Adaptive Frame Selection In Two Dimensional Convolutional Neural Network Action Recognition

Alireza Rahnama
Department of Electrical and Computer Engineering
Faculty of Engineering
Kharazmi University
Tehran, Iran
Alireza.rahnama@khu.ac.ir

Alireza Esfahani
Department of Electrical and Computer Engineering
Faculty of Engineering
University of Science and Technology
Behshahr, Iran
A.esfahani@Mazust.ac.ir

Azadeh Mansouri
Department of Electrical and Computer Engineering
Faculty of Engineering
Kharazmi University
Tehran, Iran
A_mansouri@khu.ac.ir

***Abstract*—We presented a technique in this research for dynamic frame selection to achieve robust features. This situation results in less redundancy and useful input for the network. Because it uses fewer processing resources and offers adequate accuracy, the suggested technique is appropriate for real-time applications. The network becomes more efficient and maintains adequate accuracy when informative frames are chosen and computation is minimized. The framework is tested on UCF101 as one of the large and realistic datasets. The experiments show acceptable results employing both Resnet-50 and Mobilenet pre-trained features.**

Keywords— action recognition, convolutional neural networks, real-time applications, dynamic frame selection,2D-CNN

I. INTRODUCTION

The importance of video-understanding tasks is demonstrated by the rising demand for video-based applications. Due to the expansion of video applications, video transmissions account for the majority of internet traffic. Activity detection has evolved into one of the most crucial aspects of computer vision due to the usability of video in numerous projects such as visual surveillance [1], industrial environments to smart homes [2], and sports analysis. However, action recognition from RGB data is often challenging, owing to the variations of backgrounds, viewpoints, scales of humans, and illumination conditions. Besides, RGB videos have generally large data sizes, leading to high computational costs when modeling the Spatio-temporal context for Human Action Recognition [3]. For various uses, such as surveillance video analysis is necessary. Processing each of the next frames separately would be inefficient because films of this type contain an infinite number of frames that are very identical to one another In summary, the main contributions of this paper are as follows: In this study, the frame selection scenario is employed instead of frame by frame processing. It is also recommended to employ the similarity measure mechanism to provide a dynamic frame selection strategy for obtaining reliable and long-term feature vectors. In this scenario, far fewer frames are processed and operate more effectively.

II. RELATED WORK

Since the main goal of the proposed method is selecting informative frames, this section mainly focuses on frame selection and ROI identification as the preprocessing of HAR systems. The techniques for choosing frames can be categorized into three categories: deep learning techniques, handmade feature extraction techniques, and hybrid techniques. The following most recent approaches are explained.

To choose a small set of frames to produce accurate predictions and reduce the computational cost, AdaFrame, a dynamic technique was put forth by Wu et al [4] to eliminate extraneous video data for video recognition. The most essential frames were chosen by the authors using an LSTM network. ResNet [5] is one of the CNN-based techniques that is trained to automatically select frames for video summaries. [6] The author of [7] suggests an adaptive visual tracking method based on keyframe selection and reinforcement learning (RL). Since in an input video, not all of the frames are beneficial for the performance of the prediction. In [8] the authors mainly propose an activity prediction method called key frame selection (KFS) by exploiting the keyframes which are represented by concurrent mid-level action units. The method presented in [9] captures neighborhood information around the point where maximum motion is identified. The presented approach provides acceptable results for two-person interaction, according to the experiments. HAR's video recognition system performs well when using deep learning techniques such as LSTM-based approaches, but they demand a complex training phase. Moreover, In terms of applications, and particularly in terms of the kinds of activity undertaken in HAR systems, the hybrid, and handcrafted techniques are typically restricted.

In [10] neural networks exclusively applied to accumulated residuals in the compressed domain for accelerated performance. In this scenario, the similarity of the residuals is calculated in order to perform better accumulation and frame augmentation. The idea of using augmented frames is employed also in [11]. The informative frames are selected in this paper

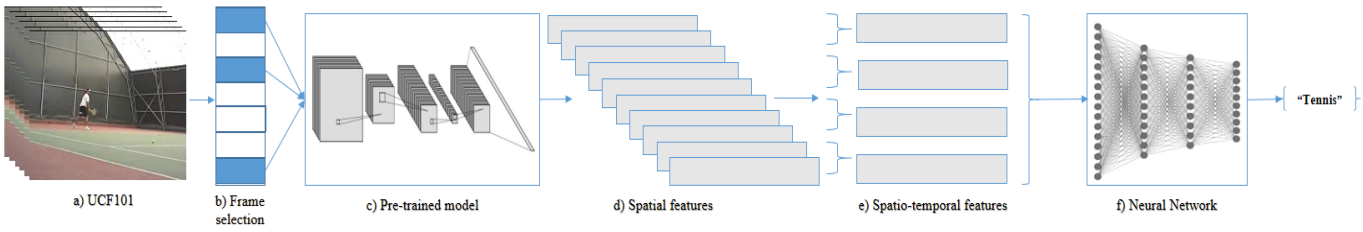

Fig. 1. This image illustrates all parts of our framework. (a)UCF101 dataset (b) frame selection algorithm (c)ResNet-50 and MobileNet (d)Spatial features are extracted by pre-trained model (e)temporal features are extracted with max-pooling (f) neural network for classification

in RGB frames. The recommended method incorporates adaptive thresholding for the selection of frames and uses a similarity measure criterion. The HAR methodology is trained on the chosen frames to assess the success of the suggested approach, and the outcomes are compared to a frame-by-frame scenario

III. Proposed Method

A. *Frame selection algorithm*

A huge number of video frames in the received movies are similar to their prior frames, and the change between them is negligible. These frames have little impact on the action recognition process, cause more processing and create redundant data. For instance, when the camera is fixed and does not move, a significant portion of the received frames consist just of the place's background as captured by the camera, with no action occurring there. These duplicate items may be ignored for detection and processing with the aid of this method, duplicate, and similar items can be deleted, and only different frames can be distinguished from the others in which the activity has been performed, and work with these frames. The first frame of the video is selected for this algorithm, and after that, the second frame is obtained. The similarity between these two frames is then calculated using the formula (1), and the result is the similarity matrix. By averaging this matrix, a scalar similarity index is archived, which is between 0 and 1. The closer this value is to 1, it means that these two frames have many similarities, and the closer to 0, illustrates that these two frames are very different. this similarity criterion is used as a frame selection procedure. The similarity frame statement is depicted in the following formula, in which $F_{(i)}$ is the current frame and $SF_{(i)}$ is the last selected frame [10]:

$$Frame - Similarity = \frac{2 \times F_{(i)} \times SF_{(i)}}{F_{(i)}^2 + SF_{(i)}^2} \qquad (1)$$

After selecting the second frame, the output of formula (1) is a similarity frame. We average the similarity frame obtained from the formula (1) to have a scaler as a similarity which is $FM_{(i)}$ for each selected frame and keep it in an array. For the next frames, we obtain the average of the total values in this array which is called Mean window in formula (2) to examine each new frame with the average of all the selected previous frames average in such a way that, if the value obtained from the formula(2) is smaller than the total similarity averages, it indicates that the new frame is different from the previous frames and must be selected. Even so, if the result of the formula(2) is greater than the total of the averages, this frame is deleted because it is too similar to the previously chosen frames. Hence, we choose a frame that has a sensible difference from all the frames we've already chosen.

$$Mean\ window = \frac{\sum_{i=0}^{n} FM_{(i)}}{n} \qquad (2)$$

B. *Spatio-temporal features*

Transfer learning is a technique where we use a pre-trained model with additional data to solve a different problem. A model is completely trained on a huge dataset in this manner and we utilize this model to find the features of a second problem or to solve smaller problems. Transfer training has a number of benefits, the most significant of which is an increase in model training speed because the model is only used for prediction and we do not train it completely. Transfer learning also reduces error because it considers the more crucial features according to the number of layers and parameters because we use the weights from the first problem for the second problem. The objective is to extract data from the first model that could be helpful for directly predicting the second model. in the paper we use ResNet-50 and MobileNet [12] as Pre-trained models to predict spatial features of data Then in order to extract the Spatio-temporal feature of the pooling strategy is employed to form the video level feature.

We use transfer learning to obtain the spatial features of the selected frames with the help of the algorithm and use these extracted data for the next step temporal pooling data augmentation allows us to have more information, we split up each video's vectors into 4 portions, after which we extract video-based features via temporal pooling [13]. This can provide more data for the training phase.

$$X^{\max(t^s,t^e)} = maxf\big(i(t)\big) \quad t = [t_s, \dots, t_e] \qquad (3)$$

As it is illustrated in the above formula, the video-level representation is created by combining per-frame descriptors with PoT. In essence, PoT representation enables tracking of feature descriptor changes. in order to data augmentation, four Spatio-temporal features are extracted for each video. These vectors are employed to train the classifier.

IV. Experimental Results

In this section, we analyze and classify the realistic dataset UCF101. as it is mentioned, the pre-trained networks are employed for spatial feature extraction. in fact, two efficient ResNet-50 and MobileNet are selected as two pre-trained networks. then the spatial features are pooled to prepare the video-based features for classification.

A. Dataset

UCF101 is considered for data analysis, UCF101 is an action recognition data set with more than 13000 realistic action videos, collected from YouTube, UCF101 consists of 101 action classes and 27 hours of video data. The database consists of realistic user-uploaded videos containing camera motion and cluttered backgrounds [14]

B. Results

The advantages of using frame selection are shown in Table 1. In the first line, we have put the total number of dataset frames and the total number of frames that were selected after applying the frame selection algorithm, the second line is the time required to obtain spatial features with pre-trained networks . Finally, with the help of our frame selection algorithm, by selecting 31% of important video data and discarding 69% of similar data, we have reached good accuracy in a shorter time. Additionally, in figure 2, we have chosen two videos at random from each class in the dataset to demonstrate how many frames from each video will be reduced using the method. The ability of pre-trained networks to extract the right results without a laborious training process The extracted spatial feature is considered to be the output of the layer before fully connected one in two efficient pre-trained networks Resnet-50 and Mobile net. The feature vectors of size 2048 and 1024 are extracted employing ResNet-50 and MobileNet respectively. then after dividing the length of each video into parts and extracting temporal features; We obtained 4 feature vectors for each video. in the training phase. A simple fully connected network, containing fully connected layers is used for classification. A feature vector of size N is fed into the network and two consecutive layers of size N/2 and N/8 are considered respectively. The ELU is considered as the activation function with 0.2 drop out for each layer. Ninety percent of the total number of videos per class are considered as the training and validation data and the rest are employed for the test.

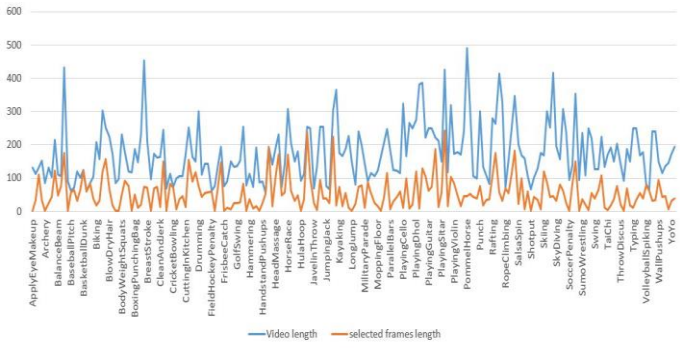

Fig. 2.Illustration of the different lengths between selected frames and the original video length

TABLE I.
COMPARISON OF TWO SCENARIOS: WITH OR WITHOUT FRAME SELECTION

Measurements	Selected frames	All frames
Total Frames	732,477	2,465,430
Time Spend Average	12.1650 Sec	37.3013 Sec
Selected Frames Average	55.1831	186.5065
percentage of the selected frame	31.2048 %	100%

TABLE II.
RESULTS WITH A DIFFERENT PRE-TRAINED MODEL ON UCF101

Pre-trained models	All frames	Selected frames
ResNet-50	98.37%	98.05%
MobileNet	97.68%	97.70%

V. Conclusion

Intelligent video-based systems require computational load, hardware cost, and memory consumption. Processing time could all be decreased instantaneously by eliminating some frames. The strategy for collecting keyframes and adaptively cropping input video for human action recognition (HAR) systems is recommended in this research based on the previous concerns. The suggested approach uses a similarity measure criterion and introduces adaptive thresholding for the frame selection. To evaluate the effectiveness of the proposed method, the HAR technique is trained on the selected frames, and compares the results achieved with frame by frame scenario. The results illustrate acceptable performance both in accuracy and computational costs.

References

[1] S. Dubey, A. Boragule, J. Gwak, and M. Jeon, "Anomalous event recognition in videos based on joint learning of motion and appearance with multiple ranking measures," *Appl. Sci.*, vol. 11, no. 3, pp. 1–21, 2021, doi: 10.3390/app11031344.

[2] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-Time Action Recognition with Enhanced Motion Vector CNNs," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2718–2726, 2016, doi: 10.1109/CVPR.2016.297.

[3] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human Action Recognition From Various Data Modalities: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 952215, pp. 1–20, 2022, doi: 10.1109/tpami.2022.3183112.

[4] Z. Wu, H. Li, C. Xiong, Y. G. Jiang, and L. S. Davis, "A Dynamic Frame Selection Framework for Fast Video Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1699–1711, 2022, doi: 10.1109/TPAMI.2020.3029425.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[6] K. Pretorious and N. Pillay, "A Comparative Study of Classifiers for Thumbnail Selection," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9206951.

[7] K. Zhao, Y. Lu, Z. Zhang, and W. Wang, "Adaptive Visual Tracking Based on Key Frame Selection and Reinforcement Learning," *Proc. - 2020 Int. Work. Electron. Commun. Artif. Intell. IWECAI 2020*, pp.

160–163, 2020, doi: 10.1109/IWECAI50956.2020.00039.
[8] H. Wang, C. Yuan, J. Shen, W. Yang, and H. Ling, “Action unit detection and key frame selection for human activity prediction,” *Neurocomputing*, vol. 318, pp. 109–119, 2018, doi: 10.1016/j.neucom.2018.08.037.
[9] M. Poonkodi and G. Vadivu, “Action recognition using Correlation of Temporal Difference Frame (CTDF)—an algorithmic approach,” *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02378-0.
[10] A. Abdari, P. Amirjan, and A. Mansouri, “Speeding Up Action Recognition Using Dynamic Accumulation of Residuals in Compressed Domain,” *SSRN Electron. J.*, pp. 1–16, 2022, doi: 10.2139/ssrn.4204346.
[11] A. A. Gharahbagh, V. Hajihashemi, M. C. Ferreira, J. J. M. Machado, and J. M. R. S. Tavares, “Best Frame Selection to Enhance Training Step Efficiency in Video-Based Human Action Recognition,” *Appl. Sci.*, vol. 12, no. 4, 2022, doi: 10.3390/app12041830.
[12] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2017, [Online]. Available: http://arxiv.org/abs/1704.04861.
[13] M. S. Ryoo, B. Rothrock, and L. Matthies, “Pooled motion features for first-person videos,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, no. Figure 1, pp. 896–904, 2015, doi: 10.1109/CVPR.2015.7298691.
[14] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” no. November, 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.