Multibranch Attention Networks for Action Recognition in Still Images

Shiyang Yan, *Student Member, IEEE*, Jeremy S. Smith, *Member, IEEE*, Wenjin Lu, and Bailing Zhang

***Abstract*—Contextual information plays an important role in visual recognition. This is especially true for action recognition as contextual information, such as the objects a person interacts with and the scene in which the action is performed, is inseparable from a predefined action class. Meanwhile, the attention mechanism of humans shows remarkable capability compared with the existing computer vision system in discovering contextual information. Inspired by this, we applied the soft attention mechanism by adding two extra branches in the original VGG16 model in which one is to apply scene-level attention whilst the other is region-level attention to capture the global and local contextual information. To make the multibranch model well converged and fully optimized, a two-step training method is proposed with an alternating optimization strategy. We call this model multibranch attention networks. To validate the effectiveness of the proposed approach on two experimental settings: with and without the bounding box of the target person, three publicly available datasets on human action were used for evaluation. This method achieved state-of-the-art results on the PASCAL VOC action dataset and the Stanford 40 dataset on both experimental settings and performed well on humans interacting with common objects dataset.**

***Index Terms*—Action recognition, contextual information, multibranch CNN, soft attention mechanism.**

I. Introduction

ACTION recognition is one of the central issues in computer vision as actions often serve as the key instrument for the semantic description of an image containing humans. Actions are also directly linked to mid-level concepts for high level tasks, such as image captioning. Despite the tremendous progresses made, there still exist many obstacles, particularly the description of the variations in human pose, the objects a person interacts with, and the scene, where the action takes place. There are two pathways to study action recognition, namely video-based and still image-based. Among the two, video-based action recognition has been relatively well investigated [1], [2]. Still image-based action recognition, on the other hand, has been studied less. The lack of motion information is arguably one of the major obstacles for still image-based action recognition.

Fig. 1. Example of similar pose leading to different actions. The left image is "brushing teeth" whilst the right image is "blowing bubbles" though they have a similar pose.

In the recent years, many methods have been proposed to tackle action recognition problems. Among them, human-object interactions have been studied as one of the important instruments toward the recognition of object related actions [3], [4]. As the human pose often plays a fundamental role in action recognition, another interesting approach is to find solutions for human pose estimation [5]. However, that approach is limited by the fact that similar poses can be associate with different actions. This is well illustrated in Fig. 1. The two children in the figure have similar poses. However, one is brushing her teeth whilst the other is blowing bubbles. The problem can be alleviated by either the introduction of contextual information, which is one of the main subjects of this paper, or an appropriate combination of pose and human-object interaction as proposed in [6], in which a conditional random field is applied to jointly model the pose and objects a person is interacting with. Other approaches for still image-based action recognition include the part-based model, with the deformable part model (DPM) [7] as the most influential one. The Poselets model [8] further developed DPM, which employs key points to build an ensemble model of human body parts, achieving improved performance in some vision tasks.

Intuitively, the solutions to human action classification hinge on the acquisition of local and global contextual information. To be more specific, local information associated with discriminative parts or objects provides detailed contextual features which would be important to action recognition. Object-related actions are associated with particular objects, which often provide key hints for recognition. Additionally, the global contextual information about the configuration of surrounding scenes is also instrumental. To summarize, the comprehensive

Manuscript received August 20, 2017; revised December 11, 2017; accepted December 12, 2017. Date of publication December 15, 2017; date of current version December 7, 2018. *(Corresponding author: Shiyang Yan.)*

S. Yan, W. Lu, and B. Zhang are with the Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: shiyang.yan@xjtlu.edu.cn).

J. S. Smith is with the Department of Electrical and Electronic Engineering, University of Liverpool, Liverpool L69 3GJ, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2017.2783944

description of an action comprises the articulation of body parts, the objects a person interacts with and the scene in which the action is performed. This can be well illustrated by the action types in sports. For example, for the action of "playing football," the poses of players, the football itself and the football pitch are all strong evidence for this action category.

Specifically, to fully consider the contextual information when recognizing actions, we exploit two types of contextual information: 1) the scene-level context and 2) region-level context, corresponding to the global and the local context, respectively. The scene-level context is to consider the surrounding scene while the region-level context is to exploit the important body parts or objects a person is interacting with. The scene-level context is coarse-grained and the region-level context is more fine-grained. In practice, given an image, the scene often means the background and region-level context are around the target person. Hence, these two kinds of context can be dealt with at the same time.

The relationship between contextual cues and visual attention has long been recognized [9]. Human perception is characterized by an important mechanism of focusing attention selectively on different parts of a scene. In natural language processing, the attention model has also been extensively studied, with applications including sequence to sequence training in machine translation [10], with the aid of two types of attention model, namely, hard attention and soft attention. Soft attention is deterministic and can be trained using back-propagation [11], which has also been extended and applied to the image captioning task [11]. Sharma *et al.* [12] used pooled convolutional descriptors with soft attention-based models for video-based action recognition and achieved good results. However, the above works on attention-based networks are all implemented with recurrent neural networks (RNNs). It would be very interesting to investigate the applicability of attention mechanism in the general CNN frameworks to which static images are the subjects to process. Though the spatial transformer networks proposed in [13] can be considered as an approach to realize soft attention in general CNN framework, our motivation is different from theirs as our model operates on both the region level and scene level. To our best knowledge, we are the first to incorporate soft attention mechanism into CNNs for action recognition from still images. For convenience, the proposed scheme is formed as multibranch attention networks. The CNN model with multibranch attention mechanism can be trained in an end-to-end way, which can be illustrated by the system diagram shown in Fig. 2.

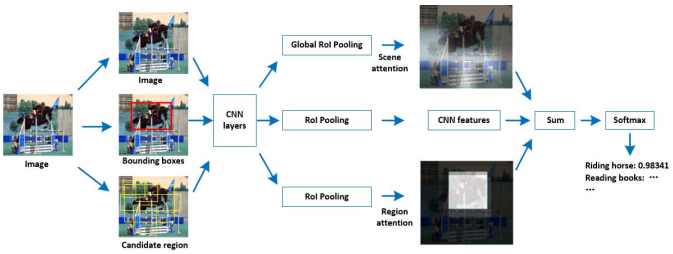

Fig. 2. System diagram of our proposed multibranch attention networks. There are a total of three branches. Top: scene attention. Middle: target person region classification. Bottom: region attention. The classification scores from three branches are summed and forwarded to a softmax layer for final classification.

II. Related Works

In this section, we review recent research on action recognition and attention models and discuss the relevance to our research.

A. Action Recognition

Video-based action recognition has been well studied. The recently published papers [14] provide good literature review. Still image-based action recognition can be roughly categorized into three groups. The first group makes use of the human body information [5], [15]. Normally the bounding box of the human is used to indicate the location of the person. For instance, Thurau and Hlavác [5] exploited human poses by learning a pose primitive for action recognition. There are also approaches making use of information from human body parts to aid the action recognition. Maji *et al.* [16] developed a body pose representation approach by learning and forming Poselets which are patches learned from body parts. Gkioxari *et al.* [15] concentrated on human body parts within a CNN model and developed a part-based approach by leveraging convolutional features, with the effectiveness demonstrated using several publicly available datasets.

The second group use the human-object interaction to discover the action categories by modeling the human-object pair and its interactions. For example, Yao and Fei-Fei [6] modeled a person's body parts and objects by a conditional random field to recognize actions from still images. Yao and Fei-Fei [4] developed Grouplet to recognize human object interactions by encoding appearance, shape and spatial relations of multiple image patches. Desai *et al.* [17] formulated the problem of action recognition as a latent structure labeling problem and developed a unified, discriminative model for human object interaction. Recently, deep CNNs have also been employed for action recognition. For instance, Gkioxari *et al.* [18] proposed an interesting method by automatically selecting the most informative regions (normally the objects) around the person bounding box and achieved promising results on several datasets.

The third group have recourse to the scene context information. The background in an image can be taken as the context or scene of an executed action. For example, Delaitre *et al.* [19] studied the efficiency of different strategies based on the bag-of-visual-words approaches. It was found that the information extracted from the background does help to boost the performance of the recognition task. Similarly, Gupta *et al.* [20] encoded the scene for action image analysis and achieved good results.

As a contrast to the previously published approaches, we considered both the objects a person is interacting with and the scene as contextual information and model them explicitly to form a unified, effective model. This is achieved with the aid of a soft attention mechanism embedded into a CNN model.

B. Attention Model

One important property of human perception is that we do not tend to process a whole scene in its entirety at once. Instead humans pay attention selectively on parts of the visual scene to acquire information, where it is needed [21]. Different attention models have been proposed and applied in object recognition and machine translation. Mnih *et al.* [21] proposed an attention mechanism to represent static images, videos or as an agent that interacts with a dynamic visual environment. Also, Ba *et al.* [22] presented an attention-based model to recognize multiple objects in images. The two above-mentioned models are all related with RNNs and with the aid of reinforcement learning strategy.

Bahdanau *et al.* [10] proposed a novel attention model for neural machine translation without the prerequisite of reinforcement learning, which can be trained end-to-end by back propagation method. It is called a soft attention model. Later, a comprehensive study for hard attention bound with reinforcement learning and soft attention for the task of image captioning was published by Xu *et al.* [11]. Followed up researches include action recognition with soft attention proposed by Sharma *et al.* [12] and video description generation [23].

A related research topic, saliency detection, is also motivated by human perceptions. However, most of the saliency detection methods [24]–[26] used low-level image features, e.g., contrast, edge, and intensity, which can be considered as fixed and bottom-up approach in contrast with the top-down approach of attention mechanism. Normally these methods cannot capture the task-specific information. Zhou *et al.* [27] applied global average pooling [28] to discriminate salient CNN features for the target object category. It is a kind of task-relevant approach. However, it is still not flexible enough to operate on the region-level as the soft attention does in this paper. The region-level context, which is a more fine-grained feature, can be captured by region-level attention easily. In short, attention mechanism is a more recent and flexible approach, which is able to learn relevant features for the specific task and plays a significant role in various vision tasks.

As an overview of the published works, soft attention models were mainly realized with the leverage of RNNs for handling sequences or time-domain information. To directly process static image, it is much desirable to implement soft attention models in the general CNN frameworks. Teh *et al.* [29] applied the soft attention mechanism in CNNs for weakly supervised object localization and achieved good results on the PASCAL VOC detection challenge [30]. They emphasized the relative importance on candidate proposals to automatically select target regions with only region-level considered.

III. Approach

In this section, we introduce the proposed approach of multibranch attention networks for action recognition. The augmented CNN system contains three branches: 1) target person region classification; 2) scene-level attention; and 3) region-level attention.

Our model was built on the VGG16 [31], which is a very effective CNN structure for large scale image analysis. According to [31], the VGG16 network has five convolutional blocks, each with three convolutional layers. We retained the structure of convolutional layers unchanged, with attention networks cascading these convolutional layers.

A. Classification of Target Person Region

Normally the benchmark action recognition datasets provide the bounding box of the target person, e.g., the PASCAL VOC 2012 action challenge [30] and Stanford 40 action dataset [32]. As our model is fully supervised, we designed this branch of CNN model to perform the classical recognition of person regions. We applied RoI pooling developed by Girshick [33] for the purpose of pooling different size regions into fixed size feature maps to facilitate the following classification. This branch is built based on Fast RCNN [33], only with some minor modifications. Specifically, we select the foreground with a overlap more than 0.5 with the target person region, and set the foreground over background ratio as 1, which indicates the framework is used for classification instead of detection. This can also be considered as a kind of data augmentation because the model samples on candidate regions instead of limiting the samples only on target person region.

B. Region-Level Attention

To exploit the fine-grained properties of a given image, we design the second branch for the CNN to explicitly capture more informative regions regarding the action performed. We take a similar strategy of selecting regions as in the R*CNN [18]. In the R*CNN, a set of regions called secondary region is selected based on the overlap ratio with the bounding box of the target region. In our research, we also set a ratio threshold to select regions for this branch. Intuitively, the regions that overlap with a certain ratio normally indicate the parts of a person or objects a person is interacting with. The regions far away from a person will be ignored based on the overlap smaller than the threshold. As a result, more related regions will be selected for further processing at the first step.

Subsequently, selected regions are aggregated with RoI pooling resulting the fixed size feature maps. In this branch, we use the fully connected features instead of convolution features because there is certain number of regions to process. Feature from fully connected layers have lower dimension and hence can largely reduce the computational burden. All the extracted features are forwarded to a linear layer to generate score map. If there are n regions each with d dimension, we can reshape the n feature maps to one feature map with a dimension of $n \times d$. Then the score map S is with a dimension of $n \times 1$. In practice, this is a fully connected layer which can be easily implemented. The soft attention model requires a region location softmax to generate the attention map which is expressed as follows:

$$A_i = \frac{\exp(S_i)}{\sum_{i=1}^{n} \exp(S_i)} \tag{1}$$

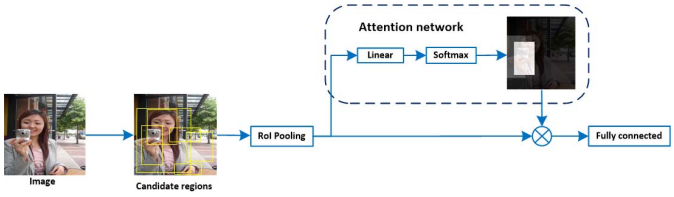

Fig. 3. Illustration of region attention: this attention branch is to allocate attentive weights on candidate regions around the target person. Regions with higher weights will provide more contributions to the final classification.

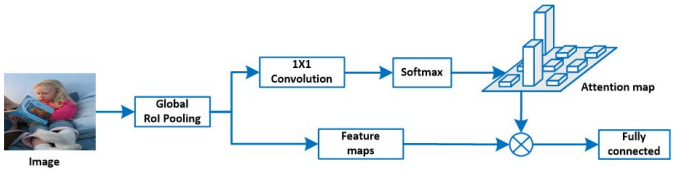

Fig. 4. Illustration of scene attention: this attention branch is to exploit the more informative regions in the scene. This is achieved by allocating attentive weights in each location of an image corresponding to different receptive fields of the original image.

where A_i is the element of the attention map for the ith region. To allocate weights on regions, this attention map is elementwise multiplied with the features F

$$\widetilde{F} = A \odot F \tag{2}$$

where $\widetilde{F}$ is the attentive feature map. To obtain a final feature representation of the regions, we sum all the weighted features into one representation

$$E = \sum_{i=1}^{n} \widetilde{F}_i. \tag{3}$$

The feature representation of all weighted regions E can be used by the fully connected layer to obtain classification scores. More details of the block diagram of region attention branch are illustrated by Fig. 3.

C. Scene-Level Attention

This branch of the CNN model is to consider the scene-level context of an action category. As previously explained, scene or background information often plays a supportive role in action recognition. However, indiscriminative extraction of all of the background would be counterproductive as some subregions of the scene may not be relevant to the action of interest. To solve this problem, we exploit the attention model to discriminatively select the most informative locations within the background. Hence, we applied soft attention over the CNN features of the scene or background as one branch to aid the action recognition.

As a scene normally means the entire image, we first pooled the original convolutional features into a fixed size feature map by a new pooling layer: global RoI pooling, which divides the entire image or feature map into several grids and then performs max pooling inside each grid. The obtained feature map will have the same size regardless of the original image size. More formally, we can pool an image with arbitrary size into a feature map F with size $w \times h \times d$, in which w, h, and d are the width, height, and channel size of the feature map, respectively.

TABLE I
NETWORK CONFIGURATION

Inputs (Images, candidate regions and labels)		
Convolution Blocks (Conv1-Conv5) derived from VGG16 [31].		
Scene-level attention	Target person region	Region-level attention
Global RoI Pooling (Pooled size: 7×7)	RoI Pooling (Pooled size: 7×7)	RoI Pooling (Pooled size: 7×7)
1×1 Convolution (Channel number: 1)	FC1 (Dimension: 4096)	Region-FC1 (Dimension: 4096)
Softmax (Over location)	FC2 (Dimension: 4096)	Linear (Dimension: 1)
Elementwise Product	Score	Softmax (Over input regions)
Sum (Over Location)		Elementwise Product
Scene-FC1 (Dimension: 512)		Sum (Over Input regions)
Scene-core		Region-FC2 (Dimension: 4096)
		Region-FC3 (Dimension: 4096)
		Region-score
Sum Scores (Dimension: Number of Categories)		
Softmax		
Cross Entropy Loss		

The pooled feature map is then convolved by a 1×1 convolution layer and the output channel of this convolution layer is also 1. A score map Z of $w \times h \times 1$ can be consequently obtained. Following the practice of soft attention in [11], the score map is further processed by a location softmax which is defined as follows:

$$A_{ij} = \frac{\exp(Z_{ij})}{\sum_{i=1}^{w}\sum_{j=1}^{h}\exp(Z_{ij})} \tag{4}$$

where A_{ij} is the element of the attention map at position (i, j). Then the attention map A is elementwise multiplied with the feature map F which can be expressed as follows:

$$\widetilde{F} = A \odot F \tag{5}$$

where $\widetilde{F}$ is the attentive feature map. To obtain the final feature representation of the scene, the attentive feature map is summed over positions which can be described by

$$E = \sum_{i=1}^{w}\sum_{j=1}^{h} \widetilde{F}_{ij}. \tag{6}$$

The feature E is subsequently forwarded to the fully connected layers to obtain the classification scores. Fig. 4 further explains the scene-level attention mechanism implemented in the multibranch attention networks.

D. Networks Architecture

The details of the CNN architecture are given in Table I. The convolutional blocks are derived from the VGG16 model [31] which includes five blocks. More detailed explanation can be referred to [31].

There are a total of three branches following the convolutional blocks starting from a global RoI pooling layer followed

by two RoI pooling layers. The global RoI pooling compresses the entire image into 7×7 feature map, which is used for scene-level attention networks. It starts from a 1×1 convolution layer with channel number of 1. A location softmax layer is connected to generate the attention map. The feature map and attention map are subsequently processed simultaneously and fused into a weighted sum of feature from each location. The following fully connected layer is named "Scene-FC1" of size 512. The "Scene-score" can be obtained based on the outputs of the fully connected layer.

The first RoI pooling (the middle column in Table I) pools the region of the target person into a fixed size feature to perform classical CNN recognition. The second RoI pooling (the right column in Table I) pools candidate regions generated with selective search algorithm into fixed size feature maps. These feature maps are then forwarded to the fully connected layers "Region-FC1" to generate feature maps with a dimension of 4096. The region softmax transfers outputs from a linear layer into an attention map over regions. The attention map is elementwise multiplied with the features and summed into a whole feature representation before another two fully connected layer. The "Region-score," "Score," and Scene-score are summed and activated by the softmax layer with cross entropy loss for the training.

E. Training Strategy

We followed the common pretraining plus fine-tuning practice of applying CNN model. Specifically, the pretrained VGG16 model on ImageNet [34] was fine-tuned for the task at hand.

The two branches of the attention mechanism can be considered as subsets of parameters toward image features, which are to be found by overall optimization for the action classification task. Such parameters on the optimization task, make the direct application of stochastic gradient descent (SGD) very challenging. Our intuition is to borrow the idea from alternating optimization [35].

More formally, we can consider the full parameter set of the CNN model as $X = \{X_1, X_2\}$, where X_1 corresponds to the parameters from the branch of the target person region classification and region-level attention and X_2 indicates the parameters from the branch of scene-level attention. The task is to optimize the CNN model which is a function of these parameters: $F = F(X)$. Alternating optimization is an iterative procedure to minimize all the variables by alternating restricted minimizations over the individual subsets of variables X, in this case, X_1 and X_2 [35]. Specifically, we propose a two-step training strategy for our networks: 1) the target person region recognition and 2) region-level attention are trained jointly at first. This is equal to optimize over the subset of X_1. Then we add the scene-level attention to the network while keeping the weights from the convolutional blocks, the target person region classification and the region-level attention unchanged. This means the optimization over subset X_2 is performed subsequently. In the first-step training, the maximum iterations were set as 40 000. Once trained, the model was added with the scene-level attention branch and further trained with other 25 000 iterations.

As indicated in [33], training all the convolutional layers of VGG16 model would be unnecessary. Instead, we kept the first two convolutional blocks unchanged and trained other layers during the first-step training.

During training, 50 candidate boxes were randomly selected based on a threshold of overlap ratio for training of region attention as we found 50 boxes can reach a balance of training efficiency and generalization capability. Five hundred candidate boxes were selected for region-level attention network when testing as 500 boxes can cover most of the meaningful regions. Further increasing this number may introduce noise and also slow the testing process.

IV. Experiments

The multibranch attention networks were implemented based on the Caffe platform [36]. The training was conducted with SGD with a batch size of 32. All the experiments were conducted with a Nividia Titan X GPU installed in a PC running the Ubuntu 14.04 operating system.

A. Experimental Setting 1 (With the Bounding Box of the Target Person)

1) PASCAL VOC 2012 Action Dataset: The PASCAL VOC action dataset serves as one of the PASCAL VOC 2012 challenges [30], which consists of ten different actions, *Jumping*, *Phoning*, *Playing an Instrument*, *Reading*, *Riding Bike*, *Riding Horse*, *Running*, *Taking Photograph*, *Using Computer*, *Walking*, as well as examples of people not performing some of these actions, which are labeled as other. The target person boxes containing the people are provided both at training and testing time. During testing, for every sample we estimate the probabilities for all actions and compute the average precision (AP).

The challenge organizers require participators to make use of the validation set for parameter optimization and the test set to report performance [30]. Hence, we first measured the performance on the validation set and then submitted results of test set to the evaluation server. When evaluating the validation set, the training set was used only for training. Both the training set and the validation set were applied for training when submitting results for the test set and evaluating performance.

The comparative experiments were conducted to optimize parameters and confirm the effectiveness of the proposed model. Table II provides the AP results on the validation set. From the table, following observations can be obtained:

a) Baseline approach: The Fast RCNN [33] was set as the baseline approach because it is generally acknowledged as a better object detection model with much improved performance than RCNN [39]. However, Fast RCNN is not limited to detection and can also be applied in action recognition from still images [18], with some modifications. Specifically, we set the foreground over background ratio in Fast RCNN as 1 during training, which indicates the model does not need to discriminate foreground from background as in the detection scheme.

TABLE II
AP RESULTS ON PASCAL VOC VALIDATION SET

Approach	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	Mean AP (%)
Image classification (VGG16 model)	78.9	64.0	91.5	71.6	88.6	92.6	83.2	71.1	89.7	53.9	78.5
Fast RCNN (single branch, no regression) [33]	82.4	69.9	90.7	72.1	93.5	97.0	84.1	82.7	87.6	65.6	82.6
Fast RCNN (single branch, with regression) [33]	87.4	70.2	91.2	75.0	**95.4**	**97.8**	85.7	81.6	85.9	72.4	84.3
Two branch (no regression, with threshold)	86.3	76.6	90.8	79.6	93.6	97.0	85.6	84.4	92.5	67.4	85.4
Multi-branch attention (no regression, with threshold)	87.8	77.0	92.3	81.4	94.4	96.5	86.2	82.8	92.2	71.3	86.2
Two branch (with regression, no threshold)	85.8	73.2	90.0	81.8	93.3	96.3	85.0	78.2	90.7	70.3	84.5
Multi-branch attention (with regression, no threshold)	85.6	72.7	91.4	81.3	93.4	96.6	84.8	79.1	90.4	70.8	84.6
Two branch (with regression, with threshold)	87.8	77.1	92.5	**81.4**	94.3	96.5	**86.3**	83.3	92.2	71.1	86.3
Multi-branch attention (with regression, with threshold)	**87.8**	**78.4**	**93.7**	81.1	95.0	97.1	86.0	**85.5**	**93.1**	**73.4**	**87.1**

For Fast RCNN without bounding box regression, 82.6% AP was reported. Adding bounding box regression can boost the AP performance to 84.3% which testifies again that multitask training with bounding box regression can boost the performance reported earlier [33].

b) Our methods: Experiments were conducted for the two-step training strategy as explained above. The AP performance from the first-step model which uses a two-branch network (target person region and region-level attention) is reported first. Borrowing the benefit of bounding box regression in Fast RCNN, we added a regression layer in the first-step training. It turns out that our model achieves better AP results than Fast RCNN, with 85.4% AP without bounding box regression and 86.3% AP when adding the bounding box regression layer. Also, the threshold for selecting candidate boxes plays a significant role in promoting performance. Specifically, boxes overlap more than 0.1 and less than 0.7 with bounding boxes of the target person were selected for the branch of region attention. The obvious improvement of AP performance when adding threshold is clearly indicated in Table II. This is reasonable because only bounding boxes that overlap with the person in a range can exploit useful context information, such as the objects the person interacts with. After this, the second-step training was conducted to train the multibranch attention model. As the weights from the branch of target person bounding box classification are kept as constants, there is no need to add bounding box regression in the second-step training. In summary, our proposed multibranch attention networks produce the best mean AP value (87.1%) among all the experimental settings which validate the effectiveness of this model.

To further evaluate our method and compare it with other newly proposed approaches, the experimental results on the test set were generated and submitted to the PASCAL VOC evaluation server for the final evaluation. We follow the training strategy explained in Section III-E considering both the training set and validation set of PASCAL VOC 2012 as training set. This is a reasonable deployment as the challenge organizers allow the validation set to be used in training when reporting results on test set. Once trained with alternating optimization of 40 000 and 25 000 iterations as the first step and second step, respectively, the model was directly used for testing. Also, current leading method, such as R*CNN [18], which will be discussed later, used a similar strategy. Hence, it is also a fair deployment. Table III shows the AP results of the proposed approach and other competing methods. Oquab *et al.* [37] trained an 8-layers network on the box of the target person to perform action recognition. Hoai [38] used an 8-layers CNN model to extract features from fully connected layers from regions at multiple locations and scales inside the image and accumulate their scores for prediction, which is more comprehensive than only training on the box from target person. The results of this method are also better than Qquab *et al.* [37]. Simonyan and Zisserman [31] combined the VGG16 and VGG19 network and retrained classifier, such as SVMs using fully connected features from the target person region and entire image.

The current top ranked method on PASCAL VOC 2012 Action dataset is R*CNN [18] which was trained on the target person region with a secondary box. The secondary box was selected using the multi-instance learning method during training and testing. Specifically, R*CNN applied the max operation on scores generated by secondary boxes and combined them with the target person region for recognition. Our methods achieved same mean AP results with R*CNN, with a 90.2% mean AP value on the testing set.

A visualization of the attention model is provided in Fig. 5. We plot the original image, region-level attention, and scene-level attention in separate three rows. The brighter a place of an image is, the more important it is for recognition. The region-level attention generates important bounding boxes while scene-level attention captures attentive regions as indicated by the figure. It is interesting to discover that normally the two attention models generate different regions which implies that they are complementary. Note that all the example images are randomly selected.

c) Analysis of each of the three branches: Table IV presents the AP results of each of the three branches and their random combinations. In single branch settings, the branch of general image features (Fast RCNN branch) yields the best results, which shows that the person regions play a fundamental role in recognition. Beside this branch, the most important branch is the second branch (region-level attention), which discover the fine-grained contextual information. From the table, it is obvious that the third branch alone (scene-level branch) cannot provide very good results. However, as discussed previously, when fused together with the other two branches, satisfactory results can be obtained, which indicates that there are little correlations between scene-level attention branch and the other two branches. This is also what we intended to accomplish by the alternating optimization initially, which is to guarantee that the scene-level attention is to capture complementary information of the other two branches. In the random combinations of two branches, the first and second branch together generates best results. This phenomenon shows consistency with the performance of single branch as

TABLE III
AP RESULTS ON PASCAL VOC TEST SET

Approach	CNN layers	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	Mean AP (%)
Oquab et al. [37]	8	74.8	46.0	75.6	45.3	93.5	95.0	86.5	49.3	66.7	69.5	70.2
Hoai [38]	8	82.3	52.9	84.3	53.6	95.6	96.1	89.7	60.4	76.0	72.9	76.3
Action Part [15]	16	84.7	67.8	91.0	66.6	96.6	97.2	90.2	76.0	83.4	71.6	82.6
Simonyan et al. (VGG16 model) [31]	16&19	89.3	71.3	94.7	71.3	**97.1**	98.2	90.2	73.3	88.5	66.4	84.0
R*CNN [18]	16	91.5	84.4	**93.6**	83.2	96.9	98.4	**93.8**	**85.9**	**92.6**	**81.8**	**90.2**
Multi-branch attention (ours)	16	**92.7**	**86.0**	93.2	**83.7**	96.6	**98.8**	93.5	85.3	91.8	80.1	**90.2** *

* The official results: http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_noeq.php?challengeid=11&compid=10

TABLE IV
COMPARISON OF EACH OF THE THREE BRANCHES AND THEIR RANDOM COMBINATIONS ON PASCAL VOC VALIDATION SET

Approach	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	Mean AP (%)
Fast RCNN alone (the first branch)	87.4	70.2	91.2	75.0	95.4	97.8	85.7	81.6	85.9	72.4	**84.3**
Region-level attention alone (the second branch)	80.7	70.0	88.8	79.7	89.6	94.4	81.3	75.4	88.8	66.3	81.5
Scene-level attention alone (the third branch)	66.3	67.0	82.5	66.9	77.9	84.4	71.4	62.5	85.2	46.5	71.0
The first and second branch	87.8	77.1	92.5	81.4	94.3	96.5	86.3	83.3	92.2	71.1	**86.3**
The first and third branch	83.2	70.0	90.3	72.7	89.5	92.6	82.0	74.4	89.7	65.3	81.0
The second and third branch	83.9	78.1	93.8	80.9	93.6	95.4	84.9	82.7	93.0	69.9	85.6
Multi-branch attention	87.8	78.4	93.7	81.1	95.0	97.1	86.0	85.5	93.1	73.4	**87.1**

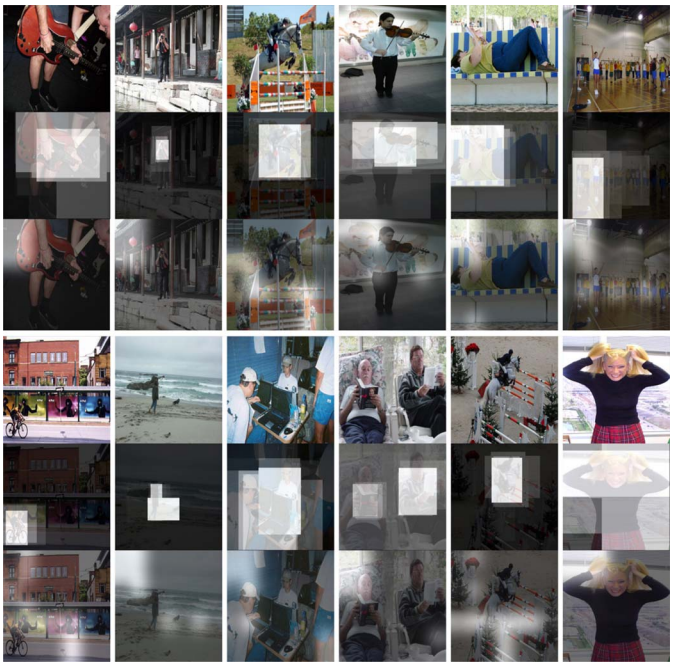

Fig. 5. Visualization of region attention and scene attention on the PASCAL VOC test set, the brighter areas of an image means the attentive regions of the image. Region attention is the row below the original image which generates attentive boxes. The scene-level attention is the row below region attention in which the brighter parts means higher importance. Note all the images in the figure are sampled randomly.

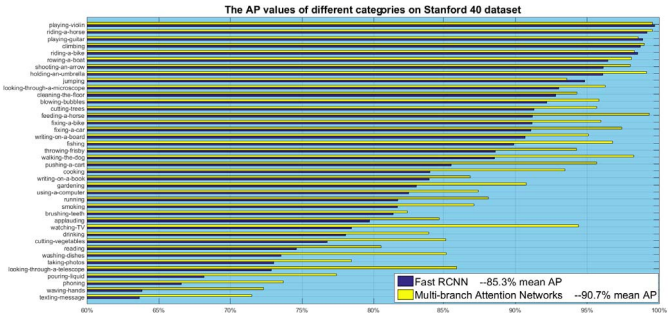

Fig. 6. AP results for different categories on the Stanford 40 dataset: the yellow bars are our method while the blue bars are the AP results of Fast RCNN.

TABLE V
AP RESULTS ON THE STANFORD 40 DATASET AND COMPARISON WITH PREVIOUS RESULTS

Method	Mean AP(%)
Object bank [40]	32.5
LLC [41]	35.2
EPM [42]	40.7
DeepCAMP [43]	52.6
Khan et al. [44]	75.4
Semantic parts [45]	80.6
VLAD spatial pyramids [46]	88.5
Fast RCNN alone [33]	85.3
Region-level attention alone	81.0
Scene-level attention alone	72.1
Two branch (ours)	90.6
Multi-branch Attention Networks (ours)	**90.7**

the first and second branch are the two most important parts of the networks.

2) Stanford 40 Dataset: The proposed method was also evaluated on the Stanford 40 dataset [32] which is a larger database containing 40 different types of daily human actions. It has 9352 images in total. The number of images for each class ranges from 180 to 300. The dataset provides the training and testing splits for each class, namely 100 images of each class for training and the rest for testing.

Fig. 6 shows the bar chart of the AP values over the 40 action categories from the Fast RCNN (the baseline) and our multibranch attention networks. It is apparent from the figure that our approach outperforms the baseline by a large margin, with a 90.7% mean AP compared with 85.3% mean AP of baseline approach.

Table V shows a comparison of our methods with alternative approaches. The model with two branches (from first-step training) shows good results. The improvement of the mean AP result by adding the branch of scene-level attention is obvious. With the multibranch attention networks, we further improved the mean AP result to 90.7%. To conclude, we achieved the best result on the Stanford 40 dataset with a 5.4% higher mean AP than Fast RCNN which is the baseline method.

TABLE VI
AP RESULTS ON PASCAL VOC VALIDATION SET (EXPERIMENTAL SETTING 2)

Approach	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	Mean AP (%)
Image classification (VGG16 model)	78.9	64.0	91.5	71.6	88.6	92.6	83.2	71.1	89.7	53.9	78.5
Ours	78.4	72.1	91.4	75.4	88.9	93.7	84.3	70.2	90.3	55.5	80.0

TABLE VII
AP RESULTS ON PASCAL VOC TEST SET (EXPERIMENTAL SETTING 2)

Approach	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	Mean AP (%)
Simonyan et al. (Image classification) [31]	-	-	-	-	-	-	-	-	-	-	79.2
Zhang et al. (Minimum annotations) [47]	86.7	72.2	**94.0**	71.3	**95.4**	97.6	88.5	72.4	88.8	65.3	83.2
Ours	**87.2**	**81.5**	89.9	**78.8**	94.4	**94.9**	**90.0**	**73.8**	**90.0**	65.3	**84.5**

B. Experimental Setting 2 (Without the Bounding Box of the Target Person)

The bounding boxes of target persons are very important during training as they provide the fundamental feature for the person to be recognized. However, they are often hard to obtain during real-world applications as the manual annotation for the bounding boxes is rather time-consuming and painful. Also, the requirements of inputting bounding boxes severely discourage further applications of the topic. Hence, in this section, we show that when the annotations of the bounding boxes of target person are not provided, our model can also perform well in the task of action recognition.

As we do not use the bounding box of the target person during training and testing, we modify the model architecture to facilitate the recognition. From results of experimental settings 1, if lacking the general CNN features of the target person, the most important branch is the region-level attention. Hence, in order to make the networks effective and simple, we set two branches in the networks for this experimental settings

1) *Image Classification Branch:* The entire image is forwarded to a global RoI pooling layer and perform general image classification. This is a fundamental branch which also provide a baseline of our two branch model.
2) *Region-Level Attention Branch:* The region attention branch is to automatically retrieve relevant regions during recognition, this is similar with the region attention branch explained previously. The only difference here is that the bounding box selection is omitted as the region of the target person is not provided.

During training, the two branches are trained jointly with 40 000 iterations under the Caffe platform. We then report the AP results on PASCAL VOC 2012 action dataset.

1) PASCAL VOC 2012 Action Dataset: As shown in Table VI, the model achieved 80.0% mean AP performance on PASCAL VOC validation dataset whilst the general image classification only achieved 78.5% mean AP result. To further validate the proposed methods, we then report the AP results from PASCAL VOC evaluation server. As shown in Table VII, the proposed model got 84.5% mean AP, which are the state-of-the-art results among the methods without training bounding boxes. This can be attribute to our region-level attention branch, which serves as a model which can automatically retrieve not only the contextual information but also the person region, in experimental setting 2.

TABLE VIII
AP RESULTS ON THE STANFORD 40 DATASET WITH EXPERIMENTAL SETTINGS 2

Method	Mean AP(%)
Image classification (VGG16 model)	81.4
Zhang et al. (Minimum annotations) [47]	82.6
Ours	85.2

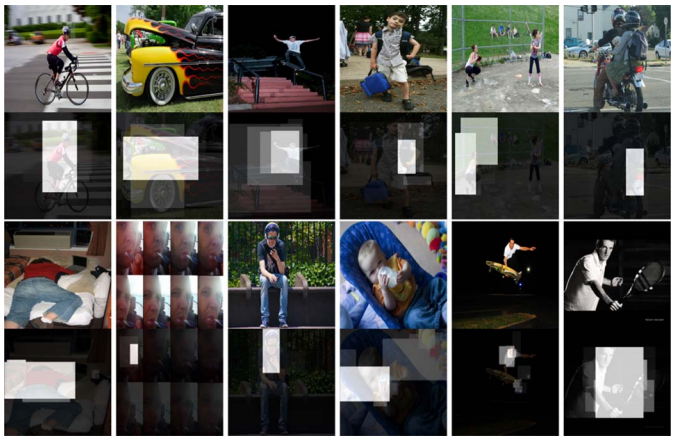

Fig. 7. Learned region-attention map of HICO dataset in the experimental setting 2: it can clearly show that the attention region contains the objects and the person region which contribute most for a certain action category. All the images are sampled randomly.

2) Stanford 40 Dataset: Table VIII provides the performance on Stanford 40 action dataset on the experimental setting 2. Our method achieves 85.2% mean AP results on the 40 action categories of the dataset, which is competitive with the mean AP results (85.3%, see Table V) of Fast RCNN (with training bounding boxes). Also, our method leads the scheme in [47], which is a recently proposed action recognition method without the training bounding boxes.

3) HICO Dataset: The PASCAL VOC action dataset and Stanford 40 dataset can be considered as medium-sized datasets. To further test the generalization capability of the proposed approach on a big dataset, we also conducted experiments on humans interacting with common objects (HICO) dataset [48]. This dataset is currently the largest one for action recognition, which consists of 50 000 images labeled to 600 human-object interaction categories. It is also related to MS COCO dataset [49] as each category in the HICO dataset

TABLE IX
MEAN AP RESULTS ON THE HICO DATASET WITH EXPERIMENTAL SETTINGS 2

Method	Mean AP(%)
AlexNet+SVM [48]	19.4
VGG16, Image classification [50]	29.4
VGG16, R*CNN [50]	28.5
VGG16, Scene-RCNN [50]	29.0
RoI and Scene fusion [50]	**33.6**
Ours	32.8

TABLE X
P-VALUE FOR THE OBTAINED RESULTS IN THE EXPERIMENTS

	Dataset	Upper-tailed p-value
Experimental Setting 1	Pascal VOC Validation Set	0.0009
	Pascal VOC Test Set	0.0052
	Stanford 40 Dataset	0.0
Experimental Setting 2	Pascal VOC Validation Set	0.0260
	Stanford 40 Dataset	0.0
	HICO Dataset	0.0310

is composed of a verb-object pair, with objects belonging to the 80 object categories from MS COCO. However, the HICO dataset does not provide human bounding boxes for a predefined action category. Hence, it is only suitable for the experimental setting 2 in this paper.

Different from PASCAL VOC action dataset and Stanford 40 dataset in which the action categories are exclusive, more than one human-object interaction category is labeled for a single instance. Actually, these action categories can be considered as mid-level features, in contrast with those as high-level actions in PASCAL VOC and Stanford 40 dataset. Hence, we treat each of the human-object interaction category as a binary classification problem and use Sigmoid as the activation function instead of softmax. As the dataset is larger, we train them with 60 000 iterations in Caffe platform and report the mean AP results of our approach.

Table IX demonstrates the mean AP results of our approach and comparison with other methods. Specifically, the baseline approach reported in [48] applied an AlexNet and SVM classifier for recognition, with only 19.4% mean AP. Reference [50] reported results of several methods. They first applied VGG16 for general image classification approach, achieved 29.4% mean AP. For R*CNN approach, they used a pretrained Faster RCNN object detector to detect human bounding boxes. With these bounding boxes, they then trained R*CNN and Scene-RCNN as in [18]. However, the mean AP results of R*CNN and Scene-RCNN is even worse than general image classification, the possible reason, as explained in [50], is that R*CNN try to find a single box using multi-instance learning, which is not able to cover all 600 action categories. This is not a problem in our method because we fully exploit region-level attention and sample 500 boxes to facilitate the recognition. As shown in Table IX, our approach achieved competitive results with the one proposed in [50] but are simpler and more efficient as we do not rely on bounding boxes at all. A visualization of learned attention region is shown in Fig. 7.

C. Testing the Statistical Significance of Experimental Results

For a more comprehensive evaluation of the proposed model, in addition to the mean AP evaluation protocol, we follow [51]–[53] to test the statistical significance of our experimental result through Fisher–Pitman permutation tests. Specifically, we apply the evaluation protocol of [54] to calculate the upper-tailed p-value of the AP results from the baseline (image classification using VGG16) and the proposed model. To test if we can reject a null hypothesis, p-value calculated using permutation tests is a suitable evaluation protocol [55].

A result has statistical significance when it has a low probability of occurring given the null hypothesis [56]. Specifically, we set the null hypothesis as that the proposed model does not bring an improvement on the performance. We then perform permutation tests on all the datasets used for both of the experimental settings 1 and 2. The results can be seen in Table X. As indicated by the results, the upper-tailed p-values from the listed datasets are close to 0. Also, all the upper-tailed p-values are smaller than 0.05, which [55], indicates we can reject the null hypothesis with statistical significance. This validates the research hypothesis that the proposed model is able to improve the performance.

V. CONCLUSION

This paper proposed a novel CNN model abbreviated as multibranch attention networks for action recognition in still images. This model incorporates a soft attention mechanism into a CNN model to explicitly exploit scene-level context and region-level context. The two context branches and target person region classifications are integrated for the final prediction. A two-step training strategy was proposed based on alternating optimization. Comprehensive experiments have been conducted for comparisons on both experimental settings with and without the bounding boxes of the target person, with results on the PASCAL VOC action dataset, the Stanford 40 dataset and HICO dataset verifying the advantages of the proposed model.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 3169–3176.
[3] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 50–65.
[4] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 9–16.
[5] C. Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.
[6] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 17–24.
[7] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2241–2248.

[8] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1365–1372.
[9] A. Goujon, A. Didierjean, and E. Marmeche, "Semantic contextual cuing and visual attention," *J. Exp. Psychol. Human Percept. Perform.*, vol. 35, no. 1, pp. 50–71, 2009.
[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[11] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.
[12] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Int. Conf. Learn. Represent. (ICLR) Workshop*, 2016.
[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[14] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," *arXiv preprint arXiv:1501.05964*, 2015.
[15] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2470–2478.
[16] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 3177–3184.
[17] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2010, pp. 9–16.
[18] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1080–1088.
[19] 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. 21st Brit. Mach. Vis. Conf. (BMVC)*, 2010, pp. 1–11.
[20] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.
[21] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[22] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.
[23] L. Yao *et al.*, "Video description generation incorporating spatio-temporal features and a soft-attention mechanism," *arXiv preprint arXiv:1502.08029*, 2015.
[24] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*. Heidelberg, Germany: Springer, 2008, pp. 66–75.
[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[26] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[28] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.
[29] E. W. Teh, M. Rochan and Y. Wang, "Attention networks for weakly supervised object localization," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2016, pp. 52.1–52.11.
[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.
[32] B. Yao *et al.*, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 1331–1338.
[33] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[34] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[35] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Proc. AFSS Int. Conf. Fuzzy Syst.*, 2002, pp. 288–300.
[36] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1717–1724.
[38] M. Hoai, "Regularized max pooling for image categorization," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014.
[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
[40] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[41] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 3360–3367.
[42] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 652–659.
[43] A. Diba, A. M. Pazandeh, H. Pirsiavash, and L. Van Gool, "DeepCAMP: Deep convolutional action & attribute mid-level patterns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3557–3565.
[44] F. S. Khan *et al.*, "Recognizing actions through action-specific person detection," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4422–4432, Nov. 2015.
[45] Z. Zhao, H. Ma, and X. Chen, "Semantic parts based top-down pyramid for action recognition," *Pattern Recognit. Lett.*, vol. 84, pp. 134–141, Dec. 2016.
[46] S. Yan, J. S. Smith, and B. Zhang, "Action recognition from still images based on deep VLAD spatial pyramids," *Signal Process. Image Commun.*, vol. 54, pp. 118–129, May 2017.
[47] Y. Zhang *et al.*, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5479–5490, Nov. 2016.
[48] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1017–1025.
[49] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[50] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 414–428.
[51] G.-P. Ge and C.-Y. Li, *Statistics for Experimenters*. New York, NY, USA: Wiley, 1978.
[52] P. R. Cohen, *Empirical Methods for Artificial Intelligence*, vol. 139. Cambridge, MA, USA: MIT Press, 1995.
[53] R. A. Fisher, *The Design of Experiments*. Edinburgh, U.K.: Oliver and Boyd, 1937.
[54] J. Kaiser, "An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples," *Stata J.*, vol. 7, no. 3, pp. 402–412, 2007.
[55] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proc. 16th ACM Conf. Inf. Knowl. Manag.*, 2007, pp. 623–632.
[56] R. R. Wilcox, *Statistics for the Social Sciences*. San Diego, CA, USA: Academic Press, 1996.

Shiyang Yan, photograph and biography not available at the time of publication.

Jeremy S. Smith, photograph and biography not available at the time of publication.

Wenjin Lu, photograph and biography not available at the time of publication.

Bailing Zhang, photograph and biography not available at the time of publication.