OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Stacked Sparse Autoencoders for Human Action Recognition

[*]Muhammed Sadiq[1],Mariofanna Milanova[2]

*(Department of Computer Science, University of Arkansas at Little Rock, USA)
** (Department of Computer Science, University of Arkansas at Little Rock, USA)
Corresponding Author: [*]Muhammed Sadiq

ABSTRACT: *Determining human actions included in video stream remains a challenging problem in computer vision world. This is due to many reasons such as variations in the performance of the same action performed by different people, different viewpoints, insufficient training set and location of the person in video frames. In this paper, we demonstrate that sparse autoencoders which are trained on a number of video frames can achieve a high accuracy in determining and recognizing the type of action done by a human using popular datasets. We propose a new human action recognition system using multiple sparse autoencoders. In the first stage, we present an unsupervised deep learning algorithm and then we perform supervised learning through fine tuning the autoencoders with the appropriate labels for different types of human actions. The system is then applied onto three different datasets (Weizmann, KTH and UCF sports). The results demonstrate a performance of 99% in the Weizmann dataset for 10 actions, 86% in the KTH dataset for 6 actions and 96.7% in the UCF sport for 8 actions.*

Keywords: *Autoencoder, background subtraction, deep learning, features, sparsity.*

I. INTRODUCTION

Nowadays, video data are increasing unbelievably every minute according to recent YouTube statistics [1]. The video content is becoming very important and related to millions of users over the internet. Therefore, the need for accurate and automatic systems for human action recognition is raising every day. Such systems play a major role in recent applications such as automatic monitoring of surveillance cameras, medical treatment, behavioral analysis, object recognition, efficient video search and video retrieval [2][3]. These systems are significantly valuable due to their ability to determine the important events or actions that occurred in the video stream. Many studies have been proposed in this active research area and the most popular way for implementing action recognition is by extracting features from video frames then choosing a suitable classifier for assigning features to the correct class of actions [2][4] [5] [6].

While deep learning (DL) is still considered as a new method in the field of image recognition and video analysis as compared to other methods in the same approach, DL can achieve higher accuracy in image recognition especially with large scale vision data. It was a turning point in the computer vision society when SuperVision team won Imagenet competition in 2012 [7]. Deep learning techniques have the ability to achieve different levels of abstractions through many layers with higher accuracy [8][9].

A considerable amount of research has been introduced in term of deep learning to perform object and human action recognition, but the issue of getting high accuracy is still at the stake for most of the studies in this field. Zhize Wu and et al [12] used stacked denoising autoencoders to learn and rebuild the depth information from joint 3D features to perform human action recognition on the video frames, features enhancement has been done by adding to the autoencoders a class and temporal constraints to catch the small details. Moez Baccouche and et al [13] proposed a 3D convolutional network to learn their features automatically and then a recurrent neural network has been trained using these features to detect the class of the video stream which is KTH dataset. A hybrid method can be found by [14] where Mahmudul Hasan and et al combined active learning with deep networks through updating the parameters of the sparse autoencoder to automatically choose the favorable features and to update the existing system. Extending the idea of using convolutional neural network in still image recognition to include video frames data was proposed by [15] and [16]. Karen Simonyan and et al [15] used Stacked Sparse Autoencoders for Human Action Recognition video frames as their input to train two stream ConvNets, while in [16] Andrej Karpathy and et al suggested using ConvNet for large scale video classification. However, their results were not satisfying.

In this paper, the idea of object recognition is extended using stacked sparse autoencoders to the point of human action recognition in any video stream data by proposing a new DL architecture through multiple

layers of representation to perform human action recognition. This architecture performs unsupervised way of learning in the first stage without supplying the autoencoders any labeled data, then we perform supervised way of learning using labeled training set (labeled video frames for each action) to update the parameters for each autoencoder to get better classification result.

The internal structure of sparse autoencoder has encoder and decoder units [10] [11], the encoder represents or compresses the input into the hidden layers which is usually smaller than the input layer, while the decoder works exactly the opposite job by trying to represent or decompress the input came from the hidden layers back to the original inputs using backpropagation algorithm [10]. After training our system with sufficient video frames from a specific dataset, we applied a multinomial logistic regression classifier (Softmax classifier) to classify the test set and calculate the result of the first stage where no labels being supplied to the autoencoders, then we attach the correct label to each video frame in the training set and perform fine tuning for the sparse autoencoders. The result of the second stage is significantly enhanced and the system achieved higher classification accuracy than the first stage. The paper is organized as follows: Section II explains sparse autoencoder methodology, in section III the proposed model is presented, section IV describes the experimental results. Finally, the conclusion is given in section V.

II. METHODALGY

Sparse autoencoder is a neural network that is capable of learning features automatically from unlabeled data (unsupervised learning) [10] [17]. It reproduces the input as the output but with different dimensionality and the learning is done by minimizing the reconstruction error between the input data at the encoder layer and its reconstruction at the decoder layer. Autoencoder performs dimensionality reduction like Principle Component Analysis (PCA). For each sparse autoencoder, all neurons in the input layer will be connected to all neurons in the hidden layer and try to find well-correlated representation by applying backpropagation algorithm to make output values closer to the input values Fig.1.

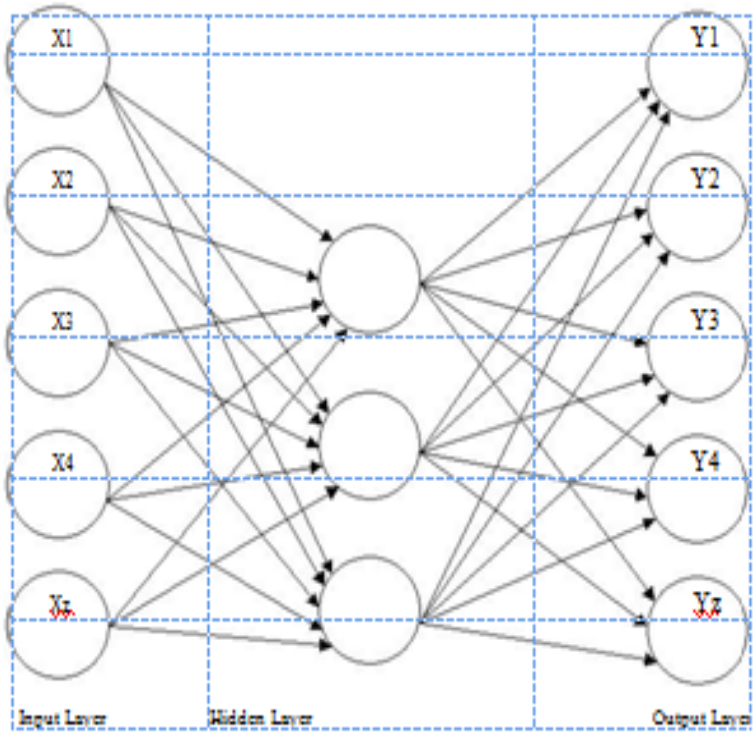

Figure.1. Internal structure of sparse autoencoder

A very useful compressed representation can be found especially if the data has correlated structure such as binary images of human actions, but if the data collected randomly, then such representation cannot be useful. The encoding procedure is done through applying a linear mapping and a nonlinear activation function to the input vector, while the decoding process maps the encoded data back to the original input, the goal is to minimize the error between encoding and decoding layers.

To encode the input vector $x \in R^{D\times 1}$ with weight matrix $W \in R^{N\times D}$, bias $b \in R^{N\times D}$ and N features. The logistic sigmoid function has been employed where sigm = $(1 + \exp(-x))^{-1}$

$$a = \text{sigm}(W_x + b_1) \text{---------- (1)}$$

While the decoding procedure for a is supposed to do the opposite job but W and b belong to the hidden layer.

$$z = V^2 + b_2 \text{ ---------- (2)}$$

Where the decoding matrix $V \in R^{N\times D}$, b_2 is decoding bias. The learning of the autoencoder is completely depended on minimizing the error between the two representations

$$L(X,Z) = \frac{1}{2}\sum_{i=1}^{m} \|z_i - x_i\|_2^2, \text{ ---------- (3)}$$

Where X and Z are the original input and the newly reconstructed data from encoding stage. Also, sparsity constraint has been added to penalizes the difference between the two representation, Let β be the weight of this penalty

$$L(X,Z) + \beta \sum_{j=1}^{N} KL(p|\hat{p_j}) \text{ ---------- (4)}$$

KL is Kullback-Leibler divergence as the following:

$$KL(p\|\hat{p_j}) = \sum_{j=1}^{s2} p\, log\frac{p}{\hat{p_j}} + (1-p)\, log\frac{1-p}{1-\hat{p_j}} \text{ ---------- (5)}$$

The sparsity parameter p should be very close to the average activation function $\hat{p_j}$, where

$$\hat{p_j} = \frac{1}{n}\sum_{i=1}^{n} [a_j^{(2)}(x^{(i)})] \text{---------- (6)}$$

$a_j^{(2)}$ is the activation of hidden unit j for input x.

III. THE PROPOSED MODEL

A novel framework was proposed to classify and recognize human actions from video scene, the proposed system consists of several autoencoders stacked together and the result from each autoencoder will be considered as the input for the next one Fig. 2.

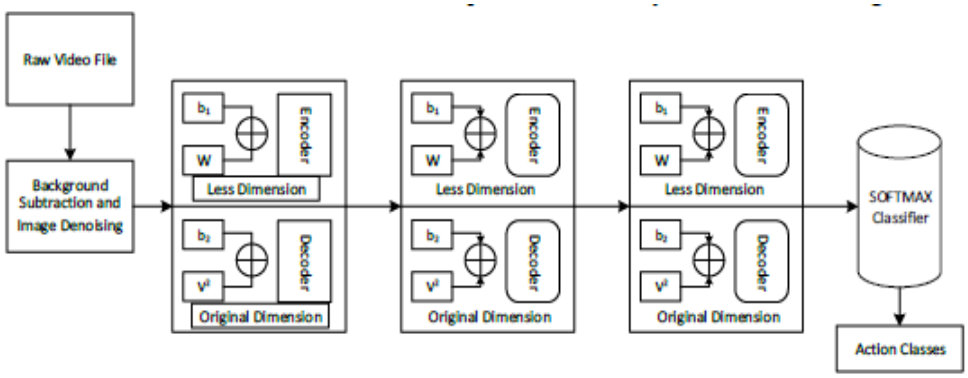

Figure. 2. The proposed system which is consist of 3 sparse autoencoders staked together then Softmax classifier.

The idea behind stacking several autoencoders is to achieve better dimensionality reduction from a big number of input neurons to a small number of output neurons. The first step of our system is to do a background subtraction for all input videos (Fig. 3).

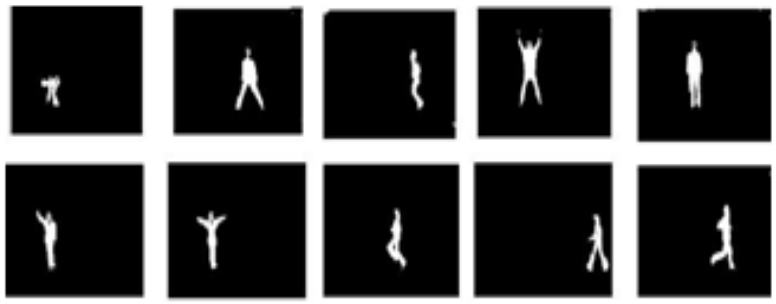

Figure. 3 a. Example of Weizmann video frames after background subtraction.

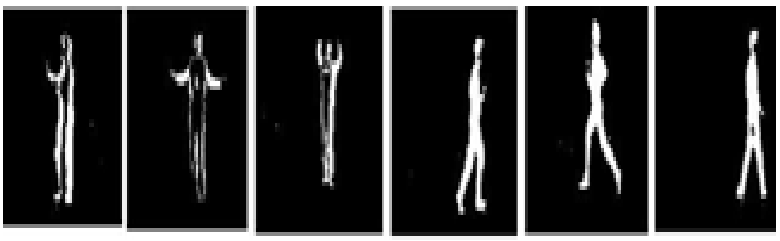

Figure. 3 b. Example of KTH video frames after background subtraction.

Figure. 3 c. Example of UCF-sport video frames after background subtraction.

Stacked Sparse Autoencoders for Human Action Recognition Frame difference algorithm has been applied with an appropriate and variable threshold for every video to cut the foreground human body from the scene for. Then the foreground pictures have been changed to binary images and we performed image denoising to the binary pictures to get more clear binary images which are easier for the system to recognize and to classify the actions correctly. It is important to mention that the execution time of the learning process depends mainly on the size of the frame and also on the computer specification (RAM size and processor speed). For instance, the Weizmann dataset has (180 *144) pixel, so we will have 25,920 neurons as our input layer to the first autoencoder, this is not a good number if we want a shorter training time and also it requires more RAM to be implemented. Another issue with big size frames when the autoencoder reduces the dimensions or compress the video frame to represent the hidden layer, the difference between hidden layer and the input layer will be huge and it will be harder to get much more useful correlated representation in this case. Therefore, a proper resizing process was done to the video frames with respect to the quality and clearness of the binary images.

The first autoencoder will have M*N neurons in its input layer depending on the size of the input video frames, the output of the first autoencoder will be taken as input to the second autoencoder, and the output of the

second autoencoder will be used as input to the third and last autoencoder. Finally, a multinomial logistic regression classifier or Softmax classifier was applied to the output of the third autoencoder to classify every action to the correct class. No labels were used until this point, but we can perform much better by attaching the correct labels to the action classes and perform fine tuning to the test set will increase the accuracy of our system.

IV. EXPERIMENTAL RESULT

In this section, we assessed the performance of our proposed method that was applied to multiple datasets (Weizmann, KTH and UCF sports), then we calculated the accuracy of recognizing every action in the datasets, the first results were based on unsupervised learning where no labels been used during the learning of the sparse autoencoder and it was usually less accuracy than the supervised stage. However, the second results obtained after attaching appropriate labels to the training set and evaluate the accuracy for each set of actions.

A. Weizmann Dataset

Consist of 90 videos for 10 actions done by 9 different people using static camera and same environment conditions for all videos, the actions that have been done are (walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack, skip), we took all the actions to train our system then we randomly select video frames for each actions for testing purposes, the result for the first unsupervised stage was 59% and for the supervised stage was 99% which is very competitive result (Fig.5.a).

Figure. 5 a. The result of our proposed system using Weizmann dataset for 10 actions.

B. UCF Sport Dataset

This dataset was mainly collected from sports television broadcast, it has higher video resolution with 720 x 480 pixels making the video frames more clear and understandable. Some videos clips captured using moving camera. We selected 8 actions as the following (diving, golf swing, kicking, lifting, riding Horse, running, skateboarding and walking). It took more time to learn our system more than the other 2 datasets because of the frame quality. The result was 56.7% for the first stage and 96.7% for the second stage (Fig. 5.b).

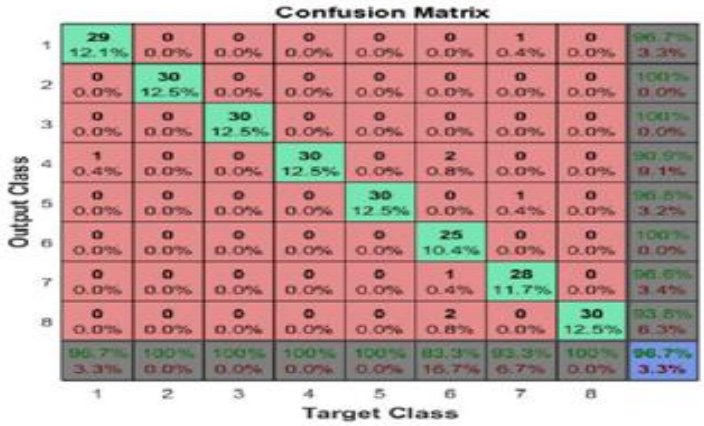

Fig. 5 b. The result of our proposed system using UCF- Sports dataset for 8 actions

C. KTH Dataset

It has 6 actions as the following (walking, jogging, running, boxing, hand waving, and hand clapping) done by 25 persons using static camera in different environments and viewpoints. The accuracy of recognizing all actions was 65.3% accuracy and 86.5% for the second stage (Fig. 5.c) which is less than Weizmann and UCF-sport datasets. Since we are classifying and recognizing video frames, there are several reasons behind getting less accuracy than the other two datasets, first, some video frames has only a small part of the human body for instance legs only not a complete human body (Fig. 6), while in the same video clip and after several frames the clip shows a complete human body. Second, many clips have totally different viewpoints for the same action. Third, the position of the person doing the action is relatively far from the camera especially at the beginning of the action. All these reasons make it harder for perfect classification.

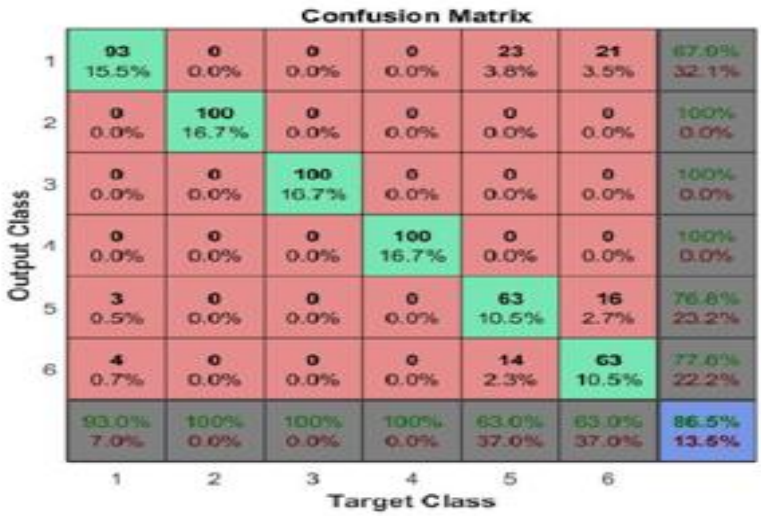

Figure. 5 c. The result of our proposed system using KTH dataset for 6 actions.

Figure. 6 Many frames in KTH dataset have only part of the human body while some have a full human body.

V. CONCLUSION

Stacking several sparse autoencoders as one neural network with an appropriate classifier at the end of the network can result in excellent recognition of human actions in video files. The training time of each autoencoder mainly depends on the input size, the number of epochs and the computer specifications, while the testing process is much faster and can be applied in real-time application. Sparse autoencoder has the ability to form a very useful representation if the data has correlated structures. Without correlated structure, it is impractical to use sparse autoencoder on the data. By applying background subtraction techniques, all the unwanted details from the video frames are removed and the focus remains on the moving human body in the frame. This process facilitates the classification and support the autoencoder in finding correlated structures easily.

REFERENCES

[1]. YouTube statistics http://www.youtube.com/yt/press/statistics.html.
[2]. Al-Ali, Salim, et al. "Human Action Recognition: Contour-Based and Silhouette-Based Approaches." Computer Vision in Control Systems-2. Springer International Publishing, 2015. 11-47.
[3]. Poppe, Ronald. "A survey on vision-based human action recognition." Image and vision computing 28.6 (2010): 976-990.
[4]. Sadek, Samy, et al. "Chord-length shape features for human activity recognition." ISRN Machine Vision 2012 (2012).
[5]. Yang, Weilong, Yang Wang, and Greg Mori. "Human action recognition from a single clip per action." Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, 2009.
[6]. Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." Proceedings of the IEEE International Conference on Computer Vision. 2013.
[7]. Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.
[8]. Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." IEEE transactions on pattern analysis and machine intelligence 35.1 (2013): 221-231.
[9]. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.
[10]. Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72.2011 (2011): 1-19.
[11]. Olshausen, Bruno A., and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?." Vision research 37.23 (1997): 3311-3325.
[12]. Wu, Zhize, et al. "Discriminative Feature Learning with Constraints of Category and Temporal for Action Recognition." International Conference on Image and Graphics. Springer International Publishing, 2015.
[13]. Baccouche, Moez, et al. "Sequential deep learning for human action recognition." International Workshop on Human Behavior Understanding. Springer Berlin Heidelberg, 2011.
[14]. Hasan, Mahmudul, and Amit K. Roy-Chowdhury. "Continuous learning of human activity models using deep nets." European Conference on Computer Vision. Springer International Publishing, 2014.
[15]. Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems. 2014.
[16]. Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.
[17]. Shin, Hoo-Chang, et al. "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1930-1943.

*Muhammed Sadiq1. “Stacked Sparse Autoencoders for Human Action Recognition.” International Journal Of Modern Engineering Research (IJMER), vol. 07, no. 08, 2017, pp. 77–83.