Action Recognition from a Single Web Image Based on an Ensemble of Pose Experts

Peihao Zhang, Xiaoyang Tan, Xin Jin

Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract. In this paper, we present a new method which estimates the pose of a human body and identifies its action from one single static image. This is a challenging task due to the high degrees of freedom of body poses and lack of any motion cues. Specifically, we build a pool of pose experts, each of which individually models a particular type of articulation for a group of human bodies with similar poses or semantics (actions). We investigate two ways to construct these pose experts and show that this method leads to improved pose estimation performance under difficult conditions. Furthermore, in contrast to previous wisdoms of combining the output of each pose expert for action recognition using such method as majority voting, we propose a flexible strategy which adaptively integrates them in a discriminative framework, allowing each pose expert to adjust their roles in action prediction according to their specificity when facing different action types. In particular, the spatial relationship between estimated part locations from each expert is encoded in a graph structure, capturing both the non-local and local spatial correlation of the body shape. Each graph is then treated as a separate group, on which an overall group sparse constraint is imposed to train the prediction model, with extra weight added according to the confidence of the corresponding expert. We show in our experiments on a challenging web data set with state of the art results that our method effectively improves the tolerance of our system to imperfect pose estimation.

1 Introduction

Human action recognition is an extremely important and active research field in computer vision [1–4]. Its purpose is to recognize what a person is doing or what the posture means. Human action recognition has many interesting and important applications, for example, surveillance, entertainment, human-computer interaction, image and video retrieval. Nowadays, most of the previous work in this area focused on recognizing human actions from videos, those work [5–7] mainly use motion cues and a lot of progress has been made in the recent years. However, compared with videos, human action recognition from static images is a relatively less-researched field. In fact, the analysis of human action in still images is very important. This can be very useful for image understanding and retrieval. Besides, it will not only help us to understand and analyze human

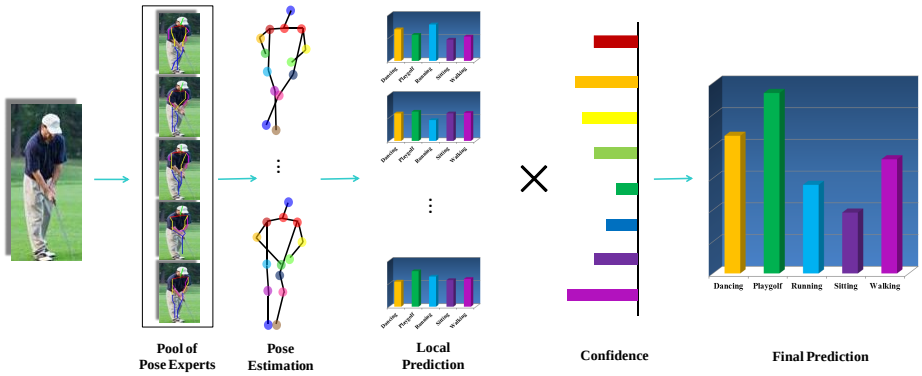


Fig. 1. The stages of our action prediction pipeline. For a test image shown in the leftmost, we use an ensemble of pose experts to extract pose cues from it, which are then respectively fed into the corresponding action predictor to evaluate its posterior probability distribution over action types. Finally, the conditional expectation is calculated for each action type over the pose experts involved, based on which the final decision is made. Technically, we use a discriminative framework with group sparse constraints to jointly train this series of action predictors (detailed in Section 3.2). This essentially allows each pose expert to play different roles in different action prediction tasks according to their specificity, and the strength of each experts is thus adaptively combined for the final action prediction.

actions under certain situations, but also can help us analyze and recognize human behaviors.

In this paper, we present a new method for recognizing human actions from still images. Our contribution is two-fold. First, we propose a global mixture of pose experts for more accurate articulation modeling. In contrast to a previously-proposed state of the art method [8] which uses a single pictorial tree with mixture of small, non-oriented parts to model the non-linear and non-convexity of the pose manifold of human being, we build a pool of pose experts, each of which individually models a particular type of articulation for a group of human bodies. We investigate two ways to construct the groups, one is based on the local shape statistics from pose annotations, and the other uses the semantic similarity related to action types. We show that both methods lead to improved pose estimation performance under difficult conditions. Furthermore, the estimated poses could be used for other tasks rather than action recognition, such as image retrieval by pose.

As our second contribution, we propose a flexible strategy which adaptively integrates the output of pose experts in a discriminative framework for action prediction (c.f., Fig. 1). In contrast to previous methods of combining the output of each pose expert using some relatively simple strategy such as majority voting, our method essentially allows each pose expert to adjust their roles in prediction according to their specificity when facing samples from different action types. To

achieve this, we first use a graph structure to capture both the non-local and local spatial correlation of the body shape estimated by the pose expert. Each graph is then treated as a single unit, over which an overall group sparse constraint is imposed to train the prediction model, with extra weight added according to the confidence of the corresponding expert. We show in our experiments on a challenging web data set with state of the art results that our method effectively improves the tolerance of our system to imperfect pose estimation.

The rest of this paper is organized as follows. After a brief review of the related work in Section 2, we details our method in Section 3. Experimental results are given in Section 4. We conclude our paper in Section 5.

2 Related Work

The major challenges of action recognition from still images come from the variability of human visual appearances (possibly with highly cluttered background), many degrees of freedom in human body postures, and lack of motion cue. In this section we give a brief review on how to deal with these issues in the literature.

Particularly, these methods can be roughly categorized into two classes. The first type of methods are appearance-based, in which various invariant feature descriptors, such as SIFT, HOG [9], visual words [10], and so on, are used as cues for action recognition [11, 1, 12, 10]. Despite many successes achieved by these methods, we argue that invariant feature sets are insufficient alone for this complicated task, since most of them can only provide partial invariance - some address this type of variations and others address that but not all; and even with these feature sets, lots of prototypes are still needed to cover the huge range of the variability exhibited in the pose space of human body, not to mention such a representation is usually with high dimension. To deal with these issues, some authors proposed to enhance the stability of feature sets using various context information (if available), such as human-object context[13–15] or group context[16, 11, 1], or using a multiple cues based approach to combine the strength of different features [2]. Recently, Wang et al. introduce a method which relies on more semantically meaningful features (i.e., pose-lets) and arrange them in a hierarchical manner to improve the invariance and discriminative power of the feature representation [3], and achieves the state of the art performance on a challenging web data set with still images [17].

Alternatively, one can decompose the task into two subsequent tasks by estimating the pose first and then recognizing the actions [9, 15, 18, 19], due to the fact that the pose conveys a lot of information about the actions. These approaches can also be thought as a way to adopt a distributional representation (pose vector) as the feature sets and we call them pose-based methods. While the appearance-based methods enjoy the rich information extracted from the raw data, the pose-based ones take the advantages of more compact pose representation and higher degree of interpretableness for human beings (regarding the results of action recognition yielded).

However, it is worth mentioning that building the action classifier directly based on the output of pose estimator [18, 19] could be dangerous due to the inherent uncertainty of the articulation modeling. To alleviate this issue, Yang et al. propose to couple the task of pose estimation with the goal of actions recognition such that more discriminative poses could be learnt [9]. This method proves to be very successful. However, in some real world situations, the boundary between two actions may be not so clear (e.g., running and walking), hence the supervision information injected through the action labels could be misleading.

If motion information is available, both of the above two types of representation could be extended to their 3D versions by modeling the input sequences as a tensor, as in dense trajectory [20, 4], action bank [21], among others [22–24]. These methods are related to our method but is unfortunately beyond the scope of the current work. we note that there exist huge number of static images of human beings in the internet, and it is of interest to properly model them and infer their high-level semantics, e.g, their poses and actions.

3 The Proposed Method

The stages of our action prediction pipeline are shown in Fig. 1. In what follows, we give the details of two major components involved, ie., the pose experts and the corresponding action prediction model.

3.1 Pose Experts

A Part-based Method In this work, we adopt a variant of part based model (PBM) as our pose expert. The basic idea of PBM [25, 26, 8] is to decompose the whole human body into many local parts (the feet, hands, arms, legs, torso, etc.), modeling them separately, and assembling them in such a way that the resulting configuration satisfying well the spatial constraints imposed by the exemplars. Mathematically, this is often equivalent to fit a tree structured model on the given image. One problem of PBM, however, is how to effectively characterize a large amount of poses in a single tree. Recently Yi Yang and Deva Ramanan [8, 27] proposed a variant of PBM called flexible mixtures-of-parts (FMP), to address this issue, and successfully applied it for human pose estimation and human detection. Chen et al. [28] improve the model by incorporating the local context information in multiple scales, and achieve more accurate results.

The key idea behind FMP is to use mixture of small, non-oriented parts for articulation modeling and to learn the spatial constraints between these mixtures under a discriminative structural learning framework. Compared to the single modal Gaussian as adopted in many PBM models, the mixture structure effectively enhances the capability of PBM to represent human body with various poses. However, what the FMP learns is essentially still a tree model, which is limited in considering only the first order spatial relationship between two adjacent parts, thus ignoring the high order spatial constraints of human body.

In other words, the FMP is a flexible model to impose the complex local compatibility on the pose space but somehow lacks non-local or global compatibility (e.g., the spatial regularity between one's left leg and his right leg, as usually exhibited in some type of actions).

To deal with this problem, we propose to group the pose space according to the desired global compatibility before articulation learning, with each group consisting of samples with similar poses or semantic meaning (e.g, actions). We then train for each group one pose estimator (called pose expert in this work) specific to that group, using the implementation of Chen et al. [28] (kindly provided by the authors). When testing (e.g, performing pose estimation for a never-seen image), we simply pick up the one output by the pose expert with highest confidence.

It is worthy mentioning that this idea of pose expert is related to that of poselet [29]. The major distinction between poselet and our method, however, is that the poselet groups parts of human body while we groups human bodies with similar poses. This different methodology leads to more broader difference when using them. For example, it is straightforward to apply our pose experts in the task of unsupervised human parsing, while a poselet model is more useful in detecting the parts of human body under different poses.

Grouping the Pose Space The properties of the group have direct influence on the specificity of the pose expert trained on it. Here two methods are considered: one is based on some semantical similarity while the other is on pose similarity. For our task at hand, one straightforward way to measure the similarity between samples is their action types. Hence in the first method (called action-specific grouping), we group together those images with the same action type (i.e., walking, running, etc.), and train one pose expert for it. This is similar to [9], but the difference lies in that they have to perform a dynamic programming-based searching for the most likely latent pose for each test image, while we consider more pose candidates due to the inherent ambiguities of pose expert.

As another strategy, we consider a more generative way to construct the pose expert, by grouping the training samples in the pose space (hence called pose-specific grouping). For this we have to design a similarity metric which reliably measures the pose similarity between two pose vectors. The traditional Euclidean distance is not a good choice since the pose vectors may distribute in a rather non-linear way in the pose space and it does not take the spatial correlation between parts into consideration. To address this issue, we use shape context feature, first used in shape matching and object detection by Belongie [30], to capture such information.

In particular, consider the set of vectors originating from one part to all other parts on a pose. These vectors express both the local and non-local configuration information of the entire shape relative to the reference part, and this information is summarized by the shape context feature as a 2D histogram. Hence shape context feature sets could be used to represent well the internal structure of the parts of human pose. In our implementation, we calculate the 2D histogram for

each part, vectorize it, and concatenate all these to get a representation for the pose of a human body. Usually this could lead to a vector with high dimension (e.g, over 6,000), and one can use PCA to condense it. With these in hand, we use the K-means algorithm with Mahalanobis distance as similarity measure to perform the grouping operation.

In either ways, we obtain several groups of poses with some degree of global compatibility preserved. We then use Chen et al.'s improved FMP model [28] to construct a pose expert for each group.

Human Parsing Using an Ensemble of Pose Experts When the task of pose estimation is of interest by itself, we use a minimum error rate principle to regress for a test image the pose using the pre-trained ensemble of pose experts. This is done by simply picking up the one output by the pose expert with highest confidence as the estimation.

3.2 Action Recognition

Graph-based Action Representation The output of each pose expert is a tree with its each node corresponding to a part in a human body, we can simply vectorize this tree for action representation [8, 27]. One limitation of this representation is that the non-local information between two non-adjacent parts is ignored, while it is well known that when training samples is few, preserving as rich information as possible is of importance for the subsequent classification task. Here we use an undirected complete graph structure, so that the spatial information regarding to any two body parts is explicitly encoded. This is similar to the shape context feature we used before when grouping the poses, and in fact the shape context feature can be interpreted as a more compact or discretized version of the complete graph.

Besides these first order features, we also incorporate a subset of second-order features by calculating the angle at the center part of an ordered triple parts. This kind of high order features is usually ignored in the previous work but is proven to be discriminative for some action types. For example, the angle formed by upper arm and lower arm in the action of walking is always bigger than that in a running action. As another example, the angle between upper leg and lower leg in playing golf would be always approximately equal to 180°.

More formally, assume that our training data set have been grouped into H clusters as described in Section 3.1, based on which we learn H pose expert, denoted as e_j. Then for a given image I, the output of the j-th expert is denoted as $R_j = e_j(I)$. Further assume that each human body has K body parts, and R_j is actually a vector with its component being the location $p_k = (x_k, y_k)$ of each part estimated by the expert, denoted as $R_j = (p_1, p_1, \ldots, p_K)$. With this, we construct a feature representation x^j for each pose expert j as follows:

$$x^j = (\psi_{1,2}, \psi_{1,3}, \ldots, \psi_{1,K}, \psi_{2,3}, \ldots, \psi_{K-1,K}, \theta_1, \theta_2, ..., \theta_{K'}) \tag{1}$$

where ψ and θ denotes respectively the first and second order features (i.e., angles mentioned before). In particular, the first order feature between any two parts m and n can be calculated as follows:

$$\psi_{m,n} = \varphi\left(I, p_m, p_n\right)$$

$$\varphi\left(I, p_m, p_n\right) = \left[dx\ dx^2\ dy\ dy^2\right]$$

where $dx = x_m - x_n$, $dx^2 = \left(x_m - x_n\right)^2$ and $dy = y_m - y_n$, $dy^2 = \left(y_m - y_n\right)^2$, accounting both the relative distance and the relative orientation between these two parts. This can also be understood as modeling the negative spring energy associated with pulling part i from a typical relative location with respect to part j. Hence given H pose experts, we have for each image I a feature representation x: $x = [x^1\ x^2 \ldots\ x^H]$.

Combining Pose Experts via Group Sparse Model One of the major challenges we face when identifying action type from the output of pose expert is the inherent ambiguity in articulation modeling. In other words, we cannot assume that the pose estimated by pose experts is perfect but in fact it is noisy and weak (in terms of performance). Hence it is risky to simply rely on the pose estimated by the pose expert with the highest score for action prediction. Instead, a better choice is to follow the Bayesian idea, i.e., taking all the output of pose experts in our pool into account.

Specifically, given a feature representation x, our goal is to estimate the maximum posterior probability of action a, i.e., $p(a|x)$. For this we train a series of action predictors corresponding to each pose expert in the pool and properly combine their responses for the final decision. Particularly, for a particular action type a, denote the parameter of the action predictor corresponding to the j-th pose expert as w^j. We jointly learn all the action predictors $w = \{w^1, w^2, ..., w^H\}$ by maximizing the following objective:

$$p(w|a, x^1, x^2, ..., x^H) \propto p(a|x^1, x^2, ..., x^H, w^1, w^2, ..., w^H) \prod_j p(w^j) \qquad (2)$$

where the model parameter w is assumed to have multivariate independent and identical priors. We use the logistic regression to model $p(a|x, w)$ as,

$$p(a|x, w) = (1 + exp(-a(\sum_j (w^j)^T x^j + b))^{-1} \qquad (3)$$

and the prior of w^j is modeled as Laplace, whose energy is further scaled according the confidence of the corresponding pose expert (detailed below). Note that in the above formulation, although the behavior of each action predictor for a pose expert is independent by the prior assumption, they jointly make the final prediction by summarizing their responses before undergoing a nonlinear transformation (c.f., Eq. 3).

Now, assume that N training data points $(a_i, x_i), a_i \in \{+1, -1\}$ are available to us, we reach the following objective by Eq. 2,

$$\min_w \sum_{i=1}^{N} log(1 + exp(-a_i(\sum_j (w^j)^T x_i^j + b))) + \lambda \sum_{j=1}^{H} (\frac{1}{n_j} \sum_{i=1}^{n_j} S(I_i, e_j)) ||w^j||_2 \quad (4)$$

where b is the bias shared by all the action predictors, and the scaling factor over the energy of laplace prior is defined to the average confidence of the corresponding pose expert, i.e.,

$$\alpha_j = \frac{1}{n_j} \sum_{i=1}^{n_j} S(I_i, e_j) \quad (5)$$

where n_j is the size of group j (c.f., Section 3.1), and $S(I, e_j)$ is the score or confidence of pose expert e_j for image I, which is known to us after the pose estimation stage.

Note that Eq.4 imposes a critical constraint that the pool of action predictors should not contribute equally to the final prediction, and some of them will even be canceled with a probability related to the confidence of the corresponding pose expert. This effectively improves the robustness against ambiguity in articulation estimation. To solve Eq.4, we use an efficient implementation of proximal methods [31]. After this, we can use these action predictors to perform action recognition in the test stage, as illustrated in Fig. 1.

4 Experiments

In this section, we report our experimental results concerning two series of experiments, i.e., human parsing (Section 4.1) and action recognition (Section 4.2).

4.1 Human Parsing Using an Ensemble of Pose Experts

We test our approach on two publicly available data set: the UIUC people data set [32] and the still web image data set collected by Ikizler-Cinbis et al. [17].

UIUC People Data Set The UIUC people data set [32] contains 593 still images. Most of these images are about people playing various sports such as badminton, Frisbee, walking, jogging or standing, hence contains very aggressive pose and spatial variations (c.f., Fig. 2). We follow the commonly used evaluation protocols in this dataset with the standard data partitions (346 for training, 247 for testing). The original dataset has 14 parts location annotated on the human body in each image, but we use 26 parts model as in [8].

For performance evaluation, we use the Percentage of Correct Parts (PCP) metric [8, 28]. A part is localized correctly only if both the distances of the endpoints from their respective ground truth endpoints are less than a fraction

(usually set as 0.5) of the part length. With this, the percentage of correct parts can be calculated for each image and then be averaged across all images.

We compare with several related state-of-the-art approaches that do full-body parsing: the iterative parsing method [25], the improved pictorial structure [33], and the discriminative hierarchical part-based model [3], Poselet conditioned pictorial structures [34], the flexible mixture of parts model [8] and its improved version by Chen et al.[28]. Note that since the UIUC data set has no action labels, we only built our pose experts according to the pose similarity (c.f., Section 3.1).

Method	Torso	Head	Upper legs	Lower legs	Upper arms	Lower arms	Total
Ramanan [25]	44.1	30.8	9.5	25.3	11.1	25.5	21.8
Andriluka [33]	70.9	59.1	36.5	22.9	26.2	10.1	32.1
Wang [3]	86.6	68.8	56.3	50.2	30.8	20.3	47.0
Pishchulin [34]	**91.5**	85.0	66.8	54.7	38.3	23.9	54.4
Yang [8]	85.0	83.4	63.6	56.3	48.8	34.6	57.6
Chen [28]	87.9	**85.4**	64.2	57.5	49.2	**38.3**	59.1
Ours (pose-specific)	89.5	84.6	**73.1**	**63.8**	**50.2**	37.8	**63.4**

Table 1. Comparison of various part-based human parsing methods on the UIUC Peoples dataset.

Table 1 and Fig. 2 give the results. It is clear from this table that our method based on an ensemble of pose-specific experts performs best among the compared ones. In particular, it improves the previous state-of-the-art performance [28] from 59.1% to 63.4%, and achieves better accuracy on most of key parts - notably, compared to [28], the proposed method significantly improves the localization accuracy of upper legs (from 64.2% to 73.1%) and upper arms (from 49.2% to 50.2%) (c.f., Fig. 2).

Still Web Image Data Set We also evaluate our method on the still web image data set by Ikizler-Cinbis et al. [17]. This data set consists of still images from 5 kinds of human action: dancing, sitting, playing golf, walking and running. Since those images are all downloaded from Web, human poses vary greatly and lots of images have cluttered backgrounds. Compared to the UIUC data set, this data set also contains far more images (2458 images in all). Thanks Yang et al. [9] for providing us their pose annotation with 14 joints on the human body on all the images in the data set. For evaluation, we follow Yang [9] and Wang [3] by partition 1/3 of the images from each kind of action for training, and the rest of the images are used for testing. Unfortunately, both authors do not perform human parsing experiments on this data set, and here we only compare our method with the baseline method [28].

Table 2 gives the result. It reveals that both grouping methods (action-specific and pose-specific) for pose experts construction lead to improved human parsing performance than the baseline algorithm, and the pose-specific grouping method works best as expected. It is worthy noting that an ensemble of action-specific pose experts still outperform the single-tree based model [28] - this is somewhat surprising since the grouping criteria is not originally designed

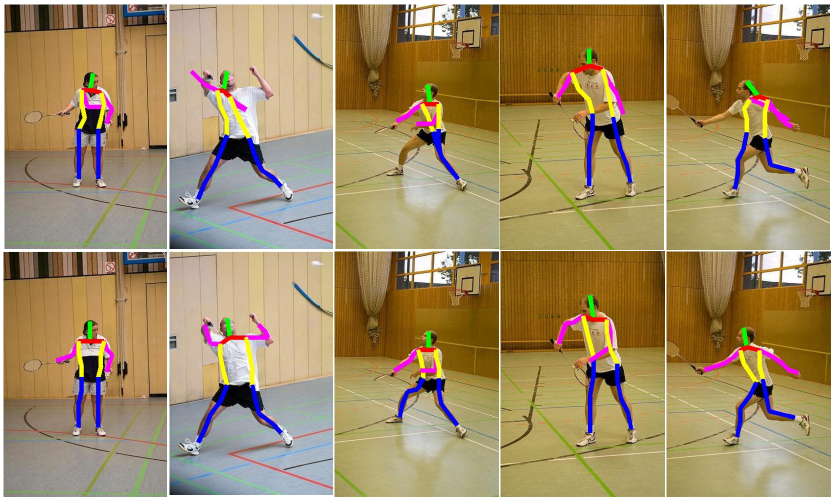


Fig. 2. Visualization of human parsing by the baseline method [28] (top row) and the proposed method (bottom row) on the UIUC people data set.

Method	Torso	Head	Upper legs	Lower legs	Upper arms	Lower arms	Total
Baseline [28]	96.6	95.3	60.1	58.7	51.0	28.3	58.8
Ours (action-specific)	**97.7**	**96.1**	62.6	59.9	54.8	31.9	61.2
Ours (pose-specific)	97.6	95.2	**68.9**	**62.4**	**60.0**	**35.4**	**64.6**

Table 2. Comparative human parsing performance of our method and the baseline method on the still web image data set.

for human parsing but for action recognition, but it can be partly explained by the conjecture that the top-level semantic information is beneficial to the task of pose estimation, as implied in Yang et al. [9]. Fig. 3 illustrates some human parsing results yielded by the compared methods.

4.2 Action Recognition from a Single Web Image

Next we report our experiments on the task of action recognition from a single web image, based on the still web image data set described in Section 4.1. As stated before, our evaluation protocol follows the ones proposed by Yang [9] and Wang [3] , i.e., using 1/3 of the images from each kind of action for training, and the rest of the images for testing.

Effectiveness of the Proposed Method To assess the effectiveness of the proposed method, we first designed several baseline algorithms by modifying one or some of its component, as follows,

– **Baseline algorithms**: We first learn a single tree-structured articulation model based on [28] from the training data. With the poses estimated by

Fig. 3. Visualization of various methods for human parsing on the still web image data set. From top to bottom, each row shows the results yielded by the baseline model [28], our method with action-specific \ pose-specific experts, respectively.

this model for the training set, we train a sparse logistic regression (LR) model as the action predictor. Two versions are implemented with different representation for action recognition, i.e., the tree-based representation and the graph-based representation (c.f., Section 3.2), as shown in the first two rows in Table 6;

- **Different grouping methods**: All the remaining variants use multiple pose experts, which are trained either in an action-specific way or in a pose specific way, as described in Section 3.1;
- **Different combining methods**: We test three kinds of ways to combine the output of pose experts for action prediction: 1) train a sparse logistic regression-based action predictor for each pose expert and combine their prediction by majority voting; 2) train the action predictors in the same as

Methods	Overall
Single Expert + Tree Rep. + Sparse LR	60.81
Single Expert + Graph Rep.+ Sparse LR	**63.79**
Action-Specific Experts + Graph Rep. + Max. resp.	62.87
Action-Specific Experts + Graph Rep. + Majority Voting	63.10
Action-Specific Experts + Graph Rep. + wGrpSparse LR	**66.08**
Pose-Specific Experts + Graph Rep. + Max. resp.	65.74
Pose-Specific Experts + Graph Rep. + Majority Voting	66.08
Pose-Specific Experts + Graph Rep. + wGrpSparse LR	**67.84**

Table 3. Performance (%) of variants of our method on the still web image data set, with both overall and average per-class accuracies reported.

previous one, but trust the one with maximum response when combining them (denoted as max. resp.); 3) jointly train those action predictors as described in Section 3.2, denoted as "wGrpSparse LR" (Weighted Group Sparse Logistic).

Table 3 gives the results. From this table we have the following observations: First, the results show that rich information is useful for more accurate action recognition - this can be seen by comparing the results shown in the first two rows - under the same settings, the method based on graph representation outperforms the one using tree-based representation. Secondly, it can be seen that the two types of action recognizers based on an ensemble of pose experts outperform the baselines. In fact, the best performer is based on a pool of pose-specific experts, achieving an accuracy of 68.84% that is over 4.0% higher than the best baseline method.

Thirdly, the table reveals that jointly training all the action predictors are effective in fusing the strength of each pose expert. By comparing the performance of jointly trained model with the ones trained independently (Max. resp. or Majority Voting), we find that the former consistently performs better than the later. To gain further understanding on this, we show for pose-specific experts-based method the energy of each action predictor assigned by the learner and the corresponding accuracy in Table 5 and Table 4, respectively. One can see from these two tables that different action predictors are good at predicting different type of actions while the energy assigned by the learner (Eq. 4) is proportion to this. For example, one can see from Table 4 that the 7-th action predictor is good at recognizing dancing and running, but not so good at playing golf and working. Accordingly, we see from the 7-th row of Table 5 that they receive large energy in both dancing and running, but will be excluded to make a prediction about playing golf and working.

Last but not least, it can be observed that the method based on the pose-specific experts work better than that based on the action-specific ones. This may be unexpected since pose-specific experts do not rely on any supervision information, while action-specific experts are trained deliberately for each type of actions. Despite this, Table 2 shows that on average action-specific experts do not perform as well as pose-specific ones in the task of human parsing (61.2%

	dancing	playgolf	running	sitting	walking	mean
group 1	68.25	56.41	78.75	46.01	52.65	63.69
group 2	55.26	47.13	80.50	73.24	57.98	65.29
group 3	64.09	41.58	73.17	57.71	62.36	62.75
group 4	59.38	63.10	81.78	70.63	32.70	63.19
group 5	59.60	44.97	80.96	41.76	59.54	61.85
group 6	56.28	54.39	79.06	69.47	60.68	65.76
group 7	66.31	37.43	71.55	44.95	36.69	55.17
group 8	49.56	46.90	76.97	36.33	60.27	58.11

Table 4. The action recognition accuracy (%) of each action predictor (each row), whose energy assigned by the learner is shown as one corresponding row in Table 5.

	Dancing	playgolf	running	sitting	Walking
group 1	1.9071	1.4831	1.0617	0.5722	0.6082
group 2	0.3056	0.7620	1.2251	1.3800	1.0011
group 3	1.0645	0.6213	0.8023	0.6384	1.0438
group 4	0.8377	1.6826	1.9567	1.7014	0.1466
group 5	0.9136	0.3988	1.0441	0.0000	1.4968
group 6	0.7858	1.6478	1.6883	0.9625	1.7677
group 7	1.9493	0.0000	0.7548	0.3556	0.0000
group 8	0.7763	0.5117	0.8766	0.0000	1.6639

Table 5. The energy of each action predictor (each row, the group number is kept the same as the corresponding pose-specific expert) as a function of action type. These energy numbers are calculated as the l_2-norm of the corresponding weight vector w^j jointly learned by optimizing objective (4). The individual accuracy of these eight action predictors is shown in Table 4.

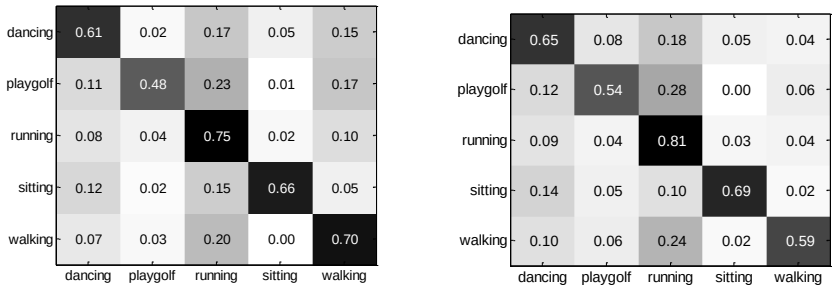
vs. 64.6%) - this implies that a less accurate pose estimator may lead to a deteriorated overall performance for action prediction.

To further understand the behavior between the two types of experts, we detail their confusion matrix in Fig. 4(a) and Fig. 4(b), respectively. By comparing these, we find that about 24.0% running actions are misclassified as walking by the approach based on pose-specific experts, while this number reduces to 20.0% by the one based on action-specific experts. This indicates that injecting high-level semantic information into the articulation model is useful to reduce the ambiguity for action prediction. Actually, since the poses of walking and running are similar to each other in many cases, images with these two kinds of actions are highly possibly to be clustered into the same group.

Comparison with the State-of-the-art Methods We compare our method with two state-of-the-art action recognition methods on still images, i.e., Yang [9] and Wang [3], and the results are given in Table 6. It can be seen from the table that our approach performs better than both methods. However, it should be noted that the accuracy numbers are not directly comparable since the training/testing data sets and features are not completely identical.

5 Conclusions

In this paper, we present a new method for human parsing and action recognition from a single still image, which is a less-studied problem. We base our method



(a) (b)

Fig. 4. Confusion matrix of the classification results on the still web image action dataset, based on (a) action-specific pose experts and (b) pose-specific experts. Horizontal rows are ground truths,and vertical columns are predictions.

Methods	Overall
Baseline	63.79
Yang [9]	61.07
Wang [3]	65.15
Ours (action-Specific)	**66.08**
Ours (pose-Specific)	**67.84**

Table 6. Comparative performance (%) of our method and two state-of-the-art methods on the still web image data set.

on a pool of pose experts, and show how to construct these pose experts and how to flexibly combine the output of these experts for improved action recognition performance. Our experiments on a challenging data set with web images indicate that 1) compared to the single expert strategy, our multiple experts approach is more effective for both tasks when the training data are relatively few; 2) our modified group sparse logistic regression learner leads to better performance than the one that trains its module independently. The importance of rich information for action recognition is also highlighted. Our current search is focused on how to extend the proposed method to the situation when motion cues are available.

Acknowledgement. The authors want to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (61073112, 61035003, 61373060), Jiangsu Science Foundation (BK2012793), Qing Lan Project, Research Fund for the Doctoral Program (RFDP) (20123218110033).

References

1. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: Computer Vision and Pattern Recognition

(CVPR), 2013 IEEE Conference on, IEEE (2013) 652–659
2. Sener, F., Bas, C., Ikizler-Cinbis, N.: On recognizing actions in still images via multiple features. In: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer (2012) 263–272
3. Wang, Y., Tran, D., Liao, Z., Forsyth, D.: Discriminative hierarchical part-based models for human parsing and action recognition. The Journal of Machine Learning Research **13** (2012) 3075–3102
4. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103** (2013) 60–79
5. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
6. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision **79** (2008) 299–318
7. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 3., IEEE (2004) 32–36
8. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1385–1392
9. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2030–2037
10. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: the 2010 British Machine Vision Conference. Volume 2. (2010) 1–11
11. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3177–3184
12. Sharma, G., Jurie, F.: Learning discriminative spatial representation for image classification. In Hoey, J., McKenna, S.J., Trucco, E., eds.: the 2011 British Machine Vision Conference, BMVA Press (2011) 1–11
13. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE (2010) 9–16
14. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31** (2009) 1775–1789
15. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 17–24
16. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS. Volume 4321. (2010) 4322–4325
17. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: the IEEE 12th International Conference on Computer Vision(CVPR 2009), IEEE (2009) 995–1002
18. Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. In: the Tenth IEEE International Conference on Computer Vision(ICCV 2005). Volume 1., IEEE (2005) 144–149

19. Ramanan, D., Forsyth, D.A.: Automatic annotation of everyday movements. In: Advances in neural information processing systems. (2003)
20. Jiang, Y.G., Dai, Q., Xue, X., Liu, W., Ngo, C.W.: Trajectory-based modeling of human actions with motion reference points. In: Computer Vision–ECCV 2012. Springer (2012) 425–438
21. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1234–1241
22. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE (2012) 20–27
23. Yuan, C., Li, X., Hu, W., Ling, H., Maybank, S.: 3D R transform on spatio-temporal interest points for action recognition. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 724–730
24. Zhu, Y., Chen, W., Guo, G.: Fusing spatiotemporal features and joints for 3d action recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, IEEE (2013) 486–491
25. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS: the Twentieth Annual Conference on Neural Information Processing Systems; 2006 December 4-7; Vancouver Canada. Volume 19., MIT Press (2006) 1129–1136
26. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61** (2005) 55–79
27. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35** (2013) 2878–2890
28. Chen, M., Tan, X.: Part-based pose estimation with local and non-local contextual information. IET Computer Vision (2014) 1–12
29. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 1365–1372
30. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24** (2002) 509–522
31. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Arizona State University. (2009)
32. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: Computer Vision–ECCV 2010. Springer (2010) 227–240
33. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1014–1021
34. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 588–595