Human Action Recognition using Multimodal Convolution Neural Network

M S Srividya[1], Anala M R[2], Ramyashree V[3], Shalini R[4], Tushar Lal[5], Aman Verma[6]

[1]Assistant Professor, [2] Professor Department of Computer Science & Engineering, Student[3-6], Department of Computer Science & Engineering[1-6], RV College of Engineering[1-6], Bangalore, India.

Abstract: This paper aims to identify the various actions and expressions portrayed by a human in the input video stream. Firstly, we extract frames from the input video stream and then perform background subtraction. The frames are then pre-processed and fed to the trained CNN model (AlexNet). The proposed method aims to build two CNN models where in the first model identifies human action and the second model identifies human expression. Finally, as a result, we integrate both the models and use the soft-max classifier to classify the identified actions and expressions accordingly to their classes.

Index Terms **-Convolutional neural network, soft-max classifier, back ground subtraction.**

I. INTRODUCTION

Human activity recognition (HAR) plays a significant role in various fields of daily life including surveillance, security, and health care. Earlier methods for HAR were statistical method. The disadvantage associated with this method is the requirement of domain knowledge about the data and separation of feature extraction part from the classification part. Hence focus is shifted towards deep learning in which there is no need for feature extraction and labelling of data. The goal of HAR is to recognize activities from video sequences and it aims to correctly classify input data into its underlying activity category.

Along with it in our project we also aim to identify the various human expressions and integrate both the models for achieving the final result of displaying the corresponding action and expression simultaneously. To achieve this result HAR uses deep learning for training the model and soft-max classifier for classifying the different human actions and Open-CV for identifying the human expressions.

II. LITERATURE REVIEW

Human action recognition has been evolving every year. There are many numbers of approaches that are defined and implemented for identifying human action and we have studied some of the analysis and we have presented it in the following survey. In [1] the author has used features trajectories method for representing the video, but since the method was not satisfactory, the author proposed using the trajectories which are dense for representing the videos. In this method, for each frame the dense points were sampled and framed and then the human was identified. In [2] the human action recognition is classified based on the actions and features. Different types of features were represented by the video data on the basis of the extracted frames. The different features that were represented here were the local, global and motion features. In global features, the localization of the body parts was not required, instead the dynamics and the body structure were used with the help of these feature detectors, they were able to identify the human and also classify his/her actions. In [3] human action recognition based on vision and prediction from videos method was suggested, where action recognition meant inferring the present state of the human action based on the execution of the complete actions and prediction of action which was used to predict the human future state of action based upon the executions of incomplete action. In [4] the author has proposed a framework for early recognition of the action and anticipating by the past features correlation, using this method the state-of-the-art result was obtained. In [5] analyzing the human action using two stream CNN model was done. This survey gives us the summary of issues as well as methods and techniques to solve each of the issues. In [6] the author has proposed the method of Histograms of Oriented Gradients [HOG] and depth motion maps wherein the features are directly extracted from the raw-data without any transformations done beforehand.

By conducting this literature survey, the idea about identifying the human action and expression simultaneously comes into mind which can be helpful in various applications such as robotics, CCTV cameras etc.

.

III. PROPOSED METHOD

In the HAR system there are two modules. They are pre-processing module and pose estimator (human action and expression). Each of the modules receives input, process the input and produce output. The pre-processing module collects the input from the video stream and cleaning the data by removing the unnecessary and empty frames. The output of the pre-processed module is then forwarded to the human action and expression recognition module. This module uses a trained CNN model to extract the required features and classify the human action into three classes namely hit, kick and standing and three expressions namely angry, happy and normal/blank.

In the input module, we have generated our own dataset. There are two datasets in this project namely Training dataset and the Testing dataset. The dataset generated by recording a video performing and portraying various actions (hit, kick, standing) and expressions (normal/blank, happy, angry) and converted them to frames.

Secondly, we have pre-processing module, in which the date is pre-processed by burning all the unwanted and empty frames and used the training dataset to train the model and testing the dataset for validation. Lastly, we have human action and expression recognition module where we build two separate CNN models where in the first model identifies the human action and the second model identifies the human expression later both these models are integrated, classified, annotate and provide the annotated video as the output.

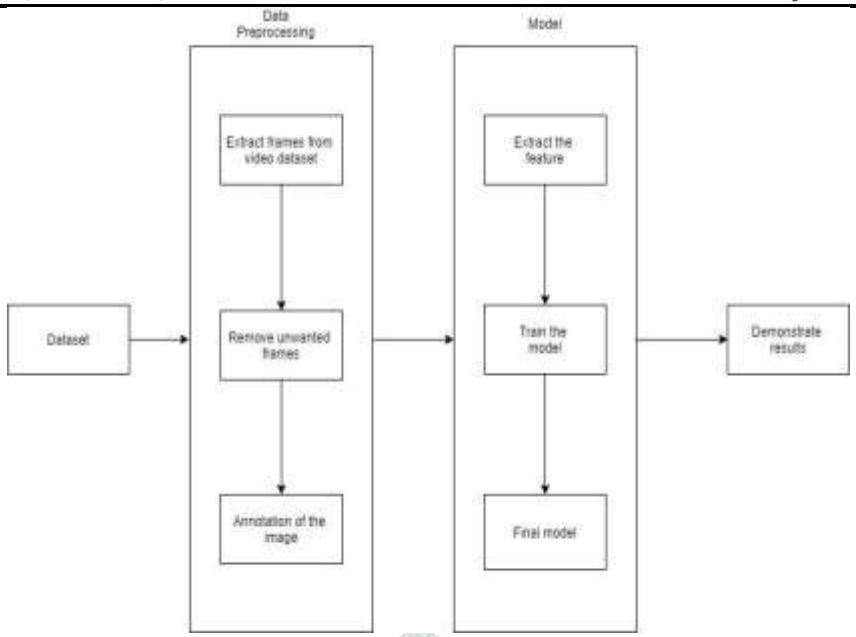

Fig-1: Module Flow Diagram

IV. METHODOLOGY

This project aims to identify the human action along with the emotions and display the result simultaneously. To achieve this, we build two CNN models. First model aims to identify the human action and second model aims to identify the human expression. Finally, we integrate both the models and perform various testing in order to achieve our expected result. We have generated our own dataset. There are two datasets in this project namely Training dataset and the Testing dataset. The dataset generated by recording a video performing and portraying various actions (hit, kick, standing) and expressions (normal/blank, happy, angry) and converted them to frames. We later pre-processed this dataset by burning all the unwanted and empty frames and used the training dataset to train the model and testing the dataset for validation.

For training our first model we made us of tools like Numpy (to load the dataset), Keras (for developing and evaluating deep learning models), Tensor flow (to define and train neural network). For training our second model we made us of tools like OpenCV (for facial recognition), Numpy (to load the dataset), Keras (for developing and evaluating deep learning models), Tensor flow (to define and train neural network).

During the training phase of both the models, we load the dataset and pre-process it and annotate them with an appropriate label corresponding to the action and expression being portrayed. So, for each of the action, expression we take a sizable amount of data say a 1000 data samples and we annotate them and finally fed them to the model for training. Later we train our model for 10 epochs. The more the number of epochs and more the accuracy and also the time consumption. Hence, we decided to train our model for 10 epochs. We have used Keras to define our CNN model and classifier used for our project is Soft-max classifier. The mathematical representation of score function for softmax classifier is shown below

$$f(y) = e^{yl} / \sum_k e^{yk}$$

And the optimizer we have used is Adam optimizer with the learning rate of 0.0004 to modify loss values with respect to improve accuracy and reduce loss. Finally, we generate a confusion matrix which describes the performance of the classification model on the test dataset.

Accuracy= Total correct predictions / Total predictions made *100

During the testing phase we create three classes each for two modules. The First module will have the classes naming hit, kick and standing. While the Second module will have the classes naming happy, angry and normal/blank. The features that we are mainly concentrating for action recognition are the postures of hands and legs and human face for identifying facial expression. We made use of HOG Descriptor, which is a feature descriptor used in image processing for the purpose of object detection and SVM Detector, to make predictions. We have also made use of background subtraction for eliminating the background objects and highlighting only the human. Then we load the trained models (human action model and human expression model). And we provide the recorded input video to compare the frames with the defined classes of the trained model wherever we predict the correct match of actions and expressions we label them and display them as a result. By following the above process, the system is able to accurately predict the various actions and expressions, hence we say that the system is built efficiently.

V. EXPERIMENT RESULTS AND ANALYSIS

We are obtaining an accuracy of 85% and model loss of 0.15% with 10 epochs. As we increase the number of epochs the time consumption for training the model increases along with the accuracy.

Table-1. Model accuracy

Contents	Training images	Testing images
Count	1224	1090
Percentage	79%	85%

Table-2. Model loss

Contents	Training images	Testing images
Count	1224	1090
Percentage	0.38	0.15

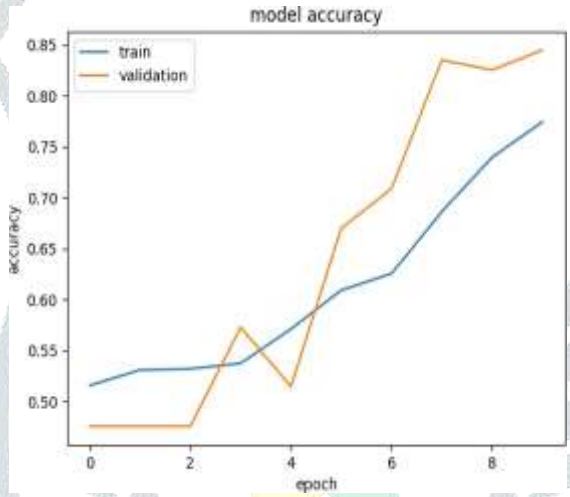

Fig-1. Model accuracy

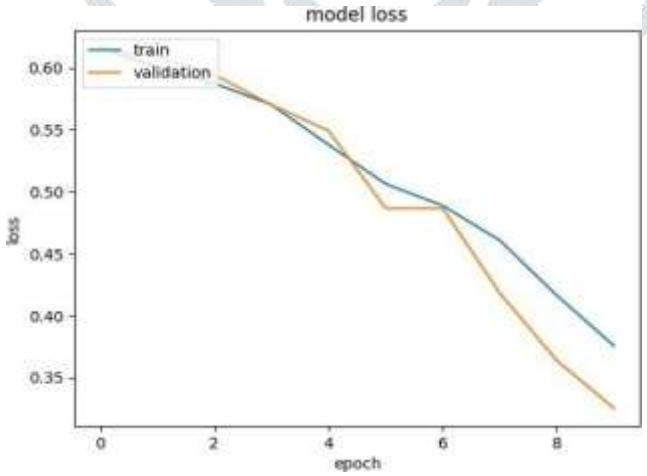

Fig-2. Model loss

Fig-3. Identified Kick action with Angry expression

Fig 4. Identified Hit action with Happy expression

Fig-5. Identified Standing action with Normal expression

VI. CONCLUSION

In this project, we have developed a method for human action and expression recognition using Multi-Model CNN Model and OpenCV tools. The proposed method there are three modules input module which excepts the input, pre-processing module which cleans the dataset and the human action recognition module which identifies and annotates the actions and expressions accordingly and classifies using soft-max classifier. The final result that we have provided is a labelled action and expression detected from the input video stream. We are obtaining an accuracy of 85% with 10 epochs. As we increase the number of epochs the time consumption for training the model increases along with the accuracy.

We have made sure that the approach will give best results in Real time and provide a research base to other researchers to carry further work in the field of image processing and deep learning. Our system could also be integrated with other things like in CCTV cameras and so on.

REFERENCES

[1] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, "Compressive sequential learning for action similarity labeling," IEEE Transactions on Image Processing, vol. 25, no. 2, pp. 756–769, 2016.

[2] Y. Kim and B. Toomajian, "Hand gesture recognition using microdoppler signatures with convolutional neural network," IEEE Access, vol. 4, pp. 7125–7130, 2016.

[3] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: A survey," Knowledge and information systems, vol. 36, no. 3, pp. 537–556, 2017.

[4] Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," IEEE Transactions on Human- Machine Systems, vol. 45, no. 1, pp. 51–61, 2018.

[5] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," in Tenth International Conference on Machine Vision (ICMV 2017), vol. 10696. International Society for Optics and Photonics, 2018, p. 106960Y.

[6] D.-X. Xue, R. Zhang, H. Feng, and Y.-L. Wang, "Cnnsvm for microvascular morphological type recognition with data augmentation," Journal of medical and biological engineering, vol. 36, no. 6, pp. 755–764, 2016.

[7] Y. Shima, Y. Nakashima, and M. Yasuda, "Pattern augmentation for handwritten digit classification based on combination of pre-trained cnn and svm," in Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEVISCMHT), 2017 6th International Conference on. IEEE, 2017, pp. 1–6.

[8] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," Multimedia Tools and Applications, vol. 76, no. 3, pp. 4405–4425, 2017. [17] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," IEEE Sensors Journal, vol. 14, no. 6, pp. 1898–1903, 2019.

[9] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE, 2017, pp. 177–186.

[10] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," Journal of real-time image processing, vol. 12, no. 1, pp. 155– 163, 2018.

[11] R. Shah and R. Zimmermann, Multimodal analysis of user-generated multimedia content. Springer, 2017.

[12] Y. Bouzouina and L. Hamami, "Multimodal biometric: Iris and face recognition based on feature selection of iris with ga and scores level fusion with svm," in Bio- engineering for Smart Technologies (BioSMART), 2017 2nd International Conference on. IEEE, 2017, pp. 1–7.

[13] A. B. Mahjoub and M. Atri, "An efficient end-to-end deep learning architecture for activity classification," Analog Integrated Circuits and Signal Processing, pp. 1– 10, 2018.

[14] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," IEEE Transactions on Human-Machine Systems, vol. 45, no. 1, pp. 51–61,

[15] F. Wang and J. Han, "Multimodal biometric authentication based on score level fusion using support vector machine," Opto-electronics review, vol. 17, no. 1, pp. 59–64, 2019.

[16] Y. Bouzouina and L. Hamami, "Multimodal biometric: Iris and face recognition based on feature selection of iris with ga and scores level fusion with svm," in Bio-engineering for Smart Technologies (BioSMART), 2017 2nd International Conference on. IEEE, 2017, pp. 1–7.

[17] A. B. Mahjoub and M. Atri, "An efficient end-to-end deep learning architecture for activity classification," Analog Integrated Circuits and Signal Processing, pp. 1–10,2020