IEICE TRANS. INF. & SYST., VOL.E101–D, NO.3 MARCH 2018

PAPER

Pose Estimation with Action Classification Using Global-and-Pose Features and Fine-Grained Action-Specific Pose Models

Norimichi UKITA[†a)], ***Senior Member***

SUMMARY This paper proposes an iterative scheme between human action classification and pose estimation in still images. Initial action classification is achieved only by global image features that consist of the responses of various object filters. The classification likelihood of each action weights human poses estimated by the pose models of multiple sub-action classes. Such fine-grained action-specific pose models allow us to robustly identify the pose of a target person under the assumption that similar poses are observed in each action. From the estimated pose, pose features are extracted and used with global image features for action re-classification. This iterative scheme can mutually improve action classification and pose estimation. Experimental results with a public dataset demonstrate the effectiveness of the proposed method both for action classification and pose estimation.

key words: *human action classification, human pose estimation, action-specific pose models*

1. Introduction

Understanding human activities is one of the main topics in computer vision. This paper focuses on two kinds of representations of human activities, namely a human **body pose** and an **action class**.

In **action classification**, each action observed in images is classified to one of predefined classes (e.g. baseball, badminton). While most of related work classify the action in videos by using temporal cues, action classification in still images [1]–[7] is not only challenging but also useful for several uses (e.g. context-based image retrieval and static cues for classification in videos).

For action classification, both global and local features provide informative cues. Global scene features are useful because human actions are relevant to the surrounding scene and related objects [8]. These features are ambiguous for action classification but can be obtained without human region localization. While human localization is not trivial, local features extracted from the region of a person of interest are beneficial for action classification. For videos, for example, several kinds of bag-of-words representations are tested in [9] in order to find discriminative local features for simultaneous human localization and action classification. Human-region localization in still images requires more discriminative features and powerful classifiers. While human detection methods such as [10], [11] can cope with severe occlusion, they mainly focus on detecting upright people as a bounding box, in which the detailed silhouette of a person is not specified. A detailed silhouette can be found by human pose estimation, which is described in what follows.

A **human body pose** in images can be defined by a deformable part model (DPM), e.g. [12], [13]. The model consists of nodes and links, which respectively correspond to a part and a geometric relationship between parts. Human pose estimation using DPMs has been extended in various aspects, e.g. appearance features between parts [14] and segmentation-based appearance features [15]. The part configuration is a crucial cue for estimating a complicated human pose in cluttered background. In addition, a human pose represents a human activity even in still images. That is, human pose and action are mutually related to each other.

The contributions of this work are (i) to improve pose estimation by multiple DPMs each of which is optimized for each action (i.e., action-specific pose models) and (ii) to employ an estimated pose as local features merged with global image features for action classification. While portions of this paper appeared in [16], this paper also presents the effects of fine-grained action-specific DPMs.

2. Related Work

Contexts for action classification and pose estimation: As well as the state of a human body, interactions between the body and objects with regard to the activity of a person are also useful for human activity recognition. For action classification, for example, mutual relationships between a human pose and objects being manipulated [17] and between action-specific body parts and objects required for the action [18], [19] are useful. As well as object features, scene features can be used for action classification [8]. The global scene features can potentially represent all objects surrounding a person. Pose estimation can be also improved by involving interacting objects in the DPM [20].

Relations between actions and poses: Action classification and pose estimation can enhance each other. Action classification can be achieved by pose matching (e.g. 2D pose-based matching [21], [22] in videos and view-invariant 3D pose matching in videos [23]–[25]). 3D pose-based matching is proved empirically to be robust to noise in [25]. Intuitive examples in still images are shown in Fig. 1 (a) and (b). These two poses are similar to each other because both persons are hitting a shuttlecock during playing badminton.

In an opposite manner, for pose tracking in videos,

Manuscript received June 24, 2017.
Manuscript revised October 23, 2017.
Manuscript publicized December 8, 2017.

[†]The author is with Toyota Technological Institute, Nagoya-shi, 468–8511 Japan.
a) E-mail: ukita@toyota-ti.ac.jp

DOI: 10.1587/transinf.2017EDP7204

action-specific model selection has been studied (e.g. switching models [26], efficient particle distribution in pose models [27], [28], unified multi-action model [29]).

Recognition in still images: Compared with recognition in videos by the aforementioned methods [23]–[29], it is more difficult to extract discriminative features from still images. For action classification, a large variety of body poses might be contained in the same action class. Examples in Fig. 1, (d–g) show significantly different 2D body poses in the same action class, baseball. The difference is caused by the following issues. (i) Class resolution: different sub actions, batting and pitching, are contained in the same class, (ii) view dependency: the same poses are captured from different viewpoints, and (iii) classification in still images: different moments (i.e. different poses) of batting are contained in the baseball class. While problems (i) and (ii) must be coped with also in videos, problem (iii) is a unique problem in still images. In this paper, in contrast to our previous work [16], the effectiveness of fine-grained action-specific models (e.g. batting and pitching rather than baseball) for the first issue is verified.

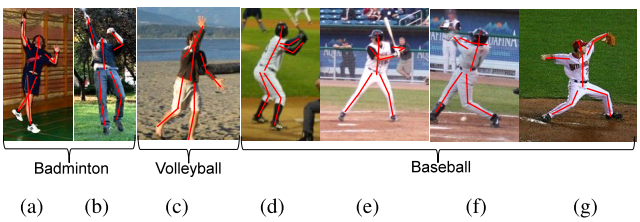

Fig. 1 Pose representation with 10 body parts. (a) and (b) are classified to *badminton*, (c) is *volleyball*, and (d), (e), (f), and (g) are *baseball*.

Furthermore, one more difficulty in recognition in still images is person localization, as tackled in [30]. This problem is more difficult than the one in videos [31], in which motion cues can be used for foreground segmentation. This difficulty is absent in classifying a scene, where each target action is performed, by using global image features. Indeed, the co-occurrence between actions and scenes is a useful clue for mutually improving their classification [32]. Unlike traditional approaches using only global features (e.g. GIST [33]), more recent ones fuse multiple features and/or classifiers; joint optimization of multiple classifiers [34], simultaneous classification and annotation using regional features [35], [36], and classification using DPMs [37]. In particular, the Object Bank [38] allows us to obtain the responses to any kinds of objects including people and objects relevant to the action, as well as background objects.

3. Overview of the Proposed Method

The overview of the proposed method is illustrated in Fig. 2. Action classification and pose estimation are iteratively performed so that (i) action classification is performed by global features and 2D pose-based features and (ii) body poses are estimated by fine-grained action-specific DPMs. An action classifier and DPMs are trained with the action labels and pose annotations of training images.

Initially (i.e., before a human pose is estimated), action classification is executed with no human pose information. Robust initialization is achieved by global features (denoted by "Object Bank $\boldsymbol{O}$" in Fig. 2) with no human localization, as described in Sect. 4.1.

The result of action classification is used for weight-

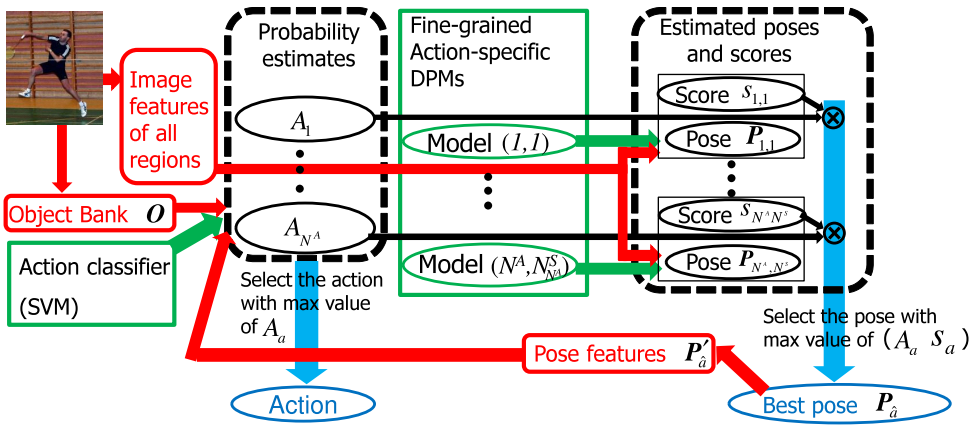

Fig. 2 Overview. Red, green, and black arrows depict the data flows of feature vectors, model parameters, and estimated values, respectively. The model parameters of SVMs and DPMs are employed with the features for action classification and pose estimation, which are performed in lefthand and righthand dotted rectangles, respectively. While the action classifier delivers the probability estimates (denoted by $A_1, \cdots, A_{N^A}$ in the figure) of N^A action classes, a DPM is prepared not for each action class but for each action sub-class. Each DPM is used for pose estimation. Among all poses estimated by DPMs, the best pose is selected so that it has the largest value of the product of the probability estimate of the action classifier and the DPM score, as indicated by $\otimes$ in the figure. A pose feature vector, which is produced from the selected pose, is fed back to be merged with global image features (i.e. Object Bank features) for iterative action classification. After iteration between action classification and pose estimation is finished, their final results are determined, as depicted by blue downward arrows.

ing the pose estimation results of action-specific DPMs (indicated by "Model(1, 1), $\cdots$, Model($N^A, N^S_{N^A}$)", where Model(X, Y) denotes a pose estimation model for Y-th sub-class of X-th action, in the figure), as described in Sect. 4.2. N^A and N^S_a denote the numbers of action classes and sub-classes of a-th action, respectively. The total number of DPMs is $N^M = \sum_a^{N^A} N^S_a$. Each DPM represents human poses observed in its respective action. By dividing a variation of possible human poses into such multiple DPMs, a variety of poses in each DPM becomes smaller. Then modeling error in each DPM is reduced. Finally, only one pose (indicated by "Best pose $\boldsymbol{P}_{\hat{a}}$" in the figure) is selected from N^M poses based on the scores of action classification and pose estimation.

In addition to action classes used in our previous work [16], human poses in each action class are further divided into sub-classes; for example, pitching and batting in the baseball class. Each sub-class has its own action-specific DPM. While this sub-classification (i) needs additional sub-class tagging and (ii) reduces the amount of training images for each sub-class, a variation in each sub-class is decreased and fine-grained classification can be achieved.

While each action sub-class has its own DPM, all sub-classes within their parent class share the same model of global features for action classification. This is because sub-actions in each class are performed in similar environments where similar global features are observed. Therefore, the poses of all sub-classes in a-th action class are evaluated with the probability estimate of a-th action class in the pose selection process mentioned above.

As well as global features, pose-based features (denoted by "Pose features $\boldsymbol{P}'_{\hat{a}}$" in Fig. 2) are used for action re-classification in iterative steps, as described in Sect. 4.3. This is the difference from previous pose-based action classification [23]–[25].

In what follows of this section, base approaches for action classification [38] and pose estimation [39], which are used in the proposed method, are described.

3.1 Action Classification Using Global Features

The Object Bank [38] provides a set of high-level image features with scale-invariant response maps of a variety of generic object detectors. The response maps are normalized so that the Object Bank features have fixed dimension regardless of the image size. 177 object detectors, which are provided by its author's codes, are used in our implementation. In total, the size of the Object Bank features is 44604-dimension; 252 D/object $\times$ 177 objects = 44604 D. The Object Bank features are extracted from each image and used with a classifier (e.g. SVM) for action classification.

3.2 Pose Estimation Using DPMs

A tree-based model is defined by a set of nodes, $\boldsymbol{V}$, and a set of links, $\boldsymbol{E}$, each of which connects two nodes. Each node has its pose parameters (e.g. x and y positions, orientation θ, and scale s) that localize the respective part. By optimizing the pose parameters in accordance with a human pose in an image, pose estimation is achieved. The pose parameters are optimized by maximizing the score function below:

$$T(\boldsymbol{P}) = \sum_{i \in \boldsymbol{V}} S^i(\boldsymbol{p}_i) + \sum_{i,j \in \boldsymbol{E}} P^{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j), \tag{1}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{P}$ denote a set of the pose parameters of i-th part and a set of $\boldsymbol{p}_i$ of all parts (i.e. $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{N^V}]^T$, where N^V denotes the number of nodes).

In our experiments, a pose estimation method proposed in [39] is employed. In [39], a unary term $S^i(\boldsymbol{p}_i)$ in (1) is an appearance-based similarity score of i-th part at $\boldsymbol{p}_i$. $\phi(\boldsymbol{p}_i)$ and F^i denote a local image patch in $\boldsymbol{p}_i$ and the i-th part's score function optimized by a deep convolutional neural network (DCNN) [40], respectively. The DCNN for F^i is trained so that it outputs a greater value if its input patch $\phi(\boldsymbol{p}_i)$ is similar to the appearance of i-th part. A pairwise term $P^{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is a set of a spring-based score and an image-dependent pairwise score, both of which have a greater value if the relative configuration of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is highly probable. $P^{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is jointly optimized by the DCNN used in optimizing F^i.

These two functions, $S^i(\boldsymbol{p}_i)$ and $P^{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ are trained properly with sample images where the pose annotations of all parts are given. This training is achieved by the structured SVM [39], [41] in our implementation so that it optimizes the pose annotations of the training data.

4. Iterative Action Classification and Pose Estimation

4.1 Initial Action Classification with High-Dimensional Global Features

Initial action classification is achieved by employing only Object Bank features. This initial classification is performed with the original dimension of the Object Bank features (denoted by $\boldsymbol{O}$) for achieving classification as accurate as possible. A SVM classifier [42], [43] gives us not only the action class of $\boldsymbol{O}$ but also its probability estimate A_a denoting the probability to belong to a-th action as proposed in [44].

Note again that all action classification processes, including this initial one as well as iterative ones, are achieved in terms of action classes rather than their sub-classes.

4.2 Pose Estimation by Fine-Grained Action-Specific DPMs and Action Probability

The result of action classification is employed for pose selection from human poses estimated by multiple action-specific DPMs. The action-specific DPM is defined for each action sub-class (i.e. fine-grained class). The representation of each fine-grained action-specific DPM is completely same with that of a general DPM described in Sect. 3.2.

However, the learning scheme of fine-grained action-specific DPMs is modified from original one [39] so that all DPMs in an action class share a step for initializing the

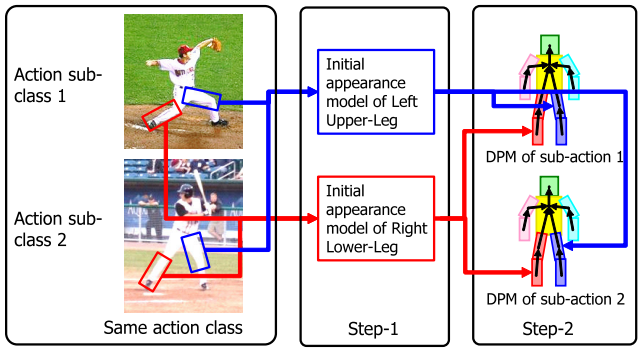

Fig. 3 Learning DPMs in an action class. Initially, the appearance of each part is trained with images of all sub-classes in Step-1. In Step-2, the DPM of each sub-class is modeled independently.

parameters of part-similarity functions, $S^i(\boldsymbol{p}_i)$ in score (1). Originally in [39] and our previous work [16], these parameters in each DPM are initialized with training images of only this DPM. While this training works well if a number of training images is given to this DPM, each action sub-class has a smaller number of training images in the proposed method. Our assumption to solve this problem is that (i) the appearance of parts is not significantly changed between action sub-classes but (ii) the configuration of parts (i.e., human pose) has a large variation between action sub-classes. Based on this assumption, (i) the training images of all sub-classes in each action are used for initializing the appearance parameters in score (1), $S^i(\boldsymbol{p}_i)$, as illustrated in Step-1 in Fig. 3. Then, (ii) for the DPM of each sub-class, all parameters in score (1) are jointly optimized using the initialized $S^i(\boldsymbol{p}_i)$ with training images of only this sub-class, as illustrated in Step-2 in Fig. 3.

Assume that N^M action-specific DPMs (denoted by "Model(1, 1), $\cdots$, Model($N^A, N^S_{N^A}$)" in Fig. 2), where N^M denotes the number of all action sub-classes, are given. For each test image, the proposed method obtains N^M poses, each of which has the top score (denoted by s_{a_f} in f-th sub-class model of a-th action) in each model. With s_{a_f} and the probability estimate of a-th action, A_a, the best pose (denoted by "Best pose $\boldsymbol{P}_{\hat{a}}$" in Fig. 2) is selected so that:

$$\hat{a} = \arg\max_{a,f}(s_{a_f} A_a) \tag{2}$$

While previous clustered models [45], [46] have no weights between different models, the proposed method has the benefit that the probability estimate of an action gives the weight to each model as shown in Eq. (2).

4.3 Action Classification with Global and Pose Features

For action classification, the absolute positions of parts in an image ($\boldsymbol{P}_{\hat{a}}$), which are obtained by pose estimation, should be changed to be invariant to human location and scale in an image. In our method, $\boldsymbol{P}_{\hat{a}}$ is changed to the following expression, $\boldsymbol{P}'_{\hat{a}}$ (denoted by "Pose features $\boldsymbol{P}'_{\hat{a}}$" in Fig. 2):

$$\boldsymbol{P}'_{\hat{a}} = [x_{<\hat{a},1>} - C_x,\ y_{<\hat{a},1>} - C_y,\ \cdots,$$
$$x_{<\hat{a},N^V>} - C_x,\ y_{<\hat{a},N^V>} - C_y]^T, \tag{3}$$

where $x_{<\hat{a},i>}$ and $y_{<\hat{a},i>}$ denote x-y positions of i-th part in $\boldsymbol{P}_{\hat{a}}$. (C_x, C_y) is the center of all $x_{<\hat{a},i>}$ and $y_{<\hat{a},i>}$. $\boldsymbol{P}'_{\hat{a}}$ represents the normalized relative positions of parts with respect to (C_x, C_y). In the body model with 26 parts, $\boldsymbol{P}'_{\hat{a}}$ is $26 \times 2 = 52$D. The pose features (3) are robust to a partial change in the pose of the whole body because the relative position between a parent and its child parts is independent from that between other parent and child parts.

The pose features (3) are used with the global features (i.e. Object Bank features, $\boldsymbol{O}$) for action re-classification. Based on experimental results in our previous work [16], these two features are combined so that they are concatenated, $[\boldsymbol{O}^T\ \boldsymbol{P}'^T_{\hat{a}}]^T$. This feature vector is then employed for action classification and probability estimation by multi-class SVM [42], [43], as with initial action classification.

With the newly estimated probabilities, pose estimation is executed again. Iteration between pose estimation (Sect. 4.2) and action classification (Sect. 4.3) is performed like the hard EM algorithm, where action classification and pose estimation are respectively regarded as the E and M steps, observed data are global features, latent variables are action classification probabilities and action-specific models, and unknown parameters are the pose parameters of the whole body.

5. Experiments

We tested the proposed method with the LEEDS sports (LSP) dataset [45] and the LSP extended training dataset [46], in which 2000 and 10000 pose-annotated images are included, respectively. In all comparative experiments in this section, 1000 images in the LSP were used for evaluation. All images were automatically collected from Flickr based on keywords related to the following sports: athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball. Since the keywords were provided by Flickr users as they like, several images are not related to any of the above eight sport classes: for example, a sitting audience at a stadium. These non-sport images have a ninth action class, "general" in our experiments, because the objective of action-specific models is to precisely represent the pose variation triggered by each action. Each image was manually annotated by one of the nine action classes[†]. Figure 4 shows examples of the nine classes. The number of sample images clustered to each action class is listed in Table 1.

Action classes in Table 1 are divided into sub-classes as shown in Table 2. While body poses included in gymnastics and parkour classes can be classified to several sub-classes, these poses are captured from a variety of orientations (e.g. from below and above a person) and the configurations of $x-y$ body joint coordinates are not clustered even in such a sub-class. So all images of gymnastics and parkour classes are included in one sub-class, "Others".

[†]The action annotations are available in the author's website.

Fig. 4 Sample images of nine action classes. The cropped window of a target person is shown.

Table 1 The number of images of each action class in training and test data. While the training dataset includes 1000 LSP images and 10000 LSP extended images ("LSP + LSP extended" in the Table), the test dataset consists of 1000 images of the LSP.

	Athletics	Badminton	Baseball	Gymnastics	Parkour	Soccer	Tennis	Volleyball	General
Train	33+542	130+0	159+101	29+2904	74+5643	127+15	98+1	76+4	274+790
Test	46	127	137	40	88	125	93	87	257

Table 2 The number of training and test images of sub-classes.

(1) Athletics	Running	Jumping	Others	(5) Parkour	–	–	Others
Train	29+100	3+388	1+54	Train	–	–	74+5643
Test	45	0	1	Test	–	–	88
(2) Badminton	Waiting & Downswing	Upswing & Serve	Others	(6) Soccer	Running	Shot	Others
Train	72+0	37+0	21+0	Train	65+0	34+1	28+8
Test	84	32	11	Test	73	14	38
(3) Baseball	Batting	Pitching	Others	(7) Tennis	Waiting & Receiving	Serve & Smash	Others
Train	38+4	45+38	76+59	Train	55+1	22+0	21+0
Test	35	38	64	Test	45	34	14
(4) Gymnastics	–	–	Others	(8) Volleyball	Waiting & Receiving	Serve, Smash & Blocking	Others
Train	–	–	29+2904	Train	16+4	32+0	28+0
Test	–	–	40	Test	34	35	18
(9) General	–	–	Others				
Train	–	–	274+790				
Test	–	–	257				

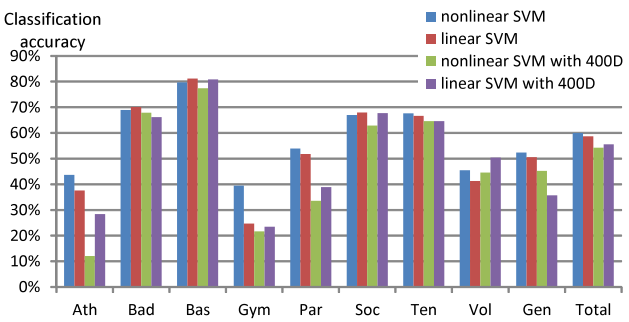

Fig. 5 Comparison of classification performance of different classifiers in initial action classification with high-dimensional and compressed Object Bank features.

The following two models were obtained in training:

- Multi-class SVM for action classification: Nonlinear [43] and linear [42] SVMs were tested.
- Action-specific DPMs of all action sub-classes: Each DPM was implemented based on [39]. The appearance feature of this base model is trained with DCNNs.

The results of initial action classification only with the Object Bank features are shown in Fig. 5. For each of the nonlinear and linear SVMs, high-dimensional and compressed Object Bank features were tested. The dimension of the compressed features was 400, which was around 1% of the high-dimensional one. While the compressed features could get nice results as with the results of scene classification reporter in [38], the high-dimensional features were better than the compressed one.

After action classification, the probability estimate of each action class is used in pose selection from multiple DPMs. As shown in Fig. 5, the results of action classification were not enough high to identify the action class of each test image. That is, in many test images, the probability estimate of a correct class was not the max score among all classes. However, even if the probability estimate of a correct class is not the max score, it gives a useful clue to pose selection if the following two conditions are satisfied:

1. The probability estimate of a correct class is not much lower than the max score.
2. The probability estimate of a correct class is relatively higher than the scores of other classes.

The evidences of the first condition mentioned above is shown in Fig. 6. Figure 6 shows the mean of p^{cor}/p^{max}, where p^{cor} and p^{max} denote the probability estimate of a

Table 3 Quantitative comparison using test data in the LSP dataset and the strict PCP metric [47], [48]. Our method used [39] for each DPM. We used the person-centric annotations given in [46].

	Torso	Upper-Leg	Lowe-Leg	Upper-Arm	Lower-Arm	Head	Total
(a) Ours (final iteration)	97.5	90.6	85.5	83.9	71.2	88.9	84.9
(b) Ours (first iteration)	97.1	89.1	82.2	80.7	68.9	88.6	82.8
(c) Our previous [16]	87.3	74.7	68.5	54.1	36.9	77.9	63.4
(d) Chen & Yuille [39]	96.0	77.2	72.2	69.7	58.1	85.6	73.6
(e) Wei et al. [49]	96.9	85.6	81.6	78.1	66.4	94.0	81.4
(f) Yu et al. [50]	98.0	93.1	88.1	82.9	72.6	83.0	85.4
(g) Rafi et al. [51]	97.6	87.3	80.2	76.8	66.2	93.3	81.2

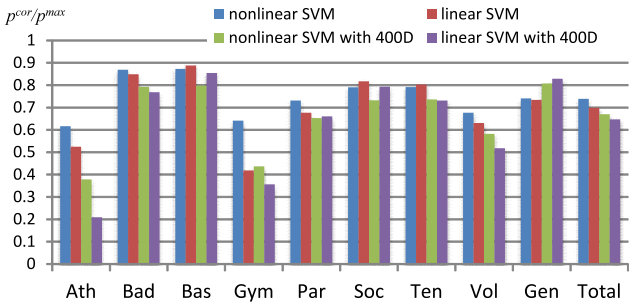

Fig. 6 Comparison of probability estimates of correct classes versus classes having the max scores in initial action classification.

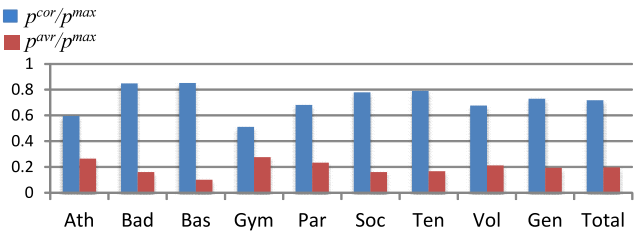

Fig. 7 Comparison of probability estimates of correct classes versus other classes in initial action classification. Nonlinear SVM with high-dimensional features was used. The mean probability of the other classes is shown in the graph (indicated by red bars).

correct class and the max score of all probability estimates in each image, respectively. It can be seen that (i) nonlinear SVM with high-dimensional features (indicated by blue bars) were superior to other classifiers and (ii) in many classes, p^{cor} was not much lower than p^{max} (at least, 60% of p^{max}) by using nonlinear SVM with high-dimensional features. Therefore, nonlinear SVM with high-dimensional features were used for obtaining s_a in Eq. (2) in the following experiments.

The second condition mentioned above is verified in Fig. 7. Figure 7 shows the mean of p^{avr}/p^{max}, where p^{avr} denotes the mean of probability estimates of all classes except a correct class in each image, estimated by nonlinear SVM with high-dimensional features. In addition to p^{avr}/p^{max} (indicated by red bars in Fig. 7), p^{cor}/p^{max} is also indicated by blue bars for comparison. It is clear that p^{cor} was higher than other scores.

The quantitative results of pose estimation are shown in Table 3. The accuracy is evaluated by the PCP (percentage of correctly estimated parts) [47]. The final result of our method was obtained after two iterations between action classification and pose estimation. The iteration was executed only twice because of the following two reasons:

- In general, repeated iterations cause overfitting in the EM-like algorithm.
- Both action classification and pose estimation almost converged in the second iteration. In particular, the iterative steps had less impact on pose estimation. This is because the probability estimate of a correct action class is significantly better than those of other classes even in the first iteration (as shown in Fig. 7). On the other hand, action classification was improved in iterations because pose features are added for this action classification.

The accuracy of the proposed method is better than our previous method (i.e., (a) > (c) in Table 3). The iterative scheme improved the performance (i.e., (a) > (b) in Table 3). It can be also seen that the proposed method outperforms or comparable with state-of-the-art methods, (e), (f), and (g). In particular, the effectiveness of our proposed method can be confirmed by comparing results between the proposed method and its base method, (d); the improvement in Total is 84.9 − 73.6 = 11.3%.

Since all images of action classes 1 to 8 were captured in sport scenarios, many people wear uniforms except the parkour class where most people wear no clothes, which result in less appearance variation. In class 9 (i.e., general class), people wear various kinds of clothes. While the uniformity and diversity of the body appearance affect the appearance score in (1), it can be seen that pose estimation accuracy in parkour class is the worst as shown in Table 3. Our interpretation of this observation is that the specific appearance variation in the sport scenarios may have less impact than the variation of a body configuration.

The effect of action classification on pose estimation is verified by improvement on pose estimation. If the pose estimation score of a correct/incorrect pose has the largest value but is exceeded by another/correct pose using the probability estimate of unsuccessful/successful action classification, its effect is regarded as negative/positive. While unsuccessful classification gives a negative impact on pose estimation accuracy (−1.3%), a positive impact by successful classification (4.2%) is larger.

Figure 8 shows several typical examples of estimated poses. The results of the base method [39] and our previous

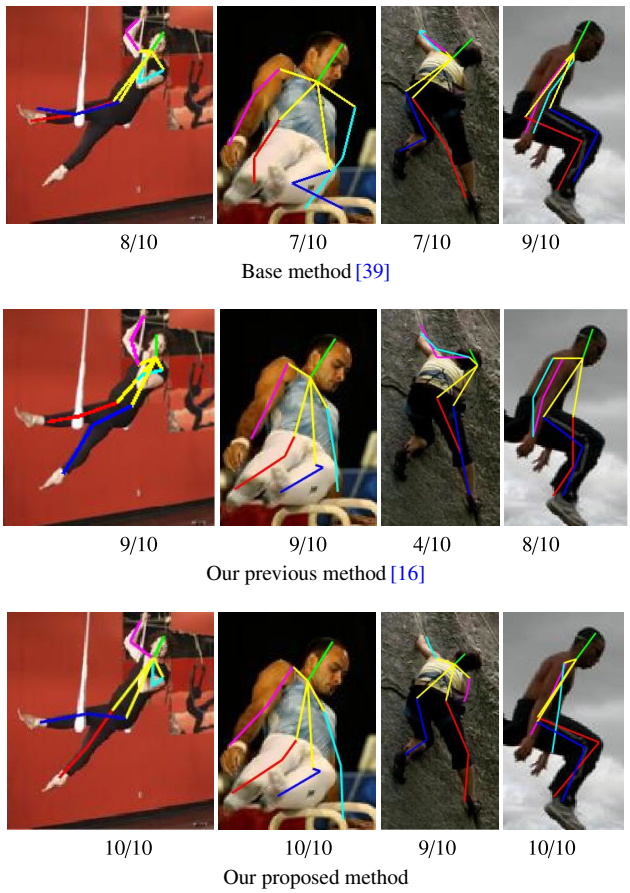

Fig. 8 Pose estimation results. The window of a target person was cropped from its original image and shown in this figure. (Top) a base method [39], (Middle) our previous method, in which action-specific DPMs are used, and (Bottom) the proposed method using fine-grained action-specific DPMs. The number of correctly localized parts is shown under each result. All of the images were selected from gymnastics and parkour classes, where human poses are significantly different from natural upright poses.

method with only action-specific DPMs [16] are also shown. As shown in these examples, our methods were successful in particular in gymnastics and parkour in contrast to the base method [39]. This is because the body poses of these actions are significantly different from those of other actions, but the variety of poses in each action was represented well by the proposed action-specific DPMs.

Finally, the results of action classification using Object Bank features with pose features are shown in Fig. 9. Red and green bars show the results of our previous work [16] and the proposed method with action sub-classes, respectively. The results were obtained after two iterations between action classification and pose estimation. For comparison, the initial classification results of the proposed method are also indicated by blue bars. From Fig. 9, it can be seen that the proposed method improved the classification accuracy from our previous work [16]. This improvement must be obtained due to better pose features because the global features used in our previous work [16] and the proposed method were identical.

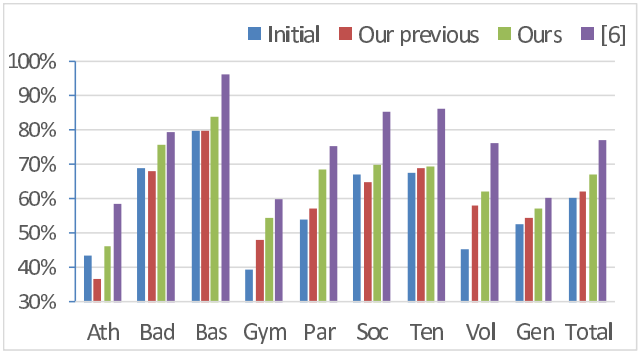

Fig. 9 Comparison of action classification performance of different classifiers. While initial classification ("Initial" in the figure) was achieved only with the Object Bank features, other two results were obtained by pose features with Object Bank features. These results were obtained by nonlinear SVM after twice-iterated action-and-pose recognition.

For comparison, the classification results of the state-of-the-art using DCNNs [6] are also shown in Fig. 9. In our experiments, top 1000 regions extracted by selective search [52] were used as region proposals for [6]. While this method [6] achieves a high performance gain in most action classes, the gain in the general class is relatively small. This may be because of a huge variety of image appearances in the general class. Note that, this method [6] and our proposed method is not compared on a fair basis. This is because this method [6] uses the ground-truth bounding box of a target person while our proposed method is designed not to require the bounding box both in training and testing phases. However, using DCNN-based action classification should be regarded as important future work for improving not only pose estimation but also action classification by the proposed joint-utilization of image and pose features.

The contributions of the proposed method validated in the experimental results are summarized as follows:

- The effectiveness of fine-grained action-specific DPMs are proved compared not only with state-of-the-art pose estimation methods but also with our previous action-specific DPMs [16] with no fine-grained modeling, as shown in Table 3 and Fig. 8.
- Pose features improve action classification as shown in Fig. 9.

6. Concluding Remarks

This paper proposed an iterative method for human action classification and pose estimation in still images. These action classification and pose estimation enhance to each other so that action classification is improved by pose features with global appearance features, and pose estimation is augmented by fine-grained action-specific DPMs.

Future work includes (i) joint optimization of multiple DPMs that share the basic structure of a human body and (ii) more discriminative pose features that are robust to the change in a viewpoint. The former is useful for improving pose estimation even if a small number of training images

are given in each action sub-class. The latter enables more correct action classification. While the proposed method requires manually-given action labels, unsupervised learning of these action labels (e.g. [53]) is an interesting topic.

As described in the summary of experiments, using DCNNs for more discriminative appearance features in action classification as well as in pose estimation is also an important research direction.

A known disadvantage of the proposed method is that the number of pose estimation models (i.e., PSMs) increases as the number of action classes increases. While all pose estimation models work in parallel, this parallel estimation is impossible if we have a huge number of action classes. In particular, only a limited number of pose estimation models can be used in parallel if the models employ DCNNs. This is because the memory on a GPU used by DCNNs is limited. To cope with this problem, exploring efficient pose estimation models using less memory is also an important future work.

References

[1] C. Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," In CVPR, pp.1–8, 2008.
[2] N. Ikizler-Cinbis, R.G. Cinbis, and S. Sclaroff, "Learning actions from the web," In ICCV, pp.995–1002, 2009.
[3] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," In CVPR, pp.2030–2037, 2010.
[4] S. Maji, L.D. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," In CVPR, pp.3177–3184, 2011.
[5] F.S. Khan, R.M. Anwer, J. van de Weijer, A.D. Bagdanov, A.M. López, and M. Felsberg, "Coloring action recognition in still images," International Journal of Computer Vision, vol.105, no.3, pp.205–221, 2013.
[6] G. Gkioxari, R.B. Girshick, and J. Malik, "Contextual action recognition with r*cnn," In ICCV, pp.1080–1088, 2015.
[7] Y. Zhang, L. Cheng, J. Wu, J. Cai, M.N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," IEEE Trans. Image Processing, vol.25, no.11, pp.5479–5490, 2016.
[8] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," In ECCV (1), vol.6311, pp.494–507, 2010.
[9] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," In ICCV, pp.2003–2010, 2011.
[10] R. Xu, B. Zhang, Q. Ye, and J. Jiao, "Cascaded l1-norm minimization learning (clml) classifier for human detection," In CVPR, pp.89–96, 2010.
[11] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis, "Human detection using partial least squares analysis," In ICCV, pp.24–31, 2009.
[12] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, 2010.
[13] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," In CVPR, pp.1385–1392, 2011.
[14] N. Ukita, "Articulated pose estimation with parts connectivity using discriminative local oriented contours," In CVPR, pp.3154–3161, 2012.
[15] N. Ukita, "Part-segment features with optimized shape priors for articulated pose estimation," IEICE Transactions, vol.99-D, no.1, pp.248–256, 2016.
[16] N. Ukita, "Iterative action and pose recognition using global-and-pose features and action-specific models," In Workshop on Understanding Human Activities: Context and Interactions, pp.476–483, 2013.
[17] A. Gupta, A. Kembhavi, and L.S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.10, pp.1775–1789, 2009.
[18] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and F.-F. Li, "Human action recognition by learning bases of action attributes and parts," In ICCV, pp.1331–1338, 2011.
[19] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," In NIPS, pp.1503–1511, 2011.
[20] B. Yao and F.-F. Li, "Modeling mutual context of object and human pose in human-object interaction activities," In CVPR, pp.17–24, 2010.
[21] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.7, pp.1310–1323, 2011.
[22] K.N. Tran, I.A. Kakadiaris, and S.K. Shah, "Part-based motion descriptor image for human action recognition," Pattern Recognition, vol.45, no.7, pp.2562–2572, 2012.
[23] P. Natarajan, V.K. Singh, and R. Nevatia, "Learning 3d action models from a few 2d videos for view invariant action recognition," In CVPR, pp.2006–2013, 2010.
[24] V.K. Singh and R. Nevatia, "Action recognition in cluttered dynamic scenes using pose-specific part models," In ICCV, pp.113–120, 2011.
[25] A. Yao, J. Gall, G. Fanelli, and L.V. Gool, "Does human action recognition benefit from pose estimation?," In BMVC, pp.67.1–67.11, 2011.
[26] J. Chen, M. Kim, Y. Wang, and Q. Ji, "Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition," In CVPR, pp.2655–2662, 2009.
[27] J. Gall, A. Yao, and L.J. Van Gool, "2d action recognition serves 3d human pose estimation," In ECCV, vol.6313, pp.425–438, 2010.
[28] N. Ukita, "Simultaneous particle tracking in multi-action motion models with synthesized paths," Image Vision Comput., vol.31, no.6-7, pp.448–459, 2013.
[29] N. Ukita and T. Kanade, "Gaussian process motion graph models for smooth transitions among multiple actions," Computer Vision and Image Understanding, vol.116, no.4, pp.500–509, 2012.
[30] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," In ECCV, vol.7575, pp.158–172, 2012.
[31] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori, "Similarity constrained latent support vector machine: An application to weakly supervised action classification," In ECCV, vol.7578, pp.55–68, 2012.
[32] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," In CVPR, pp.2929–2936, 2009.
[33] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol.42, no.3, pp.145–175, 2001.
[34] C. Li, A. Kowdle, A. Saxena, and T. Chen, "Towards holistic scene understanding: Feedback enabled cascaded classification models," In NIPS, pp.1351–1359, 2010.
[35] C. Wang, D.M. Blei, and F.-F. Li, "Simultaneous image classification and annotation," In CVPR, pp.1903–1910, 2009.
[36] L.-J. Li, R. Socher, and F.-F. Li, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," In CVPR, pp.2036–2043, 2009.
[37] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," In ICCV, pp.1307–1314, 2011.
[38] L.-J. Li, H. Su, E.P. Xing, and F.-F. Li, "Object bank: A high-

level image representation for scene classification & semantic feature sparsification," In NIPS, pp.1378–1386, 2010.
[39] X. Chen and A.L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," In NIPS, 2014.
[40] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," In NIPS, pp.1106–1114, 2012.
[41] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," In ICML, ICML '04, New York, NY, USA, p.104, ACM, 2004.
[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, vol.9, pp.1871–1874, 2008.
[43] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," ACM TIST, vol.2, no.3, p.27, 2011.
[44] T.-F. Wu, C.-J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," Journal of Machine Learning Research, vol.5, pp.975–1005, 2004.
[45] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," In BMVC, pp.1–11, 2010.
[46] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," In CVPR, pp.1465–1472, 2011.
[47] V. Ferrari, M.J. Marín-Jiménez, and A. Zisserman, "Progressive search space reduction for human pose estimation," In CVPR, pp.1–8, 2008.
[48] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," In CVPR, pp.3178–3185, 2012.
[49] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," In CVPR, pp.4724–4732, 2016.
[50] X. Yu, F. Zhou, and M. Chandraker, "Deep deformation network for object landmark localization," In ECCV, vol.9909, pp.52–70, 2016.
[51] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov, "An efficient convolutional network for human pose estimation," In BMVC, pp.109.1–109.11, 2016.
[52] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders, "Selective search for object recognition," International Journal of Computer Vision, vol.104, no.2, pp.154–171, 2013.
[53] J.C. Niebles, H. Wang, and F.-F. Li, "Unsupervised learning of human action categories using spatial-temporal words," International Journal of Computer Vision, vol.79, no.3, pp.299–318, 2008.

Norimichi Ukita is a professor at the graduate school of engineering, Toyota Technological Institute, Japan (TTI-J). He received the Ph.D. degree in Informatics from Kyoto University, Japan, in 2001. After working for five years as an assistant professor at NAIST, he became an associate professor in 2007 and moved to TTI-J in 2016. He was a research scientist of PRESTO, JST during 2002–2006. He was a visiting research scientist at Carnegie Mellon University during 2007–2009. His main research interests are multi-object tracking and human pose and activity estimation.