Expanded Parts Model for Human Attribute and Action Recognition in Still Images

Gaurav Sharma, Frédéric Jurie, Cordelia Schmid

Expanded Parts Model for Human Attribute and Action Recognition in Still Images

Gaurav Sharma[1,2], Frédéric Jurie[1], Cordelia Schmid[2]
[1]GREYC, CNRS UMR 6072, University of Caen Basse-Normandie
[2]LEAR, INRIA Grenoble Rhône-Alpes
firstname.lastname@{[1]unicaen, [2]inria}.fr

Abstract

*We propose a new model for recognizing human attributes (*e.g. *wearing a suit, sitting, short hair) and actions (*e.g. *running, riding a horse) in still images. The proposed model relies on a collection of part templates which are learnt discriminatively to explain specific scale-space locations in the images (in human centric coordinates). It avoids the limitations of highly structured models, which consist of a few (*i.e. *a mixture of) 'average' templates. To learn our model, we propose an algorithm which automatically mines out parts and learns corresponding discriminative templates with their respective locations from a large number of candidate parts. We validate the method on recent challenging datasets: (i) Willow 7 actions [7], (ii) 27 Human Attributes (HAT) [25], and (iii) Stanford 40 actions [37]. We obtain convincing qualitative and state-of-the-art quantitative results on the three datasets.*

1. Introduction

The focus of this paper is a semantic description of humans in still images using attributes and actions. Given the daily growing amount of human centric data (*e.g.* on photo sharing and social networking websites or from surveillance cameras), analysis of humans in images is more important than ever.

Most recent work on human attributes or action recognition either rely on, accurate or approximate, estimation of human pose *e.g.* [32, 35] or use general non-human-specific image classification methods *e.g.* [7, 25, 26, 37]. It has been demonstrated that state-of-the-art action recognition can be achieved without solving the difficult problem of pose estimation [7, 11, 26, 32]. Interestingly, several recent methods propose to model interactions between humans and the object(s) associated with the actions [8, 10, 15, 24, 34, 35]. While modelling *interactions* between humans and contextual objects is an interesting problem, we explore here the broader problem of modelling *appearance* of humans for attribute and action recognition. Such modelling is critical in the numerous cases where there are no associated objects (*e.g.* actions like running, walking) and/or the pose is not immediately relevant (*e.g.* attributes like long hair, wearing a tee-shirt).

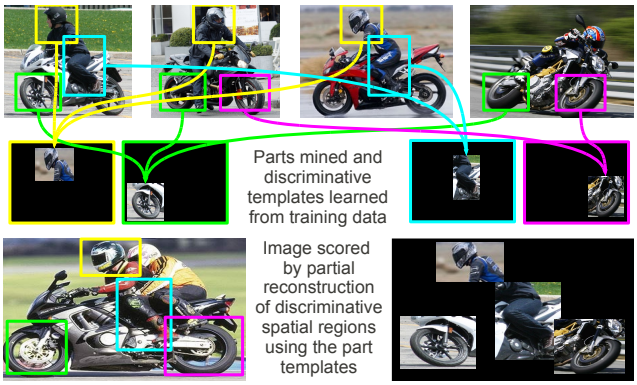

Figure 1. Illustration of the proposed method (see text).

In this paper we present the Expanded Parts Model (EPM) (Fig. 1), which provides a rich discriminative description of the appearance of humans. We work in human centered images *i.e.* we assume that the human positions in form of bounding boxes are available (*e.g.* from a human detection algorithm). Our model is a collection of part templates, each of which can explain specific scale-space regions in the images. At test time, our model scores an image by representing it with the learnt part templates. As human attributes and actions are often localized in space *e.g.* shoulder regions for 'wearing a tank top', we aim to explain the images partially with the most discriminative regions, *i.e.* the model selects sufficiently discriminative spatial evidence for the class and does not include the non-discriminative background regions (Fig. 1). In our model the parts compete to explain each image individually, which is in contrast with traditional part based discriminative models where all parts are used for every image. We propose a learning algorithm based on regularized loss minimization and margin maximization. Our learning algorithm allows us

to mine out important parts for the task, and learn their discriminative templates from a large pool of candidate parts obtained by dense random sampling of the training images. We validate our method on three publicly available datasets of human attributes and actions, and show promising qualitative and state-of-the-art quantitative results.

1.1. Related work

Models without parts. Generic image classification algorithms, which have been quite successful in human action recognition [11], generally learn a discriminative model for each class. In the Spatial Pyramid method (SPM) [16] images are represented as a concatenation of bag-of-features (BoF) histograms [5, 27] with pooling at multiple spatial scales over a learnt codebook of local features, like SIFT [19]. A discriminative class model $\mathbf{w}$ is, then, learnt using a margin maximizing classifier [16]. A new image is scored based on its match with the learnt class model, quantified as the dot product $\mathbf{w}^T\mathbf{x}$ between the image vector $\mathbf{x}$ and the class model $\mathbf{w}$. The use of histograms destroys 'template' like properties due to loss of spatial information and makes visualization difficult. Although SPM has never been viewed as a template learning method, methods using histogram of oriented gradients (HOG) [6] features have been presented as such and the recent literature is full of visualizations of templates (class models) learnt with HOG-like features *e.g.* [6, 13, 22]. Both of these methods have been applied to the task of human analysis [7] and we build on them and formulate our model in a discriminative template learning framework. We differ in that we learn a collection of templates instead of a single template.

The recently proposed Exemplar SVM (ESVM) [21] learns discriminative templates for each object instance of the training set independently and then combines their calibrated outputs on test images as a post-processing step. In contrast, we work at a part level and use all templates together during both training and testing.

More recently, a 2-level approach for image representation has been proposed [31]. Similar to our approach it involves sampling image regions and, then, vector quantizes the region descriptors, whereas we propose a mechanism to select discriminative regions and build discriminative part-based models from them.

Part-based structured models. Generative or discriminative part-based models (*e.g.* the constellation model [14] and the DPM model [13]), have led to state-of-the-art results for objects that are rigid or, at least, have a simple and stable structure. In contrast humans involved in actions can have huge appearance variations due to both cosmetic changes (*e.g.* clothes, hair style, accessories) as well as articulations or poses. Furthermore, their interaction with the context can be very complex and case dependent. Probably because of the high complexity of such a task, DPMs do not perform well for human action recognition [7]. Increasing the model complexity, *e.g.* by using a mixture of components [13], has shown to be beneficial for object detection[1]. Such increase in model complexity is even more apparent in similar models for finer human analysis *e.g.* pose estimation [9, 33, 39], where a relatively large number of components and parts are used. While the components account for global changes in aspect/viewpoint, the parts account for the local variations of the articulations. A recent study [40] recommends the design of richer models albeit with careful regularization. Here, we propose a richer, but less structured, expanded parts model.

As shown in Fig. 2 (left), in the mixture of components model the training images are usually assigned to only one component and thus contribute to training only one of the templates (and similarly in testing). This limits the capability to generate novel articulations, as a sub-articulation (hands raised) in one component can not be combined with a sub-articulation (hands along the body) in another component to generate a hybrid of the two (one hand raised and one along the body). Note that the clustering and averaging within such a model are a form of regularization/complexity control enforced by the system, which involves *manual* setting the number of parts and components.

In the proposed expanded parts model (i) we neither enforce nor forbid averaging *a priori* and (ii) we allow the model to have a large number of 'parts' (up to the order of the number of training images) if found necessary despite sufficient regularization (Fig. 1 & 2). While in part-based deformable models the parts initialization is either based on heuristics (*e.g.* initialization with regions with high average energy [13]) or available annotations [9], our method systematically explores parts at all possible locations, scales and atomicities and selects the ones best suited for the task.

Part-based loosely structured models. Our model belongs to a family of models which use parts but do not assume that all possible variations and articulations can be captured by a few averaged, spatially constrained, templates of parts. Our model has similarities with poselets [3, 4, 20] which are compound parts consisting of multiple anatomical parts, highly clustered in 3D configuration space *e.g.* head and shoulder together. Each poselet casts a vote independently for an object hypothesis. Poselets are shown to improve performance and are trained separately from specifically annotated images (in 3D). In contrast, our method tries to mine out such 'parts', at the required atomicity, with a task specific focus and from given training images. Fig. 6 (top right) shows some of our parts for the 'female' class which show some resemblances with poselets, though are not as clean.

While poselets learn discriminative templates, meth-

[1] See the results of different versions of the DPM software http://people.cs.uchicago.edu/∼rgb/latent/ which, along with other improvements, steadily increase the number of components and parts.

Figure 2. Left – Illustration of a two component model vs. the Expanded Parts Model. Right – Example 'reconstructions'.

ods such as those derived from Similarity by Composition [1], Naive Bayes Nearest Neighbors (NBNN) [2], Implicit Shape Models [17] and Collaborative Representation [38], try to reconstruct images from patches. However, their learning approaches are generally based on the reconstruction error *i.e.* are generally generative while here we aim to mine out good patches and learn corresponding discriminative templates with the direct aim of achieving good classification. Moreover, such models have not been previously applied to human attribute and/or action recognition.

2. Approach

In the following, we address a supervised classification setting where we are given a set of training images $\mathcal{I}_t = \{I_i \in \mathcal{I} | i = 1 \dots m\}$ with their corresponding binary class labels $y_i \in \{-1, 1\}$. Our goal is to learn a scoring function $s : \mathcal{I} \rightarrow \mathbb{R}$ which takes an image and assigns a score reflecting the membership of the image to the positive class. We define (the parameters of) our model to be a collection of discriminative templates with an associated scale space location and the image scoring as a process of partially 'reconstructing' the important (task specific) regions in the images from these discriminative templates.

Models based on HOG-like features [6] suffer an important limitation as they rely on shape while somewhat ignoring appearance. Shape preference seems to work, and perhaps to help, for human pose estimation [9, 33, 39] but seems to be a probable reason for the disappointing performance of DPMs on human action recognition [7]. Hence, in the present work we choose to use the bag-of-features (BoF) representation instead of HOG-like shape features in order to obtain a better appearance description. As a result of this choice, our BoF derived discriminative models $\mathbf{w}$ (similar to [16]) can not be called templates as (i) an immediate method to convert them to plausible natural images is not clear and (ii) even if such a method exists it will not lead to a unique image (as trivially, any other image obtained by jumbling around a given image's local features also has the same BoF as the starting image). However, we continue to use the word template to denote the corresponding concept in BoF space. We later explain how our model provides an approximate way for visualizing the reconstructions; Fig. 2 (right) shows examples of such visualizations. Note, however, that the proposed method can be used with any arbitrary feature space.

2.1. Regularized loss minimization

Our model is defined as a collection of discriminative templates with associated locations *i.e.* $M = \{(\mathbf{w}, \mathbf{l}) | \mathbf{w} \in \mathbb{R}^{Nd}, \mathbf{l} \in \mathbb{R}^{N \times 4}\}$ where N is the number of parts, d is the size of BoF codebook, $\mathbf{w}$ is the concatenation of N part templates (each of dimension d) and $\mathbf{l}$ is a matrix of their scale-space positions, with each row specifying a bounding box *i.e.* $l_p = (x_1, y_1, x_2, y_2)$ where x and y are fractional multiples of width and height respectively.

We propose to learn our model as a regularized loss minimization optimization with the objective

$$L(M) = \frac{\lambda}{2}||\mathbf{w}||_2^2 + \frac{1}{|\mathcal{I}_t|} \sum_{I_i \in \mathcal{I}_t} \max(0, 1 - y_i s(I_i, M)), \quad (1)$$

with $s(\cdot)$ being the scoring function (Sec. 2.2). Our objective is the same as that of linear SVMs with hinge loss. The only difference is that we have replaced the linear, $s_l(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$, score function with our score function (Eq. 2). The parameter λ sets the trade-off between model regularization and the loss minimization (cf. SVM).

Algorithm 1 Stochastic gradient descent for learning EPM

1: Initialize: $M = (\mathbf{w}, \mathbf{l})$, $r = 1$, $k = 100$ and $\lambda = 10^{-5}$
2: **for** iter $= 1, \ldots, 10$ **do**
3: $r(1) = r \times N^{-}/N$ and $r(-1) = r \times N^{+}/N$
4: **for** npass $= 1, \ldots, 5$ **do**
5: $S \leftarrow$ *rand_shuffle*($\mathcal{I}_t$)
6: **for all** $I_i \in S$ **do**
7: Solve Eq. 2 to get $s(I_i, M)$ and $\boldsymbol{\alpha}$
8: $\delta_i \leftarrow$ *binarize*$(y_i s(I_i, M) < 1)$
9: $\mathbf{w} \leftarrow \mathbf{w}(1 - r(y_i)\lambda) + \delta_i y_i r(y_i) \sum \alpha_p f(I_i, l_p)$
10: **end for**
11: **end for**
12: `parts_image_map` $\leftarrow$ *note_image_parts* (M, $\mathcal{I}$)
13: $M \leftarrow$ *prune_parts* (M, `parts_image_map`)
14: **if** iter $= 5$ **do** $r \leftarrow r/5$ **end if**
15: **end for**

2.2. Scoring function

Our scoring function is inspired by the method of image scoring with learnt discriminative templates and that by image reconstruction. We want to score the image with the part templates which are capable of reconstructing it well while penalizing high overlap. The discriminative information for human actions and attributes is often localized in space *i.e.* for 'riding horse' only the rider and the horse is discriminative and not the background and for 'wearing shorts' only the lower part of the image is important. Hence, we aim to reconstruct the image partially (in space) with the most important parts only (*e.g.* see Fig. 2).

Formally, we define the scoring function as

$$s(I, M) = \frac{1}{k} \max_{\alpha} \sum_{p=1}^{N} \alpha_p \mathbf{w}_p^T f(I, l_p) \tag{2a}$$

$$\text{s.t. } ||\boldsymbol{\alpha}||_0 = k, \quad O_v(\boldsymbol{\alpha}, \mathbf{l}) \leq \theta, \tag{2b}$$

where, $\mathbf{w}_p = [0, \ldots, 0, w_{(p-1)d+1}, \ldots, w_{pd}, 0, \ldots 0]^T$ *i.e.* a vector of same dimension as $\mathbf{w}$ with the discriminative template for the p^{th} part at the corresponding location with other components set to zero[2], $f(I, l_p)$ is the feature extraction function which calculates the BoF histogram of the image I for the patch specified by l_p and zero-pads it similar to $\mathbf{w}_p$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]$ are the binary coefficients which specify if a model part is used to score the image or not, $O_v(\boldsymbol{\alpha}, \mathbf{l})$ calculates overlap between the parts selected to score the image. The ℓ_0 norm constraint on $\boldsymbol{\alpha}$ enforces the use of exactly k parts for scoring while the second constraint encourages coverage in reconstruction by limiting high overlaps.

[2]Such zero padding is like a masking operation which ensures that the current part interacts only with the similarly located image patch.

2.3. Solving the optimization problem

We propose to solve the model optimization problem using stochastic gradient descent. We use the stochastic approximation to the sub-gradient w.r.t. $\mathbf{w}$ given by,

$$\nabla_{\mathbf{w}} L = \lambda \mathbf{w} - \delta_i \frac{1}{k} \sum_{p=1}^{N} \alpha_p f(I_i, l_p) \tag{3}$$

where, α_p are obtained by solving Eq. 2 and $\delta_i = 1$ if $y_i s(I_i, M) < 1$ otherwise $\delta_i = 0$. Alg. 1 gives the pseudo-code for our learning algorithm. The algorithm proceeds by scoring (and thus calculating the $\boldsymbol{\alpha}$ for) the current example with $\mathbf{w}$ fixed, and then updating $\mathbf{w}$ with $\boldsymbol{\alpha}$ fixed.

Initialization. In the initialization we intend to generate a large number of part candidates, which are subsequently refined by pruning. To achieve this, we randomly sample the positive training images for patch positions *i.e.* $\{l_p\}$ and initialize our model parts as $\mathbf{w}_p = [2\mathbf{x}_p, -1]^T$, where $\mathbf{x}$ denotes a BoF histogram (we are abusing notation here as actually $\mathbf{w}_p$ is zero padded to make its dimension equal to $\mathbf{w}$). Throughout our method, we append 1 at the end of all our BoF features *i.e.* $\mathbf{x}_b = [\mathbf{x}, 1]^T$ to account for the bias term (cf. SVM *e.g.* [23]). This leads to a score of 1 when a perfect match occurs ($\mathbf{w}_p^T \mathbf{x}_b = 2 \times 1 - 1 = 1$) and a score of -1 in the opposite case ($\mathbf{w}_p^T \mathbf{x}_b = 2 \times 0 - 1 = -1$), as the BoF features are ℓ_2 normalized. We sample 10^5 patches from the whole dataset and, for each class, use patches from respective positive images. For the learning rate, we follow recent work [23] and fix a learning rate, which we reduce once for annealing by a factor of 5 half way through the iterations (step 1 and 13, Alg. 1).

Scoring function. The ℓ_0 norm constraint in the scoring function makes it NP-hard. In our current implementation we use an approximate greedy solution. At any given instant we greedily select the best scoring part (and assign corresponding $\alpha_p = 1$) which does not overlap appreciably with the currently selected parts which were generated from the same training image as that of the part under consideration. We observed, on initial experiments on validation sets, that if a false positive was similar to one train image, then it would take numerous overlapping parts which were generated from that image. The previous condition is to avoid such double counting. The overlap criteria is 1/3 intersection by union [11]. While training, we score each training image from the rest of train set *i.e.* we do not use the parts which were generated from the same training image.

Training data imbalance. Usually large databases are highly unbalanced *i.e.* they have many more negative examples than positive examples (of the order of 50:1). To handle this we use asymmetric learning rates proportional

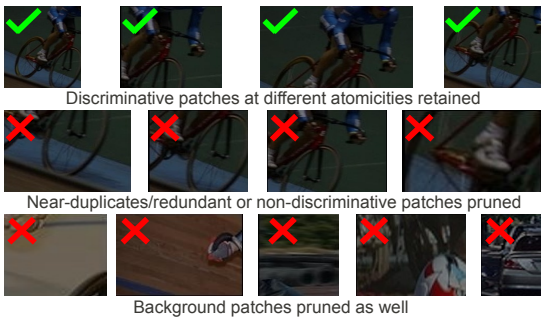

Figure 3. Example patches illustrating pruning (riding bike class).

to the number of examples of other class[3] (step 3, Alg. 1).

Parts mining by pruning. After each iteration (*i.e*. 5 passes over randomly shuffled training data) we prune the set of parts. We keep a record of which part is being used to reconstruct which images while training and then simply prune the parts which have not been used to reconstruct even a single image. Such parts only contribute to the $||\mathbf{w}||_2^2$ term and not to the loss term, hence removing them trivially minimizes the objective. Pruning in this way removes near-duplicate and non discriminative parts (Fig. 3) which were considered because of the random sampling and allows us to mine out the discriminative parts.

Regularization and number of parts. We follow [23] and fix the regularization constant $\lambda = 10^{-5}$. For fixing the number of parts we did preliminary experiments on the validation set of Willow actions database [7]. The performance increased by 10% (absolute) as k went from 10 to 100. With this we concluded that we need a sufficiently high number of parts and fixed $k = 100$ for all experiments.

Nonlinearizing using feature map. Until now, we have described linear version of our algorithm. To have non linearity we use explicit feature map [29]. We use map corresponding to the Bhattacharyya kernel *i.e*. we take dimension-wise square roots of our ℓ_1 normalized BoF histograms obtaining ℓ_2 normalized vectors which we then use with our algorithm.

Relation with latent SVM. In our model, $\boldsymbol{\alpha}$ can be seen as latent variables per image, and the whole model can be seen as a latent SVM [13]. In such cases we should train keeping in mind the semi-convexity [13] of the objective function – training as we propose is not guaranteed to reduce the objective. However, in practice we see that if the learning rate is not aggressive, training as proposed leads to reasonably good convergence (Fig. 4) and performance, and hence we continue to use our implementation.

Visualization of reconstructions. Since we initialize our parts with the BoFs of patches from training images, we can use the initial patches to visualize the reconstructions.

[3] [23] achieve the same effect by biased sampling from the two classes.

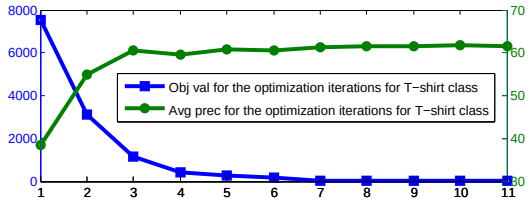

Figure 4. Convergence of our algorithm.

This is clearly a loose association as the part templates $\mathbf{w}_p$ are modified in the the learning process, but we found it to give reasonable visualizations (Fig. 2).

3. Experimental results

We evaluate our method on three challenging publicly available databases: (i) Willow 7 human actions [7], (ii) 27 human attributes (HAT) [25], and (iii) Stanford 40 human actions [37]. We first present implementation details of our approach as well as our baseline and, then, proceed to present and discuss our results on the three databases.

Implementation and baseline details. Like previous work [7, 26, 37] we densely sample grayscale SIFT features at multiple scales. We use a fixed step size of 4 pixels and use square patch sizes ranging from 8 to 40 pixels. We learn a vocabulary of size 1000 using k-means and assign the SIFT features to the nearest codebook vector (hard assignment). We use the VLFeat library [28] for SIFT and k-means computation. We use a four level spatial pyramid with $\{c \times c | c = 1, 2, 3, 4\}$ cells [16] as baseline. We use the explicit feature map [29] corresponding to the Bhattacharyya kernel, *i.e*. dimension-wise square root of ℓ_1 normalized vectors, to be comparable to our method. The baseline results are obtained with the liblinear [12] library.

Immediate context. The immediate context around the person, which might contain partially the associated object (*e.g*. horse in riding horse) and/or correlated background (*e.g*. grass in running), has shown to be beneficial for the task [7, 26]. To include immediate context we expand the human bounding boxes by 50% in both width and height.

Full image context. The context from the full image has also been shown to be important [7]. To use it with our method, we add the scores from a classifier trained on full images to scores from our method. The full image classifier uses 4 level SPM with an exponential χ^2 kernel.

Performance measure. The performance is evaluated with average precision (AP) for each class and the mean average precision (mAP) over all classes.

3.1. Qualitative results

We present qualitative results to illustrate how our reconstruction works in practice. Fig. 2 (right) shows some examples, *i.e*. composite images created by displaying the part

Table 1. Results on Willow 7 actions database (Sec. 3.2)

	Inter. [8]	Dsal [26]	SPM [16]	Ours (EPM)	Ours + context
mAP	64.1	65.9	63.7	66.0	**67.6**

Table 2. Results on Human attributes database (Sec. 3.3)

	DSR [25]	SPM [16]	Ours (EPM)	Ours + context
mAP	53.8	55.5	58.7	**59.7**

patches with non-zero alphas. The observed results are convincing, *i.e*. the method focuses on the relevant parts, such as torso and arms for 'bent arms', shorts and tee-shirts for 'wearing bermuda shorts', and even computer (left bottom) for 'using computer'. Interestingly, we observe that for both riding horse and riding bike classes, the person gets ignored but the hairs and helmet have been partially reconstructed. This seems to stress the discriminative nature of the learnt models. As the persons in similar pose might confuse the two classes, it focuses on the more discriminative aspects.

3.2. Willow actions database

Willow actions[4] [7] is a challenging database for action classification on unconstrained consumer images downloaded from the internet. It has 7 classes of common human actions *e.g*. 'ridingbike', 'running'. It has at least 108 images per class of which 70 images are used for training and validation and the rest are used for testing. The task is to predict the action being performed given the human bounding box. Tab. 1 shows the results of our method (with and without context) along with our baseline SPM and some competing methods. We achieve a mAP of 66% which goes up to 67.6% by adding the full image context. We perform better than the current state-of-the-art method [26] on this dataset on five out of seven classes and on average. As demonstrated by [7], full image context plays an important role in this dataset. It is interesting to note, that even without context, we achieve 3.5% absolute improvement compared to a method which models person-object interactions [8] and uses extra data to train detectors etc.

3.3. Database of human attributes (HAT)

HAT[5] is a database for learning semantic human attributes. It contains 9344 unconstrained human images obtained by applying a human detector [13] on images downloaded from the internet. It has annotations for 27 attributes based on sex, pose (*e.g*. standing, sitting), age (*e.g*. young, elderly) and appearance (*e.g*. wearing a tee-shirt, shorts). The database has train, validation and test sets. The models are learnt with the train and validation sets and the performance is reported on the test set. Tab. 2 shows our as well as other results on this dataset. Our baseline is already higher than the results reported by the dataset creators [25], because we use denser SIFT and more scales. Our method improves over the baseline by 3.2% (absolute) and increases further by 1% when adding the full image context. Fig. 5 shows our results (without full image context) along with the baseline. Our method outperforms the baseline for 24 out of the 27 attributes. Among the different human attributes those based on pose (*e.g*. standing, arms bent, running/walking) seem to be easier than those based on appearance of clothes (*e.g*. short skirt, bermuda shorts). The range of performance obtained is quite wide, from 24% for crouching to 98% for standing. The performances of the classes close to the bottom of Fig. 5 indicates that recognizing human attributes is far from solved.

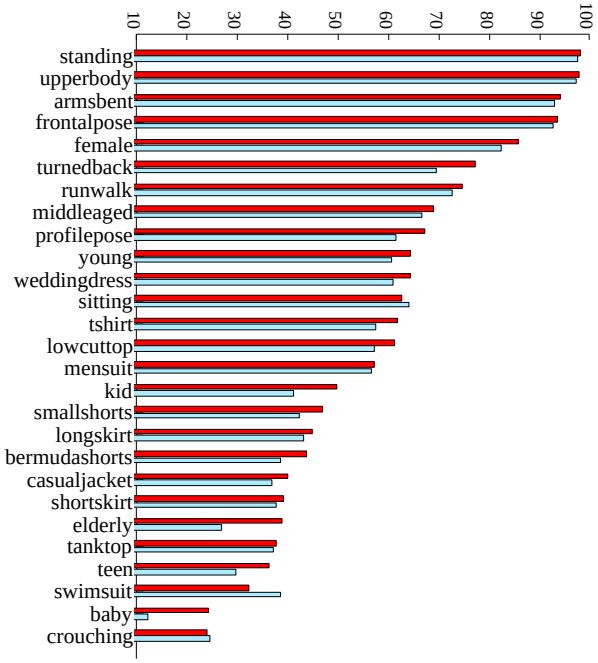

Figure 5. The per attribute performance (AP) of the proposed methods (red/dark) and the baseline SPM [16] (blue/light) on the database of Human attributes (HAT) [25].

3.4. Stanford 40 actions

Stanford 40 actions[6] [36] is a database of human actions with 40 diverse daily human actions *e.g*. brushing teeth, cleaning the floor, reading book, throwing a frisbee. It has 180 to 300 images per class with a total of 9352 images. We used the suggested train and test split provided by the authors on the website, with 100 images per class for training and the rest for testing. Tab. 3 shows our results along with results of the baseline and other methods. Our method performs better than the baseline by 5.8% (absolute) at 40.7%. We also perform better than Object bank [18] and Locality-

[4] http://www.di.ens.fr/willow/research/stillactions/
[5] http://sharma.users.greyc.fr/hatdb/
[6] http://vision.stanford.edu/Datasets/40actions.html

Table 3. Results on Stanford 40 actions database (Sec. 3.4)

	Object bank [18]	LLC [30]	SPM [16]	Ours (EPM)	Ours + context
mAP	32.5	35.2	34.9	40.7	**42.2**

constrained linear coding [30] (as reported in [36]) by 8.2% and 5.5% respectively. With context our method achieves 42.2% mAP which is the state-of-the-art result using no extra training data. The best result reported on this dataset is 45.7% by [36], who solve action recognition using bases of attributes, objects and poses. To derive their bases they use pre-trained systems for 81 objects, 45 attributes and 150 poselets, using large amount (comparable to the size of the database) of external data. Since they use human based attributes also, arguably, our method can be used to improve their generic classifiers and improve performance further *i.e*. our method is complementary to theirs.

3.5. The learnt parts and training/testing times

Fig. 6 shows the distribution of the ℓ_2 norm of the learnt part templates, along with top scoring patches for selected parts, with norms across the spectrum for three classes. The first image in any row is the patch with which the part was initialized and the remaining one are its top scoring patches. The top scoring patches give an idea of what kind of appearances the learnt templates $\mathbf{w}_p$ captures. We observe that, across datasets, while most of the parts seem interpretable, like face, head, arms, horse saddle, legs *etc*., there are a few parts which seem to correspond to random background (*e.g*. row 1 for 'climbing'). This is in line with a recent study [40]: in "mixture of template' like formulations, there are clean interpretable templates along with noisy templates which correspond to background.

We also observe that the distribution of the ℓ_2 norm of the parts follows a heavy tailed distribution. Some parts are very frequent and the system tries to tune them to give high scores for positive vectors and low scores for negative vectors and hence give them a high overall energy. There are also parts which have smaller norms, either because they are consistent in appearance (like the head and partial shoulders on clean backgrounds in row 4 of 'female' Fig. 6, or the leg/arm in the last row of 'climbing') or occur in few images. However, they are discriminative none the less. To determine a clear relation between the statistics of templates and their contribution to the overall performance is an interesting question, which we leave as future work. It is critical to control the trade-off between time efficiency vs. accuracy of a learnt model.

The training is significantly slower compared to a standard SPM/SVM baseline, i.e. by around two orders of magnitude. This is due to the fact that there is SVM equivalent cost (with a larger number of vectors) at each iteration. Testing is also a bit slower compared to an SPM, as it is based on a dot product between longer vectors. *E.g*. on Stanford dataset testing is 5 times slower (cf. SPM) and takes about 35 ms/image (excluding feature extraction).

4. Conclusion

We have presented a new Expanded Parts Model (EPM) for human analysis. The model learns a collection of discriminative templates which can appear at specific scale-space positions. It scores a new image by reconstructing it using the available part templates. We proposed a stochastic sub-gradient based learning method. The algorithm is capable of exploring a large number of candidate parts and mining out the discriminative parts best suited for the current binary classification. We validated our method on three challenging publicly available datasets for human attributes and actions. We obtained good qualitative and state-of-the-art quantitative results, when no external data is used.

We analysed the learnt parts with statistics of their discriminative templates and plan to pursue this direction further to gain additional insight. Applying the model to articulated object detection is also a natural extension.

Acknowledgements. This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation and by the ANR, grant reference ANR-2010-CORD-103-06.

References

[1] O. Boiman and M. Irani. Similarity by composition. In *NIPS*, 2006. 3

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008. 3

[3] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011. 2

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 2

[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Intl. Workshop on Stat. Learning in Comp. Vision*, 2004. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2, 3

[7] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In *BMVC*, 2010. 1, 2, 3, 5, 6

[8] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 1, 6

[9] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2, 3

[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshops*, 2010. 1

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2011. 1, 2, 4

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 5

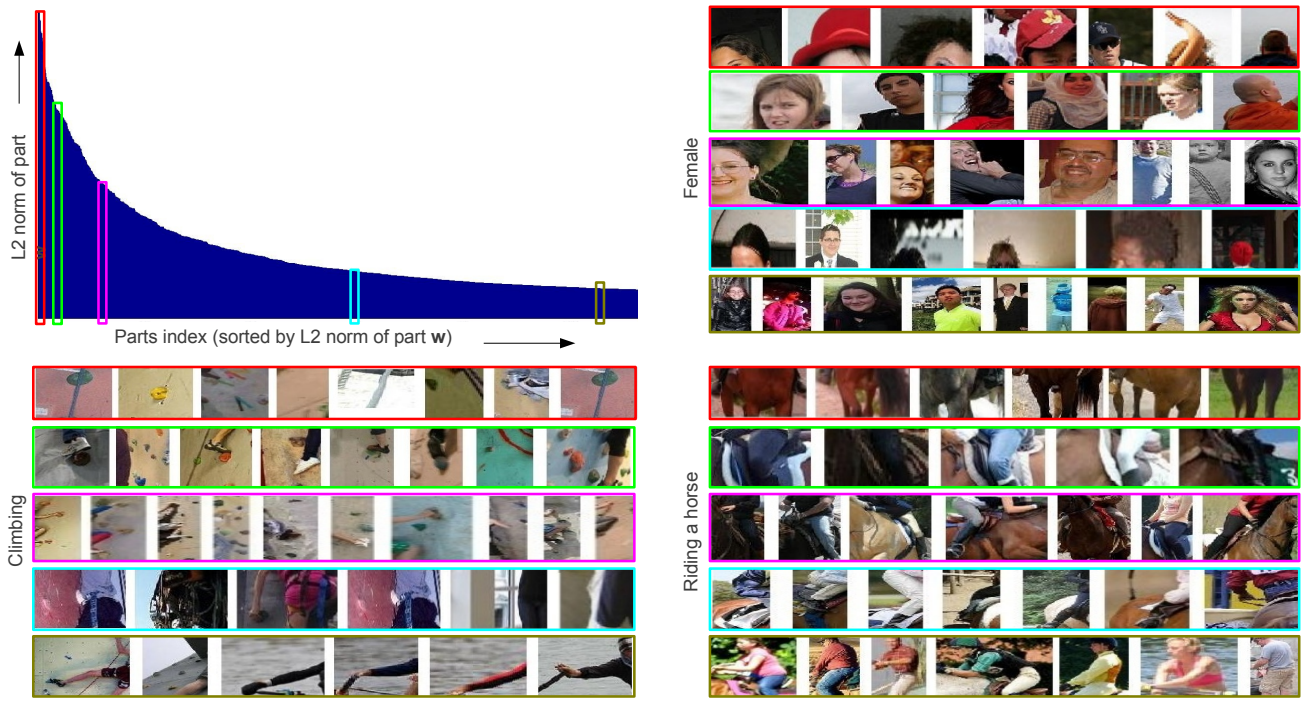

Figure 6. Distribution of the norm of the part templates and some example 'parts'. In each row, the first image is the patch used to initialize the part and the remaining images are its top scoring patches (see Sec. 3.5, best viewed in color).

[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 2, 5, 6

[14] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 71(3):273–303, March 2007. 2

[15] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 31:1775–1789, October 2009. 1

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 3, 5, 6, 7

[17] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, 2008. 3

[18] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 6, 7

[19] D. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[20] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 2

[21] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of Exemplar-SVMs for object detection and beyond. In *ICCV*, 2011. 2

[22] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 2

[23] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012. 4, 5

[24] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 2011. 1

[25] G. Sharma and F. Jurie. Learning discriminative representation for image classification. In *BMVC*, 2011. 1, 5, 6

[26] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012. 1, 5, 6

[27] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[28] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 5

[29] A. Vedaldi and A. Zisserman. Efficient additive kernels using explicit feature maps. In *CVPR*, 2010. 5

[30] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 7

[31] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li. Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification. In *ECCV*, pages 473–487, 2012. 2

[32] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 1

[33] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*. IEEE, 2011. 2, 3

[34] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1

[35] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1

[36] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 6, 7

[37] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 1, 5

[38] P. Zhu, L. Zhang, Q. Hu, and S. Shiu. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In *ECCV*, 2012. 3

[39] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. IEEE, 2012. 2, 3

[40] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012. 2, 7