Multimedia Tools and Applications (2021) 80:32567–32593
https://doi.org/10.1007/s11042-021-11215-1

Action recognition in still images using a multi-attention guided network with weakly supervised saliency detection

Seyed Sajad Ashrafi [1] · Shahriar B. Shokouhi [1] · Ahmad Ayatollahi [1]

Received: 24 August 2020 / Revised: 21 January 2021 / Accepted: 5 July 2021 /
Published online: 31 July 2021

Abstract
Action recognition in still images is an interesting subject in computer vision. One of the most important problems in still image-based action recognition is the lack of temporal information; At the same time, other existing problems such as cluttered backgrounds and diverse objects make the recognition task more challenging. However, there may be several salient regions in each action image, employing of which could lead to an improvement in the recognition performance. Moreover, since no unique and clear definition exists for detecting these salient regions in action recognition images, therefore, obtaining reliable ground truth salient regions is a highly challenging task. This paper presents a multi-attention guided network with weakly-supervised multiple salient regions detection for action recognition. A teacher-student structure is used to guide the attention of the student model into the salient regions. The teacher network with Salient Region Proposal (SRP) module generates weakly-supervised data for the student network in the training phase. The student network, with Multi-ATtention (MAT) module, proposes multiple salient regions and predicts the actions based on the found information in the evaluation phase. The proposed method obtains mean Average Precision (mAP) value of 94.2% and 93.80% on Stanford-40 Actions and PASCAL VOC2012 datasets, respectively. The experimental results, based on the ResNet-50 architecture, show the superiority of the proposed method compared to the existing ones on Stanford-40 and VOC2012 datasets. Also, we have made a major modification to the BU101 dataset which is now publicly available. The proposed method achieves mAP value of 90.16% on the new BU101 dataset.

Keywords Still image-based action recognition · Convolutional neural network · Multi-attention · Teacher-student network

✉ Shahriar B. Shokouhi
bshokouhi@iust.ac.ir

Extended author information available on the last page of the article

1 Introduction

Nowadays, computer vision is applied in many different areas [1–5]. Human action recognition is an active research field in computer vision, the purpose of which is to determine an action which is done by a human [1, 6]. Human action recognition can be divided into two categories: video and still image-based methods. The main difference between these two categories is the employment of temporal information which plays an important role in action recognition [7, 8]. However, in still image-based action recognition, there are other useful cues such as human pose, human-object interaction, salient regions, backgrounds, and objects that are usually used instead of temporal information. Several applications are known for still image-based action recognition such as image labeling [9], image search [10], and abnormal behavior recognition [11], some examples of which are shown in Fig. 1.

Human pose and action-related objects contain useful information that can be utilized in action recognition [14–17], but requires complex algorithms such as pose estimation [12] and object detection [13]. For instance, in Fig. 1, the importance of human pose and action-related objects in action recognition is shown in Fig. 1(c) and (f), and in Fig. 1(a) and (b), respectively. In addition to objects and human pose, salient regions are also useful cues in images that refer to the most important areas for action recognition. These regions in an image may include action-related objects, human body parts especially hands and legs, and human-object interaction. Examples of salient regions in action recognition images can be found in Fig. 1 (these regions are predicted via the proposed method described in section 3).

Based on the above-mentioned cues, different methods were proposed in the literature [14–27], some of which are based on "auxiliary loss" modules that have additional loss

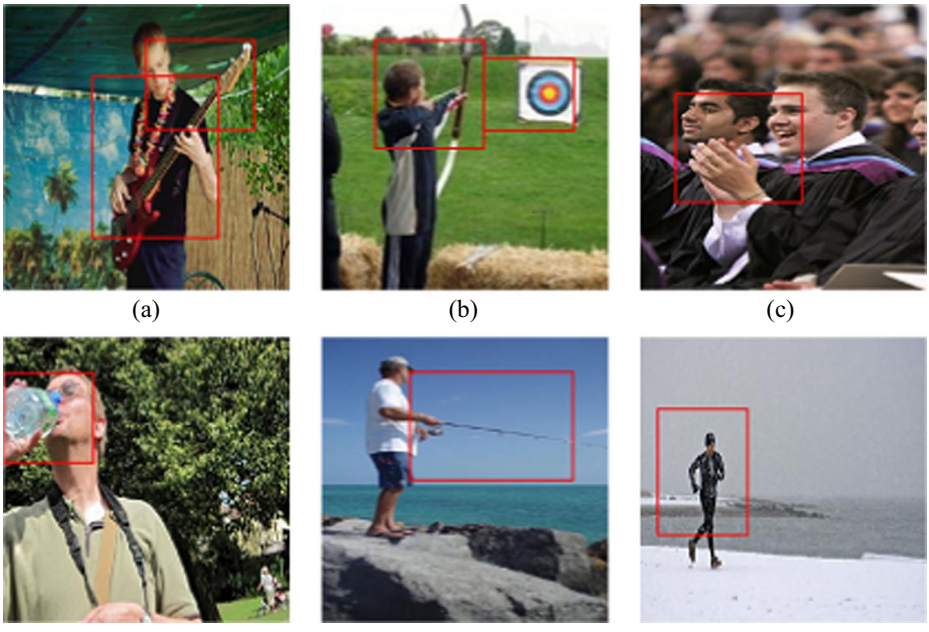

Fig. 1 A few examples of still image-based human action recognition with predicted salient regions; (**a**) playing the guitar, (**b**) shooting an arrow, (**c**) applauding, (**d**) drinking, (**e**) fishing, and (**f**) running

functions to detect a salient region such as human-related ones [14–26]. Although these methods have a significant effect on the network training, they are only focused on a single definite salient region. However, in action images, as can be seen in Fig. 1, various salient regions may exist. On the other hand, several methods were introduced based on using additional information such as region proposals [27], human body parts [23], and human bounding boxes [16] which are generally obtained by "auxiliary feature" modules such as RCNN [32], and OpenPose [33]. Making a decision with the help of auxiliary feature modules is considered as an effective method; nevertheless, it could require additional information and algorithms in the training and evaluation phases which may limit its applications and be computationally inefficient. Notwithstanding the advantages and disadvantages of the above-mentioned methods, in this paper, we propose a method based on auxiliary loss modules to detect multiple salient regions. The incentive for using the proposed module is to be able to make a decision based on various salient regions.

Multi-attention in fine-grained classification is an effective method that enables the network to find salient and discriminant regions in the subject [34–37]. In action images, there may be different salient regions that correspond not only to the human regions but also to the objects and backgrounds. It would be difficult for a network to detect all the various salient regions throughout an image and, so, it must be guided. Therefore, guiding the attentions of networks to detect best salient regions could be essential in still image-based action recognition. Hence, a weakly supervised teacher-student framework is proposed for guiding the student network to detect these salient regions, because no ground truth labels are at hand for them. The motivation for designing this teacher-student network is to guide the student network to find salient areas related to human action more effectively. Subsequently, the best salient regions are extracted from the output of the stronger network (teacher network).

In this paper, we propose a novel multi-attention guided method to detect multiple salient regions in still image-based action recognition. The block diagram of the proposed method is shown in Fig. 2. This approach has a teacher-student structure (Fig. 2(a)) where the teacher model is integrated with Salient Region Proposal (SRP) and provides the possible salient regions in each image to the student network in the training phase. The student model and the Multi-ATtention (MAT) module are then trained by these detected salient regions and output multiple feature vectors.

This research provides the following new contributions:

- A multi-attention guided network is proposed for action recognition in still images. The proposed method is based on recognizing various action-related regions of the image.
- A multi-attention (MAT) module is proposed to detect multiple salient regions based on the output feature map of the student network.
- A weakly-supervised method based on a teacher-student structure is proposed for action recognition in still images in which the teacher network provides salient regions to the student network. In the evaluation phase, only the student network is used for making predictions.
- A Salient Regions Proposal (SRP) module is designed for salient regions extraction from the output feature map of the teacher network. Depending on the input image, the SRP module may offer one or more salient regions.
- A new dataset called BU101PLUS is provided. It is a modified version of BU101 dataset [38]. Removing duplicates and adding new images are some of the modifications that have been made.

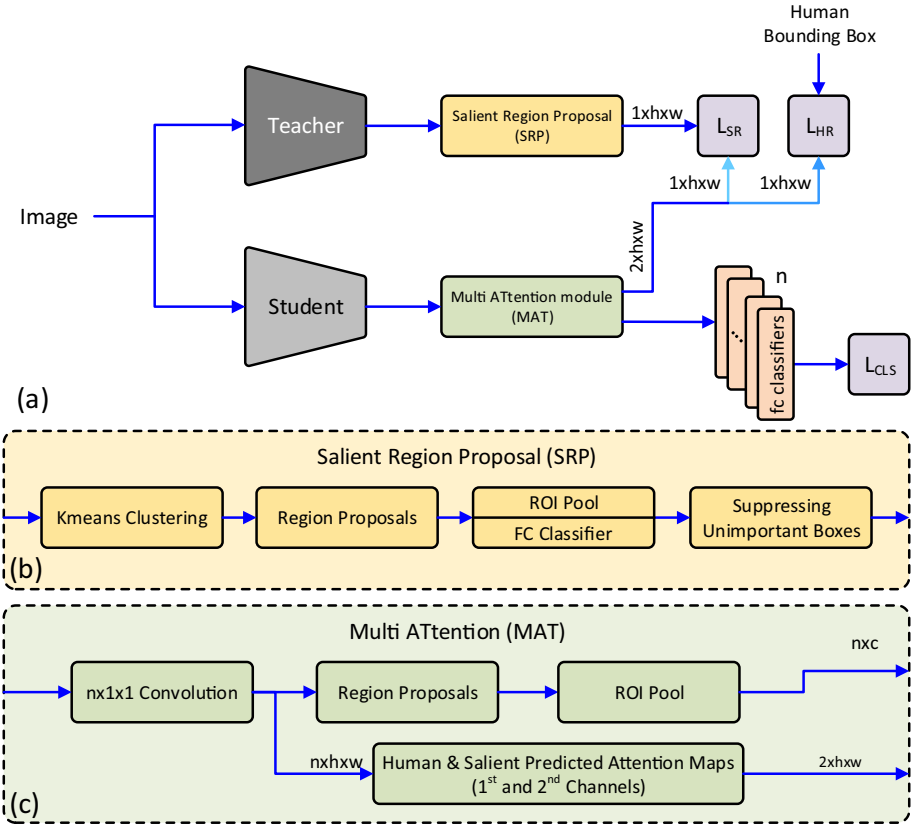

Fig. 2 An overview of the proposed method; (**a**) The block diagram of the multi-attention guided method with the teacher-student structure in the training phase. (**b**) Salient Region Proposal (SRP) module produces salient regions ground truth for each image. (**c**) Multi-ATtention (MAT) module generates the multi-attention map

The rest of the paper is organized as follows. The related works are described in section 2. The proposed method is explained in section 3 by details. In section 4 the experimental results are provided. Finally, the conclusion is presented in section 5 .

2 Related works

Still image-based action recognition methods can be divided into two subcategories: auxiliary loss and auxiliary feature methods. Each of these categories has its own advantages and disadvantages. In the following sections, the previous methods proposed in these categories are discussed. In addition, multi-attention and video-based action recognition are briefly described.

2.1 Auxiliary loss methods

Auxiliary loss-based approaches, help the network in the training phase to detect salient region [25]. The methods in this category, usually do not use any auxiliary networks such as RCNN

[32, 39, 40], and OpenPose [33]. Qi et al. [24] proposed a method based on the CNN and human pose in which an auxiliary loss function based on the human pose feature guides the network to pay more attention to the human pose. Liu et al. [25] proposed loss guided method to guide the network to detect the human-related regions. They designed a separate branch with convolutional layers for guiding the network's attention to these regions. In a recent study, Xin et al. [26] proposed a method based on auxiliary loss methods. They have introduced an entanglement loss function with the goal of reducing the within-class and increasing the between-class distances.

The main focus in all of the above-mentioned works is on the human-related regions for which auxiliary loss functions were designed and human-related regions were considered as salient regions. Although human areas are important regions to consider in action recognition, salient regions do not always fall within the human regions. The main disadvantage of auxiliary loss methods is that other action-related areas in images are disregarded and recognition is based on only human-related regions. The proposed method in this paper introduces a multi-attention technique that is not limited to human regions and can also detect different salient regions in action images as well.

2.2 Auxiliary feature methods

In the auxiliary feature-based methods, some external networks and algorithms are introduced to find human-related regions, pose, keypoints, and objects in the training and evaluation phases. Yan et al. [16] proposed a multi-branch network in which the introduced branches extract features from the whole input image, human-related regions, and the proposed regions using Region Proposal Network (RPN) [40]. In this work, RPN is an auxiliary feature module that is used in both the training and evaluation phases. Zhao et al. [23] presented a method in which the human body is divided into seven parts including head, arms, torso, hands, and lower body. This method includes two sub-networks: one for body part detection (auxiliary feature module) and the other for action prediction. Recently, Yang et al. [15] has recently proposed a method based on human pose and keypoints. This method consists of two sub-networks, fully convolutional Body Structure Exploration (BSE) and Action Classification (AC) sub-network. Similar to the other works in this category, the two sub-networks in this work are acting as auxiliary feature modules as well.

As was mentioned, in the auxiliary loss category, only the human-related areas were considered as salient regions. However, in this category, attention is paid to various regions, such as the RPN proposed regions, human pose regions, and proposed body parts. Although this could be considered an advantage for the methods in this category, the computational complexity and cost associated with these approaches are high due to the use of multi-branch networks and auxiliary sub-networks. Moreover, these auxiliary modules are also used in the evaluation process. The proposed method in this paper does not require any auxiliary network or manual ground-truth to detect multiple salient regions; it uses a teacher-student structure for this purpose.

2.3 Multi-attention approaches

Various multi-attention approaches have been proposed based on fine-grained classification and we are going to introduce some of them [34–37]. Hu et al. [34] employed one 1×1 convolutional layer after the backbone of the architecture to generate attention maps. In

another work on multi-attention approaches, Yao et al. [35] assigned the output feature maps of the network to three clusters based on the position of the maximum point in each channel. ROI pooling and classifiers are applied to the grouped features.

2.4 Video-based action recognition

Auxiliary loss [41] and auxiliary feature [26] methods have also been applied in video-based human action recognition works. Tian et al. [41] introduced a method based on the auxiliary loss function; which has a multi-task framework with an attention module. Chen et al. [26] provided a multi-CNN model for detecting human pose and motion in different frames, the outputs of which are combined using the score fusion module.

In this section, the previously proposed methods of still image-based action recognition for two categories of auxiliary loss and feature-based approaches are discussed with their advantages and disadvantages. Not needing auxiliary algorithms and information in the evaluation phase and proposing multiple regions for each image are the most important benefits of auxiliary loss and feature-based approaches, respectively. Moreover, some other research fields, such as person re-identification [28, 29] and fine-grained classification [30, 31] with multi-attention ideas were explained. In addition to the methods based on deep learning, there are methods based on hand-crafted algorithms [42–44]. To the best of our knowledge, the multi-attention technique has not been used before in still image-based action recognition.

3 The approach

The overall structure of the proposed multi-attention guided method is illustrated in Fig. 2. The output feature maps of the teacher network are passed through Salient Region Proposal (SRP) module to produce salient regions of each image which are then used in the training phase of the student model and Multi-ATtention (MAT) module. The MAT module takes the output feature map of the student model as an input and will then produce n attention and feature maps. Two of these n attention maps are then used to guide the detection of salient and human-related regions using L_{SR} and L_{HR} loss functions. Also, each of these n attention maps is then converted to n feature vectors using ROI pooling and fully connected classifiers. The proposed method consists of five main components: the teacher network (section 3.1), SRP module (section 3.2), the student network (section 3.3), MAT module (section 3.4), and loss functions (section 3.5). These components have been described below.

3.1 The teacher network

The teacher network without any additional module is trained on still image-based action recognition datasets. Figure 3(a) shows the training structure of the teacher network without using any auxiliary module. This model is based on the ResNet network [45]. After the training process, the average pooling layer is removed and the trained backbone along with the fully connected classifier is utilized in the evaluation mode (Fig. 3(b)). The trained fully connected classifier is also used in the SRP module (Section 3.2). In this case, the output of the teacher backbone network $f_t(.)$ is a feature map corresponding to each input image ($x \in \mathbb{R}^{3 \times H \times W}$):

$$y_t = f_t(x); y_t \epsilon \mathbb{R}^{c_t \times h \times w}. \quad (1)$$

In Eq. 1, w and h are $\lfloor \frac{W}{32} \rfloor$, and $\lfloor \frac{H}{32} \rfloor$, respectively. Zagoruyko et al. [46] explained that the sum of absolute values across the channel dimension of feature maps show the attention maps of the CNN model:

$$s_t = \sum_{i=0}^{c_t - 1} |y_t[i, \ldots]|; s_t \epsilon \mathbb{R}^{h \times w}. \quad (2)$$

where c is the number of channels and y_t is the output feature map of dimension $c_t \times h \times w$. Figure 3(c) and (d) show an example image and its attention map (s_t), respectively. The attention map shows the focused attention of the network on the input image. Figure 3(e) shows a few channels of the output feature map of the teacher network. As can be seen, there are various salient regions at different locations. Whereas, in Fig. 3(d) the response around the hands is much stronger than the other salient regions, e.g. blobs. Therefore, it is necessary to detect and extract different salient regions depicted in Fig. 3(e) and expose them to the student network. In this paper, the SRP module is proposed for multi salient regions extraction which is explained in the following section.

3.2 The salient region proposal (SRP) module

The SRP module is proposed to extract the salient regions from the output feature map of the teacher network. The block diagram of the SRP module is shown in Fig. 4. Every image

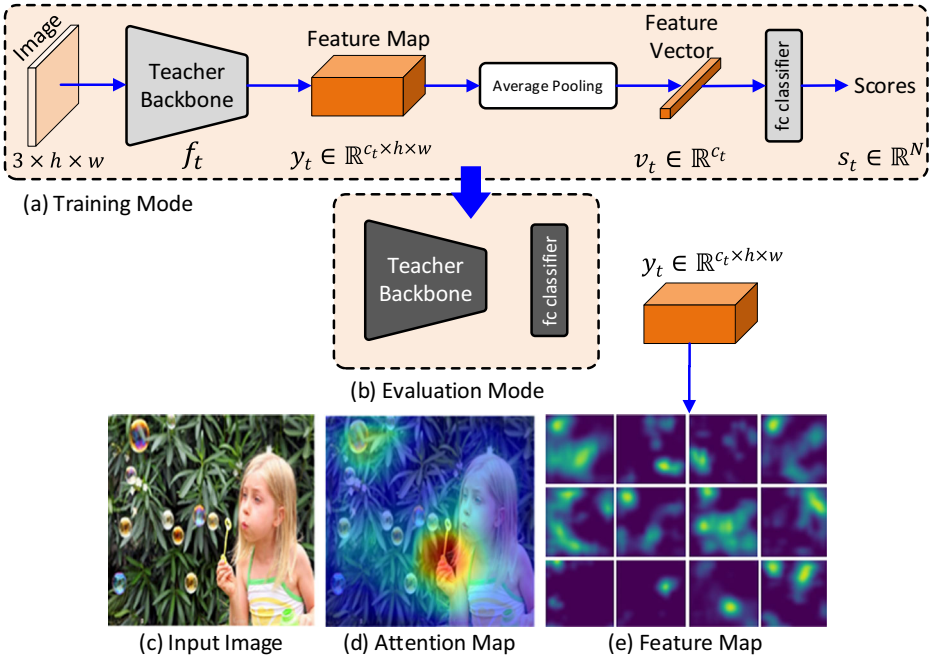

Fig. 3 The teacher network without any auxiliary module; (**a**) training mode, (**b**) evaluation mode. Both the trained teacher network and the fc classifier are used along with the multi-attention guided approach in a teacher-student structure in the proposed method

contains a different number of salient regions. The SRP module with the purpose of extracting discriminant and salient regions of each image consists of the following four steps:

1. Clustering the feature maps into k clusters based on each channel
2. Generating the bounding boxes based on the output of the clustering step
3. Applying the ROI pooling and the classifier to the feature maps based on the bounding boxes generated
4. Suppressing unimportant bounding boxes

Assume the output feature maps of the teacher network to be $y_t \epsilon \mathbb{R}^{c_t \times h \times w}$. First, the feature maps are converted to a 2D matrix $v \epsilon \mathbb{R}^{(h \times w)} \times c_t$ and then k-means clustering is applied to them. In previous works [35], clustering is performed based on the location of the maximum response in each channel of the feature map. In this paper, the clustering is done based on all points of each channel. The main reason is that the previous works [35] are usually focused on fine-grained tasks and where images only contain the subject. But in action images, there are cluttered backgrounds and different objects as well and the maximum point in each channel

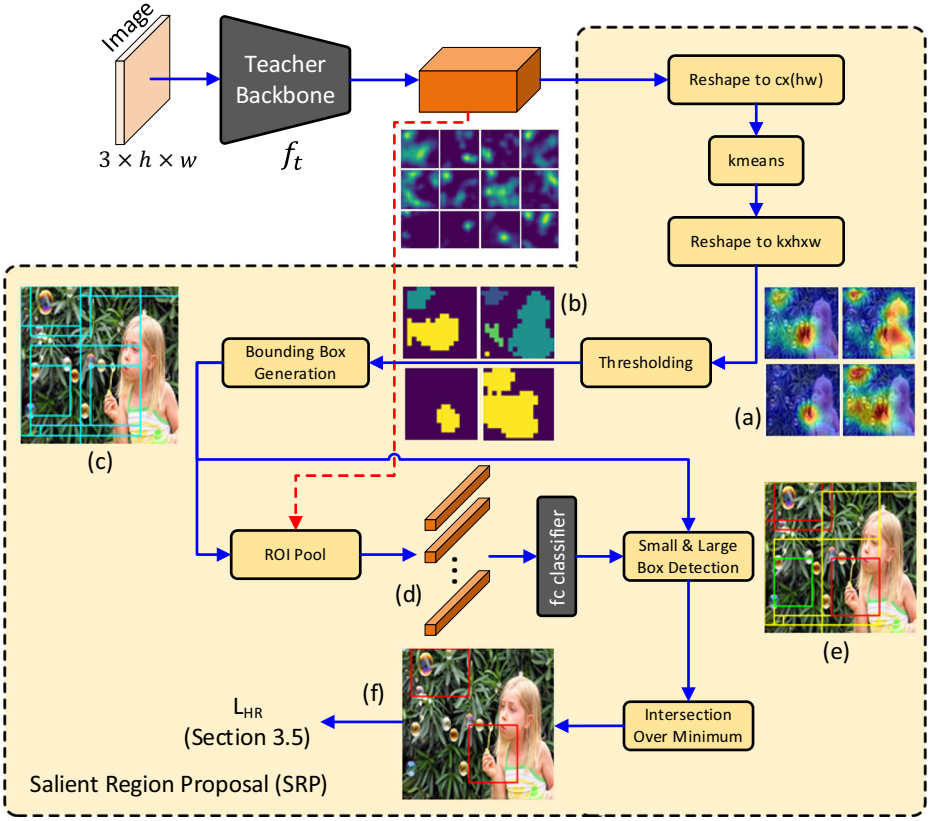

Fig. 4 Salient region extraction from the teacher model output via the SRP module; (**a**) center of clusters produced by k-means clustering $k = 4$, (**b**) obtaining binary images using thresholding ($\tau_l = 0.3$), (**c**) bounding boxes generated from binary images using morphology operations, (**d**) feature vectors obtained after the ROI pooling, (**e**) classified bounding boxes (green: false predictions; yellow: very small and large bounding boxes; red: true predictions), and (**f**) final bounding boxes remained after suppressing the unimportant ones

may not refer to the salient region. The output of the k-means clustering step is the center of k clusters that corresponds to k attention maps:

$$p_j = \frac{\sum_{i=0}^{c-1} \mathbf{1}\{l_i = j\} y_t[i, \ldots]}{\sum_{i=0}^{c-1} \mathbf{1}\{l_i = j\}}; p \in \mathbb{R}^{k \times h \times w}, j = 0 \ldots k-1. \quad (3)$$

In Eq. 3, l is the label of each channel predicted by k-means clustering. Figure 4(a), shows the centers of k clusters corresponding to the salient regions. It is clear that different salient regions are extracted from the teacher feature map.

Secondly, a set of bounding boxes are produced based on the output $p \in \mathbb{R}^{k \times h \times w}$. Each channel of p is normalized to [0, 1]. Then, using threshold $\tau_t = 0.3$ the salient regions are converted to binary images. After that, morphology operations (such as bwlabel) are applied to the binary images to find corresponding bounding boxes of each blob in the binary image. The output binary images and generated bounding boxes are depicted in Fig. 4(b) and (c), respectively.

Thirdly, ROI pooling and classification are performed for each generated bounding boxes. The output of this step is m feature vectors with dimension C_t shown in Fig. 4(d). ROI pooling is more efficient compared to average pooling when working with action images. This is because in the method proposed by Hu et al. [34], each of the generated salient regions is multiplied by feature maps to enhance salient regions and weaken the other regions, after which average pooling is applied to these enhanced saliency maps. However, in action images, salient regions constitute a small portion of the whole image and the proposed technique in [34], generates many regions that have weak responses close to zero. As a result, average pooling makes the feature vectors to be zero feature vectors.

In the final step of the SRP procedure, suppression of unimportant bounding boxes is applied and they include:

- False predictions
- Small and large size bounding boxes
- Highly overlapped boxes

Figure 4(e) shows the predictions made for each bounding box with a different color depending on its types. The green bounding boxes are related to the false predictions and must be removed. These predictions are detected by the a trained fc classifier. Small and large size bounding boxes are also removed (e.g. bounding boxes that are of a size greater than 10×10 and smaller than 3×3 in an attention map with a dimension of 14×14). This is done because large-size regions contain much of the image and small ones does not contain enough. The yellow boxes in Fig. 4(e) show the small and large bounding boxes that must be removed. Finally, overlapped bounding boxes are removed except for one of them. Intersection Over Minimum (IOM) is used here instead of Intersection Over Union (IOU). This is because the IOU does not output 1 for two nested boxes with different sizes. Whereas, in this study unlike object detection [13], one box among nested boxes must be preserved and this can be done using IOM. The remained bounding boxes are considered as salient regions for each image (Fig. 4(f)). Some of the action images and their corresponding salient regions extracted using the teacher network and the SRP module are shown in Fig. 5.

3.3 The student network

The backbone of the student network is a CNN architecture such as ResNet [45]. The average pooling and the classifier layers are removed from the student network. Therefore, the output feature map of the student network $f_s(.)$ with $x \in \mathbb{R}^{3\times H\times W}$ as the input image will be:

$$y_s = f_s(x; w_s); y_s \in \mathbb{R}^{C_s\times H\times W}. \tag{4}$$

As mentioned in the previous sections, multi-attention (MAT) module is proposed for the student network. The purpose of this module is to detect multiple salient regions in the image. In the following section, the details of the MAT module are explained.

3.4 The multi-ATtention (MAT) module

The MAT module is proposed for multi salient regions detection in action images. As can be seen in Fig. 2, The MAT module is placed right after the student network and consists of one conv1x1 layer with n filters, region proposals, and ROI pooling, the details of which are shown in Fig. 6. In the conv1x1 layer, the number of filters equals the number of desired attention

Fig. 5 Some sample images of detected salient regions using the SRP module; (**a**) input image, (**b**) image with the proposed bounding boxes (Red: Desired bounding boxes; Yellow: Large and Small bounding boxes; Green: False class predictions.), and (**c**) bounding boxes remained after suppressing the unimportant ones

maps, and its main role is to generate attention maps based on weighting the different channels of the feature map. So, the output of conv1x1 layer with $y_s \in \mathbb{R}^{C_s \times H \times W}$ feature map as its input will be:

$$s = \text{conv}_{1\times1}(y_s); s \in \mathbb{R}^{n\times h\times w}. \tag{5}$$

In Eq. 5, the output s is a set of proposed attention maps. This is similar to the teacher network, but in this case, salient regions must be extracted from each attention map as well. In Fig. 6, region proposals step includes the thresholding ($\tau_s = 0.3$) and the bounding box generation, and its output is shown in Fig. 6(c), After which each of the n channels is connected to a ROI pooling layer. As shown in Fig. 6(d), each channel may have more than one salient region. Based on the previous explanations, the existence of humans, objects, and cluttered backgrounds make the learning process difficult for the network. Therefore, it is necessary to guide the output attention maps of the conv1x1 layer so that more attention is paid to the human-related regions and the salient regions in the image.

One of the attention maps is used to concentrate on the human-related regions (Fig. 6(f)). Still image-based action recognition datasets [22, 47] usually contain human bounding boxes which are used here as ground truth. Also, one attention map is used to focus on salient regions (Fig. 6(g)), but since there is no clear definition for salient regions, it is difficult to obtain its

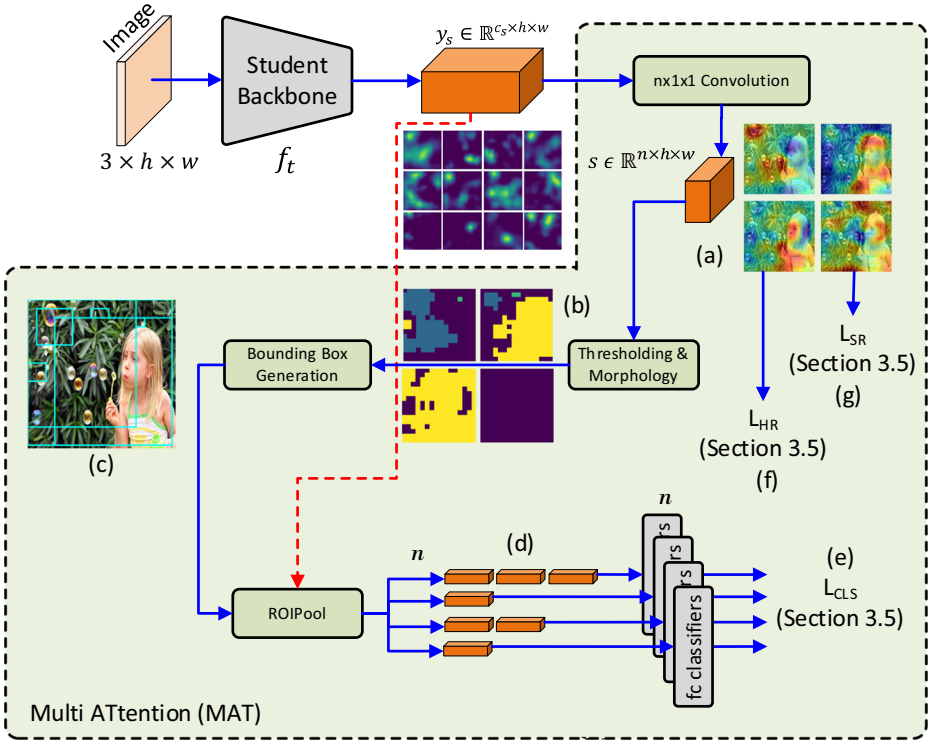

Fig. 6 The student model followed by the MAT module; (**a**) attention maps generated with $n = 4$, (**b**) binary images generated by thresholding ($\tau_s = 0.3$), (**c**) bounding boxes generated from binary images using morphology operations, (**d**) output feature vectors after the ROI pooling with $n = 4$, (**e**) classification loss (L_{CLS}), (**f**) human-related region loss (L_{HR}), (**g**) salient region loss (L_{SR})

ground truth. Therefore, the teacher network with the SRP module is used to generate weakly supervised salient region data for the student network. The other $(n-2)$ attention maps along with two other attention maps that were used in the guiding process are directly connected to the ROI pooling layer and the classifier. Therefore, only two of attention maps are being used in the guiding process. The purpose of leaving other $(n-2)$ attention maps is to let them extract general features and find possible regions that are not included in human-related and salient regions.

3.5 Loss functions

In the previous section, the two components of the human-related and salient areas were explained. A loss function is designed to guide each of these components. The proposed method has three loss functions including one for classification and the other two for guiding attention to the human-related and salient regions. Ground truths of these regions exist and loss functions are designed based on the Mean Absolute Error (MAE). The salient region loss (L_{SR}) is defined as:

$$L_{SR} = \|s_s - s_t\|_1, \tag{6}$$

where $s_s \in R^{h \times w}$ is the salient region predicted by the student network and s_t is the ground truth produced by the teacher network. If m salient regions are extracted by the teacher network, then $s_t \in R^{h \times w}$ will be a binary image obtained by the Eq. 7:

$$s_t(i,j) = \begin{cases} 1 & \text{if } (i,j) \text{ in } m \\ 0 & \text{o.w.} \end{cases}, \quad i = 0, 1, \ldots, h \ , \quad j = 0, 1, \ldots, w. \tag{7}$$

Figure 7(a) depicts an example of this generated binary image s_t. Similar to Eq. 6, the human-related region loss function (L_{HR}) is defined as:

$$L_{HR} = \|h_s - h_t\|_1, \tag{8}$$

where h_t is the already existing ground truth from datasets and h_s is the region predicted by the student network. $h_t \in R^{h \times w}$ is a binary image (1 stands for human-related regions and 0 for the others). Figure 7(b) shows an example of using human-related loss function. Finally, the cross-entropy loss (L_{CLS}) for the classification purpose is expressed as:

$$L_{CLS} = -\log\left(\frac{e^{y_t}}{\sum_c e^{y_c}}\right), \tag{9}$$

where y_t and y_c denote the label and the score of class c, respectively. The total loss considering n cross-entropy losses can be defined by the following Equation:

$$\begin{aligned} L_T &= \alpha L_{SR} + \beta L_{HR} + \frac{1}{n}\sum_{i=1}^{n} L_{CLS}^{(i)}, \\ &= \alpha \left\| s_s^{(1)} - s_t \right\|_1 + \beta \left\| h_s^{(2)} - h_t \right\|_1 - \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{y_t^{(i)}}}{\sum_c e^{y_c}}\right), \\ &= \alpha \left\| \mathrm{MAT}(f_s(x; w_s))^{(1)} - \mathrm{RSP}(f_t(x)) \right\|_1 + \beta \left\| \mathrm{MAT}(f_s(x; w_s))^{(2)} - h_t \right\|_1 - \frac{1}{s}\sum_{i=1}^{s} \log\left(\frac{e^{y_t^{(i)}}}{\sum_c e^{y_c}}\right). \end{aligned} \tag{10}$$

where α and β are hyper parameters.

4 Experiments

In this section, the evaluation results of the proposed multi-attention guided method are presented. First, the datasets are explained. Then, the experimental results on different datasets are discussed.

4.1 Datasets

The Stanford-40 action dataset [22] includes 40 daily actions, such as brushing teeth, cleaning floor, and reading book. There are 9352 images in this dataset, which are divided into two categories of training (4000 images) and testing (5352 images). Each class contains 180 to 300 images that 100 images for each class belong to the train set and the remaining are used as the test set. Also, the human bounding boxes for each image is available in the dataset.

The PASCAL VOC 2012 action dataset [47] includes 4588 images in 10 classes including jumping, phoning, playing music, riding a bike, riding a horse, running, taking photos, working with a computer, and walking. 2992 of these images are used for training and the remaining 1596 images are employed for testing. The dataset contains 140 to 220 images in each training and testing image class.

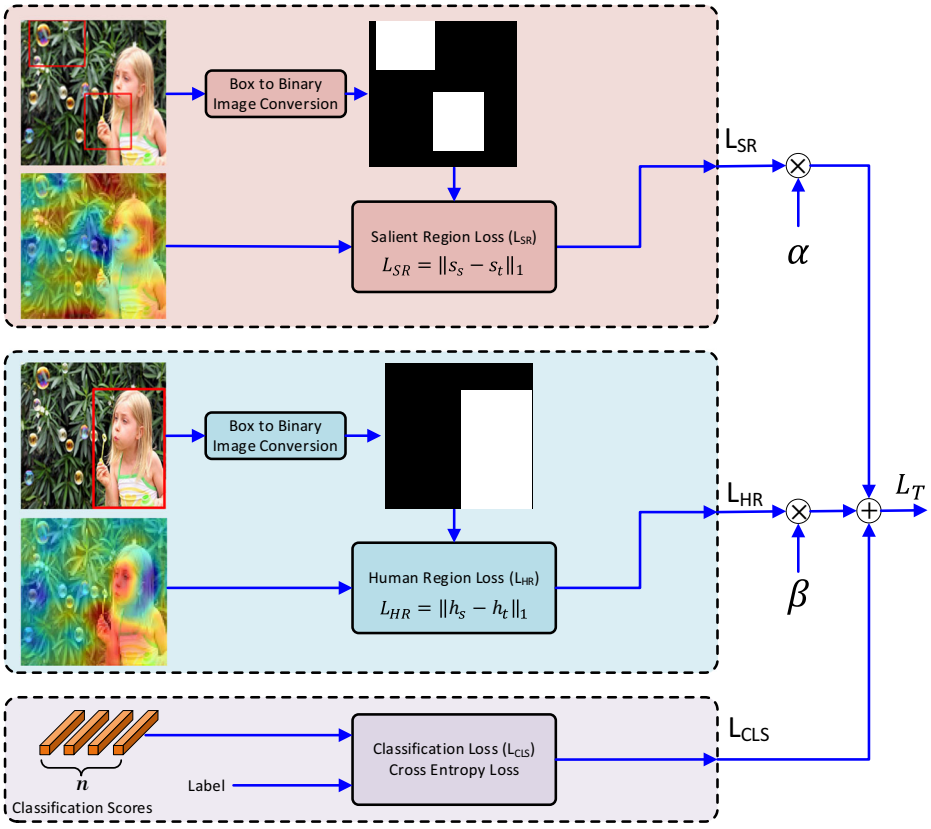

Fig. 7 The process of loss calculcation having three branches; (**a**) salient region loss, (**b**) human-related region loss, (**c**) cross-entropy losses

The BU Action dataset [38] was released by the Boston University in 2015. This dataset includes three separate datasets named BU101, BU101-unfiltered, and BU203-unfiltered. These datasets have been gathered using search engines such as Google, Flickr, etc. Currently, the BU Action is the largest dataset based on the number of images and classes that is being used in still image-based action recognition. We have used the BU101 dataset in this paper.

The BU101 dataset contains 101 image-based action classes. The 101 classes of this dataset are defined based on the UCF101 video dataset classes. This dataset contains about 23,800 labeled images, with an average of 235 images per class. The action classes in this dataset are divided into five categories: human-object interaction, body motion only, human-human interaction, music, and sport. Figure 8 shows some examples of images of each category. This dataset does not provide any additional information such as the human bounding box. The BU Actions dataset has not been used in previous studies done on still image-based action recognition.

The BU101 dataset has several problems, some of the most important of which are:

- There are many duplicate images in each class.
- There are many samples in each class showing different frames from a same video. The first row of Fig. 9 contains two examples of this problem.
- There are many images showing the same scene but taken at different angles. The second and third rows of Fig. 9 show a set of images taken at a same place but different angles.
- The number of images per class is unbalanced. Each class may have 100 to 600 images.
- Images of some classes contain only a very small part of a human. For example, most images in the eye makeup, lipstick, and typing classes include only the eye, mouth, and the hand area, respectively. The fourth row of Fig. 9 shows examples of such a problem.

Therefore, such a dataset with above-mentioned problems would not be suitable for the evaluation. These problems need to be addressed to make a valid experiment. In this paper, this dataset is modified to make standard evaluation possible. We named the modified dataset BU101PLUS. The modified dataset, which we call BU101PLUS, has the following characteristics:

- Each class in the BU101PLUS dataset has 100 images. These images are divided into two categories of 50 to be used as train and test sets.
- Images in each class are a combination of BU101 images and new images collected via Google image search.
- About 90 classes contain new images. The number of new images varies for different classes.
- The new images offer different challenges such as dealing with diverse backgrounds and unusual objects. Some examples of these new images are shown in the second and third rows of Fig. 10.
- Most of the images in the eye makeup, lipstick, blow drying hair, crawling, fencing, haircut, and typing classes are new. The first row in the Fig. 10 shows some of these mentioned classes.

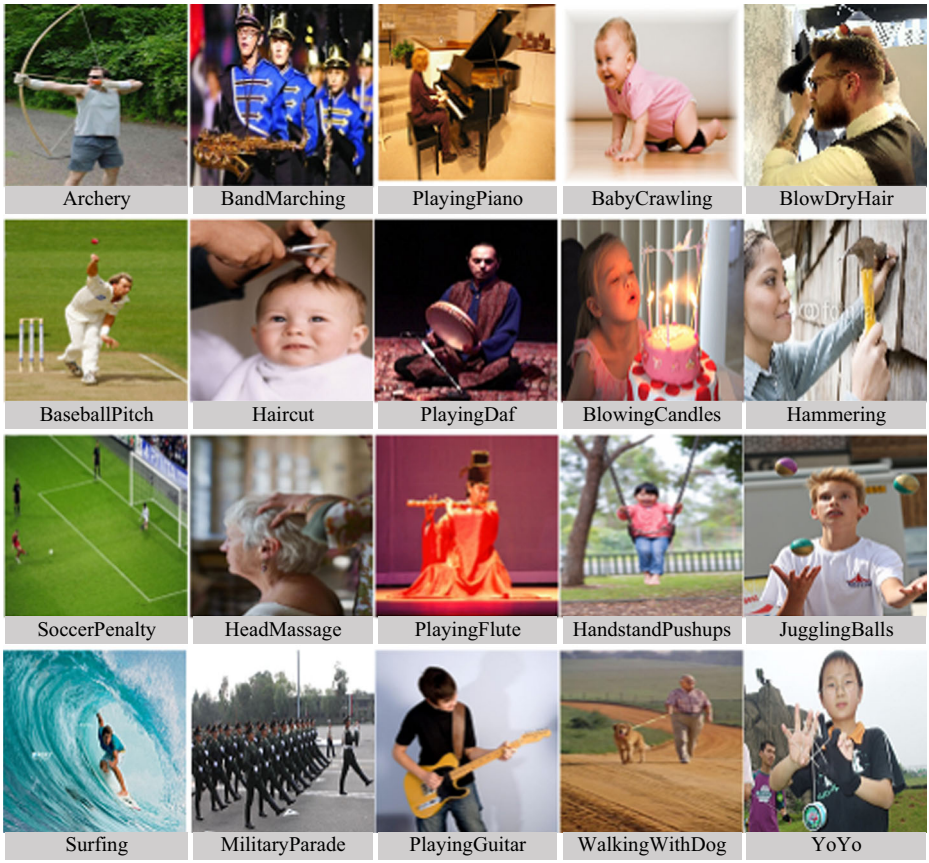

Fig. 8 Some examples of the BU101 dataset for five categories; (**a**) sports, (**b**) human-human interaction, (**c**) playing musical instruments, (**c**) body motion, and (**c**) human-object interaction

- The BU101PLUS dataset contains 10,100 images belonging to 101 classes and is publicly available.[1]

4.2 Training details

Data augmentation in most computer vision fields includes cropping, rotating, and horizontal flipping [1, 48]. In still image-based action recognition, humans play the main role. But, using conventional image cropping may remove (whole or an important part of) the human body from the input image whereas it is necessary to have at least a significant part of the human body. In this paper, the input images are cropped based on the human bounding boxes to have at least 50% of the human body in the cropped image. The data augmentation in this paper includes customized cropping, random horizontal flipping, and resizing into $3 \times 448 \times 448$ dimensions.

[1] https://github.com/seyedsajadashrafi/bu101plus-action-recognition-dataset

Fig. 9 Some examples of problems regarding the BU101 dataset

The proposed method in the training phase has a teacher-student structure. However, in the evaluation phase, only the student model is used. Both the teacher and the student models are implemented based on the ResNet network [45]. For an input image with a $3 \times 448 \times 448$ dimensions, the output feature maps of the teacher and the student models have a size of 14×14 with different channels.

Multi-attention guided method is implemented using PyTorch framework [49]. The proposed method is trained with a single Nvidia RTX2080 GPU. In the training phase, Stochastic Gradient Descent (SGD) optimization with a learning rate of 0.001, momentum of 0.9, and weight decay of 1e-5 is used. The learning rate starts at 0.001 and decays with a 0.9 factor every two epochs. The batch size, the number of attentions (A), and the values of α and β are set to 15, 4, 0.01, and 0.01, respectively.

In the evaluation phase, only the student network, the MAT module, and the classifiers are used. The output of the MAT includes n attention maps and their corresponding feature vectors. Each feature vector is fed to a separate classifier. Final prediction is achieved by averaging the outputs of the classifiers.

4.3 The Stanford-40 action dataset

Table 1, shows the evaluation results of the proposed method on Stanford-40 action dataset. ResNet-101 is used as the teacher model in this experiment.

Based on Table 1, the mAP value of the proposed method is at least 0.7% higher than the best previously works [15]. Moreover, the proposed method, unlike the others that are used for comparison, does not require any extra information and auxiliary algorithm such as human bounding boxes, RCNN, etc. in the evaluation phase.

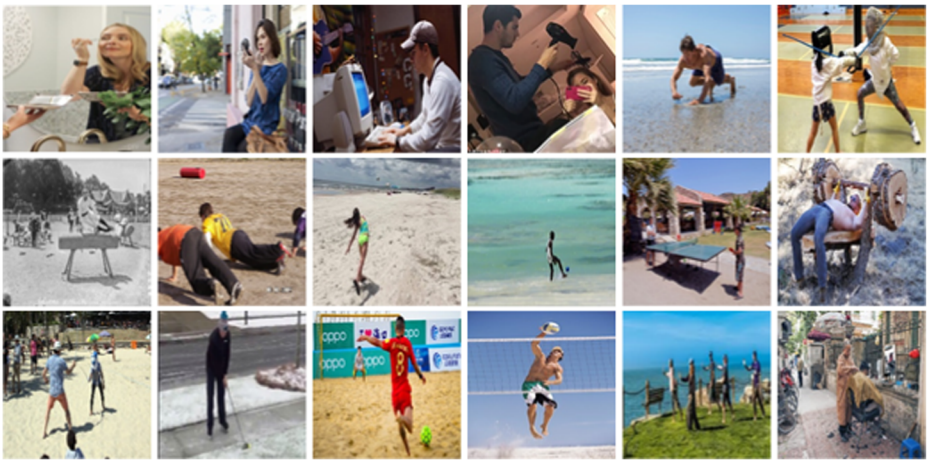

Fig. 10 Some examples of new images in the BU101PLUS

As discussed in this section, the number of attention maps is considered to be 4. Some action images with their 4 attention maps are shown in Fig. 11(a) to (e). Also, the attention map of the base ResNet-18 model is depicted in the Fig. 11(f). The base model is a ResNet model trained on Stanford-40 dataset without any additional module. Figure 11(b) represents the attention map that has been guided to detect human-related regions. For example, the human-branch detected two humans in the second row of Fig. 11. The attention map in Fig. 11(c) corresponds to the salient regions. The salient-branch of the student network learnes these salient regions given to it by the teacher network and the SRP module in the training phase. As can be seen from the last row of Fig. 11, unlike the base model, the proposed method provides correct predictions and also the camera region predicted by the proposed method is better than the base model. Attention maps of the 3rd and 4th columns of Fig. 11 have no guidance loss functions and are designed to detect any useful cues in the input image. In addition to the guided attention, multi-attention also has a positive impact on the recognition performance. In the 3rd row of Fig. 11, attention maps in columns 3 and 4 proposed different regions compared to columns 1 and 2. Attention map in Column 4 has detected the wall region which provides useful information for climbing action. In this example, in contrast to the base model, the multi-attention guided method has correctly predicted the action. A similar condition is also happening in the rowing a boat image. Multi-attention guided method has been able to detect the different and discriminant regions in the image. In this case, both the base model and the proposed method have recognized the correct class.

Table 1 Comparison of the proposed method with the previous works on the Stanford-40 dataset

Methods	mAP (%)
Hint Enhanced [24]	80.69
Loss Guided Activation [25]	91.10
Part Action Network [23]	91.20
Entanglement Loss [26]	92.20
Body Structure Cues [15]	93.80
Ours (ResNet-18)	88.36
Ours (ResNet-34)	92.45
Ours (ResNet-50)	**94.20**

The range of mAP values for the classes of the Stanford-40 dataset is wide. Based on the wide range of mAP values, classes can be divided into three categories:

1) **mAP values less than 85%:** This category includes waving hand, taking photos, texting messages and pouring liquid classes. The average mAP of these 4 classes is 79.63%. The low performance in this category is related to three challenges including the presence of small objects, lack of action-related objects, and confusion with other classes. For example, the hand waving class is mistaken for the applauding class. Moreover, this class does not have any action-related objects. The taking photos class is slightly mistaken for two other classes of looking at the telescope and the microscope. Also, texting messages class also contains small objects.
2) **mAP values between 85 and 90%:** This category includes applauding, brushing teeth, cutting vegetables, drinking, looking through a telescope, phoning, running, and washing the dishes classes. The average mAP of these 8 classes is 89.08%. This class has similar features to Category 1. There are no action-related objects in the applauding class which is similar to the hand waving class.
3) **mAP values greater than 90%:** This category often includes classes having multiple salient regions and large objects. The proposed method in this paper, detected salient regions in this category with a high precision. There are 28 classes in this category with an average mAP value of 97.75%.

4.4 The PASCAL VOC 2012 dataset

Table 2 compares the results obtained by the proposed method and the previous approaches on the VOC2012 dataset. The mAP value of the proposed method is reported using ResNet networks. The achieved mAP using ResNet-50 is higher than the previously done methods [15, 26].

The performance of the proposed method on the VOC2012 dataset is similar to the Stanford-40 dataset. The mAP value of the proposed method for classes containing salient objects (e.g., playing music, riding a bike, riding a horse, and working with computer) is much higher than the total mAP value. Running, taking photos, and phoning classes have a mAP of about 91% due to the presence of small objects and the lack of action-related objects. The lowest mAP value of 85% belongs to the walking class. This class is confused with the running class. The two classes of jumping and studying both have a mAP close to the total mAP value.

The Attention maps of the multi-attention guided method and the base model using ResNet-50 as the backbone are shown in Fig. 12. In the images of the first row, both the base and the multi-attention guided networks recognized jumping action correctly. In the base network, some parts of the background were detected as well. However, in the two guided branches of the proposed method, focus of the attention is on the whole human body rather than the background. In the images of the second to fifth rows, the base network made wrong predictions whereas the proposed method achieves correct predictions. The effect of the guiding branches on the proposed method can be clearly seen in the images of the third and fifth rows. In the base network, the salient regions in the third and fifth images are missed. Whereas in the proposed method, the guided branches focused the attention to the most important regions such as the paper and the keyboard. In the images of the fourth row, different regions including human, horse, and horse jumping obstacles have been extracted.

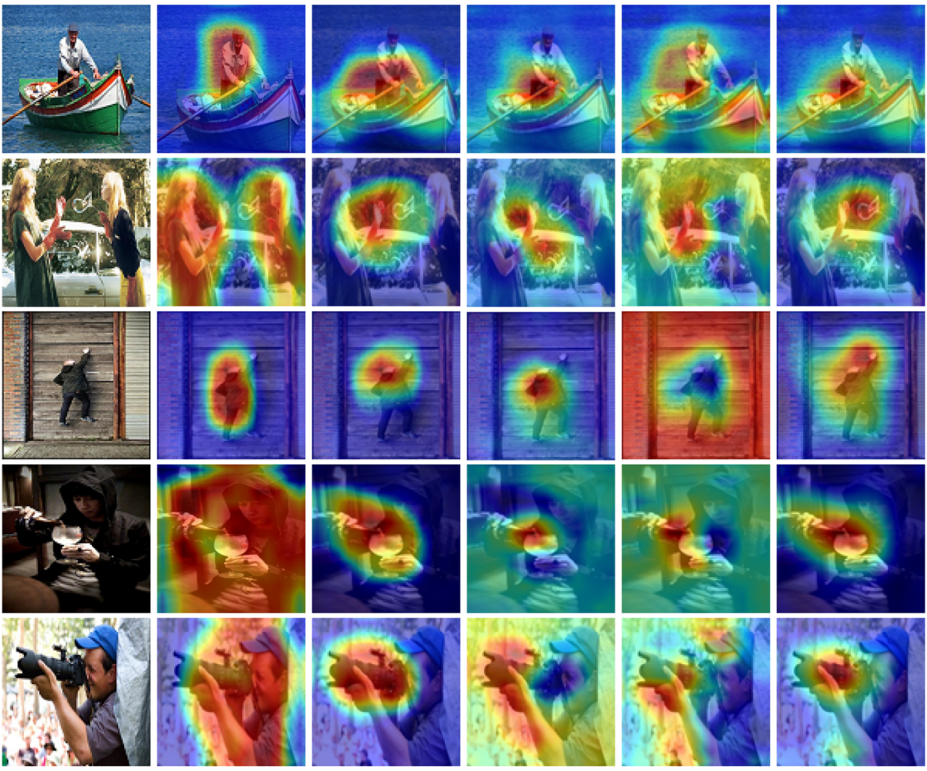

Fig. 11 The output attention map of the base model and the multi-attention guided method on Stanford-40 dataset; (**a**) input image, (**b**) the attention maps of human-branch guided loss, (**c**) the attention maps of the salient-branch guided loss, (**d**) first branch without any guided loss function, (**f**) second branch without any guided loss function, (**e**) attention maps of the base model

4.5 The BU action datasets

Table 3 evaluates the results of the proposed method and compares them with those of the base CNNs. Due to the lack of human bounding boxes in this dataset, the guided human-branch has been removed. The salient-branch focuses on the salient regions and uses the ResNet-101 as its teacher network. The mAP value for five categories of human-object interaction (HOI),

Table 2 Comparison of the proposed method with previous methods in PASCAL VOC 2012

Methods	mAP (%)
Multibranch Network [16]	84.50
R-CNN [17]	89.00
Hint Enhanced [24]	89.80
Entanglement Loss [26]	92.10
Body Structure Cues [15]	93.50
Ours (ResNet-18)	88.19
Ours (ResNet-34)	91.51
Ours (ResNet-50)	**93.80**

human-human interaction (HHI), body motion only (BMO), playing musical instrument (PMI), and sports (S) have been listed separately in Table 3.

As shown in Table 3, the proposed method has a higher mAP value than the base networks even without the human guidance branch. The results in Table 3 are reported in five separate groups, which are:

1) **Playing musical instruments:** This group consists of 10 classes. As shown in Fig. 8(c), this group contains large objects. The proposed method has achieved its best performance in this category.
2) **Sports group:** This group consists of 50 classes. The proposed method has achieved its second-best performance in this category after the music playing group. As shown in Fig. 8(a), the images in this category contain action-related backgrounds, salient objects, and human pose.
3) **Human-object interaction:** This category includes 18 classes and action-related objects are available in different sizes in this group images. The proposed network has attained a mAP value of 87.84% in this group.
4) **Human-human interaction:** This group includes 7 classes. The existence of similar classes such as head massage, haircut, eye makeup, and lipstick made the proposed method to have poor performance in this group.

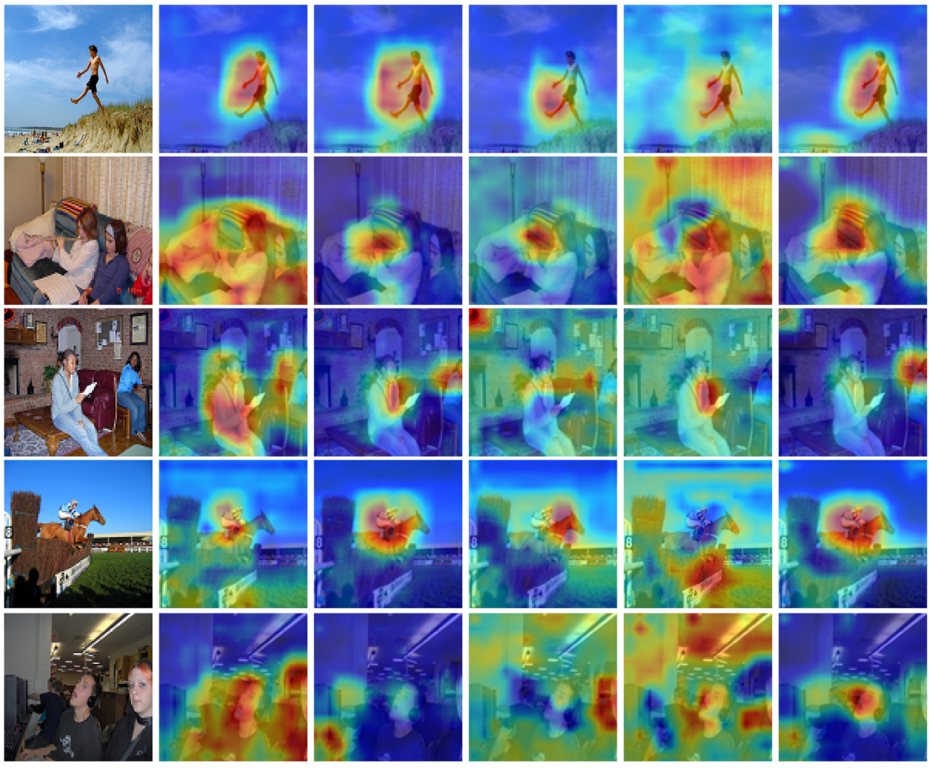

Fig. 12 The output attention map of the base model and the multi-attention guided method on PASCAL VOC 2012 dataset

5) **Body motion:** This group consists of 16 classes, which only include body movements, and several similar classes such as handstand pushups and handstand walking. Figure 13 Shows some examples of classes containing body movements-related images. The classes in each column are very similar and making predictions based on a single image would be difficult.

4.6 Ablation study

In this section, more experiments are done on the proposed method, all of which have been performed on the Stanford-40 action dataset.

4.6.1 Analysis of different modules

In this section, the effect of different proposed modules such as MAT, human-branch loss, and salient-branch loss, are discussed. Table 4 shows the results of various experiments performed to determine the effects of Such modules. Customized data augmentation (Section 4.1.1) increases the accuracy of the base model (ResNet-18) by 1.46%. By adding the MAT module and setting $n = 4$, the mAP value improves by 1.22%. Finally, by adding loss function L_T(L_{SR} and L_{HR} loss functions), the mAP value goes up significantly by 2.49%.

4.6.2 The number of attentions

The number of attentions (n) is one of the hyper-parameters of the proposed method. The experiments explained in the previous section were performed with $n = 4$ and two of the attention maps were used in the guiding process. In this experiment, the effect of the number of attentions is investigated on the ResNet-18 network by varying this number from 1 to 12. Figure 14 shows the effect of increasing the number of attentions. Also, it shows that adding both human-branch and salient-branch loss modules have positive effect on the network performance. The mAP value using two modules is better than the others. Also, by setting $n = 4$, the best mAP value is obtained and after that, the result does not change much. Also, this diagram shows that using the MAT module without any other loss module has a positive effect on the network performance.

Table 3 The mAP value comparison between the proposed method and the base networks on the BU101PLUS dataset

Methods	HOI	HHI	BMO	PMI	S	Total
ResNet-18	76.90	70.04	76.80	92.47	87.36	83.06
ResNet-34	82.69	77.47	81.99	95.74	89.98	87.07
ResNet-50	85.57	80.31	78.93	96.72	90.08	87.44
Ours (ResNet-18)	80.06	79.37	77.47	94.87	89.96	86.11
Ours (ResNet-34)	85.13	83.04	**84.56**	98.07	91.74	89.43
Ours (ResNet-50)	**87.84**	**84.98**	82.64	**98.10**	**92.60**	**90.16**

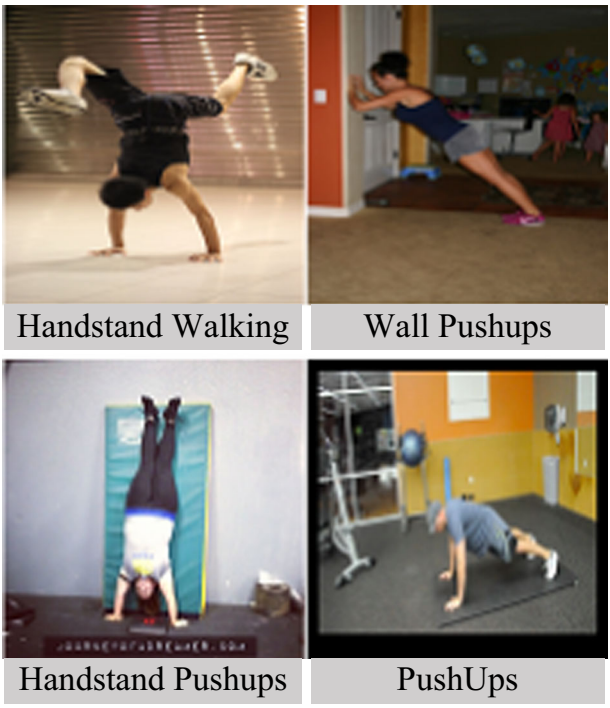

Fig. 13 Some examples of very similar classes in the body movements group

4.6.3 The SRP module

In this section, the effect of the SRP module on the performance is discussed. The SRP module is used to extract the salient regions from the teacher network. To evaluate the efficiency of the SRP module, an experiment is designed which can be seen in Fig. 15. A ResNet-18 network is provided with the SRP module and the fc classifier. In this architecture, the SRP module extracts a number of feature vectors from the output feature map of the ResNet-18 network. The number of feature vectors for each image may vary. Finally, each of the feature vectors is assigned to the fc classifier. The final prediction is the mean of all predictions.

The evaluation results in Table 5 show that the SRP module leads to an increase in the mAP value by about 2%. This increase happens because the SRP module proposes different attention maps.

Table 4 Analysis of the effect of different modules in the proposed method

Methods	mAP (%)
ResNet-18	83.23
ResNet-18 + Aug.	84.69
ResNet-18 + Aug. + MAT	85.87
ResNet-18 + Aug. + MAT+$\mathbf{L}_g$	88.36

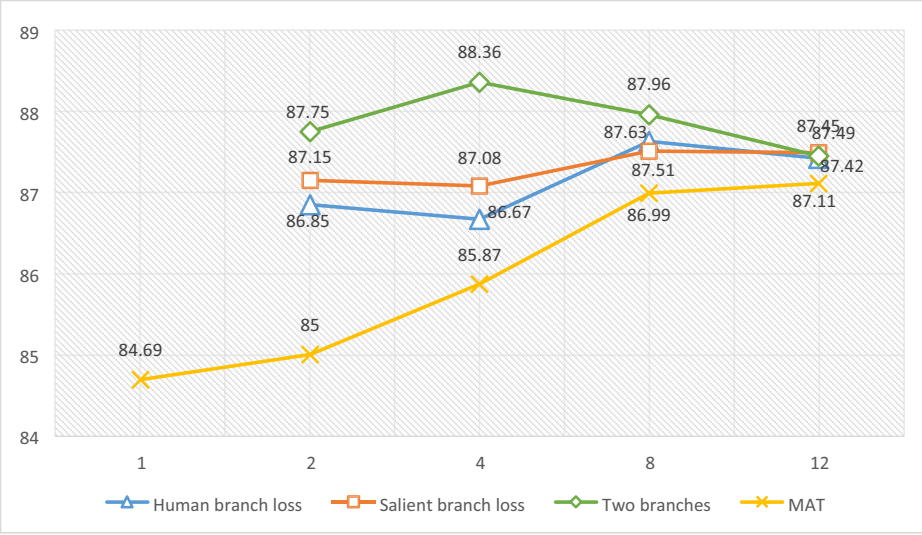

Fig. 14 Investigating the effect of the number of attention maps (n) and guiding branches on the proposed network's performance. The best mAP value is achieved by setting $n = 4$

4.6.4 The threshold parameters in the SRP and MAT modules

In the proposed method, two values of τ_s and τ_t are used in the MAT and SRP modules respectively to convert attention areas to binary images. In this section, the effect of these parameters (τ_s and τ_t) are investigated by an experiment. The experiment is performed on the validation data that contains 25% of training data. The evaluation results of changing the threshold value from 0.01 to 1 are illustrated in Fig. 16. To draw the results of changing τ_s, the value of τ_t is set to be 0.3 and vice versa. The evaluation results can be analyzed as follows:

- As the threshold decreases, the number of pixels in attention maps that have the value 1 increases. Therefore, when salient areas are obtained using the morphology step, they often get removed during suppressing unimportant bounding boxes due to their large sizes. As a result, the number of salient regions decreases causing a decrease in the performance.
- The number of pixels with label 1 diminishes with the increase of threshold value. In this case, the salient regions are reduced in size and some of them are removed in the suppressing unimportant step. Therefore, there is a decline in the network's performance.

These changes can be clearly seen for the τ_s parameter in the Fig. 16. This parameter plays a more important role on the performance compared to τ_t parameter because it helps in

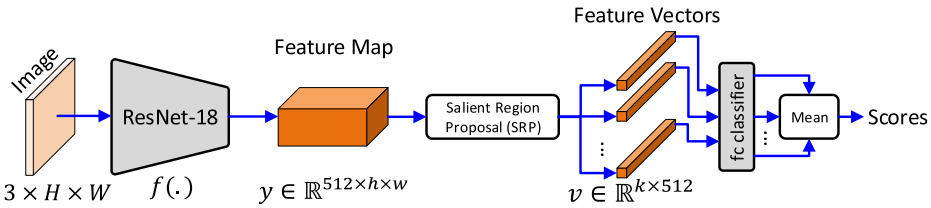

Fig. 15 The block diagram of the ResNet-18 along with SRP module

Table 5 Analysis of SRP module in mAP

Methods	mAP (%)
ResNet-18	84.69
ResNet-18+SRP	86.11

determining the salient regions before classification. The τ_t parameter has less effect (about 1.5%) on the overall performance of the network. This is because with τ_t close to 0 or 1, the salient region branch becomes ineffective and the teacher network has no effect on the student network in the training process anymore. But the τ_t diagram shows that the teacher network has a positive effect on the student network training in general.

5 Conclusion

In this paper, we proposed a multi-attention guided network to detect the salient regions and recognize the actions in still images. The three main contributions of this study are as follows: first, a multi-attention module has been proposed for making decisions on action images based on the attentions focused on the different areas of an image. Second, a more powerful network (i.e., the teacher network) has been employed for guiding the attentions to the salient and action-related regions. Third, a new dataset named BU101PLUS has been created based on the BU101 dataset.

The evaluation results on different datasets (e.g., the Stanford-40 and the PASCAL VOC2012 datasets) indicate that the proposed method achieves better results than the other relevant methods. Experiments show that the multi-attention module is a useful tool for action recognition and can be further developed in future works. Also, the new dataset (BU101PLUS) can be used as a standard dataset or even augmented and improved in future works.

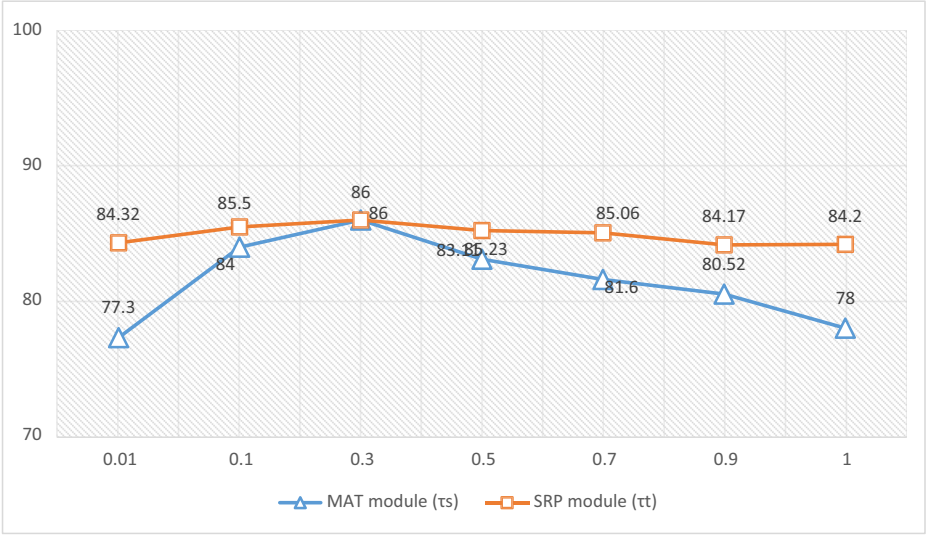

Fig. 16 Investigating the effects of threshold parameters (τ_s,τ_t) used in SRP and MAT modules

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aly S, Sayed A (2019) Human action recognition using bag of global and local Zernike moment features. Multimed Tools Appl 78(17):24923–24953. https://doi.org/10.1007/s11042-019-7674-5
2. Amirkhani D, Bastanfard A (2019) Inpainted image quality evaluation based on saliency map features. https://doi.org/10.1109/ICSPIS48872.2019.9066140
3. Beddiar DR, Nini B, Sabokrou M, Hadid A (2020) Vision-based human activity recognition: a survey. Multimed Tools Appl 79:1–47. https://doi.org/10.1007/s11042-020-09004-3
4. Bulbul MF, Islam S, Ali H (2019) 3D human action analysis and recognition through GLAC descriptor on 2D motion and static posture images. Multimed Tools Appl 78(15):21085–21111. https://doi.org/10.1007/s11042-019-7365-2
5. Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2018) OpenPose: Realtime multi-person 2D pose estimation using part affinity field. Accessed: Jun. 18, 2020. [online]. Available: http://arxiv.org/abs/1812.08008
6. Chen C, Jafari R, Kehtarnavaz N (2017) A survey of depth and inertial sensor fusion for human action recognition. Multimed Tools Appl 76(3):4405–4425. https://doi.org/10.1007/s11042-015-3177-1
7. Delaitre V, Laptev I, Sivic J (2010) Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: Procedings of the British machine vision conference 2010, pp 97.1–97.11. https://doi.org/10.5244/C.24.97
8. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4
9. Girshick R (2015) Fast R-CNN. Proc IEEE Int Conf Comput Vis 2015:1440–1448. https://doi.org/10.1109/ICCV.2015.169
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 580–587. https://doi.org/10.1109/CVPR.2014.81
11. Gkioxari G, Girshick R, Malik J (2015) Contextual action recognition with R∗CNN, Proc IEEE Int Conf Comput Vis vol. 2015 inter, pp. 1080–1088 https://doi.org/10.1109/ICCV.2015.129
12. Guo G, Lai A (2014) A survey on still image based human action recognition. Pattern Recogn 47(10):3343–3361. https://doi.org/10.1016/j.patcog.2014.04.018
13. Gupta A, Kembhavi A, Davis LS (2009) Observing human-object interactions: using spatial and functional compatibility for recognition. IEEE Trans Pattern Anal Mach Intell 31(10):1775–1789. https://doi.org/10.1109/TPAMI.2009.83
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 2016-December, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
15. Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: a survey. Image Vis Comput 60:4–21. https://doi.org/10.1016/j.imavis.2017.01.010
16. Hu T, Qi H, Huang Q, Lu Y (2019) See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. Accessed: Apr. 13, 2020. [online]. Available: http://arxiv.org/abs/1901.09891
17. Ikizler N, Cinbis RG, Pehlivan S, Duygulu P (2008) Recognizing actions from still images. https://doi.org/10.1109/icpr.2008.4761663
18. Li LJ, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. https://doi.org/10.1109/ICCV.2007.4408872
19. Li Z, Zheng Z, Lin F, Leung H, Li Q (2019) Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN. Multimed Tools Appl 78(14):19587–19601. https://doi.org/10.1007/s11042-019-7356-3
20. Li Y, Li K, Wang X (2020) Recognizing actions in images by fusing multiple body structure cues. Pattern Recogn 104:107341. https://doi.org/10.1016/j.patcog.2020.107341

21. Liao X, Li K, Zhu X, Liu KJR (2020) Robust detection of image operator chain with two-stream convolutional neural network. IEEE J Sel Top Signal Process 14(5):955–968. https://doi.org/10.1109/JSTSP.2020.3002391
22. Liu L, Tan RT, You S (2019) Loss guided activation for action recognition in still images. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 11365 LNCS, pp 152–167. https://doi.org/10.1007/978-3-030-20873-8_10
23. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2020) See more, know more: unsupervised video object segmentation with co-attention Siamese networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 3618–3627. Accessed: Dec. 22, 2020. [Online]. Available: http://arxiv.org/abs/2001.06810
24. Ludl D, Gulde T, Curio C (2019) Simple yet efficient real-time pose-based action recognition. In: 2019 IEEE intelligent transportation systems conference, ITSC 2019, pp 581–588. https://doi.org/10.1109/ITSC.2019.8917128
25. Ma S, Bargal SA, Zhang J, Sigal L, Sclaroff S (2017) Do less and achieve more: training CNNs for action recognition utilizing action images from the web. Pattern Recogn 68:334–345. https://doi.org/10.1016/j.patcog.2017.01.027
26. McAuley J, Leskovec J (2012) Image labeling on a network: using social-network metadata for image classification. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 7575 LNCS, no. PART 4, pp 828–841. https://doi.org/10.1007/978-3-642-33765-9_59
27. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 9912 LNCS, pp 483–499. https://doi.org/10.1007/978-3-319-46484-8_29
28. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognitiona review. IEEE Trans Syst Man Cybern Part C Appl Rev 42(6):865–878. https://doi.org/10.1109/TSMCC.2011.2178594
29. PyTorch. (2016) https://pytorch.org/ (accessed September 1, 2016).
30. Qi T, Xu Y, Quan Y, Wang Y, Ling H (2017) Image-based action recognition using hint-enhanced deep neural networks. Neurocomputing 267:475–488. https://doi.org/10.1016/j.neucom.2017.06.041
31. Raja K, Laptev I, Pérez P, Oisel L (2011) Joint pose estimation and action recognition in image graphs. In: Proceedings - international conference on image processing, ICIP, pp 25–28. https://doi.org/10.1109/ICIP.2011.6116197
32. Redmon J, Divvala S, Girshick R, Farhadi A (2015) You only look once: unified, real-time object detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit vol. 2016-December, pp. 779–788. Accessed: Apr. 12, 2020. [Online]. Available: http://arxiv.org/abs/1506.02640
33. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
34. Sadeghi H, Raie AA (2019) Histogram distance metric learning for facial expression recognition. J Vis Commun Image Represent 62:152–165. https://doi.org/10.1016/j.jvcir.2019.05.004
35. Szegedy C et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 07–12-June-2015, pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
36. Thurau C, Hlaváč V (2008) Pose primitive based human action recognition in videos or still images. https://doi.org/10.1109/CVPR.2008.4587721
37. Tian D, Lu ZM, Chen X, Ma LH (2020) An attentional spatial temporal graph convolutional network with co-occurrence feature learning for action recognition. Multimed Tools Appl 79(17–18):12679–12697. https://doi.org/10.1007/s11042-020-08611-4
38. Wang Y, Jiang H, Drew MS, Li ZN, Mori G (2006) Unsupervised discovery of action classes. Proc IEEE Comput Soc Confer Comput Vis Pattern Recog 2:1654–1661. https://doi.org/10.1109/CVPR.2006.321
39. Xin M, Wang S, Cheng J (2019) Entanglement loss for context-based still image action recognition. In: Proceedings - IEEE international conference on multimedia and expo, vol. 2019-July, pp 1042–1047. https://doi.org/10.1109/ICME.2019.00183
40. Yan S, Smith JS, Zhang B (2017) Action recognition from still images based on deep VLAD spatial pyramids. Signal Process Image Commun 54:118–129. https://doi.org/10.1016/j.image.2017.03.010
41. Yan S, Smith JS, Lu W, Zhang B (2018) Multibranch attention networks for action recognition in still images. IEEE Trans Cogn Dev Syst 10(4):1116–1125. https://doi.org/10.1109/TCDS.2017.2783944
42. Yang W, Huang H, Zhang Z, Chen X, Huang K, Zhang S (2019) Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 2019-June, pp 1389–1398. https://doi.org/10.1109/CVPR.2019.00148

43. Yao B, Fei-Fei L (2010) Modeling mutual context of object and human pose in human-object interaction activities. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 17–24. https://doi.org/10.1109/CVPR.2010.5540235
44. Yao B, Jiang X, Khosla A, Lin AL, Guibas L, Fei-Fei L (2011) Human action recognition by learning bases of action attributes and parts. In: Proceedings of the IEEE international conference on computer vision, pp 1331–1338. https://doi.org/10.1109/ICCV.2011.6126386
45. Yao H, Zhang S, Hong R, Zhang Y, Xu C, Tian Q (2019) Deep representation learning with part loss for person re-identification. IEEE Trans Image Process 28(6):2860–2871. https://doi.org/10.1109/TIP.2019.2891888
46. Zagoruyko S, Komodakis N (2016) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer, 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track proc. Accessed: Jun. 19, 2020. [Online]. Available: http://arxiv.org/abs/1612.03928
47. Zhao Z, Ma H, You S (2017) Single image action recognition using semantic body part actions. In: Proceedings of the IEEE international conference on computer vision, vol. 2017-October, pp 3411–3419. https://doi.org/10.1109/ICCV.2017.367
48. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision, vol. 2017-October, pp 5219–5227. https://doi.org/10.1109/ICCV.2017.557
49. Zhou W, Li H, Tian Q (2020) Recent advance in content-based image retrieval: a literature survey, Jun. 2017. Accessed: Jun. 20, 2020. [Online]. Available: http://arxiv.org/abs/1706.06064

Affiliations

Seyed Sajad Ashrafi [1] · Shahriar B. Shokouhi [1] · Ahmad Ayatollahi [1]

Seyed Sajad Ashrafi
s_ashrafi@elec.iust.ac.ir

Ahmad Ayatollahi
ayatollahi@iust.ac.ir

[1] Electrical Engineering Department, Iran University of Science and Technology (IUST), Tehran, Iran