc. Recently, there have been attempts at analyzing the contribution of these design choices more closely [2]. We delve further into this analysis, performing a study of the effects of restricting the deformation model, the use of single-resolution filters versus multi-resolution filters, and the application of energy "dropout" while sampling parts. Our results indicate that a better performance can be achieved with simpler cost-free deformation models.

1 Introduction

In [3], Dalal and Triggs proposed using Histograms of Oriented Gradients (HOG) orientations as feature descriptors for human detection. The results from this work showed that HOG features were reasonably successful at detecting humans in complex backgrounds with variable appearance and lighting conditions. While a HOG detector will generally capture a static shape appearance, significant variations in appearance and pose are more challenging.

Felzenswalb, et al. [1] sought to address this with the introduction of a mixture model to deal with strong pose variations (e.g. a horse head vs. a side-view of a horse) and deformable parts to deal with a range of articulated poses (e.g a side-view standing horse vs. a side-view rearing horse). Each component in the mixture model uses a low-resolution root filter (based on the Dalal and Triggs model [3]) and high-resolution part filters, to be able to detect objects in a range of articulated poses. Despite the model's popularity, few attempts have been made at understanding which design choices contribute most to performance.

More recently, in [2], Divvala, et al. made the suggestion that part deformation in DPM may not be as necessary as originally thought, where it can be "turned off" and still yield detector performance comparable to the original DPM. Instead of utilizing several deformable parts, a single part encompassing the full extent of the object was used. The reported results indicate that deformations are not as critical to performance as assumed.

We build on this work by carefully performing a set of experiments which analytically deconstruct the DPM to understand better how its design choices contribute to overall performance. In this work, we have observed a progression of experimental results which give deeper understanding to which factors contribute the DPM performance and to what degree. We show that better results can be obtained by simply allowing parts to deform within a neighborhood at no cost.

2 Analysis of the DPM

We conducted our research in phases to study the different aspects of the DPM model, in order to tease out the effect of each on overall performance. The DPM models in Figure 1 for class bicycle summarize our experimental design approach, which is detailed in the following sections.

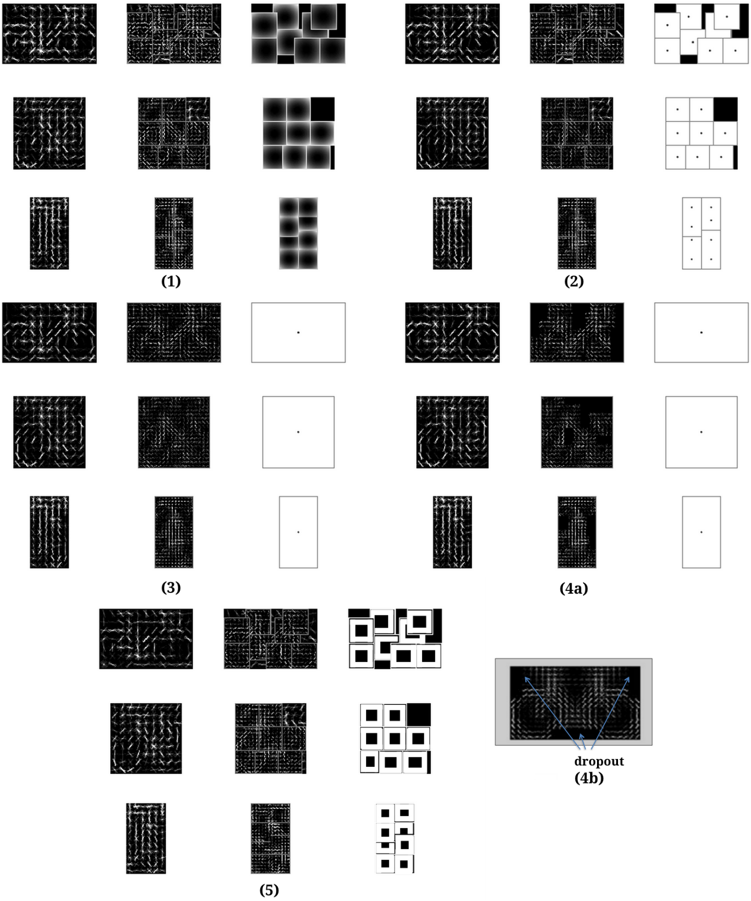

Fig. 1: Various configurations of DPM models for class bicycle used in experimentation: (1) original DPM [4] model with deformation, (2) 1X+2X with “rigid” parts (deformation removed), (3) 1X+2X, (4a) 1X+2X with dropout (shown in detail in (4b)) applied to low-energy regions, and (5) 1X+2X with “jittered” parts (parts deform within a neighborhood at no cost).

Multi-resolution Rigid Models In the primary phase of our study, we try to further the idea of “turning off’ part deformation as was done in [2] and understand the impact of deformation on overall model performance. To begin, we observed that the experiments conducted in [2] in fact do not disable deformations: instead of using several high resolution parts, the authors relied on a single high resolution part whose spatial extent matches that of the root filter yet that part is left free to deform with respect to the root[1]. We first reproduced the experiment in [2]. Next, we ran a separate experiment where the effects of deformation were entirely removed (Figure 1(3)).

Energy-Based Feature Selection We also investigated energy-based feature selection by applying dropout[2] to each model component ((4a) and (4b) in Figure 1). A dropout mask was created by greedily selecting the areas of highest energy from the component's root filter at twice resolution (with 80% coverage). Defining regions of dropout in the 2X filter mimics more closely the DPM [1] where parts cover the highest energy areas of the feature vector and the remaining areas ignored.

Single Resolution Models In this phase, we ran experiments using both a 1x (single) root filter and 2x (double) resolution root filter without any parts. The motivation here is to understand the effect of the DPM's [1] multi-scale representation on detection performance.

Bounded Cost-Free Deformation In this phase of our study, we try to further the idea of disabling the learning of a deformation model and understand the impact on overall performance. We conducted a series of experiments where the deformation parameters were not learned, but parts were rather allowed to move at no-cost by varying degrees. Specifically we define a "jitter" parameter j and allow parts to deform at no cost within their $2j+1 \times 2j+1$ neighborhood (measured in HOG bins). During both training and testing, parts simply select their optimal placement according to location of maximum response, similar to the idea of max-pooling. We allowed j to vary on $\{3, 2, 1, 0\}$. At $j = 0$, the parts essentially become rigid as they are not allowed to deform: this can be seen in Figure 1(2), where the deformation cost is set to infinity (cost is scaled black to white, lowest to highest) as opposed to the case of $j = 2$, which can be seen in Figure 1(5).

3 Results

The experimental analysis was performed using the the PASCAL VOC 2007 dataset [4] and and using the codebase from [5]. The results are summarized in Table 1. In the original DPM model (row 1), the mAP was 32.3% for the standard root and parts model (three components, eight parts) with deformation. Row 2 (1X+2X [2]) reproduces the experiment in [2] where a single high resolution part whose spatial extent matches that of the root filter is used (yet that part is left free to deform with respect to the root). We were able to confirm the reported drop of approximately 5% mAP to 27.3% mAP. Row 3 (1X+2X) shows our own version of the experiment in [2] where the deformation is bypassed as intended: the nearly identical mAP of 27% likely indicates that the unintended deformations in [2] are minimal. Row 4 (1X+2X-dropout) shows the

[1] confirmed by correspondence with the authors of [2].

[2] dropout is defined as HOG cells where features are zeroed out at train time.

result obtained by applying feature selection via dropout: performance is very similar to both rows 2 (1X+2X [2]) and 3 (1X+2X), indicating that feature dropout doesn't significantly impact performance. In each of these three experiments part movement has been restricted, demonstrating that allowing parts to move and deform bolsters overall performance.

	Experiment	aero	bike	bird	boat	botl	bus	car	cat	chair	cow	table	dog	horse	mbike	pson	plant	sheep	sofa	trn	tv	mAP
1	DPM v4 [2]	30.7	59.5	10.0	15.3	25.5	49.4	58.3	19.3	23.1	25.2	22.0	10.9	56.8	49.2	42.5	12.6	18.0	32.5	44.4	41.6	32.3
2	1X+2X [2]	25.9	46.8	9.7	13.8	18.6	42.8	39.8	13.0	17.1	21.1	16.4	10.3	55.3	42.6	36.2	11.5	16.3	27.7	42.4	38.2	27.3
3	1X+2X	25.8	52.4	5.2	14.5	17.7	41.4	51.3	13.9	16.6	20.6	16.8	2.5	53.2	41.8	35.1	11.4	15.5	25.5	42.5	37.0	27.0
4	1X+2X-dropout	25.3	53.1	9.8	14.4	17.6	40.5	51.2	14.6	15.9	21.3	16.5	3.0	53.5	42.1	35.2	10.7	15.7	24.1	42.6	35.9	27.2
5	1X	24.0	51.2	6.7	11.6	17.3	42.9	46.3	7.0	16.6	21.2	15.4	5.3	47.1	37.5	31.2	11.0	13.7	24.2	39.7	34.0	25.2
6	2X	21.5	39.1	1.8	10.0	10.2	39.7	41.7	5.9	10.5	13.6	15.1	4.9	51.7	35.3	31.0	9.9	9.6	20.3	36.9	25.9	21.7
7	1X+jitter pts-3	30.2	**60.1**	**10.3**	**15.5**	23.6	**51.7**	56.4	**21.0**	21.3	**25.2**	**29.6**	**12.0**	**56.9**	46.5	38.5	**13.4**	**19.1**	32.3	**46.0**	39.5	**32.5**
8	1X+jitter pts-2	30.0	**60.5**	**10.2**	14.3	24.5	**49.5**	57.8	**22.4**	22.8	**25.4**	**28.3**	**11.5**	**58.0**	47.7	41.8	**13.3**	**19.9**	35.4	46.3	40.0	**33.0**
9	1X+jitter pts-1	28.7	58.1	**10.3**	14.7	24.0	**49.4**	56.0	**19.5**	21.3	24.9	**23.6**	**11.1**	55.1	46.8	41.5	**12.7**	**19.0**	**33.2**	**45.5**	**42.2**	31.9
10	1X+jitter pts-0	22.1	53.1	9.5	14.9	18.7	41.9	50.5	14.6	16.3	21.4	17.5	6.5	52.0	40.8	35.6	11.1	15.7	25.6	40.8	36.8	27.3

Table 1: Deconstructing the DPM. Row 1: DPM. Row 2: Proposed method in [2] (DPM with a single high resolution part) . Row 3: Our own version of [2] (less the unintended residual deformations). Row 4: Applying dropout (keep only max energy features) to Row 3. Row 5: A single low-resolution filter. Row 6: A single high resolution filter. Rows 7-10: Standard DPM where parts are allowed to move at no cost within a local neighborhood.

The results in rows 5 and 6 clearly indicate that multi-resolution HOG models (rows 3 and 4) perform better than single resolution models, motivating DPM's current design. Interestingly, a low resolution filter (1X) outperformed a higher resolution filter (2x) by 3.5% mAP, and was only 2% lower than the multi-resolution models with deformation removed (rows 3 and 4). This result suggests that simpler low-resolution detectors suffice for certain classes, and in part, supports [2]'s finding that parts are not as significant as proposed in [1].

The third section of the Table 1 (rows $7-9$) shows the results of allowing for cost-free deformations within the $2j+1 \times 2j+1$ of each part. Allowing 1 HOG bin of jitter was only 0.4% less than the original DPM model (row 1 'DPMv4'). Allowing 2 HOG bins of jitter exceeded the original DPM result by a full 1% mAP and 3 HOG bins by 0.5% mAP. Interestingly, the performance for $j = 0$ matches that of row 3 (1X+2X).

These experiments support the idea that high resolution parts are significant if they are allowed to move, but do not require the presence of a trained deformation model. Note that Cross-validating the neighborhood jitter parameter j per category should yield an mAP of 36.2%, assuming the best parameter is selected per category: a very significant increase over the standard DPM [1].

4 Conclusions

To conclude, we have found that decoupling parts from deformation has interesting consequences, including that parts alone become somewhat inconsequential but are significant if allowed to move. More interestingly, we have found that the absence of a learned deformation model outperforms the original DPM model [1]. We have also observed a clear case for multi-resolution models over the original single resolution model of [3], regardless of deformation or parts.

5 References

1. P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010.

2. S. Divvala, A. Efros, and M. Hebert. 2012. How important are ”Deformable parts” in the deformable parts model?. In Proceedings of the 12th international conference on Computer Vision - Volume Part III (ECCV'12), Andrea Fusiello, Vittorio Murino, and Rita Cucchiara (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 31-40.

3. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on , vol.1, no., pp.886,893 vol. 1, 25-25 June 2005.

4. P. Felzenszwalb, R. Girshick, D. McAllester: Discriminatively trained deformable part models, release 4. (http://people.cs.uchicago.edu/ pff/latent- release4/).

5. Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vision 88, 2 (June 2010), 303-338.