Feature Fusion Based Human Action Recognition in Still Images

Abdul Sattar Chan[1], Kashif Saleem[2], Zuhaibuddin Bhutto[3], Mudasar Latif Memon[4], Murtaza Hussain Shaikh[5], Saleem Ahmed[6], and Ahsan Raza Siyal[7]

[1]Eletrical Engineering Dept. Sukkur IBA University, Sukkur
[2]Telecommunication Engineering Department, Dawood University of Engineering & Technology, Karachi, Pakistan
[3]Department of Computer Systems Engineering, Balochistan University of Engineering & Technology, Pakistan
[4]IBA Community College Naushehro Feroze, Sukkur IBA University, Pakistan
[5]Department of Computer Systems Engineering, Kyungsung University, Busan, South Korea
[6]Electronics Engineering Department, Dawood University of Engineering & Technology, Karachi, Pakistan
[7]Computer System Engineering Department, Dawood University of Engineering & Technology, Karachi, Pakistan

Summary
Recognizing human actions based on still-images is a challenging task involving predictions on human interaction with objects and body postures. In this paper, a novel method is proposed in which three networks are used to determine human pose, most relatable object in the scene and the overall scenario that includes actors and all objects around him. Before testing the proposed method the performance of the conventional transfer learning method is evaluated by four popular pre-trained convolutional neural networks for feature extraction and classification is performed by the Support vector machine, only principal components of extracted features are passed through SVM for predicting human action in the scene. To evaluate the proposed model Stanford40 dataset is used, the dataset contains images of 40 human actions and every image has a bounding box of the person performing the action. There is a total of 9532 images with 180-300 images per class, for the experiment only 10 classes of the dataset are used for proposed model evaluation. Experimental results show a proposed method in the paper achieves high robustness and accuracy.

Ke ywords:
convolutional neural networks, transfer learning, support vector machine.

1. Introduction

Human action recognition based on videos has been comparatively considered as an active research area in the computer vision [1] [2]. On the other hand, human action recognition still, image has not been in highlights and not being focused by modern researchers. Lately, the research community has increased attention and making efforts to set up benchmarks and sort out issues like PASCAL VOC action recognition [3]. Other than based on videos where image sequences play a vital role [4]. In still image-based action recognition, the main idea is predicted based on action labels providing an interpretation of human actions and their contact with the objects present in the scene [5].
The convolutional neural network (CNN) has emerged as a key development in the computer vision that is replaced by a conventional computer vision field. The CNN or ConvNets models improve not only image classification accuracy, but they are employed to extract features in the field of depth estimation, semantic segmentation, and object detection [6] [7]. Since CNN has a higher computational cost and memory requirements to train and deploy the model, hardware with high specifications is also essential. A system to be deployed for human action monitoring or in order to automate surveillance system, thievery detection and warning system in banks, and malls, requires a real-time processing capability even in an embedded board having the comparatively less computational power and memory. Unlike the desktop PC, embedded boards have limitations in terms of computing power, memory and power consumption due to these stated reasons deployments of deep neural network-based algorithms and systems that require extensive computations restricted by embedded systems. for that reason, it is needed to carry out a study into the optimization of convolutional neural network technology to overcome such limitations.
Therefore, in order to tackle such limitations, this paper proposes a method for detecting human actions in still images with similar performance to state of the art methods but with improved accuracy and less memory weight. Feature extraction is carried out by four different popular pre-trained networks for performance evaluation and principal component analysis reduces the dimensionality of the feature matrix and then support vector machine classify the action in the scene.

2. Related Work

Action recognition based on videos has been well established over the years with a long list of literature [1] [27], [28]. For still image-based action recognition, there are different parameters that have been investigated and experimentally tested for efficient human action recognition with high accuracy and less computational power consumption. The group of existing methods can be categorized into three categories.

Manuscript received November 5, 2019
Manuscript revised November 20, 2019

The first scheme is based on human poses that apply human part detectors to detect the parts of the human body and encode them into the pose for action recognition [8]. In [9], the author performs the training of a convolutional neural network for the estimation of human poses.
The second scheme is based on the situation or circumstances. This category not only consideration human poses but also human-object interactions as an aid to perform human action recognition. In [10], the author creates pairs of human poses and objects human is interacting and picks discriminative ones for human action recognition. Yao in [11] considered multiple interactions in a scene that include human poses, human-object interaction, as well as the affiliation amongst objects. In [12], pre-trained object detectors are deployed to detect most related objects to the person in the scene.
The third approach is a part-based method. In [13], the use of local patches of an image as parts in order to train the model which similar to classifier for action recognition [14]. In [15], human action in a scene is recognized by only using image labels in order to locate humans in a scene. The multiple detectors are used to detect the human upper body and face. After the detection of humans, the most related objects are then detected on the bases of relative locations.

3. Proposed Method

In machine learning, transfer learning or knowledge transfer is a method that utilizes previously learned knowledge to solve a new problem. For training the models with a small dataset, transfer learning using pre-trained deep conNets are very useful because of conNets face overfitting problem with small size dataset. However, overfitting can be avoided by increasing the size of the dataset costing high annotations and require high computations which can increase the complexity. In this case, the transfer learning method is used by utilizing pre-trained deep representations for the construction of new architecture [16]. In this paper, we have employed four popular pre-trained models Resnet18 [17], VGG16 , VGG19 [18] and googlenet [19].
Resnet-18 is a pre-trained convolutional neural network on more than a million images of 1000 different kinds of categories of ImageNet dataset [20]. The network consists of total 18 layers with an input layer of size 224 by 224 and having the ability to classify 1000 different categories like keyboard, mouse, pencil due to this extensive learning of feature representations for a wide range of images. Both VGG16 and VGG19 are pre-trained convolutional neural networks on the ImageNet dataset [20]. The networks consist of 16 and 19 layers respectively and having an input size of 224 by 224. Googlenet is a convolutional neural network that is a pre-trained having 22 layers of depth. It is trained on ImageNet dataset [20] and capable of classifying images into 1000 categories, such as mouse, pencil, keyboard and many animals. The network has an input size of 224 by 224.
In the proposed approach, the features are extracted by pre-trained models and output from the network is extracted from the 5th pooling layer. The principal component analysis is performed on the extracted features from the pre-trained models to reduce computations and followed by a support vector machine (SVM) classifier for action recognition. The block diagram of the proposed method is shown in figure 1 which gives the overview of the conventional transfer learning system, the first row indicates the source architecture and the second row shows the target.

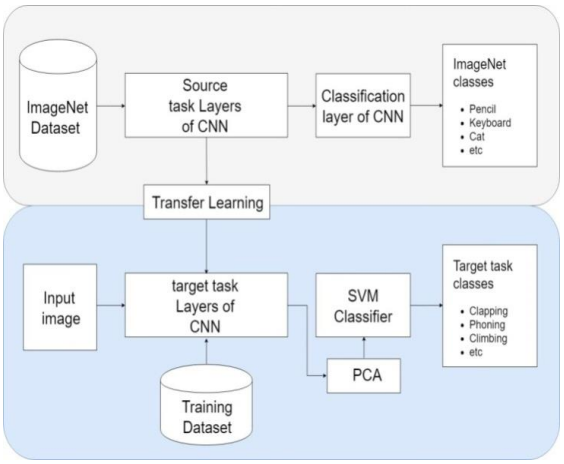

Fig. 1 Overview of the conventional transfer learning system.

In the proposed method, three major factors that constitute an action, human pose, most relatable object within the scene and overall scenario are considered. In order to include these factors three parallel networks are used followed by the feature fusion and convolutional neural network, classification is performed by SVM classifier.

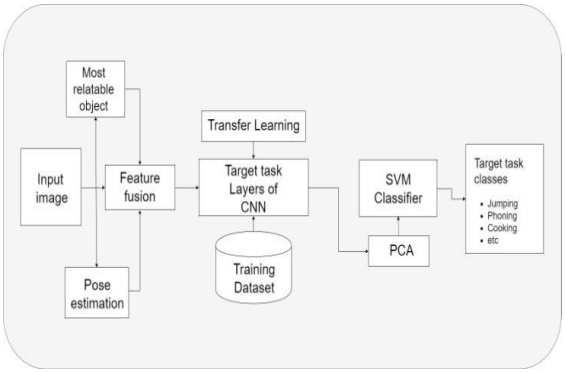

Fig. 2 The overview of the proposed method, feature fusion followed by CNN and SVM classifier.

Initially, the given input image Im , we use different networks to detect all humans, their poses and most relatable objects in the scene, creating a detected set of bounding boxes $D_b = (D_B^1, \dots, D_B^N)$ where N represents the total number of detected bounding boxes. The detected boxes for human and for the objects are represented as, D_{bH} and D_{bO} respectively and the detection confidence score for both are represented as S_h and S_o respectively. Human pose estimation and recognition for matching action are obtained by transfer learning from datasets [22][23].

The action prediction of the given image S_i is calculated for each candidate action, where a with dimension A includes all action classes, given each human-object-scenario bounding boxes (D_{bH}, D_{bO} and D_{bS}), where D_{bS} represents scenario bounding box which includes actors and all other objects to give an overall aspect of the scene a chance to play in the prediction score. S_i depends firstly on the individual confidence score of the actor S_h^i and object S_o^i , secondly human-object-scenario confidence score $S_{h,o,sc}^i$ and thirdly on pose feature representations S_p^i. The action prediction score is given as,

$$S_i = S_{h,o,sc}^i \cdot S_p^i \cdot (S_h^i + S_o^i + S_{sc}^i) \quad (1)$$

The sigmoid activation is utilized for classification to avoid competition between predicted classes. The training objective is to minimize the binary cross-entropy loss between action labels y and the predicted score $S_{ij} \in s$.

$$\mathcal{L}(s, y) = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{A}\left[y_{ij}\log(S_{ij}) + (1 - y_{ij})\log(1 - S_{ij})\right] \quad (2)$$

$$\mathcal{L}_T = \mathcal{L}(S_{h,o,sc}^i, y) + \mathcal{L}(S_p^i, y) + \mathcal{L}(S_h^i, y) + \mathcal{L}(S_o^i, y) + \mathcal{L}(S_{sc}^i, y) \quad (3)$$

Where $\mathcal{L}$ and $\mathcal{L}_T$ represents average cross-entropy loss on M sample batch and total cross-entropy loss respectively. y_{ij} is the action class for the ith action in the jth prediction and S_{ij} represents prediction score for the ith action. Figure 2 describes the proposed method with feature fusion followed by CNN network and SVM classifier.

4. Experiments and Results

In this section, we discuss the experimental setup, training process and results of the proposed method. The proposed method is tested on open source Stanford40 dataset [21]. The dataset contains 40 different human action images approximately 180 to 300 images per class, each image in the dataset has a bounding box of the person performing the action. In this paper for experimental purposes, only 10 classes are used to evaluate the proposed method on four different pre-trained models. Some samples from the dataset with classes that are used in the experiment are shown in figure 3.

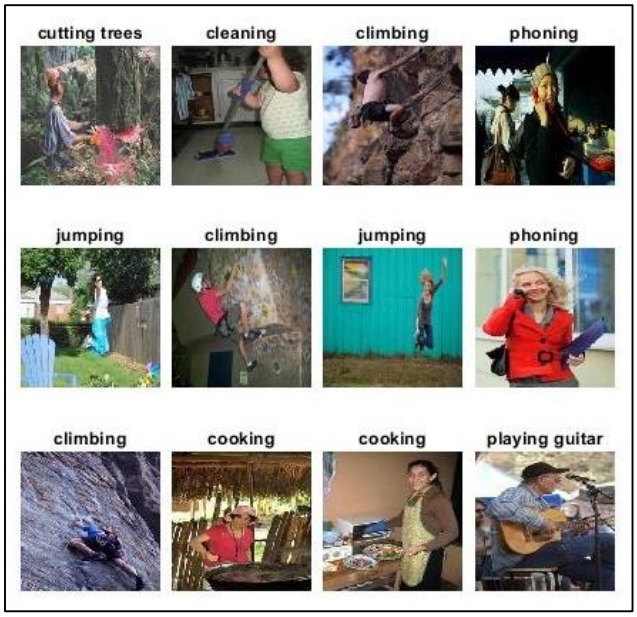

Fig. 3 Some Sample images from Stanford 40 dataset.

The feature extraction is performed by four different pre-trained networks having the same input layer size of 244 by 244 then principle component analysis performed on extracted features followed by SVM classifier to classify actions between 10 classes. The experimental results pre-trained model-wise are shown in table 1.

Table 1: Comparison of classification results on Stanford40 dataset

Methods	Resnet18	VGG16	VGG19	googLenet
Mean AP(%)	87.132	85.748	83.657	84.387

Now, the proposed method is tested on the same dataset, firstly the input image is processed through three different networks to find bounding boxes for human detection, pose estimation and object detection. The distance between all the objects detected in the scene and the human detected bounding box is calculated and the object which has minimum distance will be declared a most relatable object in the scene. Another network is used to estimate human pose to participate in action prediction score and the network utilizing previously learned knowledge is used followed by SVM classifier to detect overall scenario which includes actor and all the objects in the scene. Then finally all scores are interrelated in the decision fusion to provide a final decision. It is found that our method has performed better than conventional transfer learning methods and provide better accuracy of 86.413% Figure 4 shows some of the recognized actions by the proposed method. The mean AP comparison of the proposed method is shown in table 2. It illustrates that the proposed method achieves better results compared to the other existing methods.

Fig. 4 Classified actions from test dataset with true labels.

Table 1: Comparison of classification results on Stanford40 dataset

Methods	Mean AP(%)
Khan [24]	75.4
Semantic parts [25]	80.6
Image classification (VGG16 model)	81.4
Zhang [26]	82.6
Proposed method	87.1

5. Conclusion

In this paper, the human action recognition method is proposed based on three networks utilizing transfer learning by pre-trained Convolutional neural network architecture and SVM classifiers. The architectures of the pre-trained networks are used to determine human pose estimation, objects in the scene and overall scenario. Then followed by decision fusion where confidence scores of three different networks are related and the final decision is produced. It was established and demonstrated that transfer learning can be effectively used to utilize already learned knowledge for a new task in case of the small training dataset. Training of the deep learning model from scratch is computationally very high and time-consuming which can be avoided by using the transfer learning method. The performance of the proposed method was evaluated on stanford40 dataset and achieved 87.13% overall accuracy based on resnet18 pre-trained deep network.

References

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

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

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

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

[5] G. D. Guo and A. Lai, “A survey on still image based human action recognition”, Pattern Recognition, vol. 47, no. 10, pp. 3343–61, 2014.

[6] C.Szegedy,W.Liu, Y.Jia, and P.Sermanet, “Going deeper with convolutions,” aeXiv Prepr., 2014.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrel, “DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” Icml, vol. 32, pp. 647-655, 2014.

[8] S. Maji, L. Bourdev, and J. Malik, “Action recognition from a distributed representation of pose and appearance”, IEEE Int’l Conf. on Computer Vision and Pattern Recognition, 2011, pp. 3177–3184.

[9] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, “Efficient object localization using convolutional networks”, IEEE Int’l Conf. on Computer Vision and Pattern Recognition, 2015, pp. 648–656.

[10] V. Delaitre, J. Sivic, and I. Laptev, “Learning person-object interactions for action recognition in still images”, Advances in Neural Information Processing Systems, 2011

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

[12] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R*CNN”, IEEE Int’l Conf. on Computer Vision, 2015, pp. 1080–1088.

[13] G. Sharma, F. Jurie, and C. Schmid, “Expanded parts model for semantic description of humans in still images”, arXiv:1509.04186, 2015.

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

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

[16] Y. C. Su, T. H. Chiu, C. Y. Yeh, H. F. Huang, “Transfer Learning for Video Recognition with Scarce Training Data for Deep Convolutional Neural Network”, arXiv preprint arXiv:1409.4127, 2014

[17] H. Kaiming, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016

[18] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556 (2014).

[19] Szegedy, C,hristian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." *IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015

[20] *ImageNet*. http://www.image-net.org

[21] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei, “Human Action Recognition “, Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011

[22] R. Girshick, I. Radosavovic, G. Gkioxari, P. Doll´ar, and K. He, “Detectron,” https://github.com/facebookresearch/detectron, 2018.

[23] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in NIPS, pp. 199–207, 2015.
[24] 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", IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4422–4432, 2015.
[25] Z. Zhao, H. Ma, and X. Chen, "Semantic parts based top-down pyramid for action recognition", Pattern Recognition Letters, vol. 84, pp. 134–141, 2016.
[26] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts, IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5479–5490, Nov 2016.
[27] A. R. Siyal, Z. Bhutto, K. Saleem, A. S. Chan, M. L. Memon, M. H. Shaikh, S. Ahmed, "Ship detection in satellite imagery by multiple classifier network", International Journal of Computer Science and Network Security (IJCSNS), vol. 10, no. 8, pp. 142-148, Aug. 2019.
[28] Z. Bhutto, M. Z. Tunio, A. Hussain, J. Shah, I. Ali, and M. H. Shaikh, "Scaling of color fusion in stitching images", International Journal of Computer Science and Network Security (IJCSNS), vol. 10, no. 4, pp. 61-64, Apr. 2019.