Single Image Action Recognition using Semantic Body Part Actions

Zhichen Zhao[1], Huimin Ma[1]*, Shaodi You[2,3]
[1]Tsinghua University, [2]DATA61-CSIRO, [3]Australia National University
{zhaozc14@mails., mhmpub@}tsinghua.edu.cn, shaodi.you@data61.csiro.au

Abstract

In this paper, we propose a novel single image action recognition algorithm based on the idea of semantic part actions. Unlike existing part-based methods, we argue that there exists a mid-level semantic, the semantic part action; and human action is a combination of semantic part actions and context cues. In detail, we divide human body into seven parts: head, torso, arms, hands and lower body. For each of them, we define a few semantic part actions (e.g. head: laughing). Finally, we exploit these part actions to infer the entire body action (e.g. applauding). To make the proposed idea practical, we propose a deep network-based framework which consists of two subnetworks, one for part localization and the other for action prediction. The action prediction network jointly learns part-level and body-level action semantics and combines them for the final decision. Extensive experiments demonstrate our proposal on semantic part actions as elements for entire body action. Our method reaches mAP of 93.9% and 91.2% on PASCAL VOC 2012 and Stanford-40, which outperforms the state-of-the-art by 2.3% and 8.6%.

1. Introduction

Single image action recognition is a core computer vision task which aims to identify the human action in still images where location prior is provided. It enables better performance of image captioning [27], image and video analysis [23], human-computer interactions [3] and *etc*.

Early single image action recognition methods exploit cues such as interactive objects [11], part appearance [10, 14], template matching [5, 28] and spatial relationships [32]. Among them, part-based methods [10, 14, 32] are most successful, which extract appearance features from body parts. Recently, benefiting from deep neutral networks [22, 12], part-based methods have obtained promising results.

However, there exists a semantic gap between part appearance and body actions. Most existing methods use deep neural network as a black box and bridges such gap. Unfortunately, part appearance might be weakly associated with body actions. We show an example in Fig.1, the hand holding a bottle makes the action be mislabeled as "drinking" by previous part-based method [32], his head appearance of "wearing glasses" can hardly correct this action to be "pouring liquid".

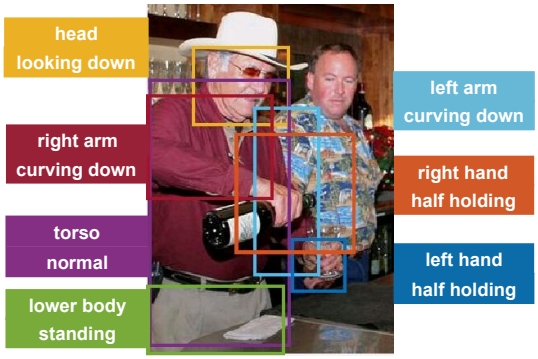

Figure 1. Inferring body action by semantic part actions. Previously part-based method [32] mis-classifies the action as "drinking" only because of the hands holding bottles. Our method, however, makes the correct prediction "pouring liquid" by noticing semantic part actions that his head is lowered and two arms are curving down.

We argue that there exists a mid-level semantic which essentially connects part appearance and body action. We name it the semantic part action. Referring to the example in Fig.1, by noticing the semantic part actions that the man's head is looking down and his arms are curving down, one might infer that he is actually "pouring liquid" rather than "drinking".

In this paper, we focus on exploiting semantic part actions to improve body action recognition. To this end, we propose a novel single image action recognition framework. As illustrated in Fig.2, first, we locate body parts (head, torso, arms, hands and lower body) using a key-point pre-

*corresponding author

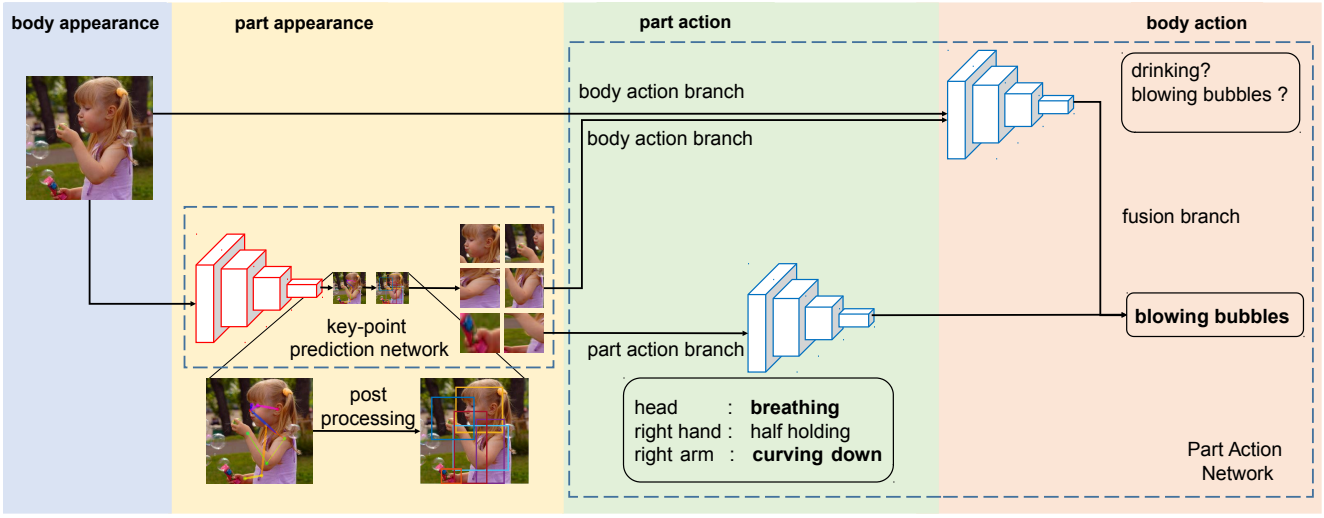

Figure 2. The proposed framework for part action prediction and body action prediction.

diction network. Second, and most importantly, body and part images are fed into a Part Action Network (PAN) to predict body actions. The proposed Part Action Network is composed of multiple branches: two body action branches that respectively receive body and part images as input and perform as common classification networks to predict body actions, a part action branch that predicts part actions, and a fusion branch that learns to combine part actions and body actions[1]. For the part action branch, we define a set of semantic part actions, *e.g.*,"head: looking up", "hand: supporting" , and collect annotations.

We evaluate our method on two popular but challenging dataset: (1) PASCAL VOC 2012 [7] and (2) Stanford-40 [29]. Our method reports improvements from the state-of-the-art [11, 32, 31] by 2.3% and 8.6% (mean average precision, mAP).

The contributions of this paper are three-fold: first, we propose the concept that human action can be inferred by local part actions, which is a mid-level semantic concept. Second, we propose the methodology which combines body actions and part actions for action recognition. And finally, the proposed method provides significant performance improvement from the state-of-the-art methods.

2. Related work

Single image action recognition. There are mainly three existing strategies for single image action recognition: context-based approaches, part-based approaches and template-based approaches. For context-based approaches, cues of interactive objects are critical. Gkioxari *et al*. [11] employ object proposals [24] to find proper interactive objects. Zhang *et al*. [31] propose a method that segments out the precise regions of underlying human-object interactions with minimum annotation efforts. Template-based approaches focus on action structures. Desai and Ramanan [5] learn a tree structure for each action, treating poses and interactive objects as leaf nodes and modeling their relations. Yao and Li [28] combine view-independent pose information and appearance information, and propose a 2.5D representation.

Part-based methods. The human body parts provide rich information for action. For action recognition and fine-grained recognition, part-based methods have shown promising results [18, 30, 26, 10]. A typical approach to combine global appearance and part appearance is concatenating their features and then use a custom classifier to predict [14]. In [10], parts are supervised by body actions, and specific networks are trained to distinguish them. In [6] the relationship of visual attribute and recognition has been studied, the concept of attribute can be also considered as one kind of variety of appearance. Zhao *et al*. [32] detect semantic parts within the bounding box, and arrange their features in spatial order to extend inter-class variance. These previous part-based methods can be represented by a model as shown in Fig.3(b).

Additional annotations. In [17] the authors use "attribute" to help recognize actions. Their attributes are mainly proposed to describe the whole body and motion scenarios, e.g. "torso translation with arm motion". However, our part actions are "atoms" or bases which describe actions of fine parts. In this way, body actions are decoupled with part actions, and it is more possible to describe numerous body actions by a finite part action set.

Pose estimation and key-point localization. To distinguish part actions, it is important to localize fine parts. In this paper, we employ methods on human pose estimation [1, 33, 21, 19]. Given key-point locations, it is conve-

[1]In our implementation, these branches share convolutional layers, see Sec.4.2.

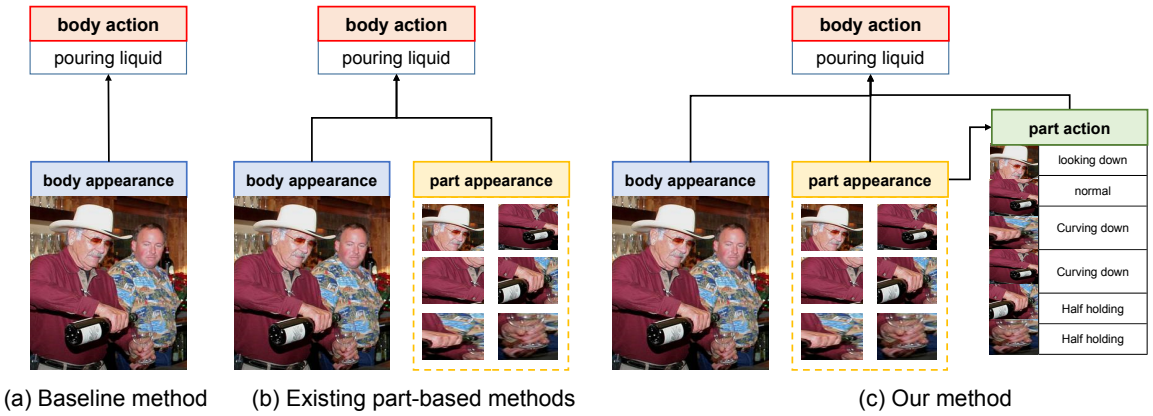

Figure 3. An illustration of baseline, existing part-based methods and our method.

Figure 4. Entire body actions as combination of semantic part actions.

Figure 5. Examples of semantic part actions. Images are from Stanford-40 dataset [29].

nient and accurate to generate part bounding boxes.

3. Semantic Part Actions

3.1. Semantic part action as mid-level semantics

Body action recognition aims to infer the high-level semantics from low-level body appearance, as illustrated in Fig.3(a). With the recently development of deep neural networks, one might get a reasonable performance by directly linking body appearance and action as a black box.

Most existing part-based methods, however, consider a break-down of the human body, and learn connections between part appearance and body actions (Fig.3(b)). "Poselet" [18] is a typical method that learns body parts by clustering algorithm, which is mainly based on part appearance. Similarly, [10] can be seen as one of "existing part-based methods" which mainly rely on part appearance. However, without supervision, part appearance is not always strong associated with the final body action. For example, a man's head appearance of "wearing glasses" can hardly reflect the man's action of "writing on a book", while the man's head action of "looking down" is more relevant.

We argue that the entire human body action is not only a direct combination of body and part appearance, but there exists a mid-level semantic, local part actions. As shown in Fig.3(c), part actions are transformed from part appearance, and used as mid-level semantics to help to infer body actions. Semantic part actions provide strong cues for body actions. For example, if part actions are "head: looking down", "torso: bending", "arms: curving down", "hands: fully holding" and "lower body: crouching", even without seeing the image, we can guess the entire action is "fixing something". In Fig.4 we show more examples.

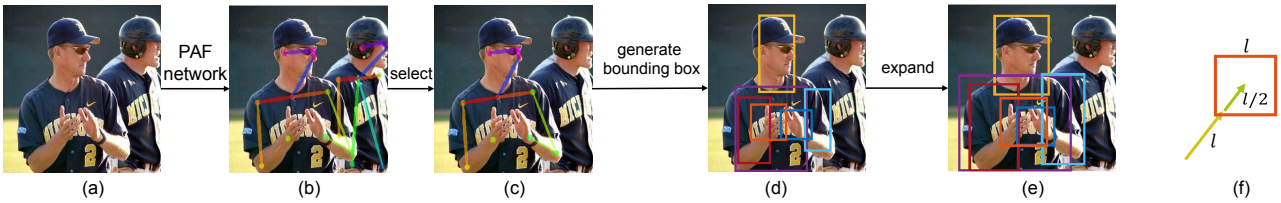

Figure 6. The pipeline of generating part bounding boxes.

Table 1. List of part actions.

head	breathing drinking, eating laughing looking up looking down looking through normal speaking brushing teeth	hand	cutting merging printing proping slack supporting washing waving writing half holding fully holding
torso	bending lying fading away normal	lower body	standing walking crouching forking running sitting
arm	curving (up) curving (down) straight (up) straight (down)		

3.2. Part action definition

As far as we know, there is no existing works on defining and identifying semantic part actions. And thus, we try to define a set of frequently appeared semantic part actions.

First of all, we define seven body parts: head, torso, lower body, arms and hands. Each of them has some semantic actions, as illustrated in Fig.5. For example, the head can be "laughing", "looking through", "looking up" *etc*.

As for part action definition, we aim at balancing diversity and compactness. We try to use as less part actions as possible (compactness) to compose as many body actions as possible (diversity).

For compactness, we try to make the part action set finite and minimize effect of objects. For example, for "hand: holding" we only propose "half holding" and "fully holding" to reflect the sizes of interactive objects roughly [1]. For diversity, if semantic of a part is truly different, we add a category to describe it, e.g. "hand: writing". A full list of possible actions is provided in Tab.1. For each part we enumerate common and meaningful part actions, based on which many body actions can be described (as shown in Fig.4).

[1] "Half holding" presents a hand that holds big object and half-clenched, such as bottles, buckets, tennis balls *etc*. While "fully holding" presents a hand interacting with narrow objects like sticks and ropes.

Since there are no part action annotations off-the-shelf, we collect annotations from the training set of Stanford-40 [29] which are manually labeled by volunteers. Despite that our part action set is constructed from a single dataset, we find it generalizes well in other datasets (see Sec.5.2), which also confirm our assumption on decomposing body action into part actions. We will release our annotations, models and codes.

4. Action Recognition

In this section, we introduce our body action prediction framework and the proposed Part Action Network (PAN).

As illustrated in Fig.2, first, a key-point prediction network is used to localize human joints, then bounding boxes of our defined parts can be generated by simple post-processing. Second, a Part Action Network is used to identify part actions and body actions.

4.1. Body part localization

We employ a key-point prediction network to efficiently localize multiple body parts. The reasons why we choose such a network are two-fold: (1) The key-points have essentially shown the locations of parts, with which part bounding boxes can be generated by post-processing. (2) There are abundant annotations and datasets [2, 16] for the key-point prediction task, which is also known as pose estimation.

Even though person bounding boxes are provided for action recognition, sometimes there are multiple people within one bounding box. Among various pose estimation methods [19, 33] we choose the Part Affinity Fields Network (PAF, [33]), which can handle multi-person tasks, to predict key-points. We find that a PAF network pre-trained on MS-COCO [16] performs surprisingly well on other datasets like Stanford-40 and PASCAL VOC 2012.

The pipeline of generating part bounding boxes is shown in Fig.6. The PAF network receives an person bounding box image (a) as input and produces all possible landmarks locations (b). By a greedy algorithm provided by [33], landmarks are grouped into multiple people. We choose the largest one (c) and generate part bounding boxes by post-processing (d). In details, most part bounding boxes are computed as the minimum bounding boxes enclosing the related key-points. For example, we can generate a bound-

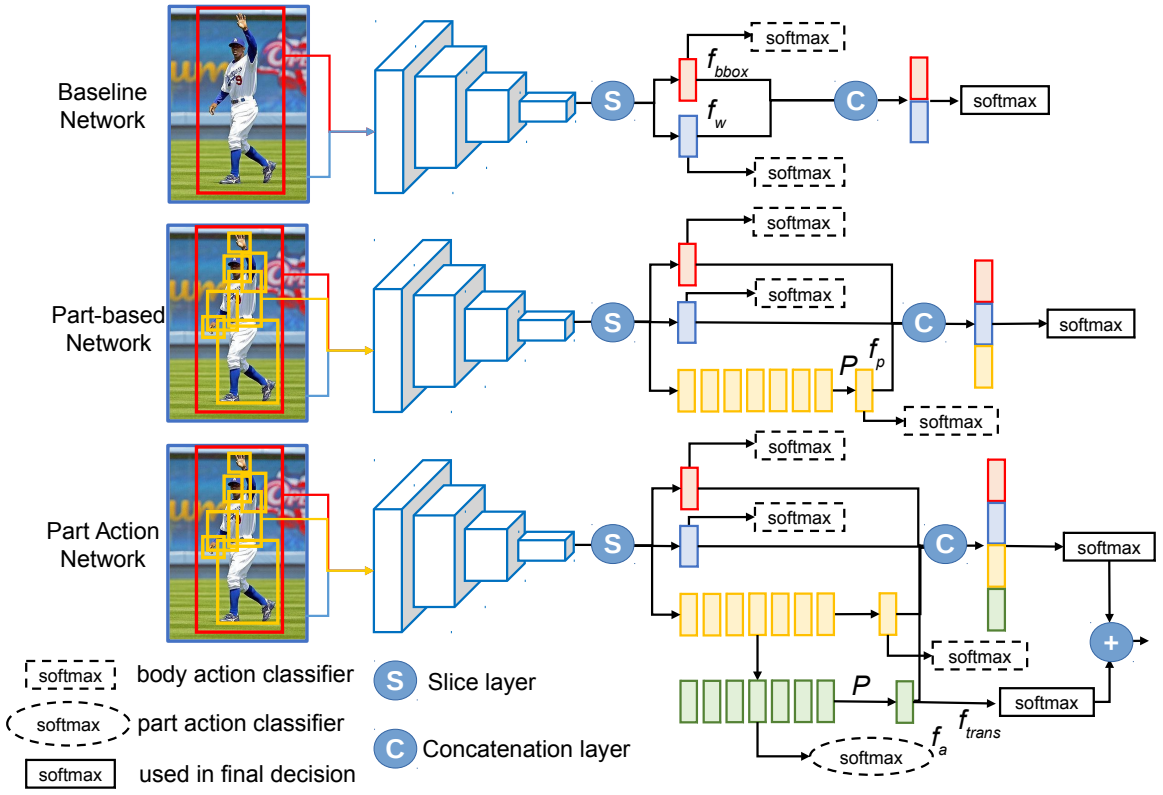

Figure 7. Networks architectures of Baseline Network, Part-based Network and Part Action Network, all of which are modified from a front-end network. Each colored block denotes a feature vector, Red, blue, yellow and green blocks denote features of I_{bbox}, I_w, part appearance and part actions. "P" denotes pooling. Except for black arrows connected with concatenation layer, slice layer and sum operation, others denote that features are transformed by fully connected layers.

ing box for torso using key-points of shoulders and hips. Since there are no key-points describing the top of head and hands, their bounding boxes are treated differently: for the head, we place a bounding box with center at the same vertical coordinate of the nose landmark and width constrained by the ears/eyes landmarks. For hand we extend the line from elbow to wrist by half, setting the endpoint as the bounding box center, and set its width and height to be the length of the forearm (f). All part bounding boxes are expanded by 50% to cover some context as (e).

In case of part localization fails we define some rules: if no landmarks can be localized of a certain part, we use a blank image as placeholder in the network (see Sec.4.2). If some of landmarks can be localized, we infer its location by the articulated part. In our experiments, the keypoint localization is highly accurate, so the generation of parts is also highly consistent.

4.2. Part and body actions prediction

In this section, we describe our Part Action Network (PAN) which receives both images and localized parts as input, and jointly learns body actions, part actions and fusion features for action prediction. As a comparison, we also propose two networks for baseline and existing part-based methods: 1) Baseline Network, 2) Part-based Network. We demonstrate their structures and discuss the differences.

Baseline Network. For action recognition, person bounding boxes are provided. It is common to use two images: image within the bounding box (denoted by I_{bbox}) and the whole image (denoted by I_w), as shown in Fig.7. In our implementation, we use the 50-layer ResNet [12] as a front-end convolution network. Both I_{bbox} and I_w are resized to 224×224 and fed into the front-end network, proceeding $32\times$ downsampling. Their features f_{bbox} (red block in Fig.7) and f_w (blue block) are separated via a slice layer applied on the pool5 feature map, and are concatenated as the final features for action classification. In the training phase we train three classifiers (black boxes, all are supervised by "waving hands"), while in test phase only the last classifier is used to output probability scores. In our framework I_{bbox} and I_w are treated as individual samples, an alternative way of combining them is using a ROI pooling layer from fast-RCNN[8]. However, it may be difficult for ROI pooling layer to extract features of tiny parts (e.g. hands).

The Baseline Network is a representative model of baseline method for action recognition. This network learns the

mapping relationship from body appearance to body action as demonstrate in Fig.3.

Part-based Network. Based on the Baseline Network, we add a branch to capture part appearance features (yellow blocks). Besides I_{bbox} and I_w, all parts that are localized as demonstrate in the previous section, are resized to 224×224 and fed into the network. Features of multiple parts (seven yellow blocks) are transformed into a single feature (the single yellow block) by a fully connected layer. Like conventional part-based methods [14, 10, 32], the transformed single feature (denoted by f_p) are supervised by body action categories. We concatenate f_{bbox}, f_w and f_p, and use the connected classifier to output the final scores.

The part-based network is a representative model of existing part-based methods, where yellow branch learns mapping relationship from part appearance to body actions.

Part Action Network. Our Part Action Network, with an additional branch to learn and predict part actions (green blocks), combines global body actions and local part actions. The part action branch firstly transforms part appearance features (seven yellow blocks) to part action features (seven green blocks), and then uses a fully connected layer fc_{trans} to transform part action features to body action features (the single green block, denoted by f_a). Since features before and after fc_{trans} are supervised by part actions and body actions respectively, fc_{trans} learns the relationship between part actions and body actions. The fusion branch concatenate 4 kinds of features, and makes final decisions. Especially, in the test phase, body action prediction in part action branch are also considered for the final decision. Scores of part action branch and fusion branch (two solid boxes in Fig.7) are averaged to form the final score.

To avoid conflict between body action labels and part action labels, we add a bias on part action labels. For example, in Stanford-40 dataset, there are 40 body actions, so $C_{bias} = 40$. The first part action ("head: breathing") is assigned to be the 41th category. For invisible parts, blank images are used, and we add an individual category for them. So the part action classifier (dashed circle) outputs $40 + 34 + 1 = 75$ probability scores (there are 34 defined part actions in all). Moreover, if annotation of a visible part is ambiguous, the part action label is set to be the same with body action label.

Among the mentioned networks above, Baseline Network is end-to-end trainable, and the others can be trained jointly, given part bounding boxes. Joint training has been verified to be powerful for object detection [20, 8], and help improve the performance in this paper.

In this paper we only collect annotations on the Stanford-40 dataset [29]. The procedure of using the set on another dataset is: first obtain a pre-trained part action network on Stanford-40. Then fix the weights of part action prediction branch, fine-tune other branches and finally obtain another model. Following these steps the part action network can be generally used in other datasets.

Table 2. Preformance (mAP) on the Stanford-40 dataset

method	mAP
Action-Specic Detectors [15]	75.4
VGG-16&19 [22]	77.8
TDP [32]	80.6
ResNet-50 [12]	81.2
Action Mask [31]	82.6
Ours (Baseline Network)	84.2
Ours (Part-based Network)	89.3
Ours (Part Action Network)	**91.2**

5. Experiments

We conduct intensive experiments to validate the proposed Part Action Network. The results show that our method reaches superior results compared with the state-of-the-art methods. Especially, on PASCAL VOC 2012 dataset, our performance is 2.3% better than the state-of-the-art and on Standford-40 is 8.6% better.

5.1. Experimental setup

Network. In this paper we train 3 classification networks: the Baseline Network, the Part-based Network and the Part Action Network. Each of them is modified from the 50-layer ResNet [12] pre-trained on ImageNet [4]. For training them, the learning rate is set to be 10^{-5}. We train for 5K iterations with a batch size of 20. Three kinds of data augmentation techniques are employed: flipping, random cropping and scale jittering [25, 22]. We use the caffe [13] framework to implement our networks. All the networks are trained on a single Titan X GPU.

Dataset. As common practice in action recognition, we use two challenging datasets: 1) PASCAL VOC 2012 [7] and 2) Stanford-40 [29]. The PASCAL VOC dataset contains 10 different actions. For each of the action type, 400-500 images are used for training and validation, and the rest are used for test. The Stanford-40 dataset contains 40 categories and uses 100 images for training. In Fig.8 we show some examples from the Stanford-40 dataset.

5.2. Comparison with existing methods

We compare our approach with the state-of-the-art methods on the two datasets.

Stanford-40 dataset. Tab.2 shows the comparison on Stanford-40 dataset [29]. The method of Action-specific Detector [15] employs transfer learning to learn action-specific detectors, which are used to detect human regions and replace ground truth bounding boxes. VGG-16&19 [22] combines a 16-layer CNN and a 19-layer CNN, and train SVMs on fc7 features. Zhao *et al*. [32] learn some semantic detectors, and arrange semantic parts in top-down

Table 3. Preformance (mAP) on the PASCAL VOC 2012 Action validation set

method	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	mAP
RCNN [9]	88.7	72.6	92.6	74.0	96.1	96.9	86.1	83.3	87.0	71.5	84.9
Action Mask [31]	85.5	72.1	93.9	69.9	92.2	97.2	85.3	73.3	92.3	60.7	82.2
R*CNN [11]	88.9	79.9	**95.1**	82.2	96.1	97.8	87.9	85.3	**94.0**	71.5	87.9
Whole&Parts [10]	84.5	61.2	88.4	66.7	**96.1**	**98.3**	85.7	74.7	79.5	69.1	80.4
Ours (Baseline Network)	87.8	75.4	91.7	81.6	93.3	96.7	87.0	77.4	92.1	67.8	85.1
Ours (Part-based Network)	88.2	86.1	92.9	87.4	94.5	97.8	90.4	86.5	92.4	72.2	88.8
Ours (Part Action Network)	**89.6**	**86.9**	94.4	**88.5**	94.9	97.9	**91.3**	**87.5**	92.4	**76.4**	**90.0**

Table 4. Preformance (mAP) on the PASCAL VOC 2012 Action test set

method	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	mAP
Action Mask [31]	86.7	72.2	94.0	71.3	95.4	97.6	88.5	72.4	88.4	65.3	83.2
R*CNN [11]	91.5	84.4	93.6	83.2	96.9	98.4	93.8	85.9	92.6	81.8	90.2
Whole&Parts [10]	84.7	67.8	91.0	66.6	96.6	97.2	90.2	76.0	83.4	71.6	82.6
TDP [32]	**96.4**	84.7	96.7	83.3	**99.4**	**99.2**	91.9	85.3	93.9	84.7	91.6
Ours (Baseline Network)	92.3	84.4	94.7	82.8	97.9	98.4	90.6	83.7	91.3	80.9	89.7
Ours (Part-based Network)	93.4	90.5	95.6	84.0	98.4	98.6	93.4	90.0	94.3	83.5	92.2
Ours (Part Action Network)	95.0	**92.4**	**97.0**	**88.3**	98.9	99.0	**94.5**	**91.3**	**95.1**	**87.0**	**93.9**

spatial order, which enlarges inter-class variance and obtain 80.6% mAP. Zhang *et al*. [31] propose a method that accurately delineates the foreground regions of underlying human-object interactions and reaches 82.6%.

As for our proposed networks, Compared with a "feature + SVM" framework [22, 12], the end-to-end trainable Baseline Network improves the performance significantly (+3%). The Part-based Network reaches 89.3%, which mainly benefits from accurate part locations. It captures part appearance features and sometimes interactive object cues. Our Part Action Network achieves a mAP of 91.2%, and outperforms the second best published method by 8.6%. Compared with existing part-based methods (which are presented by Part-based Network), it obtains a gain of 1.9%. Among all the 40 categories, the main improvement comes from categories that have similar part appearance and objects, but can be distinguished by part actions. For example, our method improves the performance on "drinking" and "pouring liquid" by 5.5% and 3.9% via noticing detailed differences of arms and head actions. It also obtains gains on other confusing categories, such as "phoning" (+2.0%), "texting message" (+5.2%), "applauding" (+4.3%) and "waving hands" (+5.5%). In Fig.8 we visualize more examples.

PASCAL VOC 2012 dataset. To measure the generalization of our part action set, we also test our networks on PASCAL VOC 2012 Action dataset [7] with no additional annotations. Tab.3 reports the results on PASCAL VOC 2012 Action validation dataset [7], the results on test set are shown in Tab.4. Gkioxari *et al*. [10] use deep poselets to detect head, torso and legs regions and concatenate the corresponding features.

In this dataset our method outperforms the others by 2.1% and 2.3% in validation and test sets respectively. In the test set, Part Action Network reaches the best results for 7 out of 10 categories. Compared with Part-based Network, Part Action Network improves the performance significantly on "phoning" (+1.9%), "reading" (+4.3%). In these categories, curving up arms, looking down heads, sitting lower bodies and supporting hands are critical. Note the Part Action Network implicitly used training data of Stanford-40 dataset, Baseline Network and Part-based Network are trained under the same supervision for fair comparison.

Table 5. results of predicting body actions by part actions.

datasets	PASCAL (validation set)	PASCAL (test set)	Stanford-40
mAP	59.0	52.1	49.2

5.3. How strong are part actions associated with body actions?

We have demonstrated that part actions are strong associated with body actions in Sec.3. In this section, we implement experiments to verify how strong the relationship is. In details, scores produced by part action classifier (dashed circle in Fig.7) with a size of 75×7 are used. For this experiment, predictions before C_{bias} are removed, and scores of seven parts are flattened, resulting in a $35 \times 7 = 245$ vector for each sample. We use a SVM with χ^2 kernel to map these part action predictions to body actions. Tab.5 shows the results on three tasks: 59.0%, 52.1% and 49.2%. The part actions can provide decent results on the two datasets, which confirms our assumption on using part actions to infer body actions.

5.4. Part action classification

The part-level action classification performance is critical for body action prediction. We split the annotated parts of 4000 training samples in the Stanford-40 [29] dataset into two equant subsets. One of them is used as training set and the other is test set. We train a 50-layer ResNet [12]. The top-1 accuracy is 50.6%. As demonstrated in Sec.5.3 and

Figure 8. Predictions on the Stanford-40 test set. The defined parts are marked by colored boxes, and their actions are listed. The mispredicted actions are marked by black boxes. Ground truth is listed below each sample.

Fig.8, some single part action predictions can be inaccurate, however, with the fusion of multiple part actions, body action predictions are reliable (note the above result only employs a half of the training data).

5.5. Visualization and analysis

We visualize part localization results, part action predictions and final body action predictions in Fig.8. In the first three columns we show some examples corrected by our method compared with the Part-based Network. It is shown that some part actions are strong associated with body actions, such as "hand: merging" for "applauding", "lower body: crouching" for "fixing a bike" and "arm: curving up" for "phoning". Some weakly associated part actions do not hurt the final results even wrongly predicted.

In the last column we show some mispredicted samples. They are mainly caused by 2 reasons: 1) errors on part action predictions, which are caused by limited training samples and high similarity of two fine part actions (see the "writing" and "blowing bubbles" samples). 2) lacking of mining contextual information. In the sample of "cutting trees", all parts are predicted perfectly. However, they provide limit help to distinguish this action from "cooking". We believe that by mining contextual cues like [11], our method will perform even better.

6. Conclusion

This paper proposes the idea of semantic body part actions to improve single image action recognition. It is based on the observation that the human action is a combination of meaningful body part actions. We define seven body parts and their semantic part actions. A deep neural network based system is proposed: first, body parts are localize by a key-point network. Second, for each body parts, a Part Action Network is used to predict semantic body part actions. Experiments on two dataset: PASCAL VOC 2012 and Stanford-40 reports mean average precision improvement from state-of-the-art by 2.3% and 8.6% respectively. Experimental analysis and visualization results also show the reasonability and effectiveness.

7. Acknowledgement

The work is supported by National Key Basic Research Program of China (No. 2016YFB0100900) and National Natural Science Foundation of China (No. 61171113).

References

[1] N. Alejandro, Y. Kaiyu, and D. Jia. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[3] P.-Y. P. Chi, Y. Li, and B. Hartmann. Enhancing cross-device interaction scripting with interactive illustrations. In *HFCS*, 2016.

[4] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, and L.-J. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

[6] V. Escorcia, J. C. Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012 (voc2012) results. 2012.

[8] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[10] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *ICCV*, 2015.

[11] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r*cnn. In *ICCV*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arXiv: 1408.5093*, 2014.

[14] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta. Semantic pyramids for gender and action recognition. *TIP*, 23(8):3633–3645, 2014.

[15] F. S. Khan, J. Xu, J. van de Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez. Recognizing actions through action-specific person detection. *TIP*, 24(11):4422–4432, 2015.

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr. Microsoft coco: Common objects in context. In *arXiv:1405.0312*, 2014.

[17] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[18] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[21] W. Shih-En, R. Varun, K. Takeo, and S. Yaser. Convolutional pose machines. In *CVPR*, 2016.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: Generic features for video analysis. In *ICCV*, 2015.

[24] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. M. Smeulders. Selective search for object recognition. In *IJCV*, 2013.

[25] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *arXiv:1608.00859*, 2016.

[26] X. Z. H. Xiong, W. Zhou, and Q. Tian. Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Transactions on Image Processing*, 25:878–892, 2016.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *arXiv:1502.03044*, 2015.

[28] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5d graph matching. In *ECCV*, 2012.

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

[30] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

[31] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu. Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing*, 25:5479–5490, 2016.

[32] Z. Zhao, H. Ma, and X. Chen. Semantic parts based top-down pyramid for action recognition. *Pattern Recognition Letters*, 84:134–141, 2016.

[33] C. Zhe, S. Tomas, W. Shih-En, and S. Yaser. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.