IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 30, 2021, accepted February 15, 2021, date of publication February 23, 2021, date of current version March 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061455

Autism Spectrum Self-Stimulatory Behaviors Classification Using Explainable Temporal Coherency Deep Features and SVM Classifier

SHUAIBING LIANG[1], (Graduate Student Member, IEEE), AZNUL QALID MD SABRI[1], (Member, IEEE), FADY ALNAJJAR[2], AND CHU KIONG LOO[1], (Senior Member, IEEE)

[1]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]College of Information Technology, UAE University, Al Ain, United Arab Emirates

Corresponding author: Chu Kiong Loo (ckloo.um@um.edu.my)

This work was supported in part by the Fundamental Research Grant Scheme (FRGS) grant under Grant FP069-2015A, in part by the Asian Universities Alliance (AUA)-The United Arab Emirates University (UAEU) Joint Research Grant 31R188, and in part by the Covid-19 Special Research Grant under Project CSRG008-2020ST.

ABSTRACT Autism spectrum disorder is a very common disorder. An early diagnosis of autism is essential for the prognosis of this disorder. The common diagnosis method utilizes behavioural cues of autistic children. Doctors require years of clinical training to acquire the ability to capture these behavioural cues (such as self–stimulatory behaviours). In recent years, the advancement of deep learning algorithms and hardware enabled the use of artificial intelligence technology to automatically capture self-stimulatory behaviours. Using this technique, the work efficacy of doctors can be improved. However, the field of self-stimulatory behaviours research still lacks large annotated data to train the model. Therefore, the application of unsupervised machine learning methods is adopted. Meanwhile, it is often difficult to obtain good classification results using unlabelled data, further research to train a model that can obtain good classification results and at the same time being practical will be valuable. Nevertheless, in the area of machine learning, the interpretability of the created model has to be vital as well. Hence, we have employed the Layer-wise Relevance Propagation (LRP) method to explain the proposed model. In this article, the major innovation is utilizing the temporal coherency between adjacent frames as free supervision and setting a global discriminative margin to extract slow-changing discriminative self-stimulatory behaviours features. Extensive evaluation of the extracted features has proven the effectiveness of those features. Firstly, the extracted features are classified by the k-means method to show the classification of self-stimulation behaviours in a completely unsupervised way. Then, the conditional entropy method is used to evaluate the effectiveness of features. Secondly, we have obtained the state-of-the-art results by combining the unsupervised TCDN method with optimised supervised learning methods (such as SVM, k-NN, Discriminant). These state-of-the-art results prove the effectiveness of the slow-changing discriminative self-stimulatory behaviours features.

INDEX TERMS Autism spectrum disorder, computational behavioural analysis, machine learning, temporal coherency, unsupervised deep learning.

I. INTRODUCTION

Autism Spectrum Disorder is a prevalent disorder. A recent paper published in March 2020 revealed that 1 in 54 children are identified with Autism Spectrum Disorder (ASD) according to the estimates from CDC's Autism and Developmental Disabilities Monitoring Network [1]. Studies have also proven that early diagnosis of ASD is associated with significant gains in intellectual ability, adaptive behaviour as well as reduction of symptom severity in children with ASD [2]–[4]. Using behavioural cues of autistic children is a common method of diagnosis for ASD [5]. Some of the exercising instruments which use those behavioural cues to diagnose ASD include the Autism Diagnostic Observation

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

Schedule (ADOS) [6] and the Autism Observation Scale for Infants (AOSI) [7]. Moreover, self-stimulatory behaviours are atypical behavioural cues that are assessed in these instruments for diagnosis. Autism diagnosis requires clinicians to interact with the child over multiple extensive sessions to identify the behavioural cues [8]. However, suitably trained clinicians may be unavailable and expensive in some areas. Therefore, using a computer to automatically analyse characteristics of children with autism such as self-stimulatory behaviour can help doctors to infer further diagnosis[9][10]. Some self-stimulatory behaviours such as head banging, are classified as self-injurious behavior [11] as they can cause damages to the children. Considering the random occurrence of self-stimulatory behaviour, it is impractical to observe the autistic children at all times during the day. An automatic self-stimulatory behaviour analysis system can help doctors and parents to care for children with autism.

The existing research on self-stimulatory behaviours is mainly divided into two categories, namely based on accelerometers [12], [13] and computer vision [8], [14] respectively. Since 2D cameras are cheaper and more accessible, we decided to develop a self-stimulatory behaviour classification algorithm based on video data.

A. PROBLEM STATEMENT AND HYPOTHESIS

Deep learning has made great achievements in the field of human action recognition [15]–[17]. Although there is a large amount of unlabeled video data in the public website (such as YouTube) on self-stimulatory behaviours research, we still lack large annotated real-word datasets to train an artificial neural network. Using unlabeled video data recorded in an uncontrolled environment to train unsupervised models is often difficult to obtain good classification performance. Hence, choosing and optimising the model to achieve good classification performance remains a challenge. Furthermore, understanding the internal classification mechanism is often difficult due to the nonlinear structure of artificial neural networks. This prevents our model from providing intuitive references and suggestions for researchers and doctors.

In this article, to use a large amount of unlabeled video data obtained from the public website, we decided to use an unsupervised method to automatically extract the features of the video data to save time and effort. From the published paper written by Wiskott and Sejnowski [18], we understand that the input of a camera is a quick-changing matrix. A slight change of the characters in a video will drastically affect the input matrix. Thus, if we can obtain a slow-changing or even steady feature of each autism self-stimulatory behaviour, we can classify those behaviours easily [19], [20]. Other than that, the ability of these slow-changing features to discriminate different self-stimulatory behaviours is also crucial. Till date, obtaining a slow-changing discriminative self-stimulatory behaviours feature remains a problem.

In order to understand the model's internal classification mechanism, the Layer-wise Relevance Propagation (LRP) [21] algorithm will be used to explain the Temporal Coherency Deep Networks and Support-Vector Machines (TCDN-SVM) hybrid algorithm that was created.

B. CONTRIBUTION

The main contribution of this paper is that we have extracted a slow-changing discriminative self-stimulatory behaviours feature and the experiment was able to obtain a state-of-the-art result. The unsupervised temporal coherency deep networks (TCDN) method was used for the extraction of this feature[22].

The TCDN algorithm is based on four Alexnet with the same parameters and a loss function based on Euclidean distance [23]. Next, in order to prove the efficiency of features extracted by our method, a method that combines K-means and conditional entropy is used. Thereupon, unsupervised feature extraction methods and supervised classification methods are combined to construct self-stimulatory behaviour classifiers. Multiple supervised methods are employed to improve the classification performance of our model to obtain a particularly good result. The methods yielded 98% accuracy at frame level and 98.3% accuracy at the video level. Finally, a TCDN-SVM model is constructed and interpreted using the LRP algorithm to ascertain why this model could achieve such good results. This model allows us to provide evidence for the early diagnosis of autism.

The rest of this article is organised as follows: Section II discusses some existing methods that can be used for the self-stimulatory behaviour classification and its limitations. Section III briefly introduces all the methods used in this study. Section IV introduces the evaluation methods and provides the result of the experiment. While Sections V, VI and VII discuss and summarise our research and propose future research directions.

II. RELATED WORK

A common diagnosis method for autism is using behavioural cues of autistic children [5]. One of the existing diagnostic instruments that is based on behavioural cues is the Autism Diagnostic Observation Schedule (ADOS) [6]. This instrument is a standardised and semi-structured evaluation method. It can assess autistic patients based on social interaction, communication, play and imaginative use of materials. Another example of such instrument is the Autism Observation Scale for Infants (AOSI) [7], this algorithm was developed to detect and monitor early signs of autism in high-risk infants. Self-stimulatory behaviours are atypical behavioural cues that are assessed in these instruments for diagnosis based on accelerometers or computer vision.

Westeyn *et al.* [24] used small 3-axis accelerometer modules to study the self-stimulatory behaviour. The accelerators are placed on the right wrist, the back of the waist and on the left ankle of a non-autistic person. Then, this non-autistic person was prompted to mimic autistic patient to perform self-stimulatory behaviours to collect data. Finally, the collected accelerometer data was assessed using hidden Markov models (HMMs). Since this dataset was collected

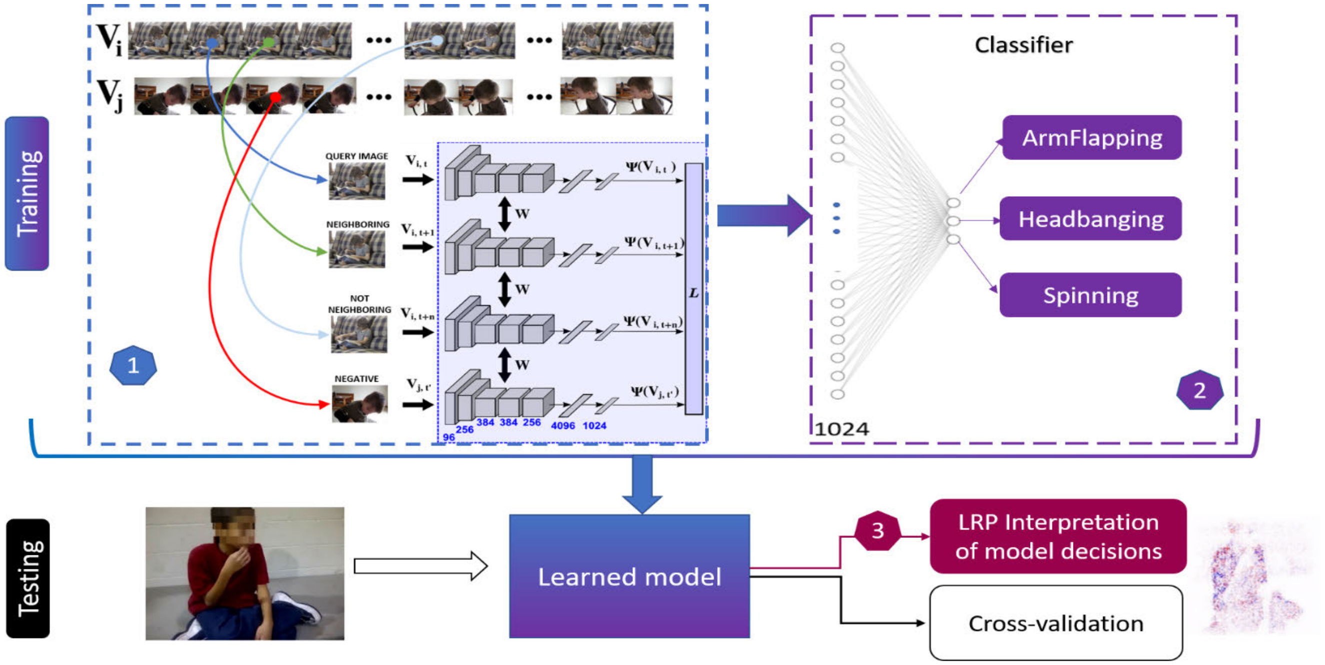

FIGURE 1. The architecture of Temporal Coherency Deep Networks and supervised classifier [22].

from a single, neurotypical adult, the model may not be well adapted to new data generated from people with autism. Other than that, Min [12] have also used accelerometer modules to assess self-stimulatory behaviours using the Time-Frequency methods to extract features together with the hidden Markov model to detect and label self-stimulatory behaviour. When self-stimulatory behaviour occurs, the system will automatically use a webcam and microphone to store the patient's video and audio data. By using this system, doctors can view the patient's video data to diagnose and treat autism. Mohammadian Rad *et al.* [13] have also used a wearable inertial measurement unit to detect the Stereotypical Motor Movement of autistic patients. The Stereotypical Motor Movement behaviour is very similar to self-stimulatory behaviours. In their published work, they have used the convolution neuron network to extract features and used LSTM to classify them. However, since the model is fully trained using supervised learning methods, the model might not be able to adapt to the new data very well.

A study published by Rajagopalan *et al.* [25] in 2013 demonstrated a standard action recognition pipeline on the new Self-Stimulatory Behaviour Dataset (SSBD). This dataset was collected from public domain websites such as YouTube. Such video format datasets from an uncontrolled environment are difficult to classify. Hence, Space-Time Interest Points (STIP) was used with the Harris3D detector in the Bag Of Words (BOW) framework to train the classifier. However, the results reported were not promising, the best accuracy was only 50.7%. Another article also published by Rajagopalan *et al.* [8] used the SELECTION OF POSELET BOUNDING BOXES method to identify the positions of autistic children to create a motion model based on the Histogram of Dominant Motions (HDM) method. In this experiment, they have achieved a state-of-the-art result (73.6%) when the 5-fold cross-validation method was employed.

III. METHODOLOGY

Referring to Figure 1, the first part used an unsupervised TCDN method to automatically extract the features of the self-stimulatory behaviour videos of children with autism. This method is adapted from a paper published by Redondo-Cabrera et. al that has introduced Quadruplet Method, it is known to achieve unsupervised classification at human action recognition task [22]. Once the features were extracted, K-means and condition entropy methods were used to verify feature effectiveness. In the second part, the performance of the identification of the self-stimulatory behaviour of autistic people is improvised. The features extracted using the unsupervised recognition method in the first part was used as input in this part to compare different supervised classification methods, such as Decision trees, Discriminant Analysis, Linear SVM, k-nearest neighbours algorithm (k-NN). Thereafter, the third part is to understand our model's internal mechanism to help humans design better models for identification of the behaviour of autistic patients and to assist doctors in making diagnoses. We then selected the interpretable Linear SVM to be combined with the TCDN algorithm to produce the TCDN-SVM algorithm. The LRP algorithm was used to interpret it. The methods used in these three parts and the results obtained will be discussed in detail below.

A. TEMPORAL COHERENCY DEEP NETWORKS(TCDN)

In the area of self-stimulatory behaviours research, self-stimulatory behaviours occur randomly. Hence, it is challenging to make an annotated video dataset, supervising the classification without labels becomes an issue as well. According to the slow feature analysis (SFA) method [18], the image signal input by the camera, such as grayscale or point, is a low-level and rapidly changing representation of the action. Even when a child with autism moves slowly, the input signal will change quickly. If a high level, slow-changing or even steady features can be extracted from the input picture signal of each type of self-stimulatory behaviour, they can be used as free supervision for classification.

In this study, we proposed an input-output algorithm which utilises the temporal coherence between contiguous video frames as free supervision to extract features. In brief, our method uses unlabeled video data to train a convolutional neural network (CNN) model to extract features. Our objective function is as follows:

$$\min_{w} \frac{\delta}{2} W^2 + \sum\nolimits_{i=1}^{T} L_u(W, U_i) \tag{1}$$

The input is a set of m unlabeled videos $V = V_1, V_2, V_3, \ldots, V_m$, and the W is the parameters of our CNN network. δ is the weight decay constant. L_u is the unsupervised regularization loss term. U_i is the representation of training tuples of video frames. The key idea of this method is to keep the temporal coherence of adjacent frames in the learned feature representation. Meanwhile, the distance between two frames separated by n frames is shorter than the distance between frames from two different videos. The length of n frames is called the temporal window.

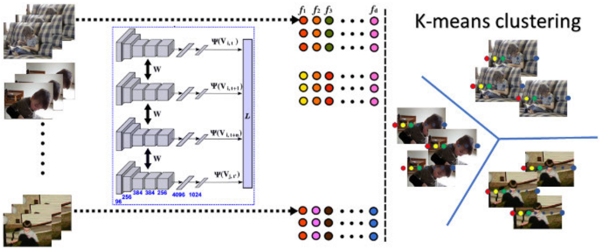

FIGURE 2. The architecture of Temporal Coherency Deep Networks (TCDN).

Figure 2 represents the structure of TCDN. The input of this architecture includes the following four frames namely $V_{i,t}$, $V_{i,t+1}$, $V_{i,t+n}$, and $\boldsymbol{V_{j,t}}$'. These four frames were extracted from two videos V_i, V_j.

The $V_{i,t+1}$ is an adjacent frame of $V_{i,t}$. There are n frames between $V_{i,t}$ and $V_{i,t+n}$. The $V_{i,t}$ and $\boldsymbol{V_{j,t}}$' originate from different videos. Then, there are four AlexNet networks that were used to process those four frames to four 1024 dimension representations (ψ). These four networks share the same parameters W. We assume that the learned feature representation ψ is a function of the learned AlexNet network parameters W. The input of this function is a frame of a video, the output of this function is a feature representation ψ (V). In order to realize our key idea, we designed a loss function L (2) based on Euclidean distance d to train this network:

$$\begin{aligned} &L_q\left(\psi\left(V_{i,t}\right), \psi\left(V_{i,t+1}\right), \psi\left(V_{i,t+n}\right), \psi\left(V_{j,t'}\right)\right) \\ &= d\left(\psi\left(V_{i,t}\right), \psi\left(V_{i,t+1}\right)\right) + max\left\{0, d\left(\psi\left(V_{i,t}\right), \psi\left(V_{i,t+n}\right)\right)\right. \\ &\quad \left. - d\left(\psi\left(V_{i,t}\right), \psi\left(V_{j,t'}\right)\right) + \alpha\right\} \end{aligned} \tag{2}$$

This loss function tries to make the feature representation of $V_{i,t}$ similar to the feature representation of $V_{i,t+1}$. However, the distance between $V_{i,t}$ and $\boldsymbol{V_{j,t}}$' must be greater than the distance between $V_{i,t}$ and $V_{i,t+n}$ by a constant α, because $V_{i,t}$ and $\boldsymbol{V_{j,t}}$' originate from different videos. Therefore, the design purpose of our loss function is to hope that the distance between two adjacent frames is as small as possible and that the distance between two frames from different videos is greater than the distance of two non-adjacent frames from the same video.

B. SUPERVISED METHOD TO CLASSIFY LABELED FEATURES

The supervised machine learning method has been known to achieve better performance as compared to the unsupervised method. Therefore, in this article, several supervised methods have been used to obtain better results. Firstly, the unsupervised temporal coherency deep networks (TCDN) is used to extract features as the input of the supervised method. As the TCDN is an unsupervised method, we were able to use all of the data to train this model, and then we can use this model to extract the feature of all frames form videos. After we obtain the features, we can use those features as the input of the supervised method. Then we can obtain a frame-level classification. Here we introduce some supervision methods provided by the Matlab classification learner app:

1) DECISION TREES

The decision tree is a straightforward and easily interpreted method[26]. The parameters of this method are modified to result in three different trees. The Coarse Tree has a few leaves to make coarse distinctions, which makes the prediction more robust. However, this method is usually unable to attain high training accuracy. Secondly, Medium Tree has a medium number of leaves. Lastly, the Fine tree has many leaves to make many fine distinctions. However, this method tends to overtrain.

2) DISCRIMINANT ANALYSIS

The Linear Discriminant method creates linear boundaries between classes[27]. The Quadratic Discriminant creates nonlinear boundaries between classes[28]

3) LINEAR SVM

The idea of SVM is to find the best hyperplane that can split data points to different classes[29]. In this article, because of the large dataset and lack of computer source, only Linear SVM was chosen.

4) K-NN

This algorithm categorises points based on their distance to points (or neighbours) in a training dataset[30]. It is a simple yet effective way of classifying new points. After evaluating the effects of using different set of parameters (e.g. number of neighbours, distance method and distance weight) on the performance of k-NN classifiers, 5 k-NN methods were chosen to classify our SSBD dataset. Fine k-NN acquired finely detailed distinctions between classes, the number of neighbours was set to 1. The distance metric employed was the Euclidean distance while the distance weight was set to equal. Next, the Medium k-NN achieved medium distinctions between classes, the number of neighbours was set to 10. Finally, Coarse k-NN produced coarse distinctions between the classes. The number of neighbours here was set to 100. On the other hand, when the Euclidean distance was changed to cosine distance, the Cosine k-NN method yielded medium distinctions between classes. The number of neighbours was set to 10, the distance weight was set to equal. At last, when the distance weight was changed to the square inverse, the Weighted k-NN yielded medium distinctions between classes. The number of neighbours was also set to 10, the distance metric used was the Euclidean distance.

C. THE EXPLAINABLE HYBRID TCDN-SVM MODEL

Once the videos are represented by TCDN method, a majority of the supervised classifications yielded good enough accuracy. The reason for such state-of-the-art performance still cannot be found. Moreover, the multiplication of the nonlinear layers in the TCDN network caused the decision process of this method to lack transparency. Considering the interpretability of the linear SVM model, we decided to use the LRP method[21] to explain the hybrid model composed of TCDN and SVM.

As depicted in Figure 3, the forward transfer process of the convolution neuron network(CNN) sends the message from the node of one layer to the node of the next layer as follows:

$$z_{ij} = x_i w_{ij} \tag{3}$$

$$z_j = \sum_i z_{ij} + b_j \tag{4}$$

$$x_j = g\left(z_j\right) \tag{5}$$

The x_i is the i-th element of the hidden layer l, weight w_{ij} links layer l with the next layer $l+1$, and the variable z_{ij} represents the forward message passed between the input neuron (i) and the output neuron (j). These forward messages were aggregated and combined after bias (b_j) was added. Then, it was input into the nonlinear activation function (g) to obtain the output (x_j). The commonly used activation function is relu $g(z_j) = \max(0, z_j)$.

Unlike forward propagation, LRP moves in the opposite direction of the layer to resolve the output of the classifier into a relevance message R. We set $R^{l,l+1}$ as relevance message which was sent from layer $l+1$ to layer l.

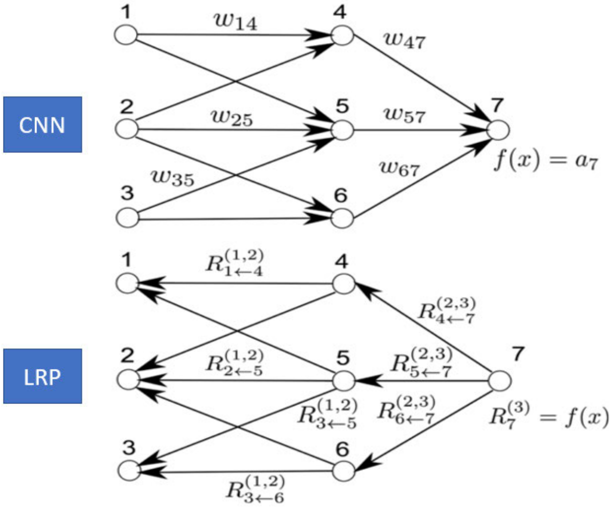
CNN
1
2
3
4
5
6
7
w_{14}
w_{25}
w_{35}
w_{47}
w_{57}
w_{67}
$f(x) = a_7$
LRP
$R_{1\leftarrow 4}^{(1,2)}$
$R_{2\leftarrow 5}^{(1,2)}$
$R_{3\leftarrow 5}^{(1,2)}$
$R_{3\leftarrow 6}^{(1,2)}$
$R_{4\leftarrow 7}^{(2,3)}$
$R_{5\leftarrow 7}^{(2,3)}$
$R_{6\leftarrow 7}^{(2,3)}$
$R_7^{(3)} = f(x)$

FIGURE 3. TOP: A neural network-shaped classifier(such as CNN)w_{ij} are weight and a_i is the activation of neuron i. Bottom: The neural network-shaped classifier during layer-wise relevance computation time. $R_i^{(l)}$ is the relevance of neural i. $R_{i\leftarrow j}^{l,l+1}$ are messages which need to be computed to ensure the relevance conservation principle [25].

A set of constraints need to be kept to ensure the relevance conservation principle of LRP is upheld during layer-wise relevance computation time:

$$R_j^{l+1} = \sum_i R_{i\leftarrow j}^{l,l+1} \tag{6}$$

$$R_j^{l} = \sum_j R_{i\leftarrow j}^{l,l+1} \tag{7}$$

$$f\,(\mathrm{x}) = \ldots = \sum_{j\in(l+1)} R_j^{l+1} = \sum_{j\in l} R_j^{l} = \ldots = \sum_{d=1}^{\dim(x)} R_d^1 \tag{8}$$

As for tasks involving image classification, the overall idea of LRP was to apprehend the impacts of each pixel from the input image on the final prediction by the classifier. In this study, we equipped three two-class SVM classifiers to complete the multi-classification task using one vs all strategy. In this strategy, for each binary learner, one of the three autistic behaviours was set as positive and all the remaining classes as negative. Then, the SVM is obtained for multi-tasking purpose. When the behavioural videos of people with autism were analysed, TCDN was used to extract the features of the frame. This input was then passed to the SVM classifier. Due to the similar structure between linear SVM and full connect layers in the AlexNet, the LRP method was used to explain TCDN-SVM, a model that is a mixture of TCDN and SVM.

This study utilised the fully connected linear LRP layer to perform the decomposition process of linear SVM because the structure of linear SVM is similar to that of the fully

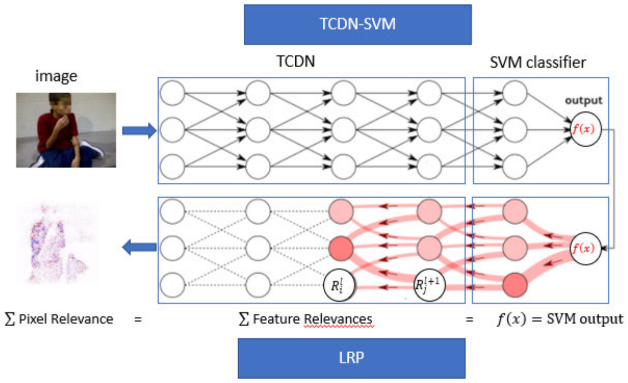

FIGURE 4. Visualization of the Layer-wise Relevance Propagation (LRP) decomposition process. In the classification step, the image is converted to a feature vector representation by TCDN and an SVM classifier is used to get a category. The LRP method decomposes the SVM output $f(x)$ into the sum of feature and pixel relevance score. The final relevances (heatmap) visualize the contributions of single pixels to the prediction.

connected linear layer.

$$z_{ij} = x_i w_{ij} \mathrm{z_j} = \sum_{\mathrm{i}} \mathrm{z_{ij}} + \mathrm{b_j} \tag{9}$$

$$\mathrm{R}^{\mathrm{l,l+1}}_{\mathrm{i}\underset{\mathrm{j}}{\leftarrow}} = \frac{z_{ij}}{z_j}\mathrm{R}^{\mathrm{l+1}}_{\mathrm{j}} \tag{10}$$

consider if the z_j is very small the relevance message $R^{(l,l+1)}_{i\leftarrow j}$ may become unbound. ε-decomposition formula was chosen, which introduces a sign-dependant numerical stabilizer ε in the formula.

$$\mathrm{R}^{\mathrm{l,l+1}}_{\mathrm{i}\underset{\mathrm{j}}{\leftarrow}} = \frac{z_{ij}}{\mathrm{z_j} + \varepsilon \cdot sign(\mathrm{z_j})}\mathrm{R}^{\mathrm{l+1}}_{\mathrm{j}} \tag{11}$$

In an ordinary image classification task, LRP usually uses the output of the softmax layer or fully connected layer in the artificial neural network as the initial input of LRP backpropagation. LRP can divide the relevance scores into positive and negative values in each layer using this step. When the LRP is propagated back into the image input layer, the relevance score at a pixel of the image becomes positive indicating that the pixel helps the model to classify the image into the correct category. Meanwhile, the colour is set to red in the heatmap. Conversely, if the correlation is negative, the pixel prevents the model from classifying the image into the correct category (the colour is set to blue). Hence, we can determine which area in the picture is important for the classification task. In our model, the output of SVM represents the distance between the sample and the hyperplane. We use this distance as the initial input of the LRP algorithm to determine the areas in each frame of the autistic patient's video that affect the classification results (distance).

IV. EVALUATION AND RESULTS

Firstly, the dataset was subjected to data preprocessing. Then, the evaluation method was introduced. The results of this study were split into three parts. In the first part, the K-means is used to classify the features and to evaluate the effectiveness of the features obtained using unsupervised learning. The results were then compared with the baseline method (random classification method) by condition entropy. In the second part, the accuracy and confusion matrix were used to evaluate the performance of the different supervised methods. Finally, the LRP output is analysed to assess the classification model.

A. DATA

In order to use real-time detection of the children's behaviour and provide early warning for parents in an uncontrolled environment, the SSBD [25] was selected as our training test dataset. The SSBD dataset contained 75 self-stimulatory behaviour videos of autistic children, which were classified into three categories. However, since these videos were obtained from public domains like Youtube, seven of the videos could not be downloaded due to copyright issues. We managed to randomly select 20 videos from each of the three classes to obtain a new dataset that contained 60 self-stimulatory behaviour videos.

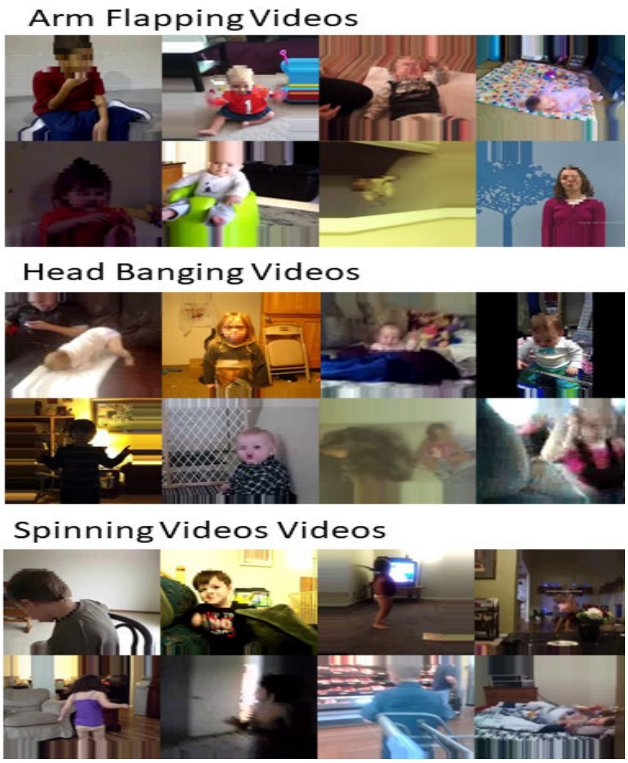

FIGURE 5. Introduction of SSBD dataset [31].

Figure 5 illustrates some snapshots of the three types of actions. The faces of the patients were masked with mosaics to protect their identities.

B. CLASSIFICATION USING AN UNSUPERVISED METHOD TO EVALUATE FEATURES

In order to ensure that our method becomes fully unsupervised, different K-means methods were used to classify the video data. The K-means is a traditional unsupervised classification method. In this study, the TCDN network was used to obtain 1024-dimensional features of each frame in the videos. We then passed them as inputs to the K-means algorithm for classification. Since the TCDN and K-means algorithms are both unsupervised algorithms, this method can be used to classify the behaviours of autistic patients completely unsupervised. For the K-means method, there are different types of initialising methods which could impact the performance of K-means classifier. Therefore, to obtain more credible results, we used three different methods to initialise K-means. The first method which is the K-means sample method randomly selects k observations from the sample set. The second method is the K-means uniform, which uniformly selects k points at random from a range of sample set. The last method selects k seeds by implementing the K-means++ algorithm for cluster centre initialisation. In step 1, one centre c_1 was chosen randomly and uniformly from the sample set. While in step 2 a new centre c_i with probability was chosen as follows:

$$f\ (\mathrm{x})\ \frac{D(x)^2}{\sum_{x\in X} D(x)^2} \tag{12}$$

$D(\mathrm{x})$ represents the shortest distance from a candidate data point x to the nearest centre which we have selected. Step 2 was repeated until k centres were considered [32]. Next, the impact of different K-means initialisation methods on K-means performance was compared. Considering that there were only three classes in our classification task, we set parameter k of K-means as 3.

Following video classification, determining a method to evaluate the performance of our model becomes a challenge. Tuytelaars et al [32] compared different methods to evaluate the unsupervised model. Based on the result of the said evaluation, Redondo-Cabrera, and Lopez-Sastre 2019 [22] have used unsupervised methods to perform human action identification, that is, using standard metrics named conditional entropy to evaluate models, as well as using a random method as the baseline. This method (conditional entropy) was used to evaluate our results. The conditional entropy method is as follows:

$$\mathrm{H}\,(\mathrm{X}\,|\,\mathrm{Y}) = \sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\log(\frac{1}{p(x|y)}) \tag{13}$$

After classification with different K-means. the random classification method was chosen as the baseline method as it provided a reasonable reference standard. In this way, the classification performance of K-means classifier can be evaluated by comparing with the baseline. The performance of any classification method should be better than the random classification method. Referring to the paper published by Redondo-Cabrera and Lopez-Sastre [22], the random classification method was used as a baseline for human motion recognition. Hence, the random classification method was employed to evaluate the efficiency of the unsupervised method on the autistic action dataset.

On the other hand, the parameters in the model were analyzed to assess the best performance. In this model, three important parameters may influence the performance of our model, namely (a) the margin α, (b) the temporal window to consider contiguous frames (w) and the non-neighbour frame index (n). However, considering that different videos may have different frame rates in the wild environment, the temporal window might not be suitable. Therefore, it is very important to evaluate this parameter in a sufficiently large dataset. In this article, we have referred to the settings of the published article [8] ($w = 1, n = 20$) because they have utilized a similar method for human motion recognition, and verified the temporal window on a larger data set (UCF101). However, considering that the movements of autistic patients are very different from that of normal people, we decided to analyse the optimal value of the margin parameter on our dataset.

TABLE 1. Condition entropy (CE) of K-means.

Method margin	CE 0.5	CE 1	CE 1.5	CE 2
random(baseline)	**1.56**	**1.56**	**1.56**	**1.56**
K-means uniform	1.12 (σ 0.17)	1.11 (σ 0.20)	1.06 (σ 0.21)	1.11 (σ 0.23)
K-means sample	1.16 (σ 0.14)	1.17 (σ 0.16)	1.15 (σ 0.17)	1.17 (σ 0.18)
K-means plus	1.13 (σ 0.17)	1.13 (σ 0.18)	1.09 (σ 0.19)	1.15 (σ 0.18)

Table 1 lists the conditional entropy and standard deviation obtained from various classification methods as the TCDN algorithm adopted different margin parameters. According to Table 1, the conditional entropy of random classification was 1.56, very close to the maximum conditional entropy $\log_2(3) = 1.58$. If the conditional entropy of a classifier results in a maximum conditional entropy, this classifier can be considered completely futile (such as a random classifier). Therefore, we can prove our baseline method (random classify) is absolutely random. However, each of our K-means methods with different K-means initialization methods was better than the baseline. In order to intuitively illustrate the impact of different margin parameters and the K-means algorithm on the classification effects, Figure 6 was plotted. For the self-stimulatory behaviour classification task, using margin 1.5 and K-means uniform method, our proposed model classified the different autistic behaviours very well.

In general, when our unsupervised temporal coherency deep network method and K-means uniform method were combined, the remaining uncertainty on the real autism behaviour categories was reduced from a random classification of $2^{1.56} = 2.95$ to $2^{1.06} = 2.08$ (our method). This result indicates that the proposed unsupervised method is useful for

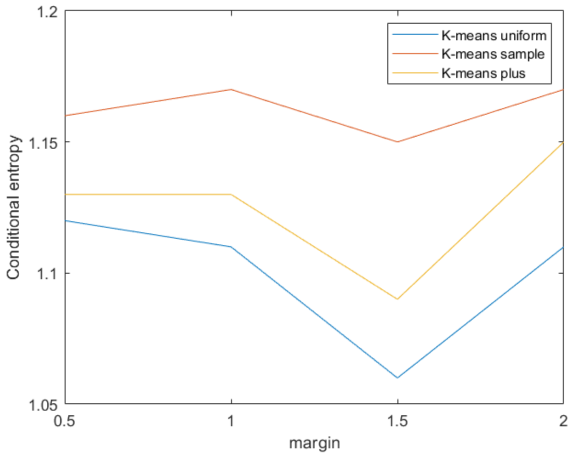

FIGURE 6. The influence of the margin.

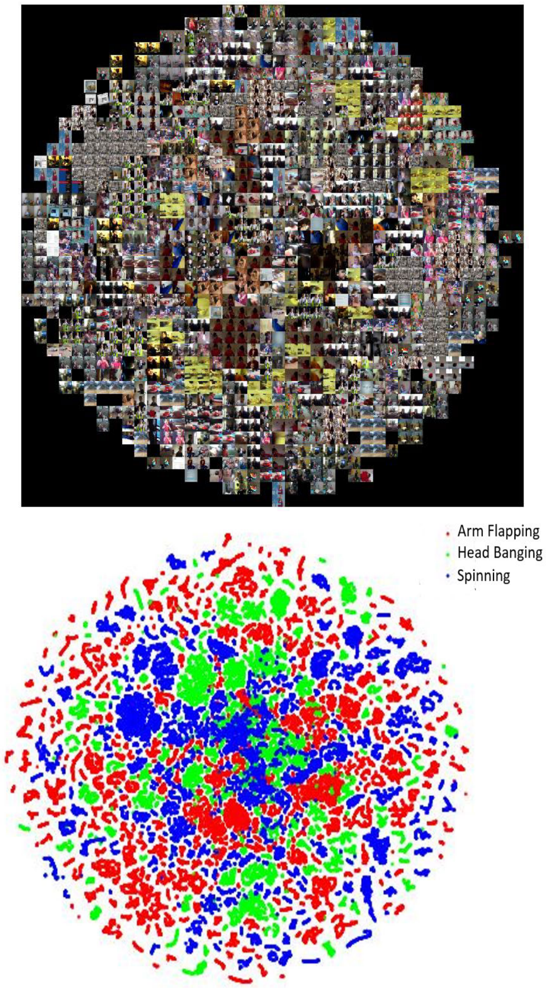

FIGURE 7. Barnes-Hut t-SNE 2-dimensional embedding with our 1024-dimension fc7-features TOP: Draw with frames BOTTOM: Draw with scatter.

the autistic data as the features extracted using this method are very effective.

Figure 7 depicts a 2-dimensional embedding using the Barnes-Hut t-SNE method. This method reduced the 1024- dimensional features data in this study to two dimensions by arranging pictures with similar features closely. This figure has shown the utilization of two different ways to show the clustering results of all self-stimulatory behaviours.

In this section, all of our K-means methods used the five-fold cross-validation. In order to test the validity of our features more rigorously, we then repeated our experiment 20 times. The average of these experimental results was accepted as our final result.

1) IMPLEMENTATION DETAILS

In the training process of TCDN, the mini-batch Stochastic gradient descent (SGD) method was used to train our unsupervised TCDN due to the lack of training resources and to maintain the stability of the training process. In the network, the convolutional layer of AlexNet was used as the basic structure before adding two fully connected layers on the pool of 5 layer outputs. Hence, we obtained 1024-dimensional features to calculate the loss function. During the training process, we set the batch size to 40 tuples of frames. As for the parameters of the network, the start learning rate was set at 0.001, while the temporal window was set at 20.

C. EXPERIMENTAL SETUP AND RESULTS OF SUPERVISED METHOD

In this section, the dataset from section IV is used. Initially, all the videos in the dataset are used to train the TCDN network. Then, the said network is used to extract the features of frames in all the videos. In total, 136613 features were gathered for all the frames. A 5-fold cross-validation method was used to evaluate the performance of our supervised methods. In this experiment, 11 supervised methods were used.

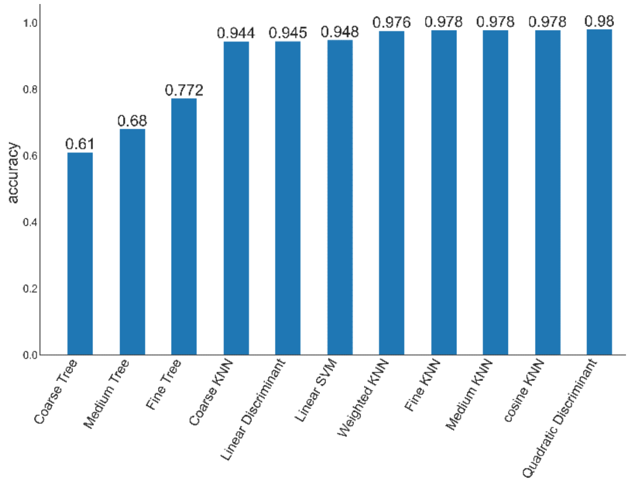

FIGURE 8. Comparison of classification accuracy at the frame level.

Figure 8 indicates the accuracy of each supervised method. Based on the observation, the Quadratic Discriminant method demonstrated the best accuracy at the frame level, up to 0.98. To comprehensively evaluate the performance of the Quadratic Discriminant method, the Confusion matrix of

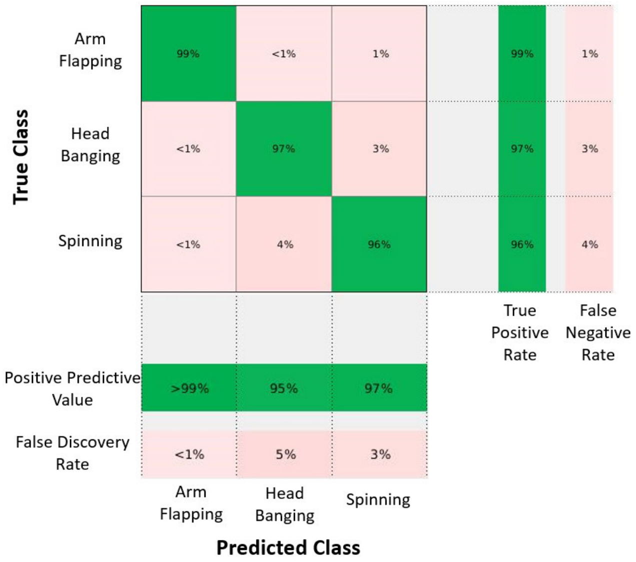

FIGURE 9. Confusion matrix of quadratic discriminant.

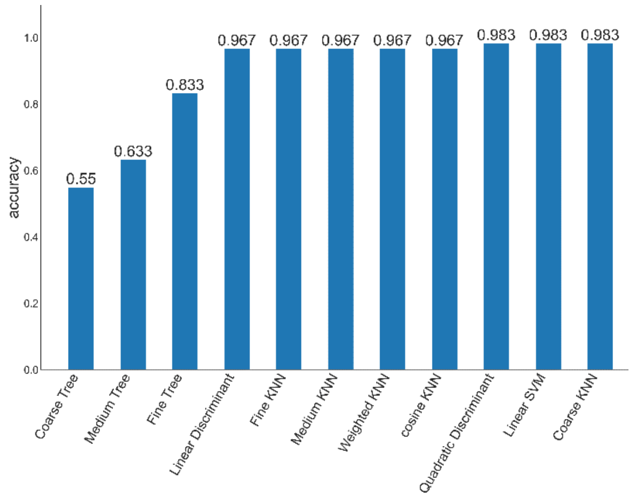

FIGURE 10. Accuracy at the video level.

our supervised methods were measured. Figure 9 represents the confusion matrix of the Quadratic Discriminant. Since our classification method is to classify by frame, the proposed method detected the movements of children with autism in real-time and they are diagnosed in real-time. However, in order to compare the results with previous studies, the accuracy of the video level was also calculated. In this study, the average classification score of all frames from a video was calculated (Figure 10), whereby the class with the largest average score was used as the classification of this video. The state-of-the-art accuracy for the Quadratic Discriminant, Linear SVM and Coarse k-NN methods were estimated at 98.3%.

Figure 11 shows the TCDN-SVM classification result of one Arm Flapping video.

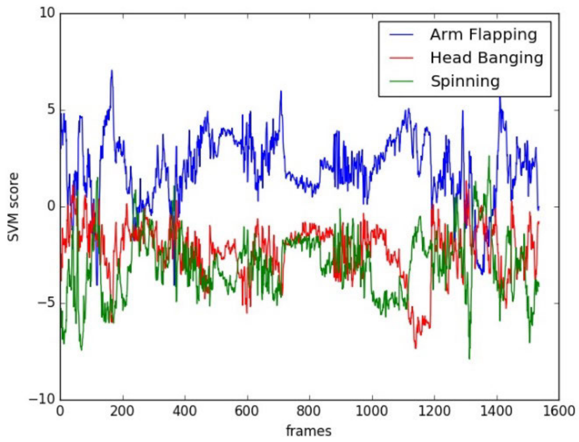

FIGURE 11. The SVM classifier output(score) in the TCDN-SVM model. This figure shows the score of each frame in an Arm Flapping video.

In the Matlab classification learner app, the three-class SVM classification task is decomposed into three binary-classification tasks. The classifier performs classification by calculating the distance between the features extracted by the TCDN algorithm and the three hyperplanes, and the classification category of the hyperplane with the largest distance is used as the final category of the feature.

TABLE 2. Comparison with the recent state-of-the-art result.

Method	accuracy
Poselet bounding box selection+ Histogram of Dominant Motions (HDM)+ discriminatory model [8]	73.6%
TCDN and Coarse KNN	98.3%
TCDN and Linear SVM	98.3%
TCDN and Quadratic Discriminant	98.3%
TCDN and Cosine KNN	96.7%
TCDN and Weighted KNN	96.7%
TCDN and Medium KNN	96.7%
TCDN and Fine KNN	96.7%
TCDN and Linear Discriminant	96.7%
TCDN and Fine Tree	83.3%
TCDN and Medium Tree	63.3%
TCDN and Coarse Tree	55%

Table 2 shows the comparison between our results and the state-of-the-art results published by Rajagopalan et al [8]. In the article published by Rajagopalan and Goecke [8], in order to track the child's body motion, they have utilized the nearest neighbour algorithm to select the postlet bounding box. In this detected body regions, a Histogram of Dominant motions (HDM) descriptors is computed and are being used to train a discriminatory model. On the other hand, we have instead, used slow-changing discriminative self-stimulatory behaviours features to train supervised models (such as SVM, k-NN, Discriminant), and to classify self-stimulatory

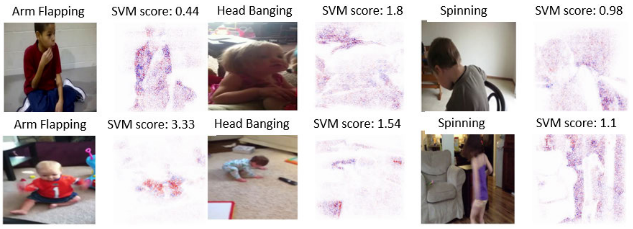

FIGURE 12. **Heatmap of the LRP output relevance(R_1).**

behaviours. Compared with the recent state-of-the-art result, we have obtained an improvement of 24.7%.

D. EXPLAINING TCDN-SVM USING LRP METHOD

In order to understand the internal mechanism of the TCDN-SVM model, we visualize the output result ($\boldsymbol{R_1}$) of the LRP algorithm and used the heatmap to represent it.

Figure 12 suggested that the basis of the model may be related to the magnitude of the action. When the model classifies arm-flapping behaviour, the model mainly focused on the effects of the arm on other parts of the body (such as occlusion). The head banging behaviour also has the upper body of the patient moving along with the head, so, the model begins to give some attention to the environment around the body. When the model begins to recognise the spinning behaviour, as the patient mainly rotates the whole body, the model recognises the human body influence on the surrounding environment (such as occlusion). Therefore, the heat map revealed that the model focused mainly on the environment near the body.

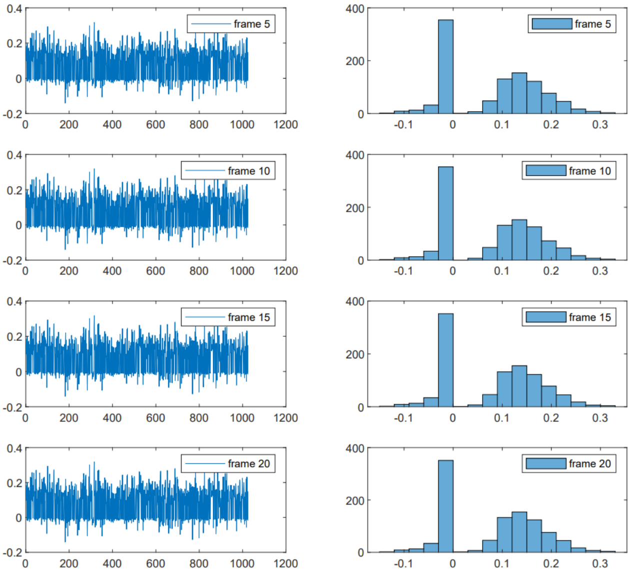

FIGURE 13. **Use line graphs and histograms to visualize features.**

As shown in Figure 13, we have visualized the features of four frames in the same video with 5 frames apart, the features extracted by the TCDN algorithm are slowly changing or even steady features, which makes it easier for the classifier to correctly classify the samples.

V. DISCUSSION

In this research, we have introduced an unsupervised feature learning method (TCDN) that can extract features from unlabeled videos. This method uses the local temporal coherence between contiguous frames as free supervision to obtain the ability of learning from unlabeled videos.

Compared with the existing pretrained CNN feature extraction method, the TCDN method can better adapt and optimize based on the researched dataset. Compared with the CNN extraction method that requires fine tune and retraining, TCDN can be trained in a completely unsupervised way to avoid annotating the dataset.

As shown in Figure 6, we have evaluated the impact of different margins on the clustering results. In order to separate the representation of different videos, a proper global discrimination margin is necessary.

As represented in Table 1, after comparing the results of different k-means classifiers with the random classification method (baseline), we revealed the effectiveness of a completely unsupervised method that combines TCDN and k-means. This gives us the ability to take advantage of many unlabeled autistic self-stimulatory behaviours videos.

Compared with the state-of-the-art result (73.6%) using HDM descriptor features and discriminant classifier, our slow-changing discriminative self-stimulation behaviours features and discriminant can achieve a higher accuracy (98.3%). This means that the features extracted by TCDN can improve the accuracy of 24.7% in autistic self-stimulation behavior classification task. The huge improvement of classification performance strongly proves the effectiveness and superiority of TCDN method.

Considering the future application in the medical field, it is necessary to understand the internal mechanism of the model. The analysis of the LRP output (Heatmap) indicated that our model classified the self-stimulatory behaviours by analysing the interaction between autistic patients and their surrounding environment. This mechanism laid a solid foundation for accurate classification of the model. The success and reasonable explanation of our model directly indicates the effectiveness of our model.

VI. CONCLUSION

In this study, we introduce an unsupervised feature learning method to extract the slow-changing discriminative self-stimulatory behaviours features from unlabeled videos.

Comparison of the conditional entropy results of the k-means classifier and the random classifier shows the efficiency of completely unsupervised classification using TCDN and k-means.

As compared with the recent state-of-the-art result (73.6%), our method is able to achieve a higher accuracy (98.3%). Considering the same classifier used in those two studies, the efficiency of the TCDN feature extraction method can be proven.

Overall, given the experimental results, we have confirmed that learning from unlabeled videos can enhance visual learning in the field of self-stimulatory behaviour research.

FUTURE WORKS

Although our proposed methods were successful, we still have some limitations. The data set used in this study was small, hence it cannot be generalized to fit nationwide data. In the future, collection of more self-stimulatory behaviours videos of autistic patients will be done to expand the dataset.

REFERENCES

[1] M. J. Maenner *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2016," MMWR Surveill. Summaries, New York, NY, USA, Tech. Rep. 4, 2020, vol. 69, no. SS-4, pp. 1–12, doi: 10.15585/mmwr.ss6904a1externalicon.

[2] A. Estes, J. Munson, S. J. Rogers, J. Greenson, J. Winter, and G. Dawson, "Long-term outcomes of early intervention in 6-year-old children with Autism Spectrum Disorder," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 54, no. 7, pp. 580–587, Jul. 2015, doi: 10.1016/j.jaac.2015.04.005.

[3] P. T. Shattuck, M. Durkin, M. Maenner, C. Newschaffer, D. S. Mandell, L. Wiggins, L.-C. Lee, C. Rice, E. Giarelli, R. Kirby, J. Baio, J. Pinto-Martin, and C. Cuniff, "Timing of identification among children with an Autism Spectrum Disorder: Findings from a population-based surveillance study," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 48, no. 5, pp. 474–483, May 2009, doi: 10.1097/CHI.0b013e31819b3848.

[4] L. Zwaigenbaum, S. Bryson, and N. Garon, "Early identification of Autism Spectrum Disorders," *Behav. Brain Res.*, vol. 251, pp. 133–146, Aug. 2013, doi: 10.1016/j.bbr.2013.04.004.

[5] J. M. Rehg, "Behavior imaging: Using computer vision to study autism," *Proc. 12th IAPR Conf. Mach. Vis. Appl.*, 2011, pp. 14–21.

[6] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, Jr., B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disorders*, vol. 30, no. 3, pp. 205–223, 2000, doi: 10.1023/A:1005592401947.

[7] S. E. Bryson, L. Zwaigenbaum, C. McDermott, V. Rombough, and J. Brian, "The autism observation scale for infants: Scale development and reliability data," *J. Autism Develop. Disorders*, vol. 38, no. 4, pp. 731–738, Apr. 2008.

[8] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1470–1474, doi: 10.1109/ICIP.2014.7025294.

[9] F. S. Alnajjar, A. M. Renawi, M. Cappuccio, and O. Mubain, "A low-cost autonomous attention assessment system for robot intervention with autistic children," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2019, pp. 787–792, doi: 10.1109/EDUCON.2019.8725132.

[10] F. Alnajjar, M. Cappuccio, A. Renawi, O. Mubin, and C. K. Loo, "Personalized robot interventions for autistic children: An automated methodology for attention assessment," *Int. J. Social Robot.*, vol. 6, pp. 1–6, Mar. 2020, doi: 10.1007/s12369-020-00639-8.

[11] K. D. Cantin-Garside, Z. Kong, S. W. White, L. Antezana, S. Kim, and M. A. Nussbaum, "Detecting and classifying self-injurious behavior in Autism Spectrum Disorder using machine learning techniques," *J. Autism Develop. Disorders*, vol. 50, no. 11, pp. 4039–4052, Nov. 2020, doi: 10.1007/s10803-020-04463-x.

[12] C. H. Min, "Automatic detection and labeling of self-stimulatory behavioral patterns in children with Autism Spectrum Disorder," *Proc. Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jul. 2017, pp. 279–282, doi: 10.1109/EMBC.2017.8036816.

[13] N. Mohammadian Rad, S. M. Kia, C. Zarbo, T. van Laarhoven, G. Jurman, P. Venuti, E. Marchiori, and C. Furlanello, "Deep learning for automatic stereotypical motor movement detection using wearable sensors in Autism Spectrum Disorders," *Signal Process.*, vol. 144, pp. 180–191, Mar. 2018, doi: 10.1016/j.sigpro.2017.10.011.

[14] J. Hashemi, T. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro, "A computer vision approach for the assessment of autism-related behavioral markers," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenetic Robot.*, Nov. 2012, pp. 1–7, doi: 10.1109/DevLrn.2012.6400865.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Int. workshop human Behav. Understand.*, 2011, pp. 29–39.

[16] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 463–469, doi: 10.1109/IJCNN.2017.7965890.

[17] D. Li, S. Yan, M. Zhao, and T. W. S. Chow, "Spatiotemporal tree filtering for enhancing image change detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8805–8820, Aug. 2020, doi: 10.1109/TIP.2020.3017339.

[18] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[19] F. Dawood and C. K. Loo, "Developmental approach for behavior learning using primitive motion skills," *Int. J. Neural Syst.*, vol. 28, no. 04, May 2018, Art. no. 1750038.

[20] F. Dawood and C. K. Loo, "Incremental episodic segmentation and imitative learning of humanoid robot through self-exploration," *Neurocomputing*, vol. 173, pp. 1471–1484, Jan. 2016.

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Muller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.

[22] C. Redondo-Cabrera and R. Lopez-Sastre, "Unsupervised learning from videos using temporal coherency deep networks," *Comput. Vis. Image Understand.*, vol. 179, pp. 79–89, Feb. 2019.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 1, 2012, pp. 1097–1105.

[24] T. Westeyn, K. Vadas, X. Bian, T. Starner, and G. D. Abowd, "Recognizing mimicked autistic self-stimulatory behaviors using HMMs," in *Proc. 9th IEEE Int. Symp. Wearable Comput. (ISWC)*, Oct. 2005, pp. 164–167, doi: 10.1109/ISWC.2005.45.

[25] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 755–761, doi: 10.1109/ICCVW.2013.103.

[26] G. De'Ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000, doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO.2.

[27] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[28] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1277–1305, Jun. 2007.

[29] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[30] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[31] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, New Orleans, LA, USA, 2006.
Can use this cite : SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms January 2007, pp 1027–1035

[32] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 284–302, Jun. 2010, doi: 10.1007/s11263-009-0271-8.

SHUAIBING LIANG (Graduate Student Member, IEEE) received the bachelor's degree in bioinformatics from Harbin Medical University, in 2018. He is currently pursuing the master's degree in computer science with the University of Malaya. His research interests include artificial intelligence, machine learning, and bioinformatics.

AZNUL QALID MD SABRI (Member, IEEE) is a graduate of the prestigious Erasmus Mundus Master in Vision and Robotics (ViBot), a Master program jointly coordinated by three different universities, such as the University of Burgundy, France, the University of Girona, Spain, and Heriot-Watt University Edinburgh, U.K. He received the master's degree by performing a research internship program from the Commonwealth Scientific Research Organization (CSIRO), Brisbane, QLD, Australia, focusing on medical imaging, and the Ph.D. degree in human action recognition (completed with distinction, très honorable), under a program jointly offered by a well-known research institution in France, Mines de Douai (a research lab) and the reputable University of Picardie Jules Verne, Amiens, France. He is currently a Senior Lecturer with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology (FCSIT), University of Malaya, Malaysia. He is an active researcher in the field of artificial intelligence, having published in multiple international conferences as well as international journals. His main research interests include computer vision, robotics, and machine learning. He is a part of the pioneering members of FCSIT's COVIRO (Cognitive, Vision and Robotics) Research Group. He is currently the Principal Investigator of multiple research grants.

FADY ALNAJJAR received the M.Sc. degree in artificial intelligence systems and the Ph.D. degree in system design engineering from the University of Fukui, Japan, in 2007 and 2010, respectively. From 2010 to 2016, he worked as a Research Scientist with the Brain Science Institute (BSI), RIKEN, Japan, where he has been a Visiting Researcher, since 2016. He has joined the College of Information Technology, UAE University, in 2016, where he has been the Head of AI and Robotics Laboratory, since 2016. He is working in neuro-robotics study to understand the underlying mechanisms for embodied cognition and mind. He is also exploring the neural mechanisms of motor learning, adaptation, and recovery after brain injury from the sensory- and muscle-synergies perspectives. His research target is to propose an advance neurorehabilitation methodology for patients with brain impairments, such as Children with autism, elderly with cognitive impairment, post-stroke patients.

CHU KIONG LOO (Senior Member, IEEE) received the B.Eng. degree (Hons.) in mechanical engineering from the University of Malaya, in 1996, and the Ph.D. degree in neurorobotics from University Sains Malaysia, in 2004. He is currently a Full Professor with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya. His main research interest includes neuroscience inspired machine intelligence. He was the IEEE Systems, Man and Cybernetics SMC Society Vice-Chairman for Malaysia Chapter, from 2013 to 2014. He was also the President of Asia Pacific Neural Network Assembly (APNNA), in 2014.