This ICCV2013 Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Iterative Action and Pose Recognition using Global-and-Pose Features and Action-specific Models

Norimichi Ukita
Nara Institute of Science and Technology
ukita@ieee.org

Abstract

This paper proposes an iterative scheme between human action classification and pose estimation in still images. For initial action classification, we employ global image features that represent a scene (e.g. people, background, and other objects), which can be extracted without any difficult human-region segmentation such as pose estimation. This classification gives us the probability estimates of possible actions in a query image. The probability estimates are used to evaluate the results of pose estimation using action-specific models. The estimated pose is then merged with the global 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 global features for initialization, action-specific models for pose estimation, and action classification with global and pose features.

1. Introduction

This paper focuses on two kinds of representations for human activities, human **body pose** and **action class**.

Most **action classification** methods classify actions in videos by using temporal cues. As the cues, a set of local features (e.g. spatio-temporal points [18], a bag of spatio-temporal words [26], a bag of spin-images [22], and a bag of motion words [35]) are widely used because of their effectiveness. The difficulty in using the local features is to extract them only from the region of a person of interest. In particular, region extraction in still images is difficult, while action classification in still images [33, 14, 37, 23] is not only challenging but also useful for several uses (e.g. context-based image retrieval and static cues for classification in videos).

In addition to the local features, the effectiveness of global scene features for action recognition has been proven [15]. While the global features are weak and auxiliary for identifying actions performed in an image, it can be extracted without human region segmentation.

(a) (b) (c) (d) (e) (f)

Figure 1. Pose representation with 10 body parts. (a) and (b) are classified to the same action, *badminton*, and (c), (d), (e), and (f) are *baseball*. Different poses are contained in the same class.

A **human body pose** in images is defined by, in general, a deformable part model. The model consists of nodes and links, which respectively correspond to a part and a geometric relationship between parts. Pose estimation is achieved so that all parts are located in an image in accordance with the body configuration of a target person (e.g. Fig. 1). The deformable part model has two kinds of parameters, namely the appearance parameters of each part and the relative geometric configuration between neighboring parts.

For detecting each part based on its appearance, image features are crucial for coping with a huge variety of part appearances. The features of each part can be divided into several clusters and then trained individually (e.g. clustering based on the configurations of 2D parts [38] and 3D parts [2]) for maintaining their discriminativity as well as generality. Discriminative training of part appearance can also improve part discriminativity [9, 1].

For parameterizing the relative configuration between parts, pictorial structure models [10], are widely used as deformable part models because of their ability to efficiently get the globally-optimal configuration of all parts.

For accurate pose estimation with a deformable part model, optimizing the above two parameters (i.e. appearance and configuration parameters) is a fundamental issue.

The contributions of this work are 1) to improve the accuracy of pose estimation by training multiple deformable part models in accordance with actions of interest and 2) to employ an estimated pose as a local feature merged with

global image features for improving action classification.

2. Related Work

Mutual action and pose recognition: While the above mentioned algorithms for action classification and pose estimation work independently, these two types of recognition can enhance each other. Action classification can be achieved by pose matching (e.g. view-invariant 3D pose matching in videos [25, 31, 39]). In an opposite manner, for pose tracking in videos, action-specific model selection has been studied (e.g. switching dynamical models [4] and efficient particle distribution in multiple pose models [12]).

An essential problem for mutual action and pose recognition is that a human pose is required for action classification achieved with the human pose. This is a chicken-and-egg problem. To cope with this problem, joint recognition of action and pose has been studied (e.g. [37, 40, 7]). Unlike this approach, this paper proposes an iterative scheme between action and pose recognition, where each recognition is simpler than joint recognition. In general, simplicity results in robustness in recognition.

Recognition in still images: Compared with video analysis by the above methods [25, 31, 39, 4, 12], 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. In examples in Fig. 1, (a–b) and (c–f) show significantly different 2D body poses in the same action classes, *baseball* and *badminton*, respectively. The difference is caused by the following problems. 1) Class resolution: different primitive actions, batting and pitching, are contained in the same class, 2) view dependency: the same poses are captured from different viewpoints, and 3) classification in still images: different moments (i.e. different poses) of batting are contained in the baseball class. While problems 1 and 2 must be coped with also in videos, problem 3 is a unique problem in still images.

Furthermore, one more difficulty in recognition in still images is person localization, as tackled in [7]. This problem is clearly more difficult than the one in videos [30], in which motion cues can be used for foreground object 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 [24]. Unlike traditional approaches using only global features (e.g. GIST [27]), more recent ones fuse multiple features and/or classifiers; joint optimization of multiple classifiers [19], simultaneous classification and annotation using regional features [34, 20], classification using deformable part based models [28], and scene representation with responses to a wide variety of objects [21]. In particular, the Object Bank [21] allows us to obtain the responses to any kinds of objects including people and objects relevant to the action, as well as background objects.

Objects interacting with a person of interest also gives important clues for pose estimation (e.g. [13, 40]). While the interacting objects might be more characteristic for identifying the human pose than a global scene, this paper focuses on the global features of the scene, which are easy to be extracted from an image for robust recognition.

3. Basic Scheme

The proposed method iteratively performs action classification and pose estimation so that 1) action classification is performed by global features and 2D pose-based features and 2) a body pose is estimated by an action-specific deformable part model optimized to the action observed in a still image. The overview of the proposed method is illustrated in Fig. 2. For this iterative framework, we have to cope with two issues; i) robust initialization of iteration between action classification and pose estimation and ii) a large variety of body poses that are observed in images of the same action class.

The proposed method achieves robust initialization by global features (denoted by "Object Bank feature $\boldsymbol{O}$" in Fig. 2), for which neither human localization nor human pose estimation are required, for initial classification.

A variety of body poses in the same class are produced by the three problems mentioned in the last section (i.e. class resolution, view dependency, and classification in still images). The pose variety in the same class makes it difficult to achieve 1) non-overlapping pose clustering among different action classes and 2) precise pose modeling. The proposed method alleviates these two difficulties as follows. 1) Inspired by a mixture of parts [38], a pose is featurized by a set of relative positions of parent and child parts. This featurization is 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. As well as the pose-based features (denoted by "Pose feature $\boldsymbol{P}'_{\hat{a}}$" in Fig. 2), global features are used for action classification in iterative steps (described in Sec. 4.3). This is the difference from previous pose-based action classification [25, 31, 39]. 2) For precise pose modeling, after training images are divided to each action class based on the ground-truth labels, poses in each action class are clustered based on their similarity (described in Sec. 4.2). Compared with clustering all possible poses [38, 16, 17], pose clustering in each action class is easier and results in precise modeling.

In what follows of this section, two base approaches for action classification [21] and pose estimation [9, 38], which are used in the proposed method, are described.

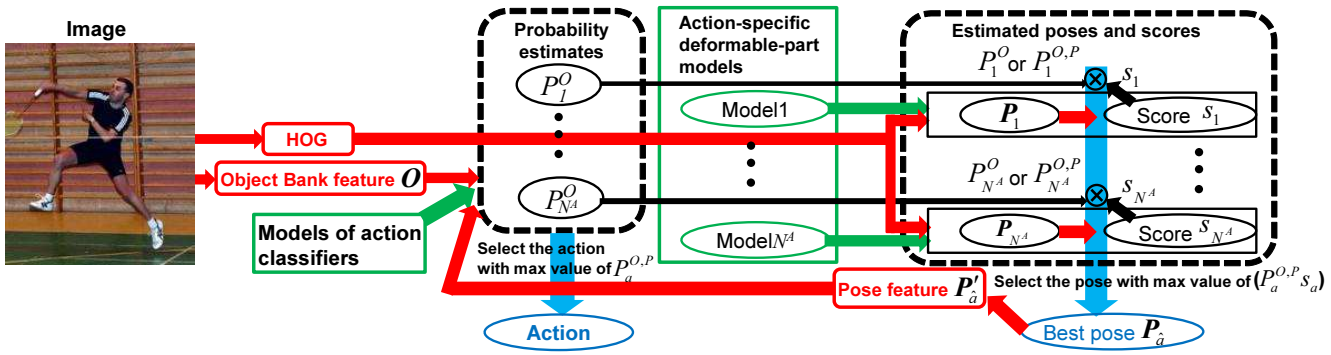

Figure 2. Overview of the proposed method. Red, green, and black arrows depict the data flows of feature vectors, model parameters, and estimated values, respectively. The model parameters are employed with the features for action classification and pose estimation, which are performed in left and right dotted rectangles, respectively. Pose features, which are produced from estimated poses, are fed back to be merged with global image features (i.e. Object Bank feature) for iterative action classification. After iteration between action classification and pose estimation is finished, their final results are determined, as depicted by blue arrows.

3.1. Action Classification using Global Appearance Features

The Object Bank [21] provides a set of high-level image features with scale-invariant response maps of a variety of generic object detectors. While the Object Bank accepts any object detectors, 177 object detectors, which are provided by its author's codes, are used in our implementation. In total, the size of the Object Bank feature is 44604-dimension. A high-level representation of the features has been demonstrated in terms of scene classification with large-scale datasets.

The huge dimensional feature can be compressed by a sparse coding regularization. The sparse coding allows us to reduce the feature dimension up to 10 % or lower while maintaining the classification accuracy.

3.2. Pose Estimation using Deformable part Models

A tree-based model is defined by a set of nodes, $\boldsymbol{V}$, and a set of links each of which connects two nodes, $\boldsymbol{E}$. One of the nodes is regarded as a root node. 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:

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

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).

A unary term $S^i(\boldsymbol{p}_i)$ is a similarity score of i-th part at $\boldsymbol{p}_i$. In our model, $S^i(\boldsymbol{p}_i)$ is the filter response using HOG features [5], each of which consists of 5×5 cells and 18 orientation bins: $S^i(\boldsymbol{p}_i) = F^{i^T} \cdot \phi(I, \boldsymbol{p}_i)$, where F^i and $\phi(I, \boldsymbol{p}_i)$ denote the filter of i-th part and the HOG extracted from $\boldsymbol{p}_i$ in image I. A pairwise term $P^{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \boldsymbol{w}^{i,j^T} \cdot \psi(\boldsymbol{p}_i, \boldsymbol{p}_j)$, where $\boldsymbol{w}^{i,j}$ is a weighted parameter, is a spring-based score between i-th and j-th parts, which has a greater value if the relative configuration of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is highly probable. In our model, $\psi(\boldsymbol{p}_i, \boldsymbol{p}_j) = [d^x_{i,j}, d^{x^2}_{i,j}, d^y_{i,j}, d^{y^2}_{i,j}]^T$, where $d^x_{i,j}$ and $d^y_{i,j}$ respectively denote $x_i - x_j$ and $y_i - y_j$, where (x_i, y_i) is the location of i-th part.

In a tree-based model proposed in [10], the globally optimized pose parameters, $\hat{\boldsymbol{P}}$, can be acquired efficiently by dynamic programming. To make this model work robustly to in-plane rotation and foreshortening of limbs, each rigid part (e.g. limb) is divided into several smaller parts in [38]. In accordance with this base model [38], 26 parts were used in our implementation; 2 for the head, 4 for the torso, 10 for the shoulders to the hands, and 10 for the hips to the feet.

4. Iterative Action and Pose Recognition

4.1. Initial Action Classification with Global Appearance Features

Initial action classification is achieved by employing only Object Bank features. The extracted feature (denoted by $\boldsymbol{O}$) is classified by linear [8]/nonlinear [3] SVM. These SVM classifiers give us not only the action class of $\boldsymbol{O}$ but also its probability estimate P^O_a denoting the probability to belong to a-th action (e.g. $P^O_1 \cdots P^O_{N^a}$ in Fig. 2) as proposed in [36].

4.2. Pose Estimation by Action-specific Deformable part Models and Action Probability

Next, pose estimation is performed independently using all action-specific models, which are trained in a training

phase. All training images are divided to those of each action based on the ground-truth action labels. Then each action-specific model is trained with the training images of its respective action.

Training data in each action-specific model are clustered in each part i. This part clustering is useful for precisely representing the model parameters (i.e. F^i and $w^{i,j}$) of i-th part. Our method adopts clustering based on the 2D configuration of parts as with [38]. In [38], each part i has its x-y location and scale parameter s as its parameters. Instead of having an orientation parameter θ, each part model consists of a mixture of *types* as follows. The training data of i is clustered depending on the relative location of i with respect to its parent part. This clustering is achieved by K-means with 5 or 6 clusters depending on the part in our experiments in accordance with the base model [38]. The ID of the cluster is called a *type*, which is denoted by t. The pose parameter of i-th part, $\boldsymbol{p}_i$, is expressed by $[x_i, y_i, s_i, t_i]$.

The above two kinds of independent modeling (i.e. action-specific models and clustering based on the 2D configuration of parts) are essentially different in terms of the number of the estimated pose(s).

Action-specific models: Given N^A action-specific models (denoted by "Model1 $\cdots$ ModelN^A" in Fig. 2), all of them are used independently for acquiring N^A poses (denoted by "$\boldsymbol{P}_1 \cdots \boldsymbol{P}_{N^A}$" in Fig. 2), each of which has the best score in each action-specific model. The action-specific modeling is effective in particular for representing the typical configurations of body parts depending on the action; for example, a handstand-like pose in athletics.

Clustering based on the 2D configuration of parts: Only one best pose of the whole body is acquired in each model, regardless of the number of types. Specifically the type having the best score is selected in each pair of parent and child parts.

The proposed method obtains N^A poses (denoted by $\boldsymbol{P}_a$ where $a \in 1, \cdots, N^A$), each of which has the top score (denoted by s_a) in a-th model. With the score s_a and the probability estimate of a-th action, P_a^O, which is obtained in action classification, the best pose $\boldsymbol{P}_{\hat{a}}$ (denoted by "Best pose $\boldsymbol{P}_{\hat{a}}$" in Fig. 2) is selected so that:

$$\hat{a} = \arg\max_a (s_a P_a^O) \tag{2}$$

While previous clustered models [16, 17] 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 model selection (2).

4.3. Action Classification with Global Appearance Features and Pose Features

To provide an estimated pose as an important clue to action classification, the absolute positions of parts in an image, $\boldsymbol{P}_{\hat{a}}$, should not be used as they are. This is because a pose feature should fit with the region of a human body in any location and of any scale. In our method, $\boldsymbol{P}_{\hat{a}}$ is changed to the following expression, $\boldsymbol{P}'_{\hat{a}}$ (denoted by "Pose feature $\boldsymbol{P}'_{\hat{a}}$"), represented by the normalized relative positions with respect to the center of all parts (denoted by (C_x, C_y)):

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

where $x_{<\hat{a},i>}$ and $y_{<\hat{a},i>}$ denote x-y positions of i-th part in $\boldsymbol{P}_{\hat{a}}$. In the body model with 26 parts, $\boldsymbol{P}'_{\hat{a}}$ is $26 \times 2 = 52$D. In our implementation, the pose feature (3) is compressed by a sparse coding regularization, as with Object Bank features [21]. This compression is for improving not efficiency but classification accuracy. The accuracy is improved by the compression because some components in (3) are completely indistinguishable among actions. The compressed dimension was empirically determined to be 22.

To leverage $\boldsymbol{P}'_{\hat{a}}$ for action classification, the following two kinds of methods are possible:

- $\boldsymbol{P}'_{\hat{a}}$ is concatenated to the Object Bank feature $\boldsymbol{O}$ for obtaining a new feature for action classification:

$$[\boldsymbol{O}^T\ \boldsymbol{P}'^T_{\hat{a}}]^T \tag{4}$$

  This feature is then employed for action classification by multi-class SVM [3, 8], as with initial action classification. Here again multi-class SVM gives us the probability estimate of each action (denoted by $P_a^{O,P}$).

- Multi-class SVM [3, 8] is applied to $\boldsymbol{P}'_{\hat{a}}$ in order to estimate the probability of being classified into action a. This probability estimate, P_a^P, is then multiplied by P_a^O as follows:

$$P_a^{O,P} = P_a^P P_a^O \tag{5}$$

  The max $P_a^{P,O}$ of all actions is detected, and a-th action corresponding to the max score is regarded as the classification result. (5) assumes that global scene features are independent of pose features, $\boldsymbol{P}'_{\hat{a}}$.

With the newly estimated probabilities, $P_a \propto P_a^{O,P}$, pose estimation is executed again. Iteration between pose estimation (Sec. 4.2) and action classification (Sec. 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. Note that, in the iterative steps, $P_a^{O,P}$ is used instead of P_a^O in Eq. (2).

Athletics Badminton Baseball Gymnastics Parkour Soccer Tennis Volleyball General

Figure 3. Sample images of nine action classes.

Table 1. The number of images of each action class in training and test data.

	Athletics	Badminton	Baseball	Gymnastics	Parkour	Soccer	Tennis	Volleyball	General
Training	32	145	157	44	70	147	103	112	190
Test	49	138	133	54	91	147	99	121	168

Table 2. Comparison of PCP. (a) our model (final iteration), (b) our model (initial iteration), (c) mixture model of non-oriented parts [38], (d) clustered pose [17], (e) parts dependent joint regressor [6], and (f) poselet conditioned pictorial structures [29].

	Torso	Head	Upper-legs	Lower-legs	Upper-arms	Lower-arms	Total
(a) Ours (final iter)	87.3	77.9	74.7	68.5	54.1	36.9	63.4
(b) Ours (initial iter)	86.9	77.9	73.3	67.8	54.1	36.4	62.8
(c) Mixture of parts [38]	84.1	77.1	69.5	65.6	52.5	35.9	60.8
(d) Clustered pose [17]	88.1	74.6	74.5	66.5	53.7	37.5	62.7
(e) Joint regressor [6]	81.6	79.2	66.5	61.0	45.1	24.7	55.5
(f) Poselet PS [29]	87.5	78.1	75.7	68.0	54.2	33.9	62.9

5. Experiments

We tested the proposed method with the LEEDS sports dataset [16], in which 2000 pose-annotated images are included. In all comparative experiments in this section, 1000 images were used for training and other 1000 images for evaluation. Each image was manually annotated by one of the following nine action classes: athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, volleyball, and "general"[1]. The class general is required because the objective of action-specific models is to precisely represent the pose variation triggered by each action, while general pose variations should be modeled by non-selective training images. Figure 3 shows examples of the nine classes. The number of sample images clustered to each action class is listed in Table 1.

In training, 1) the models of multi-class SVM for action classification and 2) action-specific deformable part models for pose estimation are acquired. For augmenting the action-specific deformable part models by discriminative training [9], negative samples were given from background images in the INRIA Person database [5].

The results of initial action classification only with the Object Bank features are shown in Fig. 4. Non-linear SVM [3] and linear SVM [8][2] with high-dimensional and compressed Object Bank features were tested. The dimension of the compressed feature was 400, which was around 1% of the high-dimensional one. While the compressed features could get nice results as reported in [21], the high-dimensional features were still better. As regards accuracy in different actions, it can be seen that a smaller number of training data resulted in lower accuracy (i.e. lower accuracy in athletics, gymnastics, and parkour).

Then the probability estimate of each action class is used in pose estimation. Even if the probability estimate of a correct class is not the max score, it gives a useful clue to pose estimation if it is 1) not much lower than the max score and 2) relatively higher than other scores. Figures 5 and 6 show evidences about these two requirements. Figure 5 shows the mean of p^{cor}/p^{max}, where p^{cor} and p^{max} denote the probability estimate of a correct class and the max score of all probability estimates in each image, respectively. It can be seen that 1) nonlinear SVM with high-dimensional features (indicated by blue bars) was superior to other classifiers and 2) in many classes, p^{cor} was not much lower than p^{max} (at least, 60% of the max score) by using nonlinear SVM with high-dimensional features. Therefore, nonlinear SVM with high-dimensional features was used for obtaining s_a in Eq.

[1]The action annotations will be available in the author's website.

[2]For SVM, default parameters given by [3] and [8] were used.

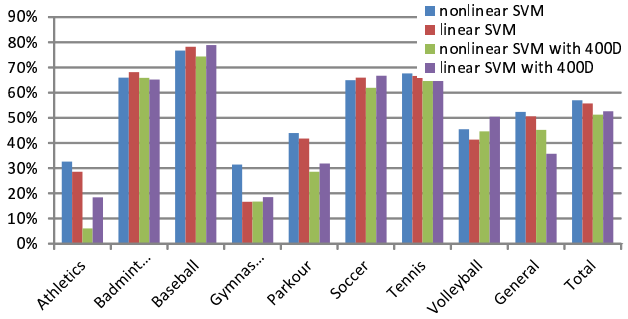

Figure 4. Comparison of classification performance of different classifiers in initial action classification with high-dimensional and compressed Object Bank features.

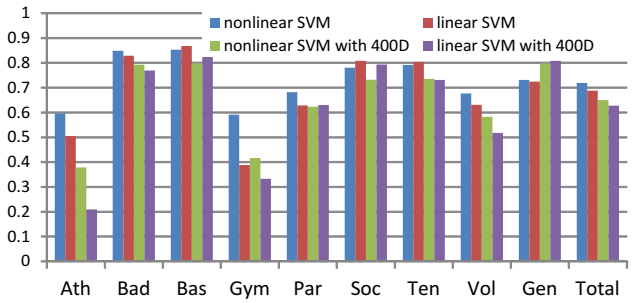

Figure 5. Comparison of probability estimates of correct classes versus classes having the max scores in initial action classification.

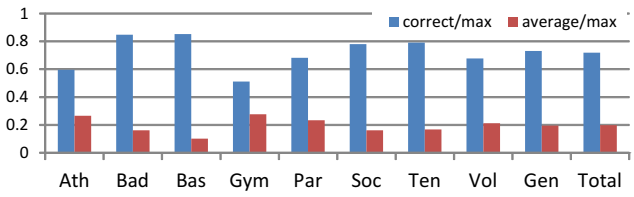

Figure 6. Comparison of probability estimates of correct classes versus other classes in initial action classification. Nonlinear SVM with high-dimensional features was used. The mean probability of the other classes is shown in the graph (indicated by red bars).

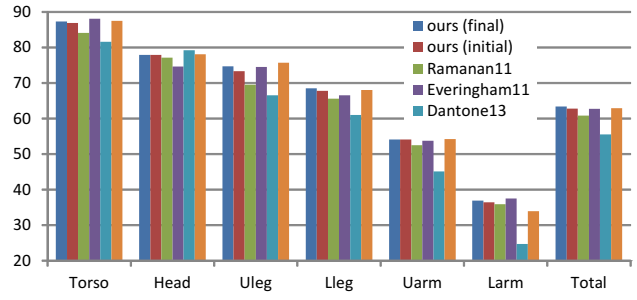

Figure 7. Comparison of pose estimation accuracy by PCP of different models. The same data is shown also in Table 2.

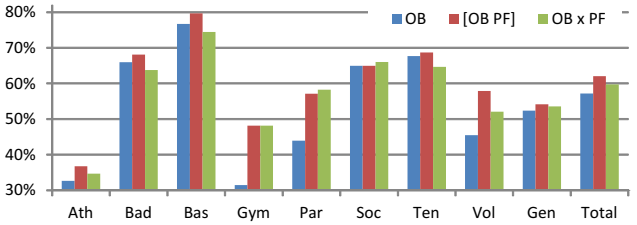

Figure 8. Comparison of classification performance of different classifiers in action classification by pose features with high-dimensional Object Bank features. The results were obtained by nonlinear SVM after twice-iterated action-and-pose recognition.

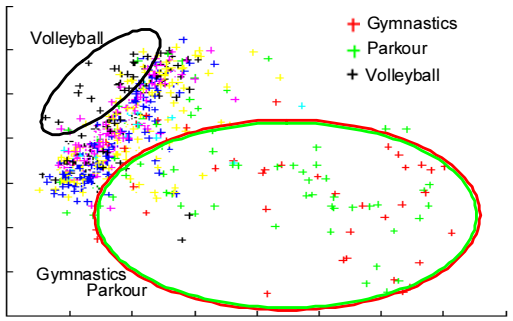

Figure 9. 2D representation of a 52D pose feature space.

(2). Figure 6 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. 6), p^{cor}/p^{max} is also indicated by blue bars for comparison. It is clear that p^{cor} was higher than other scores on average.

Initial pose estimation accuracy is shown by red bars in Fig. 7. The accuracy is evaluated by PCP (percentage of correctly estimated body parts) [11]. Note that gymnastics and parkour classes used the same action-specific model that was generated from their training images. This is because 1) the training images of these classes are fewer than those of other classes and 2) human poses in the two classes are similar. For comparison, the results of the base model (i.e. mixture of parts) [38], clustered pictorial structure models [17], and several state-of-the-arts [6, 29] are shown. Since the proposed method employs additional clues (i.e. action class), it is natural that the accuracy of the proposed method is better than others.

Next, action classification using high-dimensional Object Bank features with pose features was performed. For this classification, two kinds of methods that respectively use (4) and (5) were tested. The results of action classification after two iterations between action classification and pose estimation are shown in Fig. 8. For comparison, the initial results obtained only with the Object Bank features are also indicated by blue bars in the figure, while those obtained by (4) and (5) are indicated by red and green bars, respectively. The classification rates of gymnastics, parkour, and volleyball were improved from the initial results,

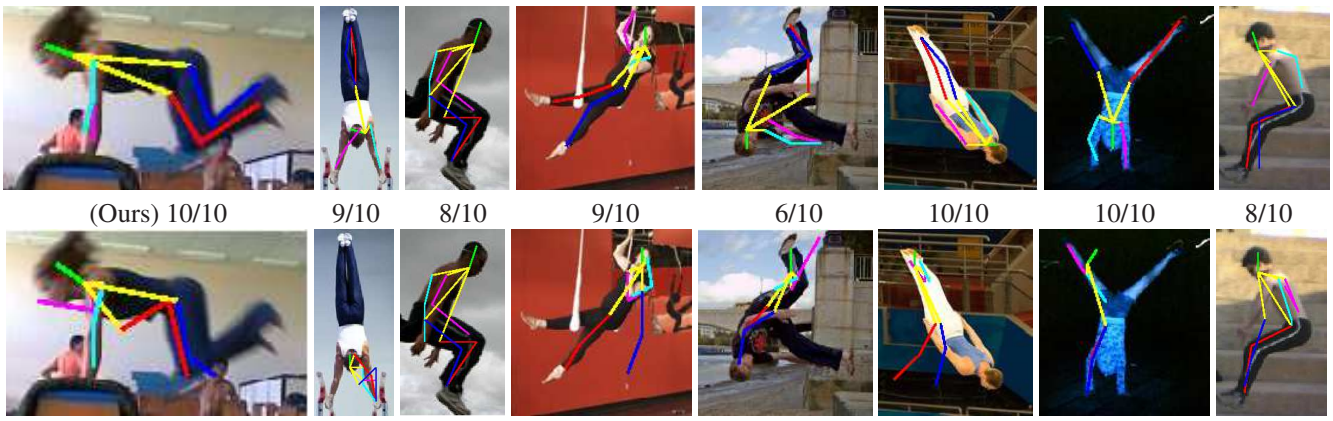

Figure 10. Pose estimation results. For each test image, two results are shown: (Top) the proposed method and (Bottom) a mixture model of parts [38], which is the base model of the proposed method. The number of correctly localized parts is shown under each result. All of the images were selected from gymnastics and parkour classes, where human poses are significantly different from natural upright poses.

while those of other classes were close enough between the initial and final results. To validate this result, the distribution of the pose features is shown in Fig. 9. Indeed it can be confirmed that many features of gymnastics, parkour, and volleyball are apart from those of others. On the other hand, most of other features are crowded. This means that the pose feature (3) should be improved so that those of different classes are distinguishable from each other.

Pose estimation accuracy after two iterations is shown by blue bars in Fig. 7. It can be seen that the accuracy was a bit improved compared with the results of the initial estimation (indicated by red bars). Figure 10 shows several examples of the estimated poses. The results of the base method [38] are also shown in the figure. As shown in these examples, the proposed method was successful in particular in gymnastics and parkour in contrast to the base method [38]. 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 models.

On the other hand, Fig. 11 shows two examples of unsuccessful results obtained by the proposed method. In the lefthand example (i.e. parkour image), both of the proposed method and the base method [38] failed completely. In the righthand example (i.e. soccer image), the proposed method was inferior to the based method [38]. An inappropriate action-specific model, which was selected due to miss action classification, caused such an inferior result.

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

- The positive effects of action-specific deformable part models are proved as shown in Table 2 and Fig. 10.

- Pose features improve action classification as shown in Fig. 8, while their impact is relatively small in contrast to the action-specific models.

Figure 11. Unsuccessful results of the proposed method. For comparison, the results of the base model [38] are also shown.

6. Concluding Remarks

This paper proposed an iterative method for human action classification and pose estimation in still images. Action classification is achieved by global appearance features with pose features, and pose estimation is enhanced by action-specific deformable part models.

Future work includes developing 1) joint optimization of multiple deformable part models that share the basic structure of a human body and 2) 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 class. For this optimization, hierarchical modeling such as [32] might be useful. The latter enables more correct action classification. Dividing each action class into sub-classes (e.g. "baseball" to "pitching" and "batting") might be also effective for more detailed deformable part modeling. On the other hand, if a more general framework, which is not limited to predefined action classes, is required, pose models should be produced for unsupervised clusters of image features such as the Object Bank.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 1

[2] L. D. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 1

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):27, 2011. 3, 4, 5

[4] J. Chen, M. Kim, Y. Wang, and Q. Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *CVPR*, 2009. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3, 5

[6] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013. 5, 6

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

[8] 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*, 9:1871–1874, 2008. 3, 4, 5

[9] 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.*, 32(9):1627–1645, 2010. 1, 2, 5

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 1, 3

[11] V. Ferrari, M. J. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 6

[12] J. Gall, A. Yao, and L. J. V. Gool. 2d action recognition serves 3d human pose estimation. In *ECCV*, 2010. 2

[13] 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.*, 31(10):1775–1789, 2009. 2

[14] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the web. In *ICCV*, 2009. 1

[15] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV (1)*, 2010. 1

[16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 4, 5

[17] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 2, 4, 5, 6

[18] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003. 1

[19] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010. 2

[20] L.-J. Li, R. Socher, and F.-F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 2

[21] 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*, 2010. 2, 3, 4, 5

[22] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *CVPR*, 2008. 1

[23] S. Maji, L. D. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 1

[24] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[25] P. Natarajan, V. K. Singh, and R. Nevatia. Learning 3d action models from a few 2d videos for view invariant action recognition. In *CVPR*, 2010. 2

[26] J. C. Niebles, H. Wang, and F.-F. Li. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. 1

[27] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 2

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

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 5, 6

[30] 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*, 2012. 2

[31] V. K. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *ICCV*, 2011. 2

[32] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, 2004. 7

[33] C. Thurau and V. Hlavác. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008. 1

[34] C. Wang, D. M. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009. 2

[35] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1762–1774, 2009. 1

[36] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004. 3

[37] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 1, 2

[38] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1, 2, 3, 4, 5, 6, 7

[39] A. Yao, J. Gall, G. Fanelli, and L. V. Gool. Does human action recognition benefit from pose estimation? In *BMVC*, 2011. 2

[40] B. Yao and F.-F. Li. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2