ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse60942020.pdf
https://doi.org/10.30534/ijatcse/2020/60942020

Human Action Classification in Still Images via Human Skin Mask and Residual Neural Network

Samar Monem[1], Shereen A. Taie [2], Mohammed Kayed [3]
Faculty of Science, Beni-Suef University, Beni-Suef, Egypt
samarmahmoud@science.bsu.edu.eg
[2]Faculty of Computers and Information, Fayoum University, Fayoum, Egypt
sat00@fayoum.edu.eg
[3]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
mskayed@gmail.com

ABSTRACT

Action recognition in still images is a challenge task as these images have no motion features like videos and the occlusions of human poses and objects in these images. This paper uses cropped person images, skin mask images derived from the cropped person images and whole images to solve this problem without any object detector. The paper proposes two different models: features-based and classification scores-based models. Two residual neural 'ResNet50' networks are trained for each of the two models. One for the cropped person images and another network for the whole images. The first model uses each residual network as a feature extraction. Then, the extracted features from the two networks corresponding to the three datasets mask skin, persons and whole images are combined in a vector which is used to train an independent Support Vector Machine classifier. In the second model, using Res-Net as a classifier, three classification scores are obtained from the three previous datasets, and then combined into a final score. The two models are validated with the datasets: Pascal VOC and Stanford 40 actions. The results show that features-based model outperforms the scores-based one. It gives mean average precisions of 86.55% and 84.6% on the two datasets, respectively.

Key words: Human action recognition; residual neural network; support vector machine; feature extraction; classification; still image, skin mask.

1.INTRODUCTION

Nowadays, as the growth of mobile phones and video cameras, the Internet has an enormous amount of images. Human action recognition in these still images is aimed to classify the action of a human whose location is usually provided. Human action recognition is an active area for researchers in pattern recognition and computer vision as its capability for providing valuable meta-data for a wide variety of applications such as recognition of gestures, image annotation, human-computer interaction, video analysis, etc.

In general, the recognition of human action in still images represents a more challenging task than static images classification as still images contain multiple humans and objects with occlusions, cluttered background, a variety of human poses and appearances. Also, another challenging task in still images is the lack of motion information. So, traditional methods that are used in action recognition for videos are not used in still images as videos contain spatial-temporal features. The recognition of human action depends on three main factors: person pose and parts variation such as 'climbing' that depends on the human part, the surrounding environment such as 'cleaning the floor' which depends on the floor, and object-person interactions such as 'playing guitar' which depends on the object guitar.

One straightforward solution for the human action recognition problem in still images uses the entire images to depict the action. It handles the problem of action recognition as just a general image classification problem [1] and [2]. These methods do not produce a high performance as the human location and features are not exploited perfectly in the classification. Recent works use the interaction between the human and the objects in the images to build the action classifier [3], [4], and [5]. However, these methods suffer from the problems resulted from false object detection. Other works rely on human parts and poses to build the action classification model [6], [7] and [8]. Also, these later methods suffer from misdetection of human parts in the image and ignore the existence of the objects around them and the manner of interacting with those objects. This paper focuses on the methods that use the neural networks as its effective role to detect human action [9] and also able to handle large data [10].

This paper proposes two models with low computational cost and better performance improvement than other related works. These two models aim to exploit the features for persons who perform the actions or skin regions in persons, and the features for the whole images that contain the objects and the surrounding environment. The paper uses the residual neural network "ResNet-50" as it achieved the first position on the classification competition of ILSVRC 2015. The two proposed models are trained with the two standard datasets: Pascal VOC action and Stanford 40 action datasets. For each dataset, another corresponding dataset of person images (annotated and provided in the dataset) is created. Also, the mask skin is applied to the

person dataset and output a third skin dataset. Preprocessing steps are applied to the person datasets, and the whole images datasets. Each of the two proposed models is trained by two ResNet-50 networks. One network for the person images (a person network) and another one for the whole images (a whole network).

In the first proposed model (a features-based one), each of the two trained networks (person and image networks) is used as a features extractor as we shall discuss later. The features resulted from the two networks for the three datasets mask skin, persons and whole images are combined in one feature vector which is used to train an independent classifier (SVM). In the second model (a scores-based one), the two networks are used as classifiers. The classification scores are calculated for the three previous datasets. The three scores are then combined according to a proposed equation to get the final classification score.

The remaining parts of this paper are structured as follows. Section II addresses several recent researches relating to recognition of human action in still images. Section III describes the details of the two proposed models. Section IV addresses the used dataset, learning setup and the experimental results. Finally, our work is concluded in section V.

2.RELATED WORKS

Several studies have been proposed to resolve the restrictions of still images to achieve reliable results of human action recognition. This section discusses a number of existing related research approaches. These approaches are discussed from two viewpoints: features extraction and required annotations. The following subsection will discuss these two viewpoints, respectively.

2.1.Feature extraction viewpoint

Image features play an important role in human recognition. Previous proposed approaches grouped into three main categories: context-based approaches, human pose-based approaches and part-based approaches.

***Context-based approaches*:** These approaches not only depend on the human himself but also the interaction between the human and the surrounding objects is considered. So, these approaches aim to detect the meaningful regions in the images that represent this interaction. The work in [4] depends on multi-scale identification of semantic regions and learners to extract the related features. First, this method detects and chooses some of the candidate patches that are likely to provide informative features for the recognition task by training SVM models. Then, the image is fed forward to get the convolutional feature map, and output all probable SVM boxes scores. Then, the CNN network is trained on each scale to classify candidates of different action classes. Multiple candidates are treated as individual samples from the same images. Finally, training and fusing a set of parameters to weight distinctive features of multi scales. The experimental results of this technique showed that the weighted concatenation exceeds the uniform concatenation.

The model in [11] had modified the classical BOW pipelines to recognize the human action. It proposed two scale coding approaches to specifically concatenate multi-scale features for human attributes and action recognition in the final image representation. The first approach, called an absolute scale coding, is depended on multi-scale representation of image with encoded image size scale. The second approach, called relative scale coding, was done by extending the coding approaches to the deep convolutional features of a pre-trained deep neural network. The final image representation was represented by combining the three constructed scale partitions small, medium and large scale features by Fisher Encodings. This method used deep convolutional features in the scale coding model instead of SIFT features in standard BOW pipelines.

Some techniques in the context-based approach eliminate the use of human bounding boxes such as [12] and [13]. The approach in [12] relied on the relations between the image superpixel classes. In the first step, the image was partitioned into a collection of superpixels and a number of pre-trained object detectors were evaluated to each superpixel to output a detector score vector which is used as measures in a graphical model. In the second step, the graphical model was used to predict the action class by using the measures of each superpixel and a fully connected superpixel class graph. The efficient greedy technique is implemented to support inference over the previous graph and this model was trained by a latent structural SVM technique. The work in [13] aims to recognize human action in images with minimum annotation efforts. First, this technique used the selective search method to build object proposals. Then, the object proposals are rotted away into finer-grained object parts to be used for delineating the precise shape of interaction regions between human and objects. Finally, the label of action is predicted by using the representation of the features obtained from the interaction regions between human and objects using an effective product quantization process.

Due to the large number of multi-scale windows generated in images by context-based approach, the computation time is increased and produces too many redundant windows especially if the dataset is large. Also, the context-based approaches suffer from the problems resulted from false context detection. The context-based approaches are useful for the action classes that depend on objects like ‘riding bike’ as the recognizing action depends on the object ‘bike’. For the action classes which depend on human parts more than objects like ‘running’ and ‘walking’, the parts of humans are more informative than the other objects in the images.

***Human poses-based approach*:** Human pose estimation is used to detect actions in still images. In [8], the combination of human poses and the selection of CNN features are used to obtain candidates person proposals. First, an oracle human detector is used to extract optimal human bounding box during training and testing time. Then, it employs learning transfer to learn action-specific detectors to detect human regions which represent a candidate bounding boxes for action recognition and replace the bounding boxes of the ground truth of humans. It significantly improves the performance of action detection. However, its main drawback is the amount of time taken to run the transfer learning methods on each person, which creates a problem of scalability when the number of classes is big. In [14], a Generalized Symmetric Parts Model (GSPM) is proposed which improves the standard bag-of-words (BoW) approach to detect semantically meaningful regions. These meaningful regions extended to the action recognition by finding generalized symmetric parts in images and learning the parameters by a max margin classifier. This model assumed that

the actions are primarily executed by hands and feet that are modeled as the generalized symmetric pairs. The remaining steps of the model are like as the general BoW classification processes like feature detection, feature extraction, encoding feature and obtaining the histograms required for classification

***Part-based approach*:** For action recognition and fine-grained recognition, part-based methods assume that any action is done by the human body parts that provide important information. In [6], a part-based method is suggested which depends on the extraction appearance features from body parts. This method included three main steps. First, detecting and using several semantic parts of human to extract informative features from them. Second, an effective detection process dividing multiple images by the same grid is evaluated. Third, a top-down spatial arrangement is evaluated simultaneously which extends the inter-class variance. Also, [7] uses body part detectors to detect human action. This technique depended on the parts: head, torso, and legs of human. After that, trains a CNN on these part regions and output the pool 5 features. Finally, it combined these features with the ground bounding box of the whole instance of the human which is supplied either by an oracle or by a human detector.

The human poses-based and part-based approaches depend on the accuracy detection of the poses and the parts of the human respectively in still images. Also, the candidate regions used in the two approaches may lead to an increase in the computational complexity especially when there are a large number of images. The two approaches only use pose and parts which is not enough for the analysis of more complex human actions that depend on the context in images. For example, for images that have a human interacting with a computer, the object computer is informative to recognize this action class label. In this case, the knowledge about the context and the interacting objects should be considered. Another complex human action from the viewpoint of the two approaches (parts and poses) occurs when the parts have almost the same structure a ‘smoking’ and ‘blowing bubbles’.

2.2.Annotation required viewpoint

Dataset annotation is a challenge task, especially with large datasets. So, this subsection discusses the techniques applied in the action classification of images from the annotation viewpoint and grouped them into three categories: no annotation, human annotation and extra annotation based approaches.

For the ***no annotation*** category, the algorithms are learned from the whole images to detect the action classes. Some algorithms depended on segmenting images into regions or proposals and applied a detector to indicate the informative regions to learn the action classification model such as [12] and [13]. Other algorithms used the entire image to depict action and handle action recognition as just a general image classification task such as [2]. In [2], a residual neural network based approach is implemented to extract deep convolutional features from the images. The model used the entire image to classify and recognize the action of human as a simple task of image classification. First, the features are extracted by using a pre-trained residual neural network of 50 layers. Then, the model classified the extracted features using SVM into the different action classes.

In [15], a different technique is proposed which learnt spatial-temporal information from whole images. This model depended on human appearance and the prediction of the future movement patterns of the human. This is achieved by predicting the temporal order of each pixel. This prediction learned by training a linear ranking system on the predicted spatial-temporal image representation tensor. Then, a transfer learning approach is applied to implement a new spatial-temporal CNN, called STCNN. It is used in classifying single human action image by fine-tuning a CNN that is explicitly pre-trained for appearance-based classification. The main disadvantage of this model is that it is trained with segmented videos to learn the hypothetically of temporal images representing a series of frames.

For the ***human annotation*** category as the scope of this paper, the bounding boxes of humans are annotated manually. Some algorithms depended on the human box only. Some of these algorithms depend on human pose-based and part-based approaches discussed previously. In [16], human annotation based model is proposed. This model is a very deep convolutional network trained for large image classification scale. It achieved the first and the second positions in the ImageNet Challenge 2014 submission. The main contribution was increasing the depth of the proposed neural network. This achieved by combining the 16 and 19 weight layers CNN. This network is applied in many classification and pattern recognition datasets. This network is also applied to detect the human action in images. It is used to extract features from the whole images and apply SVM to classify them.

For the ***extra annotation***, the algorithms depend on the human box annotation and additional annotation. The research in [17] basically used a dataset in which certain concepts were annotated to classify actions by mapping from target human images to these certain concepts with a visual sense (e.g. objects and object attributes).. So, an external annotated dataset is used, with several images labeled with a broad variety of specific concepts.. The previous mapping is then used as a representation of feature to classify a target human dataset, rather than defining the human images with directly extracted image features. This mapping allows describing an image explicitly with high-level concepts of the action classification task. This technique showed that the concepts that have been learned within each category have conceptual meaning. Also, the model in [18] used "attribute" to assist in recognizing actions. Its attributes are recommended primarily for representing the whole body and movement scenarios, e.g. "torso translation with arm movement."

The novelty of this paper is to provide two models to detect human action in still images using a pre-defined residual neural network and bounding boxes of humans. The proposed models are simple to train, do not need any segmentation or object detectors techniques in the images, do not need any additional annotation and obtain high performance as compared to the other related works. The proposed models are computationally of low cost as compared with other related works.

3.THE PROPOSED MODELS

In this section, we will discuss the two proposed models for human action recognition. As mentioned before, the two models are trained using the two neural ‘ResNet50’ networks: person residual network and whole image residual network.

Before training each network, a preprocessing step is done for the dataset images. Then, applying the mask skin to the person images. This section discusses the details of the two proposed models in the two subsections 3.4 and 3.5, respectively. Before doing that, the details of the preprocessing step, the structure of the two trained residual networks and skin mask step are provided in the subsections 3.1, 3.2, 3.3, respectively.

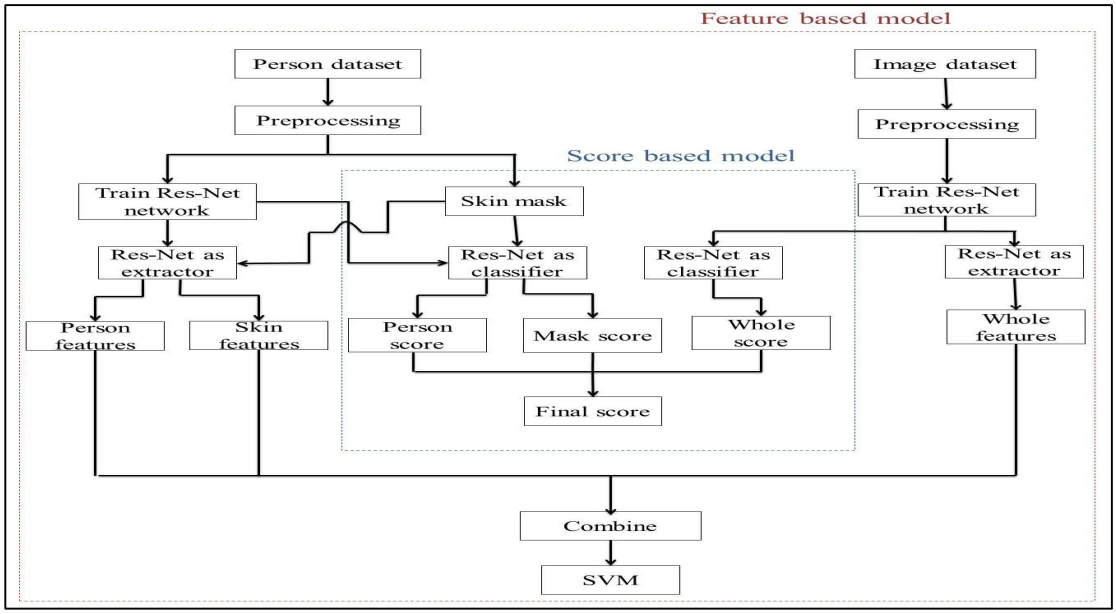

Figure 1: The structure of the two proposed models.

3.1 Preprocessing step

Given a dataset of images as input, whole image dataset, this paper assumes that the bounding boxes of the persons who performing specific actions are annotated manually in the dataset at both training and testing phases. Another dataset is formed by converting (cropping) these persons bounding boxes into images, person dataset. After that, the two dataset, person and whole images are preprocessed. In the preprocessing step, two main processes are done. The first and the main one is scaling the images to a fixed size to extract the same length of the feature vector for each image. The fixed size is 244×244×3 which is required for the residual network 'ResNet-50' that will be used in the next step. The second process is checking whether each image has three dimensions colored image or not. If not, the image is concatenated three times to generate a 3-dimensional one as the residual neural network uses 3 matrices for each RGB color channels.

3.2 Training ResNet-50 network step

In this step, two networks are trained using the two preprocessed (whole and person) datasets, one network for each dataset. In this paper, the residual neural network 'ResNet50' model is used. 'ResNet50' is a kind of deep convolutional neural networks, which use a shortcut connection to resolve the degradation problem that arose when the deep convolutional neural networks begin to converge.. The residual neural network is modified from the network block which had been trained using the ImageNet dataset. The model has 50 residual network layers (the total number of layers is 177). In short, the ResNet-50 model has five convolution layers followed by fully connected layers and a final softmax layer. The residual network architecture is shown in Figure 2. The training process uses pre-trained weights for the first 110 layers of the network, which have trained in the ImageNet dataset to give better generalization and prevent overfitting of the dataset. The last three blocks: fully connected layer, softmax layer and classification layer that have 1000 output size are replaced by the new three layers: fully connected layer, softmax layer and classification layer with output size according to the number of action dataset classes.

3.3 Skin mask step

Skin color detection can be a fundamental powerful cue in many detection like face detection, skin diseases detection, such as vitiligo and diabetes, detection of human motion, predicting pornography and nude image. The greatest difficulty for skin color detection is the wide variation in skin appearance that may occur, such as occultation effects, color, light source position, intensity, etc. Some objects that resemble skin color such as cooper, wood and certain clothes are also mistaken as regions of skin. The efficacy of skin detection depends on the color space chosen, since the color distribution of human skin depends on the color space. Most studies have focused on the pixel-based skin identification, which classifies each pixel as either skin or not skin. This paper uses what is called the "explicit skin cluster" method. It the simplest, and frequently applied method that explicitly specifies the skin cluster boundaries in some color spaces. This method is very popular as it is easy to implement and do not require a training phase. The greatest challenge in explicit skin cluster method is achieving a high rate of skin color recognition, with minimum number of false positive pixels possible. There are many different color spaces like YCbCr, RGB, HSV and HIS which can apply in the explicit skin cluster method. This paper uses the color space RGB. The person dataset is used to extract the skin mask dataset. If the RGB value of the pixel in person dataset satisfies the following conditions, they will be marked as skin color pixels:

$$\begin{cases} 45 < R < 255 \\ 34 < G < 219 \\ 30 < B < 200 \end{cases}$$

Otherwise, if the RGB values of the pixel does not satisfy the condition: it will be marked as a non-skin pixel and set to be 0.

The Figure 3 shows the difference between the whole image dataset, the person image dataset and the skin image dataset respectively.

3.4 The first proposed features-based model

In this model, the two trained networks, person and image are used to extract features from the training datasets person, skin and the whole image training datasets, respectively. These features are then combined to learn the SVM classifier. Figure 1 shows the general structure of the proposed approach. The next two subsections will discuss the details of the feature extraction step and the SVM classifier learning step, respectively.

3.4.1Feature extraction step

In this phase, given the training images dataset, each constructed residual network is used to determine high-level features from these images to form a vector of features. These features are a set of parameters that define the content of the images precisely and uniquely. The two trained ResNet-50 networks are adapted to be used as a feature extraction technique. This adaptation is done by removing the final softmax classification and the fully connected layers, and using the output of the 'avg pooling' layer as extracted residual features from the images. The last layer of each adapted residual neural network has a length of 2048.

The person and skin images are fed to the adapted person network separately. Each of them output the features which have a length of 2048. After that, these features extracted from the skin and person images are concatenated. As theses concatenated features may be redundant, the Neighborhood Component Analysis (NCA) [19] feature selection algorithm is applied to the concatenated feature vector. The NCA algorithm is a non-parametric method that estimates the feature weights to minimize the classification error. The most importance 2048 is taken from concatenated features to represent a final feature vector from skin and person image named feature selection vector.

The whole images are fed to the adapted whole image network to output the features which has a length of 2048. The whole image feature vector is concatenated with feature selection vector. The length of the final features vector which is used as input to the action classifier is 4096.

3.4.2The SVM classifier

In this step, the machine learning model: Support Vector Machine (SVM) is learned using the training dataset with the constructed and concatenated features given in the previous step. SVM is commonly used for the classification and regression of datasets in high dimensions. It gets better performance. Also to enhance the results, SVM is applied using Error-correcting output codes (ECOC) [20]. The ECOC algorithm extended binary classifiers to multi-classifiers problem. The ECOC model improves classification performance compared to other multiclass models [21]. The SVM used here aims to create a separate hyper plane from the training data that splits different classes into a high dimensional feature space by a linear mapping to be able to classify the new samples. The kernel function used in mapping the feature space into a linear space is a linear function which is given in Eq. (1).

$$K(x_i, x_j) = x_i'.x_j \qquad (1)$$

Where $K(x_i, x_j)$ is the kernel function and x_i, x_j are p-dimensional vectors representing observations i and j in X-direction, respectively.

To train SVM, one-versus-one schema is used which considers one class as positive, another one as negative, and ignores the remaining results in $N(N-1)/2$ binary leaners, where N is the number of classes. Also, 10 folds are setting in a cross-validated model.

In the testing phase, the steps are the same as in the feature extraction and classification phase. First, the testing image whole and person dataset is preprocessed according to the previously described preprocessing step section. Then the mask skin is applied to the testing person dataset to output the testing mask skin dataset. After that, the whole image is fed into the trained whole network that described in the previous training ResNet-50 network step section to get the whole image features. Also, the person image is fed to the trained person network to get the person features. Likewise, mask skin image is fed to the person network and get the mask features. The features of skin images and person images are concatenated in one feature vector. The NCA is applied to select the most informative 2048 feature. The features of whole images and feature selection vector are concatenated in one feature vector. The final feature vector is fed to the trained SVM that is described in the previous learn the SVM classifier step section to get the action class label.

3.5 The second Proposed scores-based model

In this model, the mask skin is applied to the person dataset and gets the skin dataset as discussed in previous mask skin section. Then, the classification scores for the three datasets are combined according to a suggested equation as discussed in the following subsection.

3.5.1classification score combine

After preprocessing of the person and whole datasets, skin dataset is extracted from person dataset and training the person and whole networks. The performance of the two trained networks is measured with the test person, mask skin and whole images datasets to get the three classification scores. These scores are combined according to a suggested equation to obtain the final classification score. In the score based model, the two ResNet-50 trained networks mentioned before

are treated as classifiers. In the testing phase, the testing dataset of preprocessed whole imagesis fed to the image network to get the first classification score fwhole. Likewise, the preprocessed person images are fed to the person network to get the second classification score fperson. Finally, the skin image dataset is fed to the person network to get the third classification score fskin. The final classification ftotal is calculated using Eq. (2), which combined the third calculated classification scores.

$$ftotal = \alpha\ fwhole + \beta \max(fperson, fskin);\quad \alpha + \beta = 1 \qquad (1)$$

Where α and β are random variables that are calculated based on the number of informative classes (person or whole images) in the dataset. If the dataset contains many class labels that depend on person feature more than whole image feature (e.g. running, walking and applauding), it prefer to set the β larger than α. Also, if the dataset contains class labels that based on whole image feature especially when contain an informative object, more than person image features (e.g. fixing a car, using a computer and riding a horse), it prefer to set the α value larger than the β value.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The model is evaluated on two challenging widely open databases: Pascal VOC 2012 [22] and Stanford-40 [23].

In this section, the details of databases and evaluation measures are described. The experimental setup for the trained networks is then described. Finally, the experimental results of the two proposed methods on the two target database are showed and discussed.

4.1Dataset

The PASCAL VOC Action dataset contains of 10 different actions, Jumping, Phoning, Playing Instrument, Reading, Riding Bike, Riding Horse, Running, Taking Photo, Using Computer and Walking with total 4588 images. For training, the training and validation set that specified in [20] are used, and utilize the same testing set. The ground-truth boxes of humans are given in both train and test time.

The Stanford-40 Action dataset is more complicated as it consists of 40 diverse daily actions of human like cooking, holding an umbrella, walking the dog, smoking, etc. The total of images in database is 9532 images. The training and testing set used as proposed in [21]. Also, ground-truth boxes of humans are given in both train and test time. To validate the performance of the two proposed models, the Average Precision (AP) is calculated. The AP is specified as the area under the precision-recall curve. AP is shown in the following Eq. (3).

$$AP = \int_0^1 p(r)\, dr \qquad (3)$$

Where P represents precision and r represents recall. Also, mean Average Precision (mAP) which is the average of AP over classes is measured.

4.2Learning details

In this paper two residual neural networks are trained: the whole network and the person network. Each of them is modified from the ResNet-50 as illustrated in 'section III-subsection 1) A)'. To train them, the learning rate is set to be 3e-5 and the weights are optimized by the stochastic gradient descent momentum (SGDM). The training process is performed for 12K of 5 batch size. The training process uses pre-trained weights for the network's first 110 layers that have been trained in the ImageNet dataset to speed up networks training and avoid overfitting of the dataset. Some types of techniques for data augmentation are applied like horizontal and vertical translation with a value picked randomly from range [-30 30], random reflection in the left-right direction with 50% probability, and vertical and horizontal scale with a value picked randomly from range [0.9 1.1]. Before each training epoch, shuffle the training data. If the size of the mini-batch does not evenly divide the number of training samples, then trained network discards the training data that does not fit into each epoch's final full mini-batch. The loss function is identified as binary cross-entropy.

4.3Result Comparison

In this research several methods have been introduced trying to select the appropriate one for predicting the various human actions. First, the model is comparing to other approaches on the Pascal VOC 2012 Action test set. Table 1 represents the results on the testing set of predicting human action using AP as an evaluated measure. The proposed features based and classification score based models achieve a high mAP of 86.55% and 85.63% respectively and outperform the second and third best published model respectively.

Compared with algorithms [8], [7] and [16] which used human box annotation only like us, the proposed model (features based and classification score based) improves the performance significantly by +2.5% and +1.6% respectively. The models [13], [11] and [14] that depend on context-based approach by applying a detector to detect multi-scale informative regions, the proposed features based and classification score based models improve the performance by +1.7% and +2.65 respectively. The best performance in the Pascal dataset is 91.55% which achieves by [6] model. The model of [6] depends on training some semantic detectors and organizes semantic parts in top-down spatial order. This approach is achieved the highest mAP because the Pascal dataset consists of three action classes from a total of 10 classes are highly depending on parts of human: jumping, walking and running.

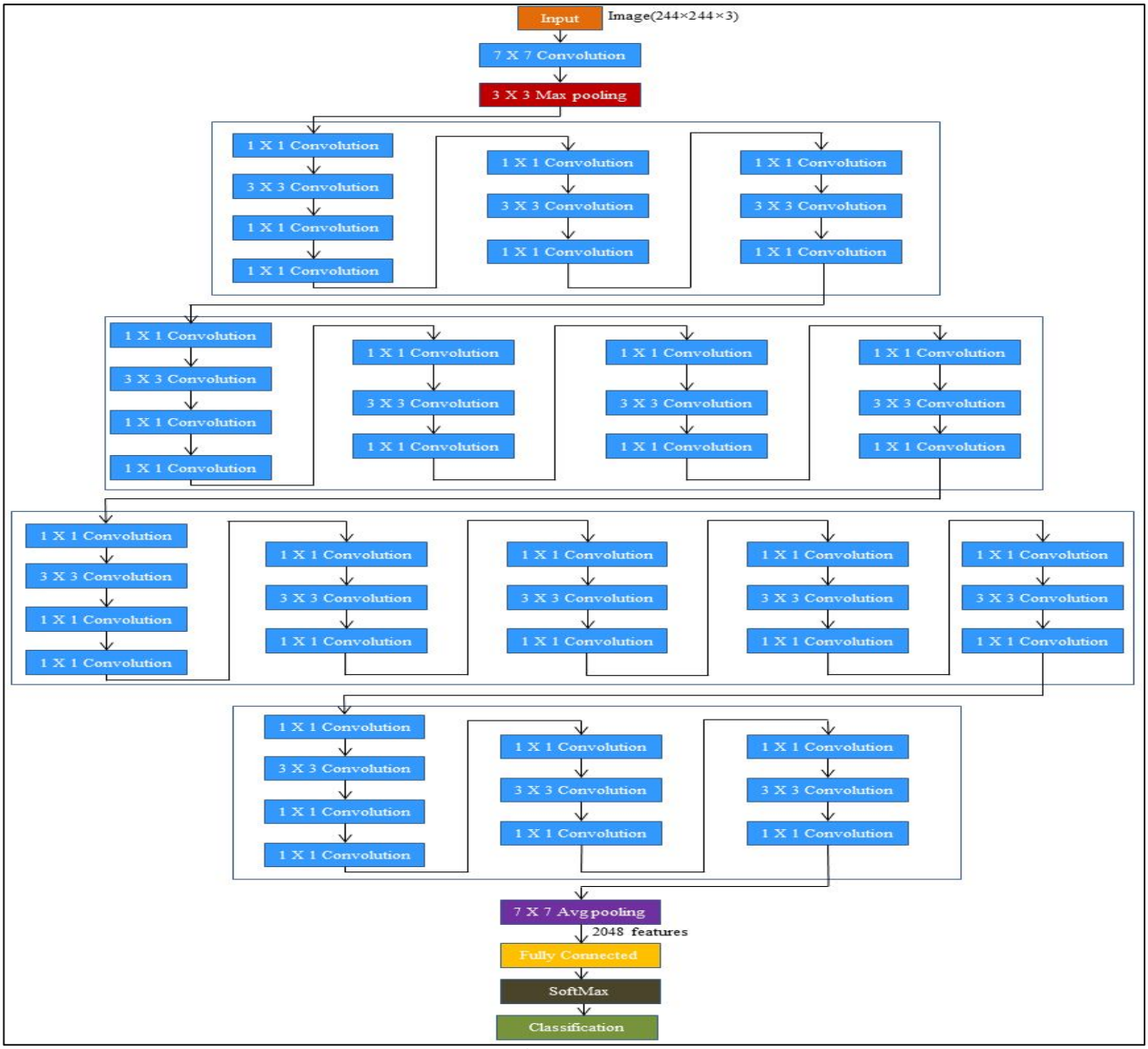

Figure 2: the structure of the ResNe-50 network.

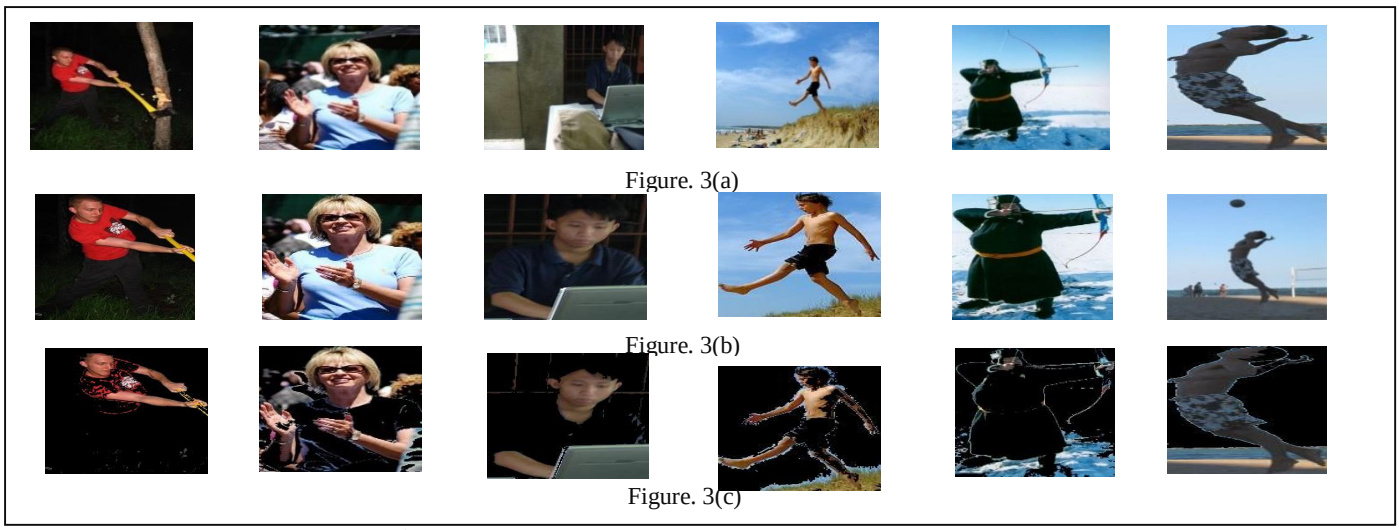

Figure. 3(a)

Figure. 3(b)

Figure. 3(c)

Figure 3: (a) a sample from the whole dataset, (b) person dataset and (c) skin dataset.

Table 1: The performance model on the Pascal 2012 test dataset

Algorithm	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	mAP
Action-Specific Person detector[8]	84.9	62.4	91.3	61.1	93.3	95.1	84.1	59.8	84.5	53.0	76.95
Wholes and Parts[7]	84.7	67.8	91.0	66.6	96.6	97.2	90.2	76.0	83.4	71.6	82.60
Very Deep Convolutional Networks [16]	89.3	71.3	94.7	71.3	97.1	98.2	90.2	73.3	88.5	66.4	84.03
Top-Down Pyramid for action recognition[6]	96.4	84.7	96.7	83.3	99.4	99.2	91.9	85.3	93.9	84.7	91.55
Minimum Annotation Efforts[13]	86.68	72.22	93.97	71.30	95.37	97.63	88.54	72.42	88.81	65.31	83.23
GSPM [14]	79.1	53.8	69.1	46.6	96.6	96.3	89.9	35.8	69.5	73.9	71.06
Scale coding bag of deep features[11]	89.5	69.7	92.4	70.8	97.2	98.0	89.8	73.8	88.4	69.4	83.90
Resnet-50 (whole)[24]	84.25	76.57	89.99	73.14	88.09	95.17	77.57	64.64	83.67	54.68	78.78
Resnet-50 (person)[24]	86.12	83.55	88.62	72.43	88.37	92.61	84.21	77.25	75.18	68.37	81.67
classification score based (0.45 fwhole+0.55 fperson)	89.89	84.81	92.1	77.73	93.81	97.05	88.02	77.53	84.17	71.18	85.63
features based	89.37	84.84	93.47	79.76	94.41	98.33	86.28	79.8	85.74	73.54	86.55

Also, the classes taking a photo and phoning partially depending on human parts. But when applied this technique [6] in the Stanford-40 dataset the mAP is going down to 80.6% when the action classes are increased to 40 classes. The increment of the action classes makes the parts of human approximate the same like (brushing teeth and blowing bubbles), (cooking and cutting vegetables) and (fixing a car and fixing a bike), etc. Finally, the best results on the Pascal VOC 2012 dataset are obtained by [6]. The second and the third best performance results are obtained from the proposed features based and classification score based models respectively. In Table 2, the proposed approach is compared with state of the art on the Stanford-40 action dataset. Among the Stanford-40 dataset, the proposed features based and simple models improve the performance of algorithms [8], [16], and [6] that use the human box annotation only like ours by +4% and +3% respectively. Also, the performance of [4], [14] and [11] is increased by the proposed features based and classification score based models by +4.6% and +3,61% respectively without using multi-scale detector. The elimination of using any annotation is very benefited which [13], [12], and [15] algorithms do, but the performance needs to improve. The proposed features based and classification score based models increase these algorithms performance by +1.96% and 0.97% respectively. In [17] model the performance is increased by using additional annotation which may be costly. The proposed features based algorithm increases the performance of [17] by 1.5% without using any additional annotations. In short, the proposed method outperforms a gain of +1.5% among all the 40 categories without using additional annotation or multi-scale detector. The proposed model achieves state-of-the-art performance on the Stanford-40 dataset. Also, the classification score based model outperforms the second best technique after the proposed features based technique.

Based on this discussion above, the two proposed models are much easier to be trained as not required any multi-scale detector techniques, no additional annotation required and have better generalization. Figure 4 illustrates the AP performance for each action class on the Standford-40 test set for the features based model and classification score based model.

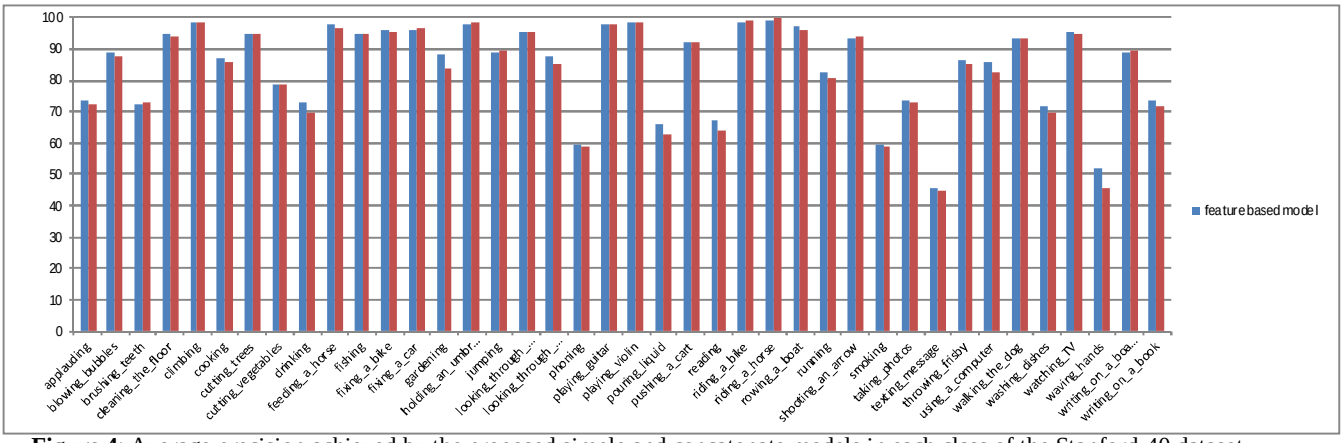

Figure 4: Average precision achieved by the proposed simple and concatenate models in each class of the Stanford-40 dataset.

Table 2: The Performance model on the Stanford-40 dataset

algorithm	mAP
Action-Specific Person detector[8]	75.4
Very Deep Convolutional Networks[16]	77.8
Top-Down Pyramid for action recognition[6]	80.6
Minimum Annotation Efforts[13]	82.64
Multi-Scale Region Candidate[4]	78.8
GSPM[14]	54.5
Scale coding bag of deep features[11]	80.0
Efficient Greedy Inference[12]	72.3
STCNN[15]	81.76
Concepts and Attributes for action[17]	83.12
Resnet-50 (whole) [24]	80.88
Resnet-50 (person) [24]	68.32
classification score based (0.6 fwhole+0.4 fperson)	83.61
features based	84.65

5.CONCLUSION

This paper proposed two models that improve the action recognition in still images. They are depend on the assumption that recognition of human action is a combination of meaningful skin or human regions and object areas. The paper used the residual neural network architecture to implement the two models. The two models have balanced among the features extracted from the whole image, the features extracted from humans who perform the action and the features extracted from skin regions in these humans. The whole image features included important information about the objects that interact with the human and the surrounding environment. The human who completes the action is informative especially if the action is totally depending on the human (e.g., running). The models are evaluated on the two datasets: PASCAL VOC 2012 and Stanford-40. As shown in our experiment, features-based model outperforms the classification score-based model. It reports a mean average precision of 86.55% and 84.6% on the PASCAL VOC 2012 and the Stanford-40 datasets, respectively. The results of experimental analysis and visualization also showed the reasonableness and efficacy.

REFERENCES

[1] B. Yao, A. Khosla, and L. Fei-Fei., **Combining randomization and discrimination for fine-grained image categorization**, *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado, pp. 1577–1584, 2011. https://doi.org/10.1109/CVPR.2011.5995368

[2] S. R. Sreela and S. M. Idicula., **Action recognition in still images using residual neural network features**, *in Proceedings of 8th International Conference on Advances in Computing and Communication*, Kochi, pp. 563-569, 2018.

[3] G. Gkioxari, R. Girshick, P. Dollár, and K. He., **Detecting and Recognizing Human-Object Interactions,** *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 8359-8367, 2018. https://doi.org/10.1109/CVPR.2018.00872

[4] Z. Zhao, H. Ma, and X. Chen., **Multi-scale region candidate combination for action recognition**, *in Proceedings of International Conference on Image Processing*, Phoenix, AZ, pp. 3071-3075, 2016.

[5] A. Prest, C. Schmid, and V. Ferrari, **Weakly supervised learning of interactions between humans and objects**, *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601-614, 2012. https://doi.org/10.1109/TPAMI.2011.158

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

[7] G. Gkioxari, R. Girshick, and J. Malik, **Actions and attributes from wholes and parts,** *in Proceedings of the IEEE International Conference on Computer Vision*, Santiago, pp. 2470-2478, 2015.

[8] F. S. Khan, J. Xu, J. Van De Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez, **Recognizing Actions Through Action-Specific Person Detection**, *in IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4422-4432, 2015. https://doi.org/10.1109/TIP.2015.2465147

[9] M. Akour, O. Al Qasem, H. Alsghaier, and K. Al-Radaideh, *The effectiveness of using deep learning algorithms in predicting daily activities*, in *Int. J. Adv. Trends Comput. Sci. Eng.*, 2019 doi: 10.30534/ijatcse/2019/57852019.

[10] N. A. Zavalko, N. L. Krasyukova, L. A. Plotitsyna, A. G. Gladyshev, and A. N. Boyko, **Neural network system for processing large-volume diagnostic data**, *in Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3211–3215, 2020, doi: 10.30534/ijatcse/2020/113932020

[11] F. S. Khan, J. van de Weijer, R. M. Anwer, A. D. Bagdanov, M. Felsberg, and J. Laaksonen, **Scale coding bag of deep features for human attribute and action recognition**, *in Machine Vision and Applications*, vol. 29, pp. 55-71, 2018.

[12] S. Abidi, M. Piccardi, and M. A. Williams, **Static action recognition by efficient greedy inference**, *in Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Lake Placid, NY, pp. 1-8, 2016.

[13] Z. Yu, C. Li, J. Wu, J. Cai, M. N. Do, and J. Lu, **Action Recognition in Still Images with Minimum Annotation Efforts**, *in IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479-5490, 2016.

[14] Z. Zhao, H. Ma, and X. Chen., **Generalized symmetric pair model for action classification in still images**, *in Pattern Recognition*, vol. 64, pp. 347-360, 2017. https://doi.org/10.1016/j.patcog.2016.10.001

[15] M. Safaei and H. Foroosh, **Still image action recognition by predicting spatial-temporal pixel evolution**, *in Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Waikoloa Village, HI, pp. 111-120, 2019.

[16] K. Simonyan and A. Zisserman, **Very Deep Convolutional Networks for Large-Scale Image Recognition**, *in Proceedings of 3rd International Conference on Learning Representations*aaaaa, San Diego, CA, 2015.

[17] A. Rosenfeld and S. Ullman, **Action Classification via Concepts and Attributes**, *in Proceedings of 24th International Conference on Pattern Recognition*, Beijing, pp. 1499-1505, 2018.

[18] J. Liu, B. Kuipers, and S. Savarese, **Recognizing human actions by attributes**, *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, pp. 3337-3344 2011. https://doi.org/10.1109/CVPR.2011.5995353

[19] W. Yang, K. Wang, and W. Zuo, **Neighborhood component feature selection for high-dimensional data**, *in Journal of Computers*, vol. 7, no1, pp. 161-168, 2012.

[20] S. Escalera, O. Pujol, and P. Radeva, **On the Deoding Process in Ternary Error-Correcting Output Codes**, *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 120-134, Jan. 2010, doi: 10.1109/TPAMI.2008.266.

[21] J. Fürnkranz, **Round Robin Classification**, *in Journal of Machine Learning Research*, Vol. 2, pp. 721–747, 2002

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, **The {PASCAL} {V}isual {O}bject {C}lasses {C}hallenge 2012 {(VOC2012)} {R}esults**, *in* http://www.pascalnetwork.org/ *challenges/VOC/voc2012/workshop/index.html.*

[23] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, **Human action recognition by learning bases of action attributes and parts**, *in Proceedings of the IEEE International Conference on Computer Vision*, Barcelona, pp. 1331-1338, 2011.

[24] K. He, X. Zhang, S. Ren, and J. Sun, **Deep residual learning for image recognition**, *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 770-778, 2016. https://doi.org/10.1109/CVPR.2016.90