Still Image Action Recognition Using Ensemble Learning

Hojat Asgarian Dehkordi, Ali Soltani Nezhad, Seyed Sajad Ashrafi, Shahriar B. Shokouhi*
School of Electrical Engineering, Iran University of Science and Technology, Tehran 16846-13144, Iran
h_asgariandehkordi@elec.iust.ac.ir, alisoltani7596@gmail.com, s_ashrafi@elec.iust.ac.ir, bshokouhi@iust.ac.ir

2021 7th International Conference on Web Research (ICWR) | 978-1-6654-0426-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICWR51868.2021.9443021

***Abstract*— In recent years, human action recognition in still images has become a challenge in computer vision. Most methods in this field use annotations such as human and object bounding boxes to determine human-object interaction and pose estimation. Preparing these annotations is time-consuming and costly. In this paper, an ensembling-based method is presented to avoid any additional annotations. According to this fact that a network performance on fewer classes of a dataset is often better than its performance on whole classes; the dataset is first divided into four groups. Then these groups are applied to train four lightweight Convolutional Neural Networks (CNNs). Consequently, each of these CNNs will specialize on a specific subset of the dataset. Then, the final convolutional feature maps of these networks are concatenated together. Moreover, a Feature Attention Module (FAM) is trained to identify the most important features among concatenated features for final prediction. The proposed method on the Stanford40 dataset achieves 86.86% MAP, which indicates this approach can obtain promising performance compared with many existing methods that use annotations.**

Keywords*— *Action Recognition, Still Image, Ensemble Learning, Feature Attention Module.

I. INTRODUCTION

In recent years, human activity classification has attracted many researchers in machine vision and pattern recognition. Human activity recognition is divided into two general categories, video activity recognition, and image activity recognition. In image-based action recognition, unlike video, the lack of temporal information creates more challenges. Therefore, developed methods in images deal with textual information [1]. Nowadays, identifying human action in still images plays a crucial role in applications such as image tagging [2], video and image analysis [3], and human-computer interaction [4].

Early works in still image action recognition utilized hand-crafted features for classification tasks [1]. In these methods, the computational costs were lower than new techniques, but they did not achieve acceptable accuracy in practice.

With the advent of Convolutional Neuronal Networks (CNNs) and their ability to extract discriminative features [5],[6], researchers have used them in different machine vision fields. Lack of training samples makes it hard to train a CNN from scratch. That is why researchers use pre-trained networks in this field.

Deep learning-based methods in still image action recognition also consist of different strategies. With the development of keypoint detection techniques, many methods categorize activity by estimating human position [7]. Some works as [8] detect objects and humans in the image to analyze their relationship. In [9], to avoid cluttered background, a human bounding box is applied to make a decision only based on the human region. Although these methods are more accurate than the early works, they always require image annotations which are time-consuming and costly to prepare. This paper presents a deep learning based approach without any annotation require

集成学习：将数据集拆分成不同的子集，每个子集预先训练一个CNN，最后通过一个注意力模块加权所有的feature,最后分类

In the proposed method, ensemble learni still image action recognition. In observatio often performs better on a subset of the dataset compare with the entire dataset. Therefore, in the first step, the dataset is decomposed into four subsets, then a lightweight CNN is trained for each one. Thus, each CNN specializes in feature extraction and classification on its own subset. Then, the final convolutional feature maps of these networks are concatenated together. In the second step, the images in subsets one to four are tagged one to four respectively and employed to train the Feature Attention Module (FAM). For an input image, the FAM generates coefficients to indicate fruitful features between concatenated features for final classification. In the third step, the concatenated features are multiplied by outputs of the FAM. These features are passed to fully connected layers subsequently.

In the continuation of this article, in section two, the related works will be reviewed. The third section describes the proposed method. The fourth section represents experimental results, and finally, this article is concluded in the fifth section.

II. RELATED WORKS

Presented methods in recognizing human activity in still image can be divided into two groups: traditional and deep learning-based methods.

A. Traditional Methods

These methods were mainly used before the development of deep learning techniques. Delaitre et al. [10] exploited features with the SIFT descriptor and classified them with the support vector machine (SVM). Yao and Fei Fei. [11] introduced grouplet technique to extract image features and employed SVM for categorization. Some of the works used HOG; for example, [12] has employed the HOG to recognize human-object interaction. X. Peng et al. [13] utilized BoVW to achieve image features and gave them to supported vector machine. Besides, many researchers have classified human action by different cues. Prest et al. [14] aggregated features from human-object interaction, scene, and human position

978-1-6654-0426-6/21/$31.00 ©2021 IEEE

and determined the action category with the support vector machine.

B. Deep Learning-Based Methods

. There is a variety of approaches in deep learning-based methods. M. Hoai et al. [15] investigated different image regions with an eight-layer network. In every action category, some of the human body parts have a unique appearance. Gkioxari et al. [16] detected body parts and captured action scores obtained from these parts. The methods presented in [17] and [18] provide human pose information to anticipate human actions. Also, some works have investigated the interaction between human and object for action recognition. Sarah Pratt et al. [19] first detected the location of human and objects. Second, used an LSTM network to recognize human-object interaction. Many works in this field have utilized multiple features to perform accurately. For this purpose, multibranch structures have been organized. Zheng et al. [20] introduced a three-branch network termed SAAM NET that generated the action mask and incorporated information from the action mask and scene.

Also, [1] represents the joint body regions through convolutional features. Besides determines the human skeleton using the LDA technique and finally predicts the action according to convolutional features and the joint areas and human skeleton. After recognizing humans and objects in the image, [21] determines the point of interaction between human and object candidates and computes the score of each candidate to final classification. A two-part network is presented in [22]. The first part divides the human body into seven parts (head, arm, etc.) and extracts each part's features. The second one is used to analyze the human body. In the end, the outputs of these parts are incorporated into action recognition. [23] considers human activity identification as a multitask procedure and utilizes the scores of image-based action recognition, attention-based action recognition, and body part-based action recognition to make the final decision.

In some works, such as the work of Mohammadi et al. [24], In order to avoid additional annotations, several networks were trained individually on the proposed dataset. Then the average of output probabilities was calculated to construct the final probability vector. One noticeable point is that the networks in this article have a lot of parameters and computational costs.

III. PROPOSED METHOD

As shown in Fig. 1, the proposed method consists of three main parts. First, the convolutional layers of four CNNs trained on a special part of the data set construct a feature extraction module. Second, a system is designed to neutralize less important features among the exploited features. Finally, the classifier will be discussed in the third part.

A. Constructing Feature Extraction Module

In this ensemble learning method, instead of training several different architectures on all dataset classes, four identical CNNs are specialized on a particular subset of the dataset. For this purpose, the dataset is split into four subsets, and each subset adopts one of the CNNs. The architecture of each CNN discussed in this stage is shown in Fig. 2.

In this architecture, the features exploited by the Efficient NET B0 [25] are first passed through Global Average Pooling (GAP) to reduce the number of network parameters. In the following, three fully connected layers with 512, 256, and 10 neurons are organized, respectively. A dropout technique with a rate of 0.4 is proposed to prevent over-fitting for each fully connected layer. At the end, a softmax layer calculates the probability of each output. These probabilities are computed by (1).

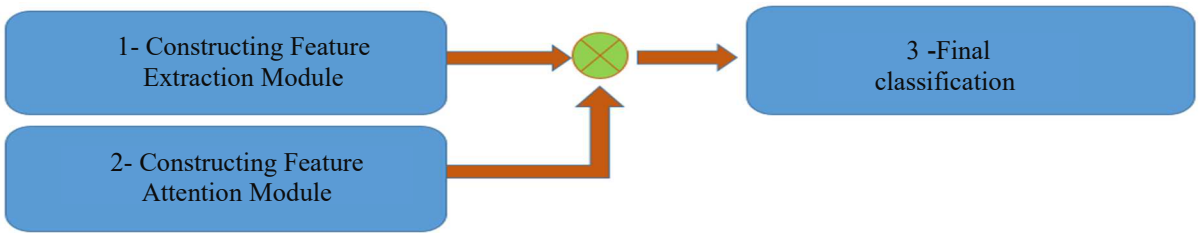

Fig. 1. Overflow of the proposed meth

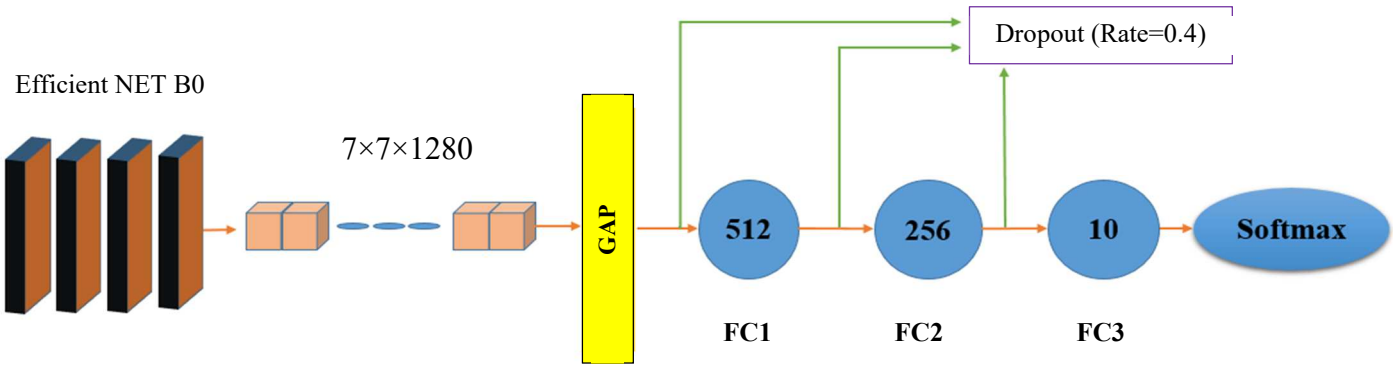

Fig. 2. The architecture of networks (networks one to four) to construct the Feature Extraction Module

$$P(Z_i) = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \qquad (1)$$

Which Z_i is the i-th output and K demonstrates the total number of outputs.

As Fig. 2 indicates, the final convolutional feature maps are tensors with the size of 1280×7×7. To ensemble these four CNNs, these feature maps are concatenated to create a tensor with the size of 5120×7×7. Passing this tensor through a Global Average Pooling (GAP) layer, it converts to a feature vector with a size of 5120×1×1. Therefore, the final classification can be done by this vector.

B. *Constructing Feature Attention Module*

According to experiments, due to the high dimensions of the feature vector (5120 dimensions) generated by four networks and the lack of training data, fully connected layers cannot categorize accurately only with this vector. To address this limitation, when the final convolutional feature maps are conca-tenated together, a coefficient is multiplied to them for neutralizing low-importance features. To generate these coefficients, the activities in subsets one to four are labeled one to four, respectively, and use for training the FAM shown in Fig. 3.

The FAM aims to distinguish the associated expert CNN for each input image among four CNNs by assigning one to the distinguished network and zero to the other networks. This module reaches 83.4% MAP on the test data. Note that the output of this module determines the coefficients of the features are used for final classification. In this case, the coefficients are assigned correctly only with a probability of 83.4%. On the other hand, if we consider the second output probability as the input image class (1, 2, 3, and 4), 11.6% MAP is obtained. Therefore, considering both outputs (first and second probability) as input image class, in 95% of test images, the image class is among the predicted classes by the FAM. Therefore, in the FAM, two outputs with more probability are set to one, and two outputs with less probability are set to zero. This procedure is demonstrated in Fig. 3.

C. *Training Classifier*

In this step, the four outputs of the FAM are multiplied by the final convolutional feature maps of the CNNs one to four, respectively. At the end, the final classification is organized by these features. The architecture of the proposed method is demonstrated in Fig. 4.

IV. Experiments

The method described in this article has been implemented in the Google Colab environment with the Pythorch library [26]. In all networks in this paper, the pre-trained Efficient NET B0 [25] with the initial ImageNet weights is used as the backbone. Evaluation of this method have been done on Stanford 40 dataset [27]. This dataset contains 9545 images in diverse 40 action classes, 4000 images for the train, and 5545 images for the test.

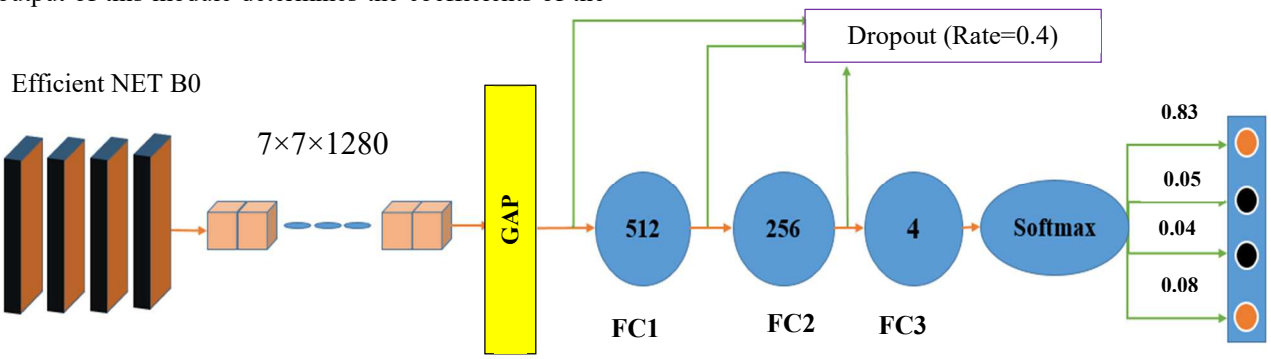

Fig. 3. The architecture of the FAM

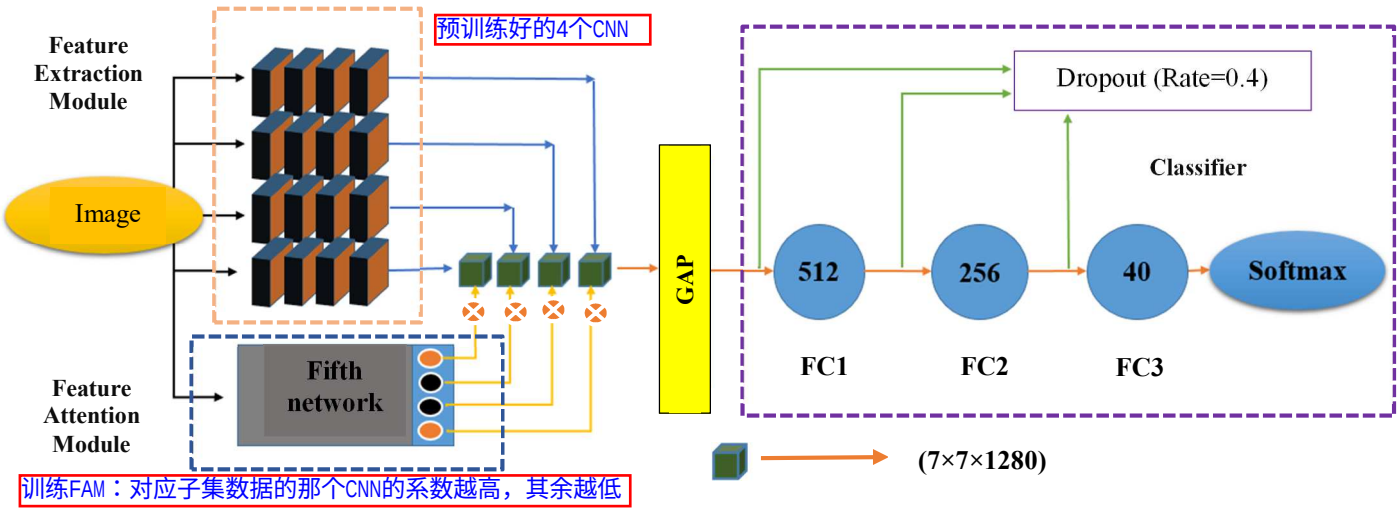

Fig. 4. The overall architecture of the proposed method.

A. Implementation Details

The images were resized to 224 × 224 pixels to implement the proposed method. In order to prevent overfitting, random resize crop and random horizontal crop were used as data augmentation techniques. The optimizer was SGD [28], and the learning rate started to change from the initial value of 0.01 and multiplied by 0.1 in every ten epochs. The value of momentum was set to 0.9. The Cross-Entropy loss function calculated the error. The formulation of this function is represented in (2).

$$Cross\ Entropy = -\frac{1}{N}\sum_j Tr_i \times \log Pa_i \quad (2)$$

where Pa,Tr, and N denote, the prediction, the ground truth, and the number of training examples, respectively.

The training process was done in 100 epochs with a batch size of 40. Note that these implementation details are valid for all of the networks presented in this paper.

B. Results

The performance of networks one to four on their respective test data is shown in Table. I. Besides, the Efficient net B0 [25] performance over the entire Stanford40 [27] is shown in this table.

As it can be seen in Table.I, the accuracy of each CNN on its own subset has increased compared to the entire dataset. Due to the low between-class difference in the dataset, a CNN can not generate discriminative features for the entire dataset. But for a subset of the dataset, the feasibility of generating discriminative features will improve.

Also, comparison result with other methods on the Stanford40 [27] is provided in Table. II. Other reported methods in this table need annotations, but our method achieves better operation without any annotation usage. A comparison of the presence and absence of the FAM is also provided. From this comparison, it can be concluded that the FAM can significantly boost the performance. It should be noted that the proposed architecture includes only 26 million parameters which is less than many presented architectures in this area.

TABLE I. PERFORMANCE RESULTS OF NETWORKS ONE TO FOUR ON EACH SUBSET OF STANFORD 40

Network	***Classes***	***MAP***
1	**1-10**	**93.9%**
2	**11-20**	**98.3%**
3	**21-30**	**97.8%**
4	**31-40**	**90.5%**
Efficient Net B0 [26]	**1-40**	**83.5%**

TABLE II. COMPARISION OF OUR RESULTS WITH PREVIOUS WORKS.

Method	***MAP%***
Zhang et al. [28]	82.64
Sharma et al. [29]	72.3
He et al. [30]	81.2
Jia et al. [31]	82.6
Girshick et al. [32]	85.3
Safaie et al. [33]	80.9
Safaie et al. [34]	81.76
Ours without FAM	81.5
Ours	**86.85**

V. CONCLUSION

In this paper, an ensemble learning method was proposed to categorize human activity in still images. Unlike most existing methods in this field, it does not require annotations such as human and object bounding boxes. To this end, four lightweight CNNs were trained on a special part of the dataset. A Feature Attention Module (FAM) was subsequently adopted to distinguish the most important features generated by these four CNNs. The results show the competitive performance of this method compared with many methods that employ annotations. In addition, as a suggestion to continue working, knowledge distillation techniques can be utilized to reduce the number of network parameters.

REFERENCES

[1] Y. Li, K. Li, and X. Wang, "Recognizing actions in images by fusing multiple body structure cues," *Pattern Recognition*, vol. 104, p. 107341, 2020, doi: 10.1016/j.patcog.2020.107341.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *in International conference on machine learning*, 2015, pp. 2048–2057.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *in Proceedings of the IEEE international conference on computer vision,* 2015, pp. 4489–4497.

[4] P.-Y. P. Chi, Y. Li, and B. Hartmann, "Enhancing cross-device interaction scripting with interactive illustrations," *in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 5482–5493.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Advances in neural information processing systems*, 2012, pp. 1097–1105.

[6] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[7] V. Belagiannis, A. Zisserman, "Recurrent human pose estimation," in: *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 468–475

[8] B. Yao, L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," *in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2010, pp. 17–24.

[9] L. Liu, R. T. Tan, and S. You, "Loss Guided Activation for Action Recognition in Still Images," *in Lecture Notes in Computer Science*, 2019, vol. 11365 LNCS, doi: 10.1007/978-3-030-20873-8_10.

[10] V. Delaitre, "Recognizing human actions in still images : a study of bag-of-features and part-based representations," 2010, doi: 10.5244/C.24.97.

[11] B. Yao, "Grouplet : A Structured Image Representation for Recognizing Human and Object Interactions." *IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA*, pp. 9–16, 2010.

[12] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing Human-Object Interactions : Using Spatial and Functional Compatibility for Recognition," *Pattern Analysis and Machine Intelligence*, vol. 31, pp. 1–15, 2009.

[13] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding,* vol. 150, pp. 109-125, 2016/09/01/ 2016.

[14] M. Hoai, "Regularized Max Pooling for image categorization," *BMVC 2014 - Proc. Br. Mach. Vis. Conf. 2014*, pp. 1–12, 2014, doi: 10.5244/c.28.32.

[15] G. Gkioxari, B. Hariharan, R. Girshick, J. Malik, and B. Berkeley, "Using k -poselets for detecting people and localizing their keypoints," *IEEE Conference on Computer Vision and Pattern Recognition,* 2015, pp. 3582–3589.

[16] C. Cao, Y. Zhang, C. Zhang and H. Lu, "Body Joint Guided 3-D Deep Convolutional Descriptors for Action Recognition," *in IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 1095-1108, March 2018, doi: 10.1109/TCYB.2017.2756840.

[17] D. C. Luvizon, D. Picard and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5137-5146, doi: 10.1109/CVPR.2018.00539.

[18] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded Situation Recognition," in: *Computer Vision – ECCV 2020*, Cham, 2020, pp. 314-332: Springer International Publishing.

[19] Y. Zheng, X. Zheng, X. Lu, and S. Wu, "Spatial attention based visual semantic learning for action recognition in still images," *Neurocomputing,* vol. 413, pp. 383-396, 2020.

[20] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *presented at the Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2019. Available: http://proceedings.mlr.press/v97/tan19a.html.

[21] Wang, Tiancai and Yang, Tong and Danelljan, Martin and Khan, Fahad Shahbaz and Zhang, Xiangyu and Sun, Jian, "Learning Human-Object Interaction Detection Using Interaction Points," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2020.

[22] Zhao, Zhichen and Ma, Huimin and You, Shaodi, "Single Image Action Recognition Using Semantic Body Part Actions," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct, 2017.

[23] B. Bhandari, G. Lee, and J. Cho, "Body-Part-Aware and Multitask-Aware Single-Image-Based Action Recognition," *Applied Sciences,* vol. 10, no. 4, p. 1531, 2020.

[24] S. Mohammadi, S. G. Majelan and S. B. Shokouhi, "Ensembles of Deep Neural Networks for Action Recognition in Still Images," 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), 2019, pp. 315-318, doi: 10.1109/ICCKE48569.2019.8965014.

[25] A. Paszke *et al*, "PyTorch : An Imperative Style , High-Performance Deep Learning Library," no. NeurIPS, 2019, http://arxiv.org/abs/1912.01703.

[26] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei, "Human Action Recognition by Learning Bases of Action Attributes and Parts," *Internation Conference on Computer Vision (ICCV)*, Barcelona, Spain. November 6-13, 2011.

[27] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177– 186.

[28] Y. Zhang, L. Cheng, 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, Nov. 2016, doi: 10.1109/TIP.2016.2605305.

[29] G. Sharma, F. Jurie and C. Schmid, "Expanded Parts Model for Semantic Description of Humans in Still Images," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 87-101, 1 Jan. 2017, doi: 10.1109/TPAMI.2016.2537325.

[30] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition" *CVPR*, 2016.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S, "Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding," In: *arXiv*: 1408.5093, 2014.

[32] R. Girshick, "Fast r-cnn," in 2015 IEEE International Confer- 811 ence on Computer Vision (ICCV), Dec 2015, pp. 1440–1448.

[33] M. Safaei, P. Balouchian and H. Foroosh, "TICNN: A Hierarchical Deep Learning Framework for Still Image Action Recognition Using Temporal Image Prediction," *25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 3463-3467, doi: 10.1109/ICIP.2018.8451193.

[34] M. Safaei and H. Foroosh, "Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2019, pp. 111-120, doi: 10.1109/WACV.2019.00019.