Action Recognition in Still Images with Minimum Annotation Efforts

Yu Zhang, Li Cheng, Jianxin Wu, *Member, IEEE*, Jianfei Cai, *Senior Member, IEEE*,
Minh N. Do, *Fellow, IEEE*, and Jiangbo Lu, *Senior Member, IEEE*

***Abstract*—We focus on the problem of still image-based human action recognition, which essentially involves making prediction by analyzing human poses and their interaction with objects in the scene. Besides image-level action labels (e.g., riding, phoning), during both training and testing stages, existing works usually require additional input of human bounding-boxes to facilitate the characterization of the underlying human-object interactions. We argue that this additional input requirement might severely discourage potential applications and is not very necessary. To this end, a systematic approach was developed in this paper to address this challenging problem of minimum annotation efforts, i.e. to perform recognition in the presence of only image-level action labels in the training stage. Experimental results on three benchmark datasets demonstrate that compared with the state-of-the-art methods that have privileged access to additional human bounding-box annotations, our approach achieves comparable or even superior recognition accuracy using only action annotations in training. Interestingly, as a by-product in many cases, our approach is able to segment out the precise regions of underlying human-object interactions.**

***Index Terms*—Action recognition, still image, without annotation**

I. Introduction

Video-based human action recognition has been a relatively established and well-regarded research problem in computer vision [1], [2], while still image-based human action recognition is a comparably less studied and arguably more challenging problem. Recently it has gained increased attentions in research community with serious efforts in establishing benchmarks and organizing challenges such as the influential PASCAL VOC Action recognition competition [3]. Different from video-based action recognition [4], [5] where temporal image sequences are available and play fundamental roles, in still image-based action recognition [6], the central theme involves predicting the action label based on interpreting human poses and their interaction with objects of the scene.

Besides image-level action labels (e.g., riding, phoning), existing works usually also require manually annotated human bounding-boxes as input [7], [8], [9], [10], [11], [12], [13] during both training and testing stages, which have played a critical role in modelling the typical human poses of different actions, and in characterizing the underlying human-object interactions. As manual annotation of these bounding-boxes is rather time-consuming and painful, this input requirement might severely discourage potential applications. Recently, Gkioxari *et al.* [14] have moved a step forward and shown that it is possible to locate and recognize human actions in testing images without using manually annotated bounding-boxes. Although they do not directly utilize manually annotated bounding-boxes, they need to train human detectors to guide their recognition at the test stage. Similarly, Prest *et al.* [15] have also studied human-object interactions in the weakly supervised environment via pre-trained human detectors. Besides, it remains an on-going research topic on how to robustly and precisely detect humans from images.

In fact, we argue that this input requirement is not very necessary. In this paper, we aim to recognize actions from still images based on minimum annotation efforts (i.e. only image label annotations in the training stage).

Consequently, a systematic pipeline was developed in our approach as follows: (1) Object proposals are first generated using selective search, and are further decomposed into finer-grained object parts. They are used to delineate the detailed shape of human-object interaction regions. (2) Finally, an action label is predicted with the help of an efficient product quantization method to encode features obtained from the human-object interaction regions.

Therefore, the main contributions of this paper are as follows: First, a systematic approach was developed for the problem of action recognition in still images with minimum annotation efforts. It entirely eliminates the un-necessary need for human bounding-boxes as input during both training and testing stages, thus opening doors to many practical applications with least demands on manual annotations. Empirically our approach is shown to perform on a par with or even better than state-of-the-art methods on benchmark datasets, even when these existing methods enjoy privileged access to human bounding-boxes annotations that our approach does not have. Second, very often, our approach is able to accurately delineate the foreground regions of underlying human-object interactions, which is also called "*action mask*" in our paper. Moreover, for the cases when our system fails to delineate a proper action mask, the predicted action labels are mostly also

Jiangbo Lu is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). Jianfei Cai is partially supported by Singapore MoE AcRF Tier-1 Grant RG138/14. This work was partly done when Y. Zhang was with the Advanced Digital Sciences Center.

Y. Zhang and L. Cheng are with the Bioinformatics Institute, A*STAR, Singapore. E-mail: {zhangyu, chengli}@bii.a-star.edu.sg.

J. Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: wujx2001@nju.edu.cn.

J. Cai is with the School of Computer Engineering, Nanyang Technological University, Singapore. E-mail: asjfcai@ntu.edu.sg.

M. N. Do is with the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. E-mail: minhdo@illinois.edu.

J. Lu is with the Advanced Digital Sciences Center, Singapore. E-mail: Jiangbo.Lu@adsc.com.sg. J. Lu is the corresponding author.

wrong, and it is also true for the other way around.

It is worth mentioning that Simonyan *et al.* [16] also attempt to predict the action label on the entire input image with VGG very deep convolutional neural network (CNN) models. Thus, the need for manual annotated bounding-boxes is removed. Their method did not distinguish the foreground and the background in images. As a result, the feature from the whole images cannot delineate different roles of the foreground actors and the environments in actions. As demonstrated in empirical experiments, by making the efforts in delineating the foreground action mask in our pipeline, the prediction performance of our approach is shown to surpass that of [16] by a large margin.

II. Related works

In this section, we review several state-of-the-art works on action recognition, image segmentation, and object detection.

A. Related works on action recognition

Video-based action recognition has been relatively well-established with a long list of literature over the years and interested readers might want to consult recent survey articles such as [1], [2].

For still image-based human action recognition [6], different action representations have been investigated for effective recognition. Existing methods can be roughly grouped into three categories. The first category is the *pose*-based methods. Maji *et al.* [7] use part detectors to detect the parts of human bodies and encode them into poselet for action recognition. Tompson *et al.* [17] learn a CNN for human pose estimation.

The second category is the *context*-based methods. They consider not only the human himself, but also the human-object interactions to aid the action recognition. Delaitre *et al.* [18] generate interacted person-object pairs and select discriminative ones for action recognition. Yao *et al.* [19] consider multi-interactions in an action including human poses, human-object interaction, as well as the relationship among objects. Gkioxari *et al.* [11] employ learned object detectors to detect the most related object to the person in an action. The third category is the *part*-based methods. Sharma *et al.* [12] propose to use image local patches as parts and learn a DPM [20]-like classifier for action recognition.

Prest *et al.* [15] recognized action images using the same setup as ours, i.e., only using image labels in all images. The difference is that Prest *et al.* used multiple human part detectors including the upper-body and face detector to locate the human in the image. According to the detected person, the related objects are then detected based on their relative locations to the person and their shared appearance across images. In contrast, our method deals with humans and objects without using any prior information (e.g., human detectors) and treat them in the same way.

With the prevalence of CNN [21], new results in action image recognition are constantly updated. Gkioxari *et al.* [14] tried to recognize action images without using human bounding-boxes in testing images. They learned an R-CNN [22] detector using the human bounding-boxes in training images, and used it to detect the human location in testing images. In this work, they only considered the human in the action. To represent the human, they proposed to use different parts like the head, torso and legs, etc. They applied CNN on the image, and extracted the pool5 (in AlexNet [21]) outputs to represent the parts.

Gupta *et al.* [10] also advocated investigating the human and related objects in action images. They proposed a new dataset with annotations on both humans and related objects, which has not been released by now.

B. Related works on image segmentation & object detection

There are vast literatures on both topics; but here we only review the most related works. Chai *et al.* [23] propose to co-segment objects in images. Their method makes use of the central area of the image as the initialization for GrabCut [24], which is useful for images with single objects. Meanwhile, the centrality prior is also widely used in image saliency detection techniques such as [25]. Unfortunately, as action images in our context usually involve multiple objects, there is no guarantee that they have to be located at the image center. As a result, this assumption is not applicable for still image-based human action recognition. In contrast, our approach is capable of automatically segmenting out action-related objects and human without relying on this assumption.

The works on multi-foreground co-segmentation [26], [27] aim to detect recurring objects in similar images. Liu *et al.* [28] segmented multi-objects in an image by using a random geometric prior forest. Objects in the training set are first organized in a tree structure. Any detected object in testing is reconstructed from its neighborhood in the tree. These works are quite different as in our context, only action related objects are of interest. Moreover, the objects in action images are more diverse.

Zhang *et al.* [29] achieved weakly supervised image segmentation by learning the distribution of spatially structured superpixel sets from image-level object labels. Liu *et al.* [30] proposed a multi-class video segmentation in a weakly supervised manner. Ren *et al.* [31] used multiple instance learning to solve image segmentation with only image-level labels. The works of weakly supervised image segmentation and object detection methods e.g. [32], [33], [34], [35], [36], [37], [38], [39] are also related but do not directly apply to our problem: During the training stage of weakly supervised image segmentation and object detection, multiple object tags are provided at the image level. For example, for an image with a person riding on a horse, it will be tagged explicitly with “person” and “horse”. However, in our context, an image may just be labeled as a unique label of “riding”. Thus, given an action label, the techniques developed for weakly supervised image segmentation and object detection might fail to segment action related objects.

III. Our Approach

Following the divide-and-conquer principle, the challenging task of still image-based human action recognition is decomposed into two subproblems here, as also illustrated in Fig. 1.

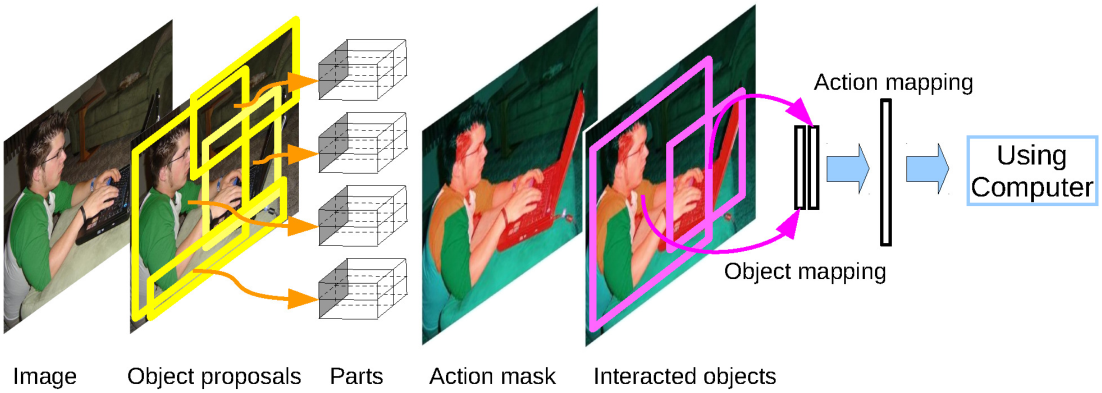


Fig. 1. System overview. Without using human bounding-boxes in either training or testing images, we first extract object proposals and parts from the input image. Then, we learn the action mask (in red) using parts. Finally, we extract feature vectors from the objects in the action (magenta bounding regions) and encode them into an action representation for recognition. Our method recognizes the "using computer" action in this case.

The first subproblem involves on how to delineate the detailed shape of human-object interaction regions (i.e. the action mask). Subsequently the second subproblem concentrates on proper feature representation for the recognition task.

To address the first subproblem, a joint optimization scheme is proposed to obtain a meaningful foreground action mask from the input image. We adopt a part-based representation, as it has shown good discrimination ability in related topics of action recognition [14] and fine-grained image recognition [40]. Inspired by the observation that usually a small groups of distinct part patterns are present given a particular action class, we propose to learn the action-related object parts with a group sparsity constraint, which facilitate the unsupervised formation of action-specific receptive fields from the part candidates. Meanwhile, we also consider the enforcement of spatial coherence among parts using low level visual features in each image, which coincides well with the assumption in action datasets including PASCAL VOC 2012 Action [3] where action-related objects are usually spatially close. This leads to a five-step iterative pipeline for unsupervised discovery of a foreground action mask of the current image.

The second subproblem in our context focuses mainly on a good and dedicated feature representation from the action mask for recognition purpose. As the objects (including human here as human can also be regarded as a special object) discovered during the first subproblem capture rich discriminative information, they are continuously used here. On the other hand, due to the high-dimensional nature of the object related Fisher vectors, we propose to utilize an efficient product quantization (PQ) [41] method, as PQ has been shown to work well for feature compression for a broad range of vision related applications.

Notation summary: For an input image I_i, the location of an object proposal j is denoted by a binary mask $B_{i,j}$. The location of the k-th part in j-th object proposal is denoted as a binary mask $B_{i,j,k}$. The feature vector of a part is $\boldsymbol{z} \in \mathbb{R}^d$. All the parts in the i-th image is denoted by $\boldsymbol{Z}_i$, which contains stacked part feature vectors in image i. $\boldsymbol{z}_{i,j,k}$ refers to the part feature corresponding to $B_{i,j,k}$. We further denote with $\boldsymbol{\alpha}_i$ the binary action foreground mask (1 for foreground and 0 for background) of I_i, and θ^H the global part model. They are represented using different Gaussian mixture models (GMM). $\boldsymbol{\xi}^c$ denotes the class specific part model, which has the same length to our image representation. $\boldsymbol{x}_{i,m}$ represents m-th low level visual feature in I_i , while θ_i^L is used to model the distribution of low level features in I_i.

A detailed account of our two-step approach is presented in what follows.

A. *Subproblem 1: Learning to delineate the action mask from an input image*

Given an input image, we start by generating object proposals using selective search [42]. For any generated object proposal, we will introduce how parts are extracted and represented in Sec. III-A1. Then, we learn the action foreground mask in Sec. III-A2. Finally, we compute the global representation using this mask in Sec. III-B.

1) Part generation and representation: An action involves the interaction of possibly multiple humans and objects, and in our context a human can also be seen as a special object. However, direct usage of the objects in obtaining the action mask may not properly characterize the essential spatial relationship of objects, since not all parts in an object are involved in the action related interactions. For example, in the action of "riding horse", only the lower part of the human object is on the upper part of the horse object. Thus, we would need to work with fine detailed information of object parts to better characterize the spatial interactions of an action.

This observation inspires us to represent each object with parts explicitly. It is worth pointing out that part based object representation has been working very well in related topics of human action recognition [14] and fine-grained object recognition [40]. An efficient method [43] is used in our approach to produce object part: Each object is specified by a bounding-box produced by the unsupervised selective search method [42]. This is followed by a CNN model applying to the (warped) bounding-box, where the outputs of the last convolutional layer of CNN are used to compute parts. Taking the output $X \in \mathbb{R}^{S \times S \times D}$ from a bounding-box, where S is the spatial size and d is the number of filters, the $(S-M+1)^2$ parts are generated with the following multi-max pooling technique:

$$z^M_{r,s,d} = \max_{\substack{r \le r' < r+M,\\ s \le s' < s+M}} X_{r',s',d} \ , \quad (1)$$
$$\text{s.t.} \quad 1 \le M \le S, 1 \le d \le D \ ,$$

where (r, s) are the relative coordinates of the top-left corner of the part bounding-box embedded with respect to its enclosing object bounding-box, and M is the scale index of parts. As M is assigned to a range of values, then the parts of corresponding scales are generated from the object of interest, with each corresponding to a specific size of receptive fields in the object.

At this moment, parts are independently generated from different objects without considering the interactions among objects and in particular the human-object interactions that are critical to characterize an action. This drives us to consider what follows on how to delineate the detailed shape of human-object interactions for each action. This is a rather challenging task as in our context only an action label annotation is provided for one image in training, and worse, there also exist ambiguities in these action labels.

2) Discover the action mask: We solve this problem with the assumption that related objects in an action are spatially close. This assumption holds in many benchmark datasets like VOC 2012 Action. We will learn a shared visual model for the object parts in all action classes. Meanwhile, we considered the spatial coherence for parts within the same action through low level visual features (e.g., RGB values of pixels in the color image), such that the detected objects/parts can be meaningful for an action rather than isolated from each other.

Specifically, our task here can be formulated as an energy minimization problem on a Markov random field:

$$\min_{\boldsymbol{\alpha}, \theta^H, \{\boldsymbol{\theta}^L_i\}, \boldsymbol{\xi}^c} \quad \sum_i \left(\sum_m U(\alpha_{i,m}; \boldsymbol{x}_{i,m}, \boldsymbol{Z}_i, \theta^H, \theta^L_i, \boldsymbol{\xi}^c) \right) \quad (2)$$
$$+ \sum_i \left(\sum_{m,n} V(\alpha_{i,m}, \alpha_{i,n}) \right),$$

where θ^H and θ^L_i each denotes a separate Gaussian mixture model (GMM). V is the smoothness term, which evaluates the spatial labeling coherence of the objects/parts in an action image. It is defined as:

$$V(\alpha_{i,m}, \alpha_{i,n}) \quad (3)$$
$$= \delta(\alpha_{i,m}, \alpha_{i,n}) dist(m, n)^{-1} \text{exp}(-\beta \| \boldsymbol{x}_{i,m} - \boldsymbol{x}_{i,n} \|^2),$$

which is the same as that in GrabCut [24], where $dist(m, n)$ is the Euclidean distance of neighboring pixels. The unary term U is defined as:

$$U(\alpha_{i,m}; \boldsymbol{x}_{i,m}, \boldsymbol{Z}_i, \theta^H, \theta^L_i, \boldsymbol{\xi}^c) \quad (4)$$
$$= -\text{log } p(\alpha_{i,m}; \boldsymbol{x}_{i,m}, \boldsymbol{Z}_i, \theta^H, \theta^L_i, \boldsymbol{\xi}^c)$$
$$= -\text{log}\pi^L_{\alpha_{i,m}, k_{i,m}} + \frac{1}{2} \text{logdet} \Sigma^L_{\alpha_{i,m}, k_{i,m}}$$
$$+ \frac{1}{2} [\boldsymbol{x}_{i,m} - \mu^L_{\alpha_{i,m}, k_{i,m}}]^T (\Sigma^L_{\alpha_{i,m}, k_{i,m}})^{-1} [\boldsymbol{x}_{i,m} - \mu^L_{\alpha_{i,m}, k_{i,m}}],$$

where θ^L_i is a GMM learned on pixel RGB values:

$$\theta^L_i = \{ \pi^L_{\alpha,k}, \mu^L_{\alpha,k}, \Sigma^L_{\alpha,k} \ \mid \ \alpha = 0, 1; k = 1, ..., K^L \} \quad (5)$$

containing K^L Gaussian mixture components. Similarly, θ^H_i is defined as:

$$\theta^H_i = \{ \pi^H_{\alpha,k}, \mu^H_{\alpha,k}, \Sigma^H_{\alpha,k} \ \mid \ \alpha = 0, 1; k = 1, ..., K^H \}, \quad (6)$$

which is a GMM learned on part features. π, μ, and Σ are the weights, means, and covariance in each GMM, respectively. p evaluates the probability of each pixel assigned to the action foreground mask or not. θ^L_i are estimated with $\boldsymbol{Z}_i$, θ^H and $\boldsymbol{\xi}^c$, i.e., the parts in an action are selected from $\boldsymbol{Z}_i$ using θ^H and $\boldsymbol{\xi}^c$, and their receptive fields form the initial estimation for the foreground and background.

The action foreground masks are discovered through iterative optimization of Eq. (2) on part features $\boldsymbol{z}$ and low level visual features (pixel RGB values) as illustrated in Fig. 2. First, we start learning a class part model $\boldsymbol{\xi}^c$ for each action class using part features and image labels (cf. Eq. 7), since part features contain more semantic information than low level visual features. Instead of evaluating each part separately, similar parts are clustered into groups in a part feature space. One important observation is that each action class only has a few effective part groups. This inspires us to estimate $\boldsymbol{\xi}^c$ with the group sparsity constraint. Then an initial action mask $\bar{\boldsymbol{\alpha}}_i$ is obtained for the input image i. Because $\bar{\boldsymbol{\alpha}}_i$ still does not consider the spatial coherence assumption, in the next step we aim to refine $\bar{\boldsymbol{\alpha}}_i$ into $\boldsymbol{\alpha}_i$ based on low-level visual features using GrabCut where the spatial coherence of objects (including the human object) are incorporated. Finally, a global part model θ^H is introduced in either the foreground action mask or background over the entire training images, respectively. The above-mentioned steps are then iterated until the obtained action mask becomes stable.

Specifically, we first generate object proposals from each image using selective search [42]. Each object proposal is specified by a binary mask (1 for object and 0 for background) of the entire image. In the i-th image I_i, we denote its object proposals as $\{B_{i,1}, ..., B_{i,|I_i|}\}$. Parts from each object proposal are obtained according to Eq. (1). This is followed by the following iterative steps to solve Eq. (2):

Step 1 **image representation with** θ^H. For each image, a Fisher vector (FV) [44] is computed using all part features in it according to θ^H (GMM). In the initial iteration, θ^H is set to the GMM learned from part features $\boldsymbol{Z}_i$ over all training images. When computing FV, the dimension of part features is reduced from D to D' using PCA. Only the mean and variance are used in each Gaussian to compute FV following [44]. Thus, the FV length corresponding to one Gaussian is $2D'$, and the length of the whole FV is $2D'K^H$.

Step 2 **learn the part model** $\boldsymbol{\xi}^c$ **of each action class** c. A specific part model is learned for each action class. Note the dimensions in FVs are organized into groups with each group of features computed from a Gaussian component in the GMM. Using the FVs A

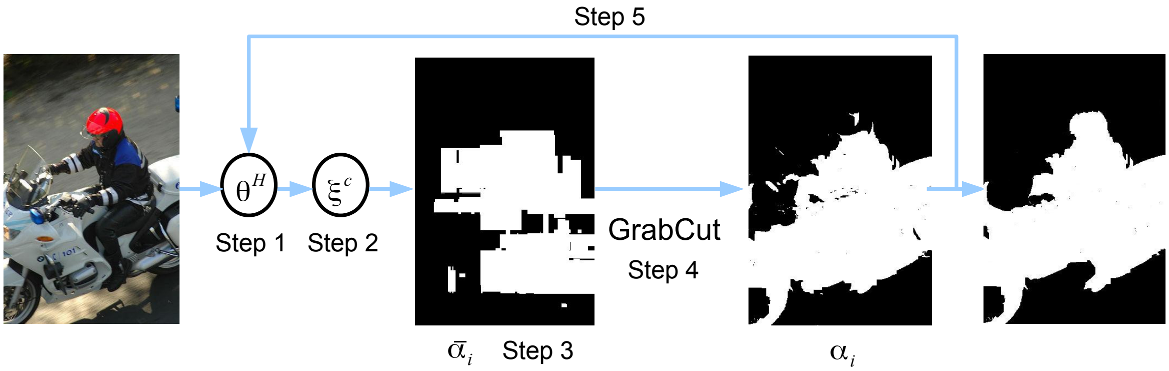


Fig. 2. The iterative process of discovering the action mask from an input image. Refer to Sec. III-A2 on details of each step.

(the FV of one image as a row in A) and the image labels $\boldsymbol{y}$ (images in class c are labeled 1 and the rest are -1) of all training images, we learn $\boldsymbol{\xi}^c \in \mathbb{R}^{2D'K^H}$ for each class c with the group sparsity constraint as:

$$\min_{\boldsymbol{\xi}^c} \quad \frac{1}{2}\|A\boldsymbol{\xi}^c - \boldsymbol{y}\|^2 + \lambda_1\|\boldsymbol{\xi}^c\|_1 + \lambda_2\sum_t \|\boldsymbol{\xi}^c_{G_t}\|_2, \tag{7}$$

where $\boldsymbol{\xi}^c$ is divided into K^H non-overlapping groups $\{\boldsymbol{\xi}^c_{G_t} \in \mathbb{R}^{2D'} | t = 1, ..., K^H\}$. Each G_t denotes a Gaussian in θ^H. Eq. (7) can be solved using SLEP [45]. Then, we choose the Gaussian components which have non-zero weights in $\boldsymbol{\xi}^c$, and name their centers as the *representative parts* R^c of action class c:

$$R^c = \{t \mid \|\boldsymbol{\xi}^c_{G_t}\|_1 > 0\}, \tag{8}$$

where R^c is a set of indexes of GMM components for class c.

Step 3 **obtain the initial action mask $\bar{\boldsymbol{\alpha}}_i$ of image i.** For each image i, we compute $\bar{\boldsymbol{\alpha}}_i$ using the learned representative parts. In each image, we select those parts which are nearest to the representative parts (in all the GMM components). Since each part corresponds to a receptive field in an object proposal, its location in the object proposal can be computed according to the filter and step sizes in the CNN model as [46], [47]. The receptive filed of the selected part $\boldsymbol{z}_{i,j,k}$ is also denoted with a binary mask in the image as $B_{i,j,k}$, which corresponds to the k-th part in the j-th object proposal of the i-th image. Then, $\bar{\boldsymbol{\alpha}}_i$ of the i-th image is computed as:

$$\hat{B}_i = \frac{1}{N_i}\sum_j\sum_k B_{i,j,k}[nn(\boldsymbol{z}_{i,j,k}) \in R^{y_i}], \tag{9}$$

$$\bar{\boldsymbol{\alpha}}_i = \mathbf{1}(\hat{B}_i > \eta), \tag{10}$$

where $\mathbf{1}$ is an indicator function, nn is the nearest neighbor operator, and $>$ is an element-wise larger operation here. η is an threshold, N_i equals to the number of selected parts in the i-th image, and y_i is the image label.

Step 4 **update $\boldsymbol{\alpha}_i$ and θ^L_i with $\bar{\boldsymbol{\alpha}}_i$ and low level features $\boldsymbol{x}_{i,m}$ in image i.** We refine $\bar{\boldsymbol{\alpha}}_i$ into $\boldsymbol{\alpha}_i$ on low-level visual features $\boldsymbol{x}_{i,m}$ using GrabCut in each image. With the initial action mask $\bar{\boldsymbol{\alpha}}_i$, we use low level features $\boldsymbol{x}_{i,m}$ (pixel RGB values) extracted from it to learn a foreground GMM, and use the rest to learn a background GMM, both of which form θ^L_i. Then, we run GrabCut on the image to obtain the refined action mask $\boldsymbol{\alpha}_i$.

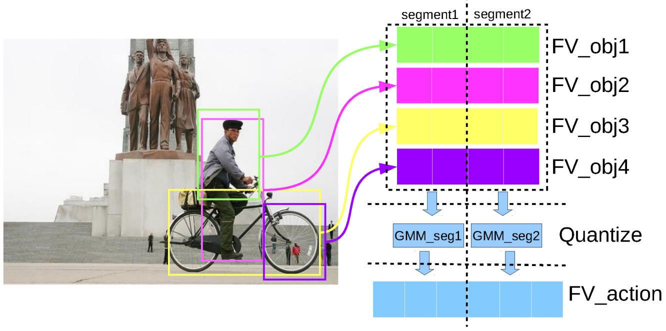


Fig. 3. Action representation. Feature vectors of objects are divided into segments. The feature vectors in all objects of each segment are encoded by a GMM into FV. The FVs of all segments are concatenated into the action representation.

Step 5 **update θ^H for all images**. With the refined action mask $\boldsymbol{\alpha}_i$, we update the global part model θ^H. We use the parts located in the foreground mask in all images to learn the foreground part GMM, and use the rest parts to learn the background part GMM. They collectively form the updated global part model θ^H.

Iterate steps 1 to 5 over all training images until:

$$\frac{1}{N}\sum_i^N \frac{\|\boldsymbol{\alpha}_i^T - \boldsymbol{\alpha}_i^{T-1}\|_2}{\|\boldsymbol{\alpha}_i^T\|_2} < \epsilon, \tag{11}$$

where $\boldsymbol{\alpha}_i^T$ is the action mask after the T-th iteration, N is the number of training images. Empirically we find 5 rounds are sufficient for the iterative process to converge. Finally, we get the shared global part model θ^H, the part model $\boldsymbol{\xi}^c$ for each class, and the foreground action mask $\boldsymbol{\alpha}_i$ for each image. They will continue to participate in addressing the next subproblem to be discussed.

TABLE I
AVERAGE PRECISION (%) ON VOC 2012 ACTION VALIDATION SET.

	Ours w/o mask	Ours	Ours+image	Action Part [14]	Action Part [14]	R*CNN [11]
Train BB	no	no	no	yes	yes	yes
Test BB	no	no	no	no	yes	yes
CNN	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16
Jumping	82.93	82.33	85.51	79.4	84.5	90.1
Phoning	66.26	69.17	72.09	63.3	61.2	80.4
Playing Instrument	89.93	91.10	93.88	86.1	88.4	95.1
Reading	68.13	67.34	69.89	64.4	66.7	81.0
Riding Bike	90.62	91.47	92.20	93.2	96.1	95.8
Riding Horse	94.2	96.04	97.23	91.9	98.3	98.6
Running	81.78	84.35	85.32	80.2	85.7	87.4
Taking Photo	66.37	71.21	73.31	71.2	74.7	87.7
Using Computer	89.95	90.48	92.34	77.4	79.5	93.4
Walking	56.19	59.98	60.67	63.4	69.1	71.4
mAP	78.64	80.35	**82.24**	77.0	80.4	**88.0**

B. Subproblem 2: Action representation

Based on the solution we have obtained for the first subproblem, here we focus on developing a good and dedicated feature representation for the action recognition purpose.

Because each action involves a variable number of objects (including humans), we will fuse the objects into an action representation rather than to directly stacking part features together to form an action vector. Besides, using objects can mitigate the possible issue of imperfect action masks. Later in Table II (cf. ours vs. VGG VD [16] w/o BB), it is empirically demonstrated that this representation is better than the simple holistic representation, where multiple CNN models are directly applied on the entire input image.

First, to encode the object-based parts, we only consider the object proposals $B_{i,j}$ that have a sufficient overlapping percentage with the action mask $\boldsymbol{\alpha}_i$ in each image i,

$$\frac{\|\boldsymbol{\alpha}_i \cap B_{i,j}\|_2}{\|B_{i,j}\|_2} > \gamma. \tag{12}$$

Empirically this filters away a large amount of irrelevant background object proposals. Now, in each of the remaining objects, its parts are encoded into a Fisher vector (FV) with θ^H. This could lead to a very high-dimensional object representation ($>$ 100 K).

Next, we aim to represent the image action by the set of objects in its action mask. As different actions may involve variable numbers of objects, to maintain a same dimensional action representation, one common strategy is to use bag-of-visual-words (BOVW) method. It is however very difficult to apply a vector quantization for very high dimensional vectors in our context. Instead we propose to utilize an efficient product quantization (PQ) method [41] to encode the set of action-related objects, as PQ has demonstrated to be suitable for high-dimensional feature compression to achieve compact storage in large-scale image retrieval and classification topics. Considering the feature vectors of objects from all actions, we first divide them into short segments or subvectors of equal lengths. A specific GMM is learned for each of such segment. Then a FV is computed based on the object features belonging to the segment of a given action using the learned GMM. Next, the FVs over segments are concatenated to represent the particular action of interest. Finally, the action class prediction is made by a learned one-vs-all linear SVM classifier [48]. The PQ process is illustrated in Fig. 3.

C. Action recognition for testing images

So far we have described the training stage. For a test image, the learned $\boldsymbol{\xi}^c$ can still be applied. By executing only steps 3 and 4 of Sec. III-A2, its action mask corresponding to each action class is obtained, and subsequently its action feature vector is acquired. The action prediction is then made by invoking the aforementioned linear SVM classification.

IV. EXPERIMENTS

Throughout empirical evaluation, the VGG16 [16] CNN model is used to extract part features from object proposals. The part features are further reduced to 256 dimensions by applying PCA to retain around 90% of the total energy. The global part model θ^H contains 256 GMMs, where half of them are learned from the foreground action mask and the rest is from background over all training images. θ_i^L contains 20 GMMs, where half of the GMMs are for foreground and background, respectively. Through empirical examination we have observed that more GMMs does not necessarily lead to improved performance but with dramatically increased computation cost. λ_1 and λ_2 are fixed to 0.01 in Eq. (7) to make the portion of selected foreground part groups in each class around $30\% - 50\%$. We set $\eta = 0.5$, $\epsilon = 0.1$, and $\gamma = 0.5$. We first encode parts into a FV in each object. The FV's dimension of each object is $256 \times 2 \times 256 = 131,072$. Then, we encode objects into an action using PQ. We divide object FVs into short segments. Each segment has a length of 512 (mean and variance part of each GMM in FV), and is reduced to 256 by PCA. For each segment, we learn a GMM with 4 components. Empirically we have also observed that more GMMs do not improve the recognition accuracy but at the costs of significantly increased storage and computation. The length of the final action representation is $256 \times 256 \times 2 \times 4 = 524,288$.

The proposed part based action representation is denoted as “ours”. We also use VGG16 to extract CNN features from the whole image in the same way as [16], and denote this representation as “image”. We evaluate the appended features “ours+image”. Besides, we evaluate the proposed part based representation without using the action foreground mask, i.e.,

TABLE II
AVERAGE PRECISION (%) OF VOC 2012 ACTION TEST SET.

	Ours w/o mask	Ours	Ours+image	VD [16]	VD [16]	R*CNN [11]	Action part [14]	RMP [8]
Train BB	no	no	no	no	yes	yes	yes	yes
Test BB	no	no	no	no	yes	yes	yes	yes
CNN	VGG16	VGG16	VGG16	VGG16,19	VGG16,19	VGG16	VGG16	AlexNet
Jumping	82.03	83.51	86.68	-	89.3	91.1	84.7	82.3
Phoning	78.12	70.57	72.22	-	71.3	83.8	67.8	52.9
Playing Instrument	90.55	92.31	93.97	-	94.7	92.2	91.0	84.3
Reading	67.19	68.67	71.30	-	71.3	81.2	66.6	53.6
Riding Bike	93.28	94.78	95.37	-	97.1	96.9	96.6	95.6
Riding Horse	94.56	96.70	97.63	-	98.2	98.4	97.2	96.1
Running	85.22	87.49	88.54	-	90.2	93.1	90.2	89.7
Taking Photo	68.92	70.72	72.42	-	73.3	84.3	76.0	60.4
Using Computer	84.55	86.31	88.81	-	88.5	90.9	83.4	76.0
Walking	62.73	64.58	65.31	-	66.4	77.9	71.6	72.9
mAP	80.72	81.57	**83.23**	79.2	84.0	**89.0**	82.6	76.4

using all object proposals to compute the action representation. It is denoted as “ours w/o mask”. These methods are evaluated on three benchmark datasets:

- **PASCAL VOC 2012 action dataset** [3]. It contains 10 different actions: Jumping, Phoning, Playing instrument, Reading, Riding bike, Riding horse, Running, Taking photo, Using computer, Walking. It has $2,296$ training images and $2,292$ validation images. Our evaluation results are obtained on the testing images through the “boxless action classification” of the publicly available competition server.
- **Stanford 40-action dataset** [49]. It contains 40 actions, and $9,532$ images in total. We use the provided split of training and testing images for evaluation.
- **Willow 7-action dataset** [50]. It has 7 classes of common human actions. It has at least 108 images per class of which 70 images are used for training and validation and the rest are used for testing.

All the experiments are carried out on a desktop computer with an Intel i7-3930K CPU, 64G main memory, and an Nvidia Titan Black GPU.

A. Results on PASCAL VOC 2012 action Dataset [3]

We evaluate the proposed method on PASCAL VOC 2012 validation and test sets in Table I and II, respectively. We first show that the proposed action detection method can help improving the classification accuracy. On the validation set, the proposed method (80.35%) leads to 1.7% better mean average precision (mAP) than that without using the action mask (ours w/o mask, 78.64%). When we append the CNN feature computed from the whole image, the accuracy is further improved to 82.24%. This shows that the proposed part based action representation has a synergy with the global image features. On the testing set, the appended action representation achieves 83.23% mAP.

We show the action mask learning process in Fig. 4. Because there is no object ground truth provided in this dataset, we evaluate qualitatively rather than quantitatively. We observe that the detected action mask often captures the interacted humans and objects accurately. For example, in the 4th image (reading), although the human is not in the center of the image, he can still be detected as part of the action mask. If we instead use GrabCut initialized only at the center, then only the newspaper in the image can be segmented into the action mask.

It is clearly demonstrated in experiments that variants of our approach always outperform existing state-of-the-arts without using human bounding-boxes. On the validation set, to our best knowledge, [14] uses the least amount of human annotations, i.e., only human bounding-boxes in training images (only train BB) are used. [14] employs a part based human representation. In contrast, our approach (i.e. ours) utilizes the part based representation of human-object interactions. Ours (80.35%) has 3.35% better mAP than [14] (77.0%), which validates that objects can provide context information to humans to improve the recognition performance. Ours also has comparable performance to [14] (80.4%) using human bounding-boxes in both training and testing images. On the testing set, [16] showed the results without using human bounding-boxes in all images. They extracted CNN from the whole images using two CNN models, which has a worse result (79.2%) than our part based representation (81.57% and 83.23%). This shows that part based representation is useful in characterizing human actions.

Moreover, variants of our approach achieve comparable results w.r.t. existing methods when they are using human bounding-boxes (both train BB and test BB) [14], [16], [8]. R*CNN [11] is shown to deliver higher mAP than other methods. This can be attributed to the fact that strong human and object detectors are learned with human annotations, which facilitates precise matches for actions.

Exploiting the interaction between humans and objects is beneficial in action recognition/detection in the weakly supervised environment. We take a closer look at the results on “ours” and “action part” [14] (without testing BB) on the validation set. Our approach (i.e. ours) detect all related humans and objects in the action, while [14] only detects humans. As shown in Table I, 8 out of 10 actions have better results in ours than in [14]. These 8 actions are: Jumping, Phoning, Playing Instrument, Reading, Riding Horse, Running, Taking Photo, and Using Computer. The detection results on the testing set are presented in Fig. 5. One can see that the related objects can be well captured in each action on the unknown images. We also observe that the detected action mask can well capture the

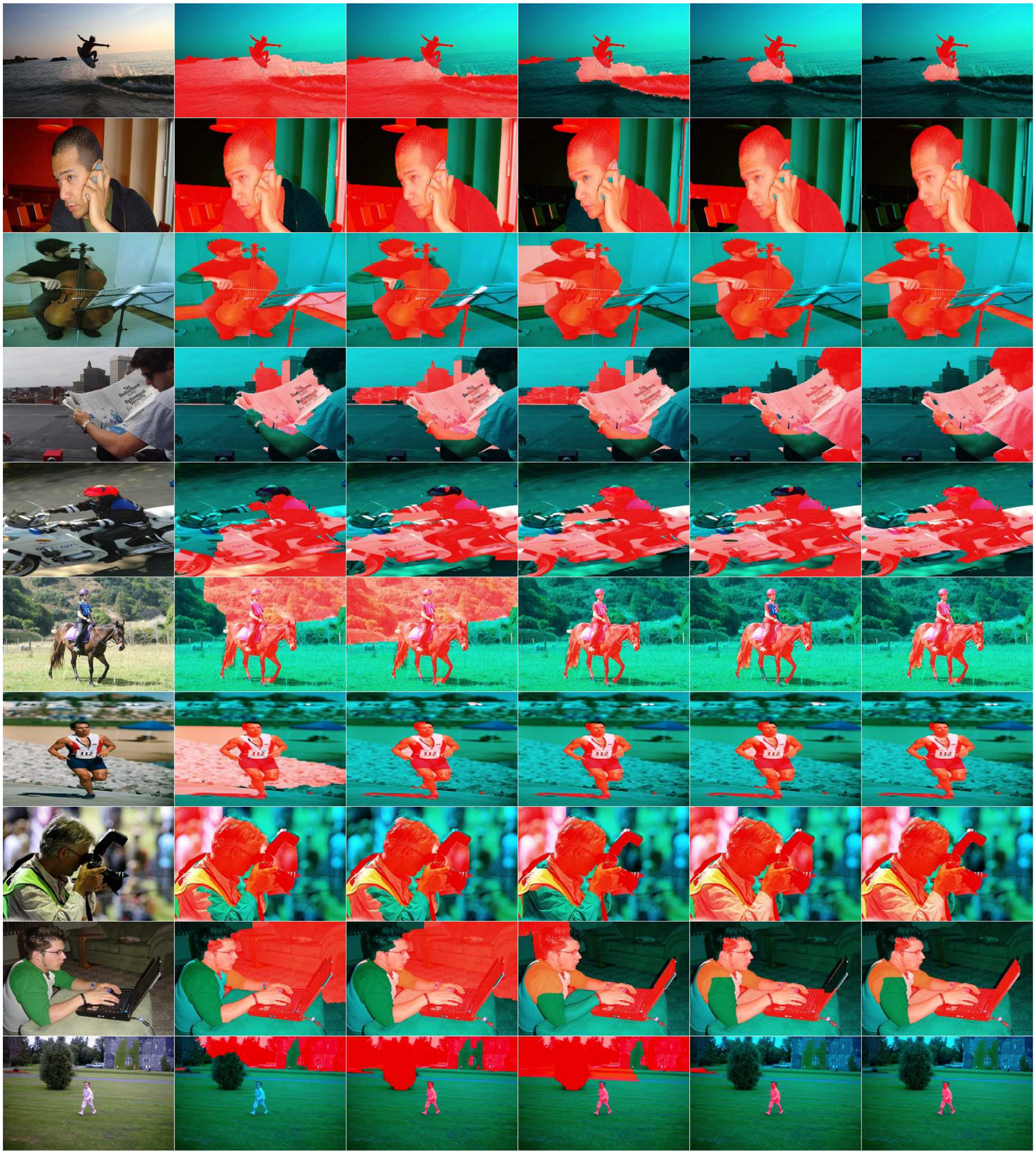


Fig. 4. Learned action masks over iterations on PASCAL VOC 2012 action dataset. The first column shows the input images. The rest five columns shown in red color are the action masks $\boldsymbol{\alpha}$ obtained (from the 1st to the 5th iteration) to delineate human-object interactions.

interacted objects in most images of the dataset. Furthermore, some failure cases of our approach (i.e. ours) is presented on the validation set in Fig. 6. Most of these failures are caused by the ambiguity of the action in the image. For example, the ground truth of the first image is "jumping". However, the human is actually "running to jump".

We also tested the time cost of our method on this dataset. Ours cost about one day training time and takes 0.5s per image testing time. In [11], training takes about 3 hrs and testing takes 0.4s per image. Although ours has longer training time, the testing time is comparable to [11].

B. Results on more action datasets

We evaluated our method on more action recognition datasets including Standford 40-action dataset [49] and Willow 7-action dataset [50].

The results are presented in Table III and IV, where variants of our approach have significantly better results than existing methods even when they have privileged access to manually annotated human bounding-boxes. EPM [12] also use a part

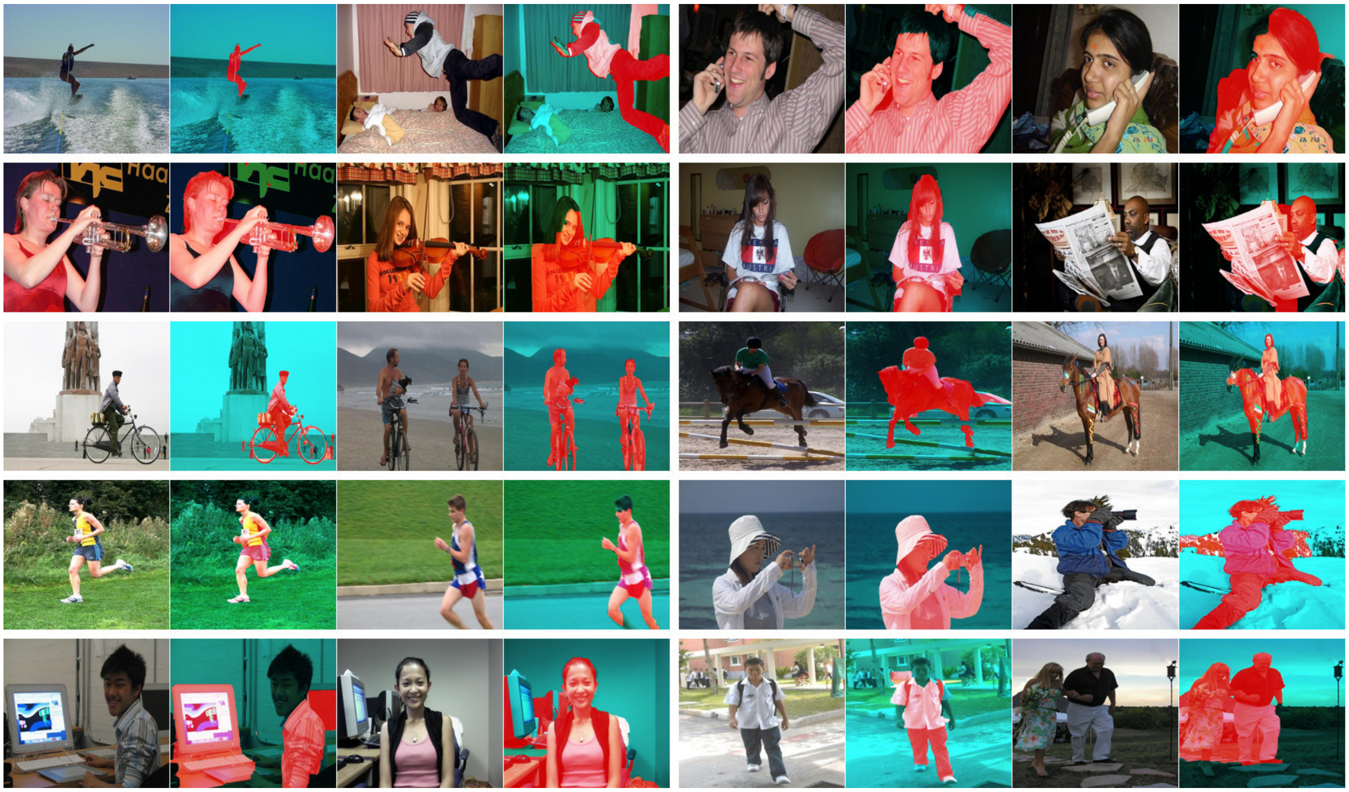


Fig. 5. Action masks of testing images in VOC 2012 action. Two samples including the input image and the acquired action masks are shown in each detected class. The classes are (from left to right, from top to bottom): Jumping, Phoning, Playing instrument, Reading, Riding bike, Riding horse, Running, Taking photo, Using computer, Walking. The red color denotes the learned action foreground mask.

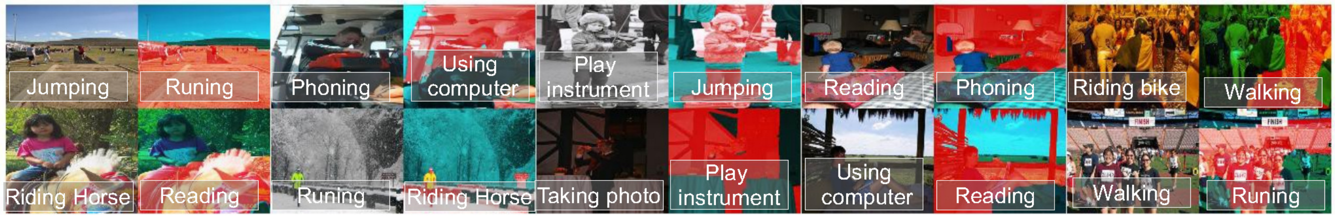


Fig. 6. Mistakes on PASCAL VOC Action val. set. We show the original image and the wrong prediction with their labels respectively.

TABLE III
AVERAGE PRECISION (%) OF STANFORD 40-ACTION DATASET.

Methods	Train BB	Test BB	CNN	mAP
Object bank [51]	yes	yes	-	32.50
LLC [52]	yes	yes	-	35.20
EPM [12]	yes	yes	VGG16	72.30
Ours	no	no	VGG16	80.77
Ours+image	no	no	VGG16	**82.64**

TABLE IV
AVERAGE PRECISION (%) OF WILLOW 7-ACTION DATASET.

Methods	Train BB	Test BB	CNN	mAP
EPM [12]	yes	yes	-	67.60
Ours	no	no	VGG16	75.31
Ours+image	no	no	VGG16	**76.96**

based representation on humans but with a lower results than those of ours. This is because they detected the useful parts from the human region for recognition. In contrast, our action representations build on all the interacted objects, which contain more contextual information.

C. Weakly-supervised object detection

Our method can be extended to weakly supervised object detection problem, where the object detector is learned with only image level labels. Our action recognition problem is similar to the weakly supervised object detection problem, where both problems only use image labels in training. The difference is that our action recognition problem is more challenging, because it involves the detection of multiple objects (without a number limit) in an action but object detection methods only need detect one object. Thus, our method can be freely applied to the weakly supervised object detection problem.

We compared our method with related methods on the PASCAL VOC 2012 detection dataset in Table V. Our method has a comparable result on this dataset. The state-of-the-

TABLE V
MEAN AVERAGE PRECISION (MAP %) OF PASCAL VOC 2012 DETECTION DATASET USING THE WEAKLY-SUPERVISED OBJECT DETECTION METHOD.

Methods	mAP
weak sup. [33]	74.5
ours	75.7

art methods [33] detected bounding boxes of objects on the whole image using image level labels. Ours used a part-based object representation and can detect the boundary of objects rather than bounding boxes, which can provide more subtle information for discriminance.

D. Discussion

In this paper, we proposed a new method for action recognition in still images without using human bounding boxes but with only image labels in training images.

Particularly, we find that:

- Bounding box annotations on humans are not necessary for action recognition in either training or testing stage (cf. Table I and Table II).
- Meaningful parts in objects can be learned using image level labels for each action class (cf. Fig. 4).
- The receptive fields of the learned parts in each action form an estimation of the the action foreground mask in each image (cf. Fig. 2).

We have provided the following methods for efficient action detection/recognition with only image level action labels in training images:

- We use an efficient multi-max pooling technique to compute multi-scale part representations from the outputs in CNN on each object (cf. Sec. III-A1).
- Part features are clustered in each image and computed into a Fisher Vector, from which an action class model with the group sparsity constraint is learned to compute the action specific parts (cf. Sec. III-A2, step 1-2).
- We use the GrabCut method to compute the action foreground mask from the receptive fields of the learned action specific parts in each image (cf. Sec. III-A2, step 3-4).
- We proposed to use product quantization to encode the objects in each action into the final representation for action recognition (cf. Sec. III-B).

In our experience, there is one issue with the proposed framework: the learning process may introduce heavy computations, when the numbers of images and action classes are very large in the dataset. Our methods generate a large number of parts for each action. It is important to research on how to reduce the number of effective parts and quickly assign them into action foregrounds and backgrounds.

V. CONCLUSION AND OUTLOOK

In this paper, we propose to recognize image-based actions with only action label annotations in training images (i.e. without using human annotations in both training and testing images). A systematic pipeline is developed in our approach to first generate object proposals using selective search, which are further decomposed into finer-grained object parts, that are subsequently used to delineate the detailed shape of human-object interaction regions. Then the action label is predicted with the help of an efficient product quantization method to encode features obtained from the human-object interaction regions. Experimental results on three benchmark datasets demonstrate the competitiveness of our approach with respect to the state-of-the-art methods even when they are allowed to use additional human annotations. Our future work includes improving the extraction of human-object interactions of different actions by using image level action labels *only*.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Comput.*, vol. 28, no. 6, pp. 976–90, 2010. 1, 2

[2] G. Cheng, Y. Wan, A. Saudagar, K. Namuduri, and B. Buckles, "Advances in human action recognition: A survey," *arxiv*, pp. 1–30, 2015. 1, 2

[3] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1, 3, 7

[4] J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2577–2584. 1

[5] T. Zhang, Y. Zhang, J. Cai, and A. Kot, "Efficient object feature selection for action recognition," in *International Conference on Acoustics, Speech and Signal Processing*, 2016. 1

[6] G.-D. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–61, 2014. 1, 2

[7] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 3177–3184. 1, 2

[8] M. Hoai, "Regularized max pooling for image categorization," in *Proc. British Machine Vision Conference*, 2014. 1, 7

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724. 1

[10] S. Gupta and J. Malik, "Visual semantic role labeling," arXiv:1505.0447, 2015. 1, 2

[11] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2015, pp. 1080–1088. 1, 2, 6, 7, 8

[12] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for semantic description of humans in still images," arXiv:1509.04186, 2015. 1, 2, 8, 9

[13] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 280–288. 1

[14] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2015, pp. 2470–2478. 1, 2, 3, 6, 7

[15] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601–614, 2012. 1, 2

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 2, 6, 7

[17] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 648–656. 2

[18] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *Proc. Advances in Neural Information Processing Systems*, 2011. 2

[19] B. Yao and L. Fei-Fei, “Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012. 2

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. 2

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Advances in Neural Information Processing Systems*, 2012. 2

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and semantic segmentation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 142–157, 2015. 2

[23] Y. Chai, E. Rahtu, V. Lempitsky, L. V. Gool, and A. Zisserman, “TriCoS: A tri-level class-discriminative co-segmentation method for image classification,” in *Proc. European Conf. Computer Vision*, 2012, pp. 794–807. 2

[24] C. Rother, V. Kolmogorov, and A. Blake, “GrabCut–interactive foreground extraction using iterated graph cuts,” in *SIGGRAPH*, 2004, pp. 309–314. 2, 4

[25] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, “Global contrast based salient region detection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015. 2

[26] G. Kim and E. P. Xing, “On multiple foreground cosegmentation,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 837–844. 2

[27] T. Ma and L. J. Latecki, “Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 1955–1962. 2

[28] X. Liu, M. Song, D. T. J. Bu, and C. Chen, “Random geometric prior forest for multiclass object segmentation,” *IEEE Trans. on Image Processing*, vol. 24, no. 10, pp. 3060–3070, 2015. 2

[29] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu, “Weakly supervised multiclass video segmentation,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 57–64. 2

[30] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, “Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 1908–1915. 2

[31] W. Ren, K. Huang, D. Tao, and T. Tan, “Weakly supervised large scale object localization with multiple instance learning and bag splitting,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 405–416, 2016. 2

[32] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, “Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation,” in *Proc. IEEE Int’l Conf. on Computer Vision*, 2015, pp. 1742–1750. 2

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free? weakly-supervised learning with convolutional neural networks,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 685–694. 2, 10

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2016. 2

[35] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath, “Weakly supervised localization using deep feature maps,” arXiv:1603.00489, 2016. 2

[36] Y. Zhang, J. Wu, and J. Cai, “Compact representation of high-dimensional feature vectors for large-scale image recognition and retrieval,” *IEEE Trans. on Image Processing*, vol. 25, no. 5, pp. 2407–2419, 2016. 2

[37] Y. Zhang, J. Wu, J. Cai, and W. Lin, “Flexible image similarity computation using hyper-spatial matching,” *IEEE Trans. on Image Processing*, vol. 23, no. 9, pp. 4112–4125, 2014. 2

[38] Y. Zhang, J. Wu, and J. Cai, “Compact representation for image classification: To choose or to compress?” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 907–914. 2

[39] L. Niu, J. Cai, and D. Xu, “Domain adaptive fisher vector for visual recognition,” in *Proc. European Conf. Computer Vision*, 2016. 2

[40] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based r-cnns for fine-grained category detection,” in *Proc. European Conf. Computer Vision*, 2014, pp. 834–849. 3

[41] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011. 3, 6

[42] J. 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. 3, 4

[43] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, “Weakly supervised fine-grained categorization with part-based image representation,” *IEEE Trans. on Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016. 3

[44] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the Fisher vector: Theory and practice,” *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013. 4

[45] J. Liu, S. Ji, and J. Ye, “SLEP: Sparse learning with efficient projections,” *Arizona State University*, 2009. [Online]. Available: http://www.yelab.net/publications/2009_slep.pdf 5

[46] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int’l Conf. on Computer Vision*, 2015, pp. 1440–1448. 5

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in *Proc. European Conf. Computer Vision*, 2014, pp. 346–361. 5

[48] 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. 6

[49] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei, “Human action recognition by learning bases of action attributes and parts,” in *Proc. IEEE Int’l Conf. on Computer Vision*, 2011, pp. 1331–1338. 7, 8

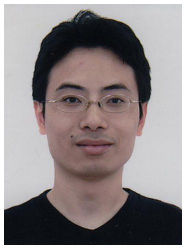
[50] V. Delaitre, I. Laptev, and J. Sivic, “Recognizing human actions in still images: a study of bag-of-features and part-based representations,” in *Proc. British Machine Vision Conference*, 2010, updated version, available at http://www.di.ens.fr/willow/research/stillactions/. 7, 8

[51] L.-J. Li, H. Su, Y. Lim, R. Cosgriff, D. Goodwin, and L. Fei-Fei, “Object bank: A high-level image representation for scene classification and semantic feature sparsification,” in *Proc. Advances in Neural Information Processing Systems*, 2011. 9

[52] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367. 9

Yu Zhang received his PhD degree in computer engineering from Nanyang Technological University, Singapore. He has worked with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore. He is currently a postdoctoral fellow in the Bioinformatics Institute, A*STAR, Singapore. His research interest is computer vision.

Li Cheng is a principal investigator with Bioinformatics Institute (BII), A*STAR, Singapore. Prior to joining BII July of 2010, He worked at Statistical Machine Learning group of NICTA, Australia, TTI-Chicago, USA, and University of Alberta, Canada, where he obtained his Ph.D.in Computer Science. His research expertise is mainly on computer vision and machine learning.



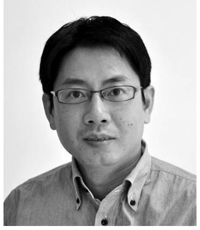
Jianxin Wu (M'09) received his BS and MS degrees in computer science from Nanjing University, and his PhD degree in computer science from the Georgia Institute of Technology. He is currently a professor in the Department of Computer Science and Technology at Nanjing University, China, and is associated with the National Key Laboratory for Novel Software Technology, China. He was an assistant professor in the Nanyang Technological University, Singapore, and has served as an area chair for ICCV, CVPR and senior PC member for AAAI. His research interests are computer vision and machine learning. He is a member of the IEEE.

Jianfei Cai (S'98-M'02-SM'07) received his PhD degree from the University of Missouri-Columbia. He is currently an Associate Professor and has served as the Head of Visual & Interactive Computing Division and the Head of Computer Communication Division at the School of Computer Engineering, Nanyang Technological University, Singapore. His major research interests include computer vision, visual computing and multimedia networking. He has published more than 170 technical papers in international journals and conferences. He has been actively participating in program committees of various conferences. He has served as the leading Technical Program Chair for IEEE International Conference on Multimedia & Expo (ICME) 2012 and the leading General Chair for Pacific-rim Conference on Multimedia (PCM) 2012. Since 2013, he has been serving as an Associate Editor for IEEE Trans on Image Processing (T-IP). He has also served as an Associate Editor for IEEE Trans on Circuits and Systems for Video Technology (T-CSVT) from 2006 to 2013.

Minh N. Do (M'01-SM'07-F'14) was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Australia, in 1997, and the Dr.Sci. degree in communication systems from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, in 2001.

Since 2002, he has been on the faculty at the University of Illinois at Urbana-Champaign (UIUC), where he is currently a Professor in the Department of Electrical and Computer Engineering, and hold joint appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, the Advanced Digital Sciences Center, and the Department of Bioengineering. His research interests include signal processing, computational imaging, geometric vision, and data science.

He received a Silver Medal from the 32nd International Mathematical Olympiad in 1991, a University Medal from the University of Canberra in 1997, a Doctorate Award from the EPFL in 2001, a CAREER Award from the National Science Foundation in 2003, and a Young Author Best Paper Award from IEEE in 2008. He was named a Beckman Fellow at the Center for Advanced Study, UIUC, in 2006, and received of a Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007. He was a member of the IEEE Signal Processing Theory and Methods Technical Committee, Image, Video, and Multidimensional Signal Processing Technical Committee, and an Associate Editor of the IEEE Transactions on Image Processing. He is a Fellow of the IEEE for contributions to image representation and computational imaging. He was a co-founder and CTO of Personify Inc., a spin-off from UIUC to commercialize depth-based visual communication.



Jiangbo Lu (M'09-SM'15) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering, Katholieke Universiteit Leuven, Leuven, Belgium, in 2009. Since September 2009, he has been working with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore, where he is leading a few research projects as a Senior Research Scientist. His research interests include computer vision, visual computing, image and video processing, robotics, interactive multimedia applications and systems, and efficient algorithms for various architectures. Dr. Lu served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) in 2012-2016. He received the 2012 TCSVT Best Associate Editor Award.