Grenze International Journal of Engineering and Technology, Jan Issue

Smart Surveillance System using Deep Learning

Mehul Kanotra[1] and Dr. Reena Gupta[2]
[1-2]University School of Information, Communication and Technology, Delhi, India
Email: mehulkanotra98@gmail.com, reena@ipu.ac.in

***Abstract*—Traditional closed circuit television (CCTV) requires that a human actively and closely monitoring the output feed of the camera 24/7, which is inefficient and costly. Therefore, there is a need for a smart system that can monitor the feed of a CCTV camera and recognize human actions in real time with little or no human intervention. In this paper, we put forward a smart surveillance system, where images from various devices (for instance a colored CCTV camera) can be fed into, and in these images, human agents are detected using an object detection algorithm and annotated with a bounding box. The human subjects in these bounding boxes are then consequently fed into a Deep Learning model to classify activities. Hence, we propose a pipeline of first detecting and then classifying human activities in images captured by surveillance devices such as a CCTV camera. Our work therefore can be divided into two parts, object detection and classification, for detection, we utilize the state of the art object detection algorithm YOLOv3 and for classifying the human activities we train a deep learning model on the Stanford 40 action dataset. The dataset includes simple human actions such as jumping and running to more complex and composite actions such as playing guitar or riding a bike. We also compare the performance in terms of accuracy of our base CNN model with other approaches, which are trained for human activity recognition on the Stanford 40 action dataset. Our deep learning model obtained the maximum classification accuracy of 87% on parent images obtained from the Stanford 40 action dataset and a classification accuracy of 66% on images where human subjects were extracted from the parent images.**

***Index Terms*— Human activity recognition; CNNs; Smart surveillance system; deep learning; transfer learning.**

I. INTRODUCTION

Human activity identification plays a crucial role in a variety of computer vision applications, for instance context based video retrieval [1], human computer interaction [2], image and video analysis [3] and image notation [4]. Hence human action recognition has been an area of active interest in the field of computer vision for many years. Despite the fact that a variety of action identification approaches have been put forward over the last decade, action recognition in images remains difficult due to viewpoint disparities, complex backgrounds, and changes in human stance.

The earlier technique for action identification in still images was the Bag of Visual Words (BoVW) [5, 6, 7] which was used to acquire a global representation of an image. Delaitre et al [8]. used a BoVW and an SVM classifier to classify human actions. For recognising human activities, some systems modeled human-object interactions. To mimic human-object interaction, Yao et al [9] employed pose information and objects. Prest et al. [10] used a weakly supervised learning scenario to learn about the interaction between items and humans.

Grenze ID: 01.GIJET.9.1.25

Some research utilized part-based techniques to recognise human motions, which blend global data with properties of distinct body parts [11, 12].
In recent times, because of the extraordinary performance of Convolutional Neural Networks (CNNs) in computer vision [13, 14, 15] and their effective feature extraction capacity from raw images, Deep Learning has emerged as a feasible technique for detecting human activities. CNNs and Poselets were utilized by Gkioxari et al. [12] to characterize human acts and attributes. By simply retraining the classification module of a pre-trained network, Oquab et al. [16] used the transfer learning technique. For identifying actions in still images, Yan et al. [17] used the VGG16 [18] network with two additional attention branches. Mohammadi et al. [19] employed an ensemble method for identifying human behaviors, relying on attention processes on CNN's output feature maps to extract more appropriate features for classification..
However, one additional thing to consider is that CNNs require a huge amount of labeled dataset for recognising actions in order to avoid overfitting, thus the lack of large enough dataset can itself be a problem. By leveraging CNNs pre-trained on ImageNet [22], we use the transfer learning strategy to address the lack of large labeled action identification datasets in this study..
One of the notable applications of human activity recognition is in the field of intelligent surveillance. In traditional closed circuit television (CCTV), a human actively and closely monitors the output feed of the camera 24/7 which is inefficient and costly. As a result, human activity recognition can be used to help with CCTV camera monitoring by recognising and identifying potentially hazardous or dangerous actions caught by the cameras. However, development of an algorithm to aid human activity recognition for the purpose of creating smart surveillance has many inherent challenges such as the requirement of a huge dataset for images. One additional challenge arises when images captured from devices such as a CCTV camera feed contains multiple human subjects performing different actions and activities, hence it becomes crucial to detect and classify each of these subjects separately and accurately.
In this paper, we propose a pipeline that can perform as an intelligent surveillance system where we firstly detect human subjects in a complex environment and then classify the actions performed by these subjects. To detect multiple human subjects separately in a single frame we use object detection algorithm. We perform object detection by leveraging state of the art YoloV3[20] object detection algorithm to distinguish and separate numerous human agents that can be present in a single RGB image. By performing this, we can then apply human activity recognition algorithm separately on these detected human agents. This has an added advantage when used for classifying human actions for a smart CCTV system which can capture multiple human subjects in a single frame. In order to classify human actions performed by detected human subjects, we train a CNN model on the Stanford 40 [21] action dataset by utilizing transfer learning and by adding a custom ANN layer to the trained CNN base. Multiple deep learning models are trained on this approach and their accuracy is tested on the test images from the dataset. Then in order to test the validity of the model being used as a backbone of a smart surveillance system, the test images were fed into the YoloV3 object detection model to detect and crop out various human agents performing various actions. These cropped out images were then fed into a previously trained deep learning model to determine the accuracy of model on such cropped out images. Hence, our contributions are three as follows:

- Put forward a pipeline that can be utilized as a smart surveillance system.
- To deal with the paucity of labeled data, we use the transfer learning technique.
- We make use of the YoloV3 object detection model to distinguish and crop out the images where the human agents are performing some activities and we execute human activity recognition on these images using the obtained deep learning models.
- We investigate performance of various deep learning as well as machine learning models on the dataset.

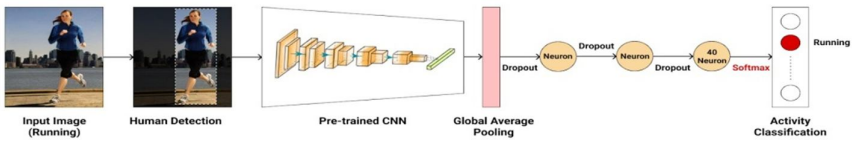

Fig. 1: The proposed method's overall pipeline. Our framework comprises of two main module: i) The Object Detection component which detects and creates a bounding box around the detected human subject and ii) the Classification role which uses the cropped image containing the detected subject as input and then proceeds to generate the probabilities for 40 classes contained in the Stanford 40 dataset. One thing to note is that all fully connected layers are removed in the Pre-trained CNN

II. Proposed Approach

The pipeline we propose can be divided into two parts as depicted in Figure 1.: First, detection of humans performing actions on an image that is captured by a simple camera or by a CCTV feed, which is done using the YOLO object detection algorithm. Once the bounded boxes of the human performing actions are extracted from the image, using the bounding box as the base the images are then cropped out to create a child image where only the human subject is present. These cropped images are then processed as an input to a trained classifier on the Stanford 40 action dataset. This ensemble method provides an additional benefit of detecting and classifying multiple and distinct human actions which are performed by numerous agents in the image. In order to obtain optimal classification performance, we aimed to train multiple models on Stanford 40 action dataset [21], then we test its performance on cropped out images which are obtained by feeding and cropping out RGB images into YOLO object detection algorithm. The Deep Learning model is trained in such a manner that firstly it is trained on Stanford 40 action dataset using transfer learning, then it is tested on images where human subjects are cropped out using YoloV3 object detection algorithm. In order to test various deep learning approaches, we firstly train the CNN model and then perform feature extraction on it to generate an array of features that can be used to train SVM algorithm.

Therefore this smart system can recognize human activities using CCTV or any other RGB camera feed, this system can have a plethora of applications. This can be used to perform in applications such as human computer interaction, security, video surveillance and home monitoring. More notably the system has a wider potential appeal as a smart surveillance system where it can detect multiple people's actions to flag violent or criminal situations without a human actively monitoring the camera feed.

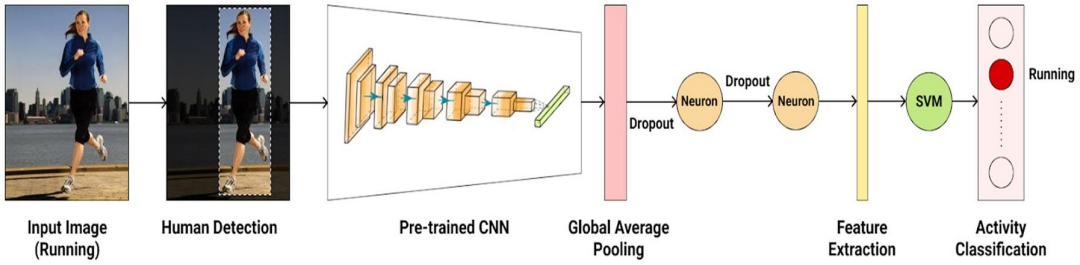

Fig. 2. Performing Feature Extraction and then using the extracted features to train a SVM on Stanford 40 action dataset

A. Training CNN Model

In order to solve the lack of large labeled activity recognition image dataset, we adopted the transfer learning approach, a method that uses CNNs that have been pre-trained to solve a different classification job, in this case the ImageNet challenge. [22]. Transfer learning uses information from source tasks (ImageNet challenge) to enhance performance in the intended tasks (activity recognition). In order to leverage transfer learning, we used InceptionResNetv2 [23] and MobileNetv2 [24] convolutional bases which were pre-trained on the ImageNet challenge. We then proceeded to add an artificial neural network on top of the convolutional base. Thus in conclusion, in the Feature Extraction stage, Global Average Pooling is used to generate a feature vector, which is then passed to a variety of fully connected and dropout layers.. Overfitting is reduced by using dropout layers. The dropout rate empirically set 0.4. Finally, the Stanford 40 dataset's probabilities for 40 classes are calculated using Softmax activation on the output of 40 neurons obtained at the end of ANN. We use the Cross-entropy loss for the action categorization job, which is written as:

$$Cross\ Entropy = -1/N \sum_{j} y_j \times \log\left(\hat{y}_j\right)$$

The ground truth is represented by y, prediction by ŷ, and number of training examples are denoted by N respectively.

B. Yolo Cropping

In order to segment out human subjects in an image containing multiple agents (both human and non-human alike), object detection is paramount, hence we utilized the already proven object detection algorithm YOLOv3. We feed the test images into the YOLO algorithm for cropping out of human subjects which are performing various activities. In order to improve cropping criteria, only human subjects were cropped out and a fixed minimum resolution size was established, below which the cropped out image was discarded. This was done to avoid creating cropped images where very small and no information could be extracted and acted upon by the deep learning model . Additionally one added benefit of this approach was that it resulted in separating out multiple agents which were performing different actions and on which the Deep Learning model could be applied separately resulting in better human activity analysis.

C. Feature Extraction

After the CNN model was trained, the last layer of CNN model is used to extract features from test and train images and consequently these features are then utilized to train machine learning algorithm - Support Vector Machine (SVM). Hence along with testing the performance of our CNN model, we also tested the performance of SVM on images that included cropped out human subjects made using YOLO algorithm. This approach is depicted in Figure 2.

III. Experimental Results

A. Implementation Details

The Google Colaboratory environment was used for our testing. Our proposed method was implemented in Keras [25], and its effectiveness was evaluated using the Stanford 40 dataset [21], which contains 40 different types of human behaviors. There are 4000 training photos and 5532 test images in this dataset. The dataset was divided into three sections: training, validation, and testing, using a 70:15:15 ratio. For training, validation, and testing, all input images were downsized to 512 512 pixels. To avoid overfitting, four types of data augmentations are utilized at random: rotation (from 0 to 10 degrees), horizontal flipping, width shifting (0 to 10%), and height shifting (0 up to 10 percent). With a learning rate of 0.0001, we utilized Adam optimizer [26] to train the suggested model.

B. Performance Of Cnn On Uncropped Dataset

We trained a total of 4 CNN models on Stanford 40 action dataset using different pre-trained CNN models which were InceptionResNetv2, MobileNetv2, consequently features extracted from InceptionResNetv2 and used to train a SVM and similarly features extracted from MobileNetv2 and used to train a SVM. Table 1 shows the results of the Stanford 40 test dataset and the number of parameters for each model. It can be observed that in comparison to our other model with pre-trained CNNs, the model built by extracting features from InceptionResNetv2 and then using them to train the SVM model displays the highest performance in terms of classification accuracy.

TABLE I. PERFORMANCE OF OUR APPROACH ON NON-CROPPED PHOTOS USING VARIOUS PRE-TRAINED CNNS (OBJECT DETECTION NOT PERFORMED)

Methods Used	**Accuracy in Action Classification (%)**	Total parameters
InceptionResNetv2 [23]	86	72,731,912
MobileNetv2 [24]	72.9	18,293,864
InceptionResNetv2 + SVM	87.2	72,690,912
MobileNetv2 + SVM	73.5	18,252,864

As it can be observed, the performance of models increases when we perform feature extraction from trained CNN models and then train machine learning algorithms such as Support Vector Machines (SVMs). This not only increases the performance but also uses a lesser number of parameters resulting in lesser computation time.

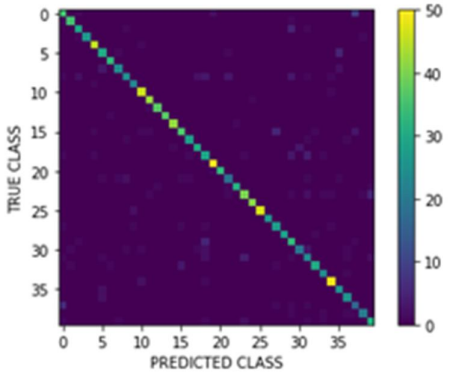

Fig. 3. Confusion Matrix of predicted class (InceptionResNetv2 + SVM) on Uncropped Dataset

C. Performance of Our Model Compared to Previous Techniques

We differentiate the capability of our model with previous techniques which have been tested in Stanford 40 Action dataset in Table 2.

As stated in the table, models such as AG-Net [30] and Ensemble method on features extracted from Channel Attention Module [29] do provide better accuracy performance on Stanford 40 action dataset, however there are some key considerations namely that in AG-Net, semantic regions are identified which could lie outside the bounding box of human subject and hence it will be rendered ineffective when performing human subject segmentation. This is similar for Channel Attention Module as that will extract features and may give more importance to objects other than human subjects and as a consequence, it could not be used perform human activity recognition when there are more than 1 human subject.

TABLE II. COMPARISON OF OUR APPROACH WITH EXISTING PRE-TRAINED METHODS TRAINED ON STANFORD 40 ACTION DATASET

Method	Accuracy in Action Classification (%)	Consideration
Action Mask [27]	82.6	
Deep Representations and Residual Learning [28]	86.5	
Ensemble method on features extracted from Channel Attention Module on the output from pretrained CNN [29]	93.17	Uses attention module to extract more specific features which may give more importance to objects other than human subjects and will fail when on segmented human subject
AG-Net [30]	97.83	Identifies semantic regions - will not work if more than two subjects are present in image
Improved Bilinear Pooling with Mask Aggregation [31]	85.24	
Discriminative Dictionary Design [32]	51.2	
Our Approach (InceptionResNetv2)	86	
Our Approach (InceptionResNetv2 + SVM)	87.2	

D. Performance of CNN on Cropped Dataset

The performance of our CNN models on unseen test images, which are obtained by firstly detecting human agents and then performing cropping on the bounded box obtained in the previous step, is shown in Table 2. As it can be observed that the performance goes considerably down when using MobileNetv2 pre-trained model as it contains a lesser number of parameters when compared to InceptionResNetv2. The performance also goes down when using pre-trained model InceptionResNetv2, however the drop is not as large as in case of MobileNetv2.

After deep analysis of the performance of our model on cropped images, we observed that our model is able to distinguish and classify actions where human hand movements and posture is unique to the action being performed when compared to the other actions included in the dataset. Our model obtained excellent accuracy

performance on actions such as - ‘blowing bubbles’, ‘cutting trees’, ‘playing guitar’, ‘playing violin’, ‘riding a bike’, ‘shooting an arrow’ and ‘using a computer’. Whereas our model performed poorly when the human poses and limb position were not distinguishable from other action and surrounding environment around the human subject was required to accurate classification, hence segmenting human subjects from their environment did not help, and such human actions were - ‘looking through a telescope’, ‘texting message’. ‘Walking the dog’ and ‘watching TV’.

TABLE III. PERFORMANCE OF OUR APPROACH ON CROPPED IMAGES USING SEVERAL PRE-TRAINED CNNS (OBJECT DETECTION PERFORMED)

Our Approach with Different Pre-trained CNNs	**Accuracy in Action Classification (%)**
InceptionResNetv2 [23]	66
MobileNetv2 [24]	29
InceptionResNetv2 + SVM	65.9
MobileNetv2 + SVM	30.2

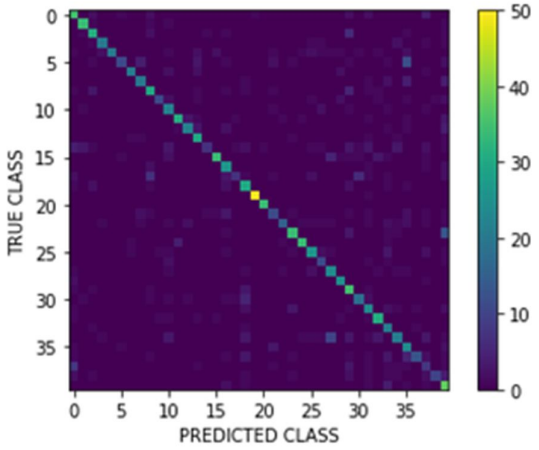

Fig. 4. Confusion Matrix of predicted class (InceptionResNetv2 + SVM) on Cropped Dataset

IV. CONCLUSION

In this paper, we put forward a novel smart surveillance system that detects and classifies human activities by reading input images from devices such as CCTV cameras. In order to create a system capable of classifying such activities, we create a pipeline, where human agents are first detected and cropped out and then fed into our CNN model trained on Stanford 40 action dataset. Our model is also capable of detecting and classifying multiple human subjects separately, hence we can conclude that our model can be utilized in environments where in a single frame, multiple human subjects are present and it is crucial to detect actions of these humans separately and accurately.

REFERENCES

[1] D. Schonfeld, "MotionSearch: Context-Based Video Retrieval and Activity Recognition in Video Surveillance," 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009, pp. 194-194, doi: 10.1109/AVSS.2009.106.

[2] 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.

[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] 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.

[5] 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.

[6] M. M. Ullah, S. N. Parizi, and I. Laptev, “Improving bag-offeatures action recognition with non-local cues.” in BMVC, vol. 10, 2010, pp. 95–1.

[7] D. Oneata, J. Verbeek, and C. Schmid, “Action and event recognition with fisher vectors on a compact feature set,” in Proceedings of the IEEE international conference on computer vision, 2013, pp. 1817–1824.
[8] V. Delaitre, I. Laptev, and J. Sivic, “Recognizing human actions in still images: a study of bag-of-features and part-based representations,” in BMVC 2010-21st British Machine Vision Conference, 2010.
[9] B. Yao and 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. IEEE, 2010, pp. 17–24.
[10] A. Prest, C. Schmid, and V. Ferrari, “Weakly supervised learning of interactions between humans and objects,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601–614, 2011.
[11] Z. Zhao, H. Ma, and S. You, “Single image action recognition using semantic body part actions,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3391– 3399.
[12] G. Gkioxari, R. Girshick, and J. Malik, “Actions and attributes from wholes and parts,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2470–2478.
[13] 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.
[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[15] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[16] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717– 1724.
[17] S. Yan, J. S. Smith, W. Lu, and B. Zhang, “Multibranch attention networks for action recognition in still images,” IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 4, pp. 1116–1125, 2017.
[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014
[19] 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.
[20] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv Prepr. arXiv1804.02767, 2018
[21] 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.
[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.
[23] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In AAAI, 2017, pp. 4278-4284.
[24] Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
[25] F. Chollet et al., “Keras: Deep learning library for theano and tensorflow,” URL: https://keras. io/k, vol. 7, no. 8, p. T1, 2015
[26] Kingma, Diederik & Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.
[27] 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.
[28] Siyal, Ahsan & Bhutto, Zuhaibuddin & Muhammad, Syed & Syed, Muhammad Shehram Shah & Iqbal, Azhar & Mehmood, Faraz & Hussain, Ayaz & Ahmed, Saleem. (2020). Still Image-based Human Activity Recognition with Deep Representations and Residual Learning. International Journal of Advanced Computer Science and Applications. 11. 471-477. 10.14569/IJACSA.2020.0110561.
[29] Mohammadi, Sina & Ghofrani, Sina & Shokouhi, Shahriar. (2020). Ensembles of Deep Neural Networks for Action Recognition in Still Images.
[30] A. Bera, Z. Wharton, Y. Liu, N. Bessis and A. Behera, "Attend and Guide (AG-Net): A Keypoints-Driven Attention-Based Deep Network for Image Recognition," in IEEE Transactions on Image Processing, vol. 30, pp. 3691-3704, 2021, doi: 10.1109/TIP.2021.3064256.
[31] W. Wu and J. Yu, "An Improved Bilinear Pooling Method for Image-Based Action Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 8578-8583, doi: 10.1109/ICPR48806.2021.9413028.
[32] Roy, A., Banerjee, B., Hussain, A. et al. Discriminative Dictionary Design for Action Classification in Still Images and Videos. Cogn Comput 13, 698–708 (2021). https://doi.org/10.1007/s12559-021-09851-8.