Stacked Deformable Part Model with Shape Regression for Object Part Localization

Junjie Yan, Zhen Lei, Yang Yang, and Stan Z. Li*

Center for Biometrics and Security Research & National Laboratory
of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
{jjyan,zlei,yang.yang,szli}@nlpr.ia.ac.cn

Abstract. This paper explores the localization of pre-defined semantic object parts, which is much more challenging than traditional object detection and very important for applications such as face recognition, HCI and fine-grained object recognition. To address this problem, we make two critical improvements over the widely used deformable part model (DPM). The first is that we use appearance based shape regression to globally estimate the anchor location of each part and then locally refine each part according to the estimated anchor location under the constraint of DPM. The DPM with shape regression (SR-DPM) is more flexible than the traditional DPM by relaxing the fixed anchor location of each part. It enjoys the efficient dynamic programming inference as traditional DPM and can be discriminatively trained via a coordinate descent procedure. The second is that we propose to stack multiple SR-DPMs, where each layer uses the output of previous SR-DPM as the input to progressively refine the result. It provides an analogy to deep neural network while benefiting from hand-crafted feature and model. The proposed methods are applied to human pose estimation, face alignment and general object part localization tasks and achieve state-of-the-art performance.

1 Introduction

This paper focuses on localizing object parts from monocular image. For human and face category, this problem is often named as "human pose esti mation" or "face alignment". Accurate part localization serves as the basis of many high level applications. For example, a recent work [9] shows that directly extracting features around reliable face parts (landmarks) achieves leading face recognition performance. As surveyed in [28], human part localization can help with action recognition and human computer interaction. For general object, reliable part localization contributes to fine-grained object recognition, as proved in [46,6]. However, this problem is very challenging due to the variations in subject level (e.g., a human can take many different poses and dresses), category level (e.g., adult and baby) and image level (e.g., illumination and cluttered background).

* Corresponding author.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 568–583, 2014.

Human pose estimation and face alignment have been extensively explored for decades and achieved much progress. The critical issue is how to model the versatile spatial deformation and plausible appearance variation. The seminal work [21] exploits the pictorial structure (PS) from [23], which uses Gaussian distribution to capture the deformation of each part and constrain the relative position of interrelated parts via a tree structure. PS is improved by strong appearance representation (e.g., [17,25,29,30]), discriminative classifier (e.g., [25,44]) and powerful structure (e.g., [41,39,42,36,38,40]), and finally it becomes the leading method in localizing human parts on challenging benchmarks. DPM [20], as one of the representative works in this category, uses structural SVM training and HOG feature in pictorial structure for object detection, and it is lately extended by [44] for human pose estimation.

PS [21] and its widely used extension DPM [20,44] , however, cannot capture the global information and have limited flexibility, due to the deformation constraint by the fixed anchor location. To break the limitation of DPM, we propose a novel approach by incorporating shape regression into DPM, namely SR-DPM. Specifically, the shape regression estimates part locations using the appearance information globally. We set the regressed shape as the anchor locations in DPM and allow the deformations of parts around them to satisfy the local appearance consistency. Compared to traditional DPM, SR-DPM is of high degree of freedom to model global and local variations sufficiently. Due to the fact that shape regression and DPM can benefit from each other, we build an objective function to jointly learn them. It is a non-convex optimization problem, and we design a coordinate descent procedure to solve it.

In addition, we show that stacking SR-DPMs could further improve the performance. The complex shape variations are often beyond the representation capacity of single DPM or SR-DPM. To fully explore the data, we propose the stacked SR-DPM (S-SR-DPM), where each SR-DPM uses the output of previous SR-DPMs as the input and progressively refines the result. Note that the SR-DPMs in different layers use different parameters. The S-SR-DPM provides a natural analogy to deep convolutional neural network (DCNN) in increasing representation capacity [5]. Compared with the end-to-end learning in DCNN, the S-SR-DPM takes advantage of well designed hand-crafted pipelines and can achieve good performance with much fewer training data.

Previous works usually only consider part localization of a special category (e.g., human and face). In this paper we show wide applications of our method on human, face and general object. For human pose estimation, we conduct experiments on challenging LSP [25]. For face alignment, we use the LFPW [4] as the testbed. It terms of general object, we use the annotations [3] of animals from Pascal VOC [19]. We compare our method with different state-of-the-art methods on these three tasks and achieve the leading performance.

The rest of the paper is organized as follows. Section 2 reviews the related work. The proposed SR-DPM and its stacked form are described in section 3 and section 4. We show experiments in section 5 and finally conclude the paper in section 6.

2 Related Work

Many works on human pose estimation are based on pictorial structure in either generative or discriminative manner. The pictorial structure [21] uses Gaussian model to capture the deformation of each part and links parts by tree structure. Inference in pictorial structure is very efficient due to the dynamic programming and distance transform [21]. The pictorial structure is lately exploited in deformable part model (DPM) [20] with HOG feature and latent-SVM learning, and it achieves great success in Pascal VOC object detection. [44] extends DPM for articulated human pose estimation by adding part subtype and using part annotations in learning. [3] proves the advantage of fully supervised learning of parts over latent learning in [20] for general objects. [40] shows that automatically learning the tree is better than hand-crafted physical connections. [29] uses Poselets [7] to capture mid-level cues to latently capture high-order dependencies for pictorial structure. Many recent works improve PS in more part levels, more global models and more part models [39,42,36,38,26,33,15,31]. A very recent work [30] combines different appearance cues under the pictorial structure framework and achieves the current leading performance.

Although being similar to human pose estimation problem, face alignment field often uses very different methods, mainly due to the stronger spatial constraint of human face than human body. The most popular models include active shape model (ASM [11]), active appearance model (AAM [10]) and their extensions. Different from the Gaussian deformation of each local part in PS, ASM/AAM captures the shape deformation globally with PCA constraint. The global PCA constraint, however, has been indicated to be very sensitive and is lately extended to be constrained local model (CLM [12,34,4,2]) by a shape constraint on appearance of local parts. [47] exploits the DPM developed in [44] for joint face detection and alignment. [45] further improves the work with optimized mixtures and a two-step cascaded deformable shape model. In very recent, face alignment is taken as a regression problem [8,14,43,37], which directly learns the mapping the appearance to shape and achieves the leading performance on face alignment benchmarks and challenges (e.g., 300-W [32]). These methods, however, are sensitive to initialization, which makes them unsuitable for more difficult human and object part localization.

We stack multiple SR-DPMs, which is related to a very recent work [35]. In [35], multiple fisher vector coding layers are stacked to get a similar performance of deep neural network for image classification task. In [16], boosting is used to estimate the shape with pose-index feature, where the features are re-computed at the latest estimation of landmark localization. In [43], linear regression are stacked for face alignment.

Compared with previous works, the main contributions of this work are summarized as follows:

- We propose SR-DPM to incorporate DPM with shape regression and show how to jointly learn them. The SR-DPM is much more flexible than DPM in handling real world object deformation.

- We stack multiple SR-DPMs to increase the representation capacity, where each layer progressively refines the part locations. As shown empirically in experiments, the stacked SR-DPM is critical for better performance.
- To our best knowledge, it is the first work to simultaneously achieve state-of-the-art performance on human pose estimation, face alignment and general object part localization.

3 Deformable Part Model with Shape Regression

The DPM is composed of the root filter β_0 and some parts. Each part has a appearance filter β_i and deformation term d_i. Given an object part configuration specified by $S = [x_1, y_1, \cdots, x_N, y_N]^T$ and object location (x_0, y_0), the DPM favors some special part configurations by:

$$s(S, I) = \beta_0^T \phi_a(x_0, y_0, I) + \sum_{i=1}^{N} (\beta_i^T \phi_a(x_i, y_i, I) - d_i^T \phi_d(x_i, y_i, a_{x_i}, a_{y_i})), \quad (1)$$

where $\phi_a(x_i, y_i, I)$ is the HOG feature of the i-th part, and $\phi_d(x_i, y_i, a_{x_i}, a_{y_i})$ is the separable quadratic function to represent the deformation. $\phi_d(x_i, y_i, a_{x_i}, a_{y_i})$ is defined based on the relative location between the (x_i, y_i) and its anchor location (a_{x_i}, a_{y_i}), which is fixed after the specification of (x_0, y_0). It is straightforward to add mixture parts [44] or mixture components [20], but we leave them out to simplify the notation.

For each sliding window in localization, only the root location (x_0, y_0) is known in advance and each part location is inferred by maximizing the part appearance score minus the deformation cost associated with displacement to anchor location. Since parts are directly attached to the root, their locations are inferred independently given the fixed root by:

$$\max_{x_i, y_i} (\beta_i^T \phi_a(x_i, y_i, I) - d_i^T \phi_d(x_i, y_i, a_{x_i}, a_{y_i})), \quad (2)$$

where (x_i, y_i) traverses all possible locations of the part. The procedure can be efficiently solved by distance transform as used in [21,44].

Our improvement comes from the anchor location of each part. In DPM, the anchor location of each part is defined according to relative position of either the root [20] or its parent part [44]. It limits the flexibility since that each part can only have a small deformation around its fixed anchor location. Additionally, the star-structure used cannot capture global information, such as the high order spatial dependencies of left-arm, right-arm, left-leg and right-leg.

In this paper, we propose to use regression to estimate the anchor locations directly from the image appearance to capture the global information and increase the flexibility. After that we allow each part to have deformation based on these adaptive anchor part locations under the constraint of DPM. Let us use $\widehat{A} = [\widehat{a_{x_1}}, \widehat{a_{y_1}}, \cdots, \widehat{a_{x_N}}, \widehat{a_{y_N}}]^T$ to specify the estimated anchor part locations. Suppose the initial shape is A^0 and ground-truth shape is A^*, we always want

that each (x_i, y_i) to have relationship with all the parts initialized by S_0 (which is the mean shape) to capture the global information. The function can be very complex, and in this paper we use a simple linear function to approximate it:

$$\widehat{A} = f(A^0, I) = A^0 + W^T \Phi(A^0, I), \tag{3}$$

where $\Phi(A^0, I)$ is the local appearance feature extracted around all parts. In this paper, we define it as the HOG feature [13] from the implementation in [20]. We concatenate feature vectors of all parts specified by A^0 to be a long vector, which has Nn_d values and n_d is the length of HOG vector for a part. The dimension of corresponding regression matrix W is $Nn_d \times 2N$. In Eq. 3, each new part location is estimated based on all the initial part locations, thus Eq. 3 encodes global information which previously cannot be captured in pictorial structure based models. No parametric shape prior, such as global shape PCA in ASM and local part Gaussian deformation in pictorial structure, is assumed in Eq. 3. It has advantage especially for real world objects, whose spatial deformation can be very complex and simple parametric prior cannot describe it well.

The above shape regression, however, is not enough for object part localization. The reason is that it cannot measure the confidence of the estimated part locations, which is very important for sliding window based scanning. Additionally, the global shape regression matrix not explicitly consider the appearance consistency of regressed part location. To this end, we further use the deformable part model to incorporate shape regression, by replacing the fixed anchor location with the shape regression output $\widehat{A}$:

$$s(S, I) = \beta_0^T \phi_a(x_0, y_0, I) + \sum_{i=1}^{N} (\beta_i^T \phi_a(x_i, y_i, I) - d_i^T \phi_d(x_i, y_i, \widehat{a_{x_i}}, \widehat{a_{y_i}})) \tag{4}$$

$$where \quad \widehat{A} = [\widehat{a_{x_1}}, \widehat{a_{y_1}}, \cdots, \widehat{a_{x_N}}, \widehat{a_{y_N}}]^T = A^0 + W^T \Phi(A^0, I).$$

For each sliding window in localization, we find the S to maximize the confidence score defined above, and take it the the estimated shape configuration of the sliding window. The deformable part model with shape regression (SR-DPM) provides the flexibility to capture large variations, but it also brings challenges, since the regression matrix W and the deformable part model parameter β are all unknown. In the following part, we present the objective function for joint learning and show the optimization method.

3.1 Model Learning

The objective function for model learning is motivated by the original DPM used in object detection, which is defined as:

$$\arg\min_{\beta, S_m} \frac{1}{2}\|\beta\|^2 + C \sum_{m=1}^{M} max(0, 1 - y_m \cdot s(S_m, I_m)), \tag{5}$$

where the first term is used for regularization and the second term is the hinge loss to punish error in detection. M is the number of training samples, and S_m

is the part configuration of the m-th image I_m. $y_m = 1$ for positive and -1 for negative. In this function, only the root location of S_m is annotated, and each part location is inferred according to Eq. 2. The loss function favors the score of positive sample above 1 and score negative sample below -1. It is a standard latent SVM problem and has many off-the-shelf solvers, such as the one used in [20]. One problem in solving is that the negative number is of combinatorial explosion, and we often use a negative sample mining step to gradually add negative samples.

In our SR-DPM for object part location, we also want to ensure that the estimated part configuration specified by S_m matches the ground truth part configuration specified by S_m^*. In this way, the objective function is extended to be:

$$\arg\min_{\beta,W,S_m} \frac{1}{2}\|\beta\|^2 + C_1 \sum_{m=1}^{M} max(0, 1 - y_m \cdot s(S_m, I_m)) + C_2\|W\|^2 + C_3 \sum_{m=1}^{M_p} \|S_m - S_m^*\|^2, \tag{6}$$

where C_1, C_2 and C_3 are used to control the relative weights of different terms. $\|W\|^2$ is used to regularize the regression matrix W. The last term $\sum_{m=1}^{M_p} \|S_m - S_m^*\|^2$ is used to measure the consistency of estimated shape S_m and ground truth shape S_m^*. In this function, the S_m is estimated according to the shape regression model parameterized by W and DPM parameterized by β in Eq. 4. Since only the shapes of positive samples are of interest, the shape loss is measured only on positive samples. The above object function provides a way to jointly learn the deformable part model and shape regression, which can benefit from each other. However, it also results in a highly non-convex problem, due to the inference procedure of S_m. We use a coordinate descent procedure to optimize them:

- When the W and S_m are fixed, the function only has the first two terms and becomes a SVM problem to learn the discriminative parameter β, and we use the solver from [44].
- When β is fixed, the optimal W is hard to solve directly since that the HOG transform is non-derivative. Instead, we find an approximation of W by relaxing the last term. We extensively search to find the part configuration $\widetilde{A_m}$, which can converge to a shape closest to ground truth shape S_m^* with regard to the DPM parameterized by β. Once we have $\widetilde{A_m}$, the regression matrix W just needs to ensure that the regressed shape is consistent with $\widetilde{A_m}$, so that we can approximately minimizing the term $\sum_{m=1}^{M_p} \|S_m - S_m^*\|^2$ by $\sum_{m=1}^{M_p} \|A_m^0 + W^T\Phi_a(A_m^0) - \widetilde{A_m}\|^2$. We concatenate shape vector $\widetilde{A_m} - A_m^0$ for $m \in [1, M_p]$ to be a matrix $\mathcal{A}$ and appearance feature vector $\phi(A_m^0, I_m)$ to be a matrix Φ_a, where $\mathcal{A} \in R^{2N \times M_p}$ and $\Phi_a \in R^{Nn_d \times M_p}$. Let I be an identity matrix in $R^{Nn_d \times Nn_d}$, the optimal W in Eq. 6 is approximated by:

$$W = \arg\min_{W_t} C_2\|W\|^2 + C_3 \sum_{m=1}^{M_p} \|A_m^0 + W^T\Phi_a(A_m^0) - \widetilde{A_m}\|^2 \tag{7}$$

$$= (\Phi_a\Phi_a^T + \frac{C2}{C_3}I)^{-1}\Phi_a\mathcal{A}^T. \tag{8}$$

– When W and β are fixed, we can use the standard inference procedure defined in Eq. 4 to find the optimal S_m.

Implementation Details. In our experience, the above procedure usually converges in 3 loop. To start the loop in learning, we need an initialization of W and S_m. The W is got by replacing the $\widetilde{S_m}$ in Eq. 7 with S_m^*, and the S_m is initialized by ground truth S_m^*. In the DPM training step, we always use the parameter got in last iteration as the "warm start", which leads to the fast convergence. We divide training samples into different views. For samples in each view, we align training shapes using similarity transform to remove the offset and normalize them into the same scale. After that, we estimate a multi-variate Gaussian distribution of the shape. For each sliding window in testing, we estimate the scale and translation of mean distribution, and then use it as the initialization A^0.

4 Stacked Deformable Part Model with Shape Regression

In this part, we further improve the part localization performance by stacking the proposed SR-DPM. The intuition comes from recent successes of deep convolutional neural networks (DCNN) in image classification [27] and object detection [24]. These works prove the representation capacity advantage of deep model for real world objects. However, to our best knowledge, no work has shown the advantage of DCNN for general object part localization, partially due to the conflict between the large variations and limited training data.

To balance the representation capacity of deep model and limited training data, we use hand-crafted feature and model for each layer and stack them to form a deep model. Compared with pure data-driven end-to-end learning, our method has much fewer parameters and benefits from reliable priors such as HOG feature and pictorial structure, while still keeps the advantage of rich representation capacity.

The SR-DPM can be taken as a map g, where the input is an image plus a shape and the output is a new shape on this image. Since the oracle map g^* is very complex, there exists an inconsistency between g and g^*. Suppose the training set is $\mathfrak{A}$, then the error on the training set is:

$$\sum_{A_i \in \mathfrak{A}} \|g^*(A_i^0, I_i) - g(A_i^0, I_i)\|^2, \tag{9}$$

where A_i^0 is the initial shape of i-th training sample. To further reduce the training error, we use a series of functions $\mathcal{G} = \{g_1, \cdots, g_K\}$, where K is number of functions. We want to approximate g^* by minimizing:

$$\sum_{A_i \in \mathfrak{A}} \|g^*(A_i^0, I_i) - g_T \circ g_{T-1} \circ \cdots \circ g_1(A_i^0, I_i)\|^2, \tag{10}$$

where each g_i is a SR-DPM, and it uses the output of g_{i-1} as the input. Since in each layer, the function g_i is nonlinear, the whole function is highly non-linear

and has strong representation capacity to approximate the complex map from image to part locations. We name this model as the stacked SR-DPM (S-SR-DPM).

SR-DPMs in the S-SR-DPM are learned sequentially. The initial shape A_0 and image are taken as the input to train the first function g_1 specified by a SR-DPM on the training set, by the coordinate descent learning described in the above section. For the following g_i, we greedily optimize it by:

$$g_i = \arg\min_{g} \sum_{A_i \in \mathfrak{A}} \|g^*(A_i^0, I_i) - g(g_{i-1} \circ \cdots \circ g_1(A_i^0, I_i))\|^2. \tag{11}$$

The map number keeps increasing until the training error does not decrease any more (typically in experiments, a 4 layer S-SR-DPM is enough). In our current implementation, we only use this layer-wise training procedure because of the limited computation resource, despite that the global training is possible. We find that just layer-wise training can significantly improve the performance.

The inference procedure of the S-SR-DPM can be divided into inference of each single layer SR-DPM, which has a global shape regression step and deformable part model step. The procedure is different from traditional iterative optimization in that in each iteration we use different model parameters. Given an image, sliding window based scanning is used, where a non-maximal suppression (NMS) is adopted to eliminate overlapping shape configurations and finally preserve the one with the highest confidence score. We show qualitative examples of S-SR-DPM inference on face alignment in Fig. 1.

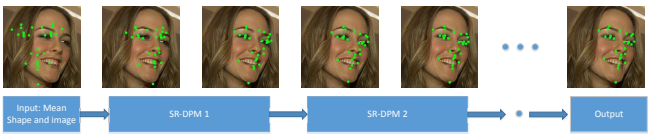

Fig. 1. Examples of S-SR-DPM inference on face alignment (best viewed in color)

5 Experiments

We conduct experiments on human pose estimation, face alignment and general object part localization task. We emphasize that our method achieves competitive performance on the three tasks, compared with different state-of-the-art methods.

5.1 Human Pose Estimation

For human pose estimation, we use the "Leeds Sport Poses" (LSP[1]) [25] to validate different settings and compare with the state-of-the-art methods. LSP is

[1] The dataset is available at `http://www.comp.leeds.ac.uk/mat4saj/lsp.html`

one of the most challenging datasets for human pose estimation, which includes 1000 sports humans for training and 1000 sports humans for testing. The performance is measured by Percentage of Correctly localized Parts (PCP) [22] on 10 object parts defined according to the 14 joints. 6 subtypes are used for each part. For all the experiments on LSP, we use the observer-centric annotations as suggested in [18].

Diagnostic Experiments. We report the mean PCP of 10 parts in different settings in Fig. 2. For the DPM, we use the code from [44] which is carefully tuned for human pose estimation. For our methods, we test the SR-DPM with independent shape regression and DPM learning, the SR-DPM with joint shape regression and DPM learning, and the S-SR-DPM whose depth is set to be among 2, 3 and 4. All these methods are trained on the training set of LSP and use the same 32 dimensional HOG feature from [20]. It can be found that adding shape regression improves a 1.2% margin over the original DPM. When the deformable part model and shape regression are jointly trained, we get a 1.9% further improvement. More improvements come from stacking multiple SR-DPMs to a deeper model. The 2-layer S-SR-DPM gets a 0.8% gain and 3-layer S-SR-DPM gets a 1.2% gain. In our final implementation, we use the 4-layer S-SR-DPM. It improves the final PCP performance by 4.5% over DPM and 1.4% over SR-DPM, which proves the advantage of our S-SR-DPM in capturing large variations for human pose estimation.

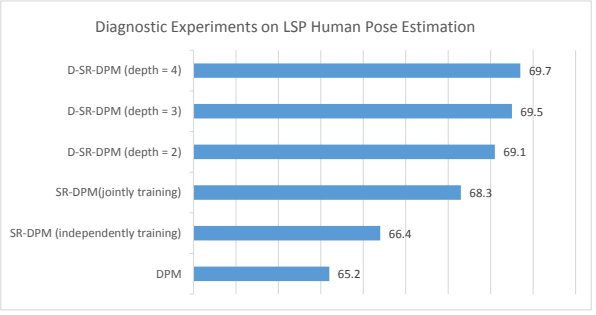

Fig. 2. Mean PCP of different settings on LSP

Comparison with State-of-the-Art Methods. We report the PCP of our methods and the state-the-art methods from recent works in Tab. 1. The "upper leg", "lower leg", "upper arm" and "fore arm" averages the left and right. The performance of our method is better than [1,44,29,18] and on par with a recent result from [30]. Note that [30] fuses multiple appearance cues such as specialized detector and mid-level Poselet, while our method only uses low-level HOG for appearance. Our method is better than [30] in localizing parts with large

deformation, such as fore arm and upper arm, which proves the advantage of our method in representation capacity. [30] provides powerful appearance cues and achieves better performance for torso and head. [30] has advantage in appearance modeling and SR-DPM is better in deformation representation, thus they can be combined for further improvement.

Table 1. Comparisons on PCP results for human pose estimation on LSP

	torso	upper leg	lower leg	upper arm	fore arm	head	mean
Andriluka et al., [1]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
Yang&Ramanan [44]	83.3	72.5	65.6	64.4	41.7	80.4	65.2
Pishchulin et al., [29]	87.5	75.7	68.0	54.2	33.9	78.1	62.9
Pishchulin et al., [30]	88.7	78.8	73.4	61.5	44.9	85.6	69.2
Eichner&Ferrari [18]	86.2	74.3	69.3	56.5	37.4	80.1	64.3
SR-DPM	85.4	75.2	68.8	67.6	45.3	83.6	68.3
S-SR-DPM	85.8	76.8	70.6	69.3	46.9	84.0	69.7

5.2 Face Alignment

For face alignment, we use the Labeled Face Parts in the Wild (LFPW[2]) from [4] as the testbed. It contains 29 face landmarks of real world faces with large appearance variations caused by expression, pose and illumination. Because some URLs are not available, we only get 811 of the 1132 training images and 224 of the 300 test images in this experiment. We only use a single component for face and a single subtype for each landmark. The Cumulative Error Distribution (CED) curve and mean error are used to report the performance. For the CED curve, we normalize the error by the inter-ocular distance to remove the influence of face scale.

The experimental comparisons are reported in Fig. 3. We first compare different settings of our method. The DPM performance is generated by a face-oriented DPM extension from [47]. By adding the shape regression, the normalized mean error has a 18.2% relative decrease. The joint learning of shape regression and DPM decreases the relative error by 14.5% over the independent learning. The stacked model is very effective for face alignment, and it has a 26.8% relative improvement over the single layer SR-DPM, and 48.8% relative improvement over the DPM.

We also compare our method with the state-of-the-art methods by CED curves. It can be found that our SR-DPM is already better than the strong method from [4]. The S-SR-DPM is better than the SDM [43] when the normalized error is below 0.061. Our method has a 0.0298 normalized mean error, while the SDM is 0.0347. The previous best result is [8], which reports a 0.0343 mean error and is sightly worse than our method. It is worth noting that the compared methods all needs reliable face bounding box for initialization, while our method can automatically find the bounding box by sliding window based scanning.

[2] `http://homes.cs.washington.edu/$\sim$neeraj/projects/face-parts/`

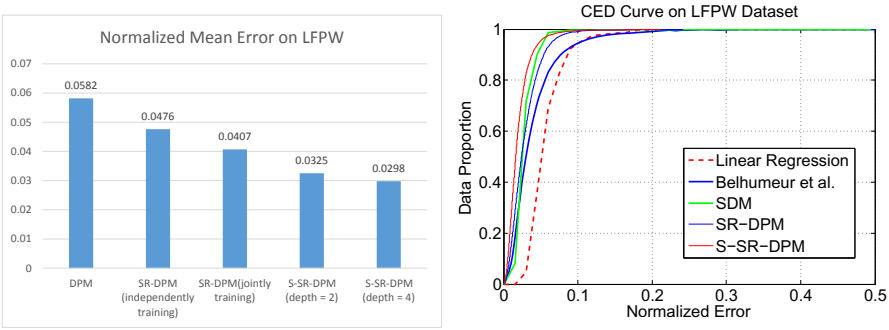

Fig. 3. Comparisons on LFPW face alignment dataset

5.3 Object Detection and Part Localization

Part localization for general object such as animal is more difficult than human and face, which is partially reflected by the detection performance on Pascal VOC. We use the images from Pascal VOC 2007 [19] and the annotation of parts from [3] [3], which includes "bird", "cat", "cow", "dog", "horse" and "sheep". Since the animals tend to be more flexible than human and face, we use the more sophisticated clustering techniques introduced in [3], where relative position, scale, aspect ratio and visibility of parts and the object are used as the feature for clustering, and finally 4 components are used for each category. We refer to [3] for the details. For our S-SR-DPM, the layer number is set to be 4, which is the same as the S-SR-DPM for human pose estimation and face alignment. The object detection is evaluated first and then the part localization.

Table 2. Average Precision of different methods for animal detection in VOC 2007

	Bird	Cat	Cow	Dog	Horse	Sheep	mAP
DPM Ver4 [20]	10.0	19.3	25.2	11.1	56.8	17.8	23.4
DPM Ver5 [20]	10.2	23.0	24.1	12.7	58.1	21.1	24.9
SS-DPM [3]	12.7	26.3	34.6	19.1	62.9	23.6	29.9
Proposed SR-DPM	14.9	27.5	35.7	21.9	64.4	25.5	31.7
Proposed S-SR-DPM	16.7	28.7	36.9	23.5	66.1	27.1	33.2

For object detection, we report the average precision (AP, defined in [19]) of each category on Pascal VOC 2007. The DPM release4[4], DPM release5 [5] and strongly supervised DPM (SS-DPM) [3] are used for comparison, as reported in Tab. 2. The part location information is important for large deformation, as

[3] www.csc.kth.se/cvap/DPM/part_sup.html
[4] http://cs.brown.edu/$\sim$pff/latent-release4/
[5] http://www.cs.berkeley.edu/$\sim$rbg/latent/

reflected by the large performance gain over DPM by strong supervised DPM and our methods. The SS-DPM and our proposed SR-DPM and S-SR-DPM use exactly the same training images and annotations. The SR-DPM improves SS-DPM by 1.8% AP and the S-SR-DPM further improves it by 1.5%. We note that the performance gain is more significant for categories with large deformations, such as bird and cat, which are the most difficult categories for current DPM based detection methods.

Table 3. Part Localization performance evaluated on PASCAL VOC 2007 animals. We report performance PCP of SS-DPM [3], the proposed SR-DPM and S-SR-DPM.

	Method	Bird	Cat	Cow	Dog	Horse	Sheep	mean per part
head	SS-DPM[3]	25.4	60.0	36.3	40.5	65.7	29.4	42.9
	SR-DPM	28.2	64.3	37.6	42.4	66.8	32.1	45.2
	S-SR-DPM	29.1	64.8	40.4	44.6	68.5	33.1	46.8
frontal legs	SS-DPM[3]	-	8.9	25.9	23.1	37.3	17.6	22.6
	SR-DPM	-	12.4	29.3	27.3	38.4	19.6	25.4
	S-SR-DPM	-	13.7	31.4	28.1	41.2	21.6	27.2
fore legs	SS-DPM[3]	12.1	-	37.1	-	39.3	10.9	24.9
	SR-DPM	14.4	-	39.1	-	42.7	12.5	27.2
	S-SR-DPM	17.9	-	41.2	-	44.5	14.6	29.3
torso/back	SS-DPM[3]	-	17.2	58.2	6.7	57.7	57.1	39.4
	SR-DPM	-	20.7	63.1	10.6	59.7	60.3	42.9
	S-SR-DPM	-	21.3	63.4	11.4	61.2	61.1	43.7
tail	SS-DPM[3]	6.1	1.7	-	0.9	32.0	2.4	8.6
	SR-DPM	10.2	4.2	-	5.9	35.0	5.7	12.2
	S-SR-DPM	11.1	6.7	-	5.4	36.1	5.3	12.9
mean per category	SS-DPM[3]	14.5	22.0	39.4	17.8	46.4	23.5	-
	SR-DPM	17.0	25.4	42.3	21.6	48.5	26.0	-
	S-SR-DPM	19.0	26.6	44.1	22.1	50.3	27.1	-

For part localization, we again use the PCP criterion [22], and compare our method with [3], which is the only available result on this setting. We report the PCP of each part in each category and mean PCP of strongly supervised DPM (SS-DPM), SR-DPM and S-SR-DPM in Tab. 3. By incorporating shape regression into deformation part model, while using exactly the same training data and parameter setting with SS-DPM, the proposed SR-DPM achieves a mean per part PCP improvement from 2.3% to 4.6% and a mean per category PCP improvement from 2.1% to 3.8%. When stacked model is used, S-SR-DPM further improves the mean per part/category PCP from 0.7%/0.5% to 1.8%/2.0%. Similar to the observations on object detection, our method has noticeably improvements for categories with large deformations such as bird, cat and dog. We show some qualitative results in Fig. 4.

Fig. 4. Qualitative results of S-SR-DPM for human pose estimation, face alignment and object part localization(best viewed in color)

6 Conclusion

In this paper, we propose two critical improvements over deformable part model to localize object parts from a single image. The first is that we extend DPM to SR-DPM, which exploits the shape regression to capture global information and provides flexible anchor locations. After that, we use the deformable part model to refine the result according to the anchor locations and measure the confidence score. We show how to learn the shape regression and DPM jointly by a coordinate descent procedure. The second improvement is that we prove stacked SR-DPM (S-SR-DPM) increases the representation capacity and leads to better localization performance. We show the advantages of SR-DPM and S-SR-DPM for human pose estimation, face alignment and object part localization, which are usually taken as three different problems.

Acknowledgement. This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. IEEE (2009)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: CVPR. IEEE (2013)
3. 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)
4. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR. IEEE (2011)
5. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning (2009)
6. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR. IEEE (2013)
7. 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)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR. IEEE (2012)
9. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: CVPR. IEEE (2013)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. PAMI (2001)
11. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. CVIU (1995)

12. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition (2008)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. IEEE (2005)
14. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: CVPR. IEEE (2012)
15. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 158–172. Springer, Heidelberg (2012)
16. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR. IEEE (2010)
17. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures (2009)
18. Eichner, M., Ferrari, V.: Appearance sharing for collective human pose estimation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 138–151. Springer, Heidelberg (2013)
19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV pp. 303–338 (2010)
20. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010)
21. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
22. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. IEEE (2008)
23. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers (1973)
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint (2013)
25. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
26. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. IEEE (2011)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
28. Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L.: Visual Analysis of Humans. Springer (2011)
29. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR. IEEE (2013)
30. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Pstrong appearance and expressive spatial models for human pose estimation. In: ICCV. IEEE (2013)
31. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR. IEEE (2011)
32. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge (2013)
33. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: CVPR. IEEE (2013)
34. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. IJCV (2011)
35. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep fisher networks for large-scale image classification. In: NIPS (2013)
36. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV. IEEE (2011)

37. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. IEEE (2013)
38. Tian, Y., Zitnick, C.L., Narasimhan, S.G.: Exploring the spatial hierarchy of mixture models for human pose estimation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 256–269. Springer, Heidelberg (2012)
39. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 227–240. Springer, Heidelberg (2010)
40. Wang, F., Li, Y.: Beyond physical connections: Tree models in human pose estimation. In: CVPR. IEEE (2013)
41. Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 710–724. Springer, Heidelberg (2008)
42. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR. IEEE (2011)
43. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR. IEEE (2013)
44. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. IEEE (2011)
45. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: ICCV (2013)
46. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)
47. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. IEEE (2012)