Neurocomputing 446 (2021) 145–155

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Vision-assisted recognition of stereotype behaviors for early diagnosis of Autism Spectrum Disorders

Farhood Negin [a], Baris Ozyer [b], Saeid Agahian [c,*], Sibel Kacdioglu [b,c], Gulsah Tumuklu Ozyer [b]

[a] INRIA Sophia Antipolis, Valbonne 06902, France
[b] Computer Engineering Department, Ataturk University, 25240 Erzurum, Turkey
[c] Computer Engineering Department, Erzurum Technical University, 25050 Erzurum, Turkey

ARTICLE INFO

Article history:
Received 9 September 2020
Revised 27 February 2021
Accepted 3 March 2021
Available online 13 March 2021
communicated by Zidong Wang

Keywords:
Action recognition
Autism Spectrum Disorder
Patient monitoring
Bag-of-visual-words
Convolutional neural networks

ABSTRACT

Medical diagnosis supported by computer-assisted technologies is getting more popularity and acceptance among medical society. In this paper, we propose a non-intrusive vision-assisted method based on human action recognition to facilitate the diagnosis of Autism Spectrum Disorder (ASD). We collected a novel and comprehensive video dataset f the most distinctive Stereotype actions of this disorder with the assistance of professional clinicians. Several frameworks as a function of different input modalities were developed and used to produce extensive baseline results. Various local descriptors, which are commonly used within the Bag-of-Visual-Words approach, were tested with Multi-layer Perceptron (MLP), Gaussian Naive Bayes (GNB), and Support Vector Machines (SVM) classifiers for recognizing ASD associated behaviors. Additionally, we developed a framework that first receives articulated pose-based skeleton sequences as input and follows an LSTM network to learn the temporal evolution of the poses. Finally, obtained results were compared with two fine-tuned deep neural networks: ConvLSTM and 3DCNN. The results revealed that the Histogram of Optical Flow (HOF) descriptor achieves the best results when used with MLP classifier. The promising baseline results also confirmed that an action-recognition-based system can be potentially used to assist clinicians to provide a reliable, accurate, and timely diagnosis of ASD disorder.
© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Autism Spectrum Disorder (ASD) is a neurobiological disorder having significant symptoms such as social communication and interaction deficiencies, repetitive behavior, limited language, and mental development. ASD cannot be detected during pregnancy, but assumed to start while in the womb and continues for a life without any assured treatment. During recent years, researchers have been investigating whether autism occurs due to genetic, biological, and environmental causes [1]. However, the exact reasons for the genetic modifications and other factors contributing to ASD still remain as a riddle. In addition, the number of worldwide cases diagnosed with ASD dramatically increases every year [2]. According to the research conducted by the US Department of Health, while in the 1990s, 1 out of 150 newborn children was diagnosed with ASD, this ratio has been increased to 1 out of 54 children over the past few years [3].

Diagnosing ASD in minors is still a big challenge for physicians as it requires long-term monitoring and examination of the patients' symptoms. Primary identification of the ASD symptoms is suggested by Leo Kanner [4] in 1943 through observation of 11 children who showed abnormalities in communication skills, social interactions, and repetitive behaviors. Physicians have been therefore focused on social and communication deficits of ASD to make diagnosis [5]. However, some communicative and behavioral symptoms associated with autism may not be unique to autism, which makes it difficult to have an early diagnosis [6]. Similar to the other health problems, early diagnosis of ASD in the initial years is very important to prevent it from developing other symptoms and minimize social interaction and behavioral disorders [7]. The special education and therapies put into practice for children diagnosed with ASD during their infancy yielded positive impacts as they get older [8]. It is a common agreement among physicians that special training and therapies for treatment should begin as early as possible, especially at a very young age [9,10]. Despite knowing the significance of a timely diagnosis, accurate diagnoses were mostly carried out not earlier than the ages of 4 or 5 that is already late for the treatment.

The repetitive behaviors are considered significant clues for the diagnosis of ASD [11]. These repetitive behaviors have been catego-

* Corresponding author.
E-mail addresses: farhood.negin@inria.fr (F. Negin), baris.ozyer@atauni.edu.tr (B. Ozyer), saeid.agahian@erzurum.edu.tr (S. Agahian), sibel.kacdioglu@erzurum.edu.tr (S. Kacdioglu), gulsah.ozyer@atauni.edu.tr (G.T. Ozyer).

https://doi.org/10.1016/j.neucom.2021.03.004
0925-2312/© 2021 Elsevier B.V. All rights reserved.

rized into five main subgroups by Kristen et al. [12]: stereotype, self-injurious, compulsive, ritualistic, and sameness behavior. The focus of this study is on stereotype behavioral acts, specified as self-stimulatory behaviors that are recurrently performed mostly without any noticeable adaptive function. Unlike other subgroups, these repetitive movements are not subjected to any rules or restrictions as in activities of daily living [12,13]. The most common stereotype behaviors of ASD include flapping arms like wings, shaking head back and forth, and spinning around itself, etc. In clinical settings, physicians monitor these patients' behaviors while they are playing with their parents or toys as if they are at their homes [14]. However, patients mostly cannot act naturally during the treatment because they consider the physicians as authority [15]. The most common method for early diagnosis of the symptoms is a report obtained by parents which is not objective [16]. Because in most cases, parents' misunderstanding and biased assessments due to emotional sensitivity to their children reduce the reliability of the information [17]. Home video recordings were therefore used to monitor a child's behavior on special days or daily lives in an uncontrolled environment [7,18]. However, the analyzes and examinations depend only on the observations obtained directly through the videos and experience of physicians.

Nowadays, computer-aided technologies are an indivisible part of the health-care systems thanks to their low cost, reliable, and accurate performances. Accordingly, vision-assisted solutions call the attention of researchers due to their contactless and non-intrusive nature. However, the performance of vision-based systems in uncontrolled environments is still limited because of the complications of camera movement, object appearance and exposure, object scale, camera viewpoint, diffused background, and lighting [19]. In addition to these challenges inherent to vision-based systems, ASD children exhibit specific characteristics of repetitive behaviors, which contribute to the complexity of analyzing such videos. For instance, the repetitive actions can be triggered suddenly and the amplitude and frequency of the action patterns can vary over time. Multiple repetitive behaviors can also be performed at the same time and can be influenced by existing objects in the scene. Besides, the video quality may be reduced due to the shaky camera held by the parents during the recordings [20,21]. Above all, there is a serious lack of data in this domain, which hinders fast-paced progress. To the best of our knowledge, there are only a few ASD datasets that are publicly available. However, most of them [22] focus on the study of facial expressions and eye movements, and disregard the importance of corporal behavior in the diagnosis of the disorder.

Motivated by the above-mentioned challenges, our aim is to automatically analyze the stereotype behavior of ASD in an uncontrolled environment based on action recognition approaches. The contributions are as follows:

- We introduce recognition of repetitive actions as a new challenge in action recognition which is not examined in the context of ASD.
- To analyze such actions, we introduce a novel ASD action recognition dataset for evaluation of the developed frameworks and for further studies in the future.
- We develop two different action recognition frameworks to analyze ASD behaviors and produce baseline results for the collected dataset. The first one follows the bag-of-visual-words approach that accepts various feature types as input. The second one receives pose sequences as input and models actions using an LSTM network.
- We evaluate the ASD actions and we report the results of extensive experiments conducted on our dataset. We also compare our results with variations of the developed frameworks and two recent fine-tuned deep neural network based frameworks.

In the rest of the paper, first, we provide a brief explanation of the state-of-the-art methods in Section 2. We present the developed and evaluated frameworks in Section 3. Detailed explanation about the collected dataset, report of the obtained results, their interpretation, and discussion are presented in Section 4. We conclude the paper in Section 5.

2. Related work

Over the past few years, there is a strong interest among researchers to automatically analyze ASD behaviors. Different machine learning and computer vision algorithms have been developed to recognize these behaviors from videos recorded in controlled and uncontrolled environments [23–29]. Li et al. [25] took advantage of facial attributes such as facial expressions, action units, arousal, and valence to classify ASD using convolutional neural networks. In [26], the authors proposed a semi-automatic and non-instructive approach to estimate body-pose based on a head motion by tracking facial features. In [27], the authors developed a video-based computational method by making use of recurrent neural networks to distinguish healthy subjects from ASD subjects by analyzing them while performing grasping gestures. Similarly in [28], different gaze patterns and track eye movements are investigated in videos. Unlike the former method, they classified ASD children by means of a multi-layer Long Short-Term Memory (LSTM) network. Overall, the main focus of these studies has been mostly on analyzing facial expression and emotions, gestures, eye movement tracking, and cognitive behavior modeling. The main shortcoming of these methods is disregarding the corporal behaviors of the subjects, which carry crucial information and are influential in analyzing and evaluation of the disorder.

The behavior of ASD children is characterized by their impaired interpersonal communication abilities, lack of response through eye contact, and head orientation [30]. Within the context of the social and communicative behavior of ASD children, authors in [31] provided the Multi-Modal Dyadic Behavior (MMDB) dataset. They tried to decode the engagement level of social dyadic interactions by analyzing mid-level behavioral cues from an activity recognition perspective. However, the recorded data can be only used to analyze the interactive skills of children rather than their characteristic physical behavior. The evolution of body poses of ASD patients over time is investigated by [32] to distinguish typical from atypical behaviors using low-level latent information. Their collected dataset included daily living activities of autistic children obtained by means of the NODA platform of Behavior Imaging Company [33]. The videos are relatively long (2 to 10 min) suited for analyzing long-term daily living behavior of subjects as opposed to short, quick, and repetitive behaviors. Recently, a novel sensitive action and emotion recognition system was developed to be used in robot-assisted therapy sessions of ASD children [34]. They used the DE-ENIGMA dataset containing multi-modal recordings of therapy sessions of ASD children assisted with robots or physicians. They investigated the physical activities and emotional behaviors of children through a 3d human pose reconstruction method. Likewise, by their dataset, they could only evaluate interactive child-robot and child-therapist activities. In another study [35], an autism dataset which was consisted of 3D skeletons was collected using the Kinect sensor, however, healthy subjects mimic stereotyped behaviors.

Lately, instead of examining interactive behaviors or relying on intricate facial expressions, researchers started to realize the impact of repetitive behaviors in the diagnosis of ASD disorder. Several groups tried to design wearable sensors to capture motion patterns of the data recorded from the wrist or other body parts [36,37]. Most of them utilized accelerometer sensors to measure the magnitude

and direction of low-frequency accelerations and to detect stereotype behaviors in daily activities. To obtain more precise measurements, some studies [38,39] integrate acoustic sensors and cameras (mounted on a static platform) with accelerometer. The primary objective of these studies is to provide a system to alert parents and help them to prevent self-injurious behaviors when the stereotype actions are triggered. Nonetheless, these devices are usually intrusive, uncomfortable, and even sometimes make allergies and rashes in the skin, which makes them discouraging to use. Moreover, for effective utilization, the user needs to follow a precise instruction, which is not always an easy task, especially for ASD children. Some other groups picked a different path and assessed repetitive behaviors by means of videos from an action recognition aspect. The one glimpse early ASD detection (O-GAD) network [40] that consists of a 3D ConvNet temporal feature extractor and a temporal pyramid network is proposed to detect ASD actions and then discriminate repetitive behaviors. Although these vision-based solutions are non-intrusive, low-cost, and effective, their usage is still limited due to the constraints of vision and action recognition systems. These limitations are mostly related to the temporal modeling of actions. In recent years, three main category of methods are introduced to address those limitations. Two-stream networks [41] (successor of single-stream networks [42]) jointly learn spatial information by encoding appearance features and temporal information by encoding short-term motion through optical flow. However, temporal information is not efficiently encoded through optical flow and the long-term motion is also ignored in this method. An obvious solution for this problem is sequential networks that uniformly sample features from video frames and feed them to a recurrent network such as LSTM [43,44]. Despite their successful performance in different scenarios, these methods fail in actions when there are subtle changes in the scene during the action. In addition, spatial and temporal operations are dissociated which disarms the model to extract intrinsic spatiotemporal patterns. 3D convolutions [45] and attention mechanism [46] are introduced to describe the temporal structure of the videos. They use spatiotemporal filters to capture local spatiotemporal information from short segments of videos (coupled by spatial inflation and inception mechanism [47]). To capture global context, attention mechanism within a CNN-RNN encoder-decoder framework is utilized. Other than having more parameters compared to 2D convolutions due to additional kernel dimension, rigid spatiotemporal kernels restrain capturing subtle motions. Besides, they have no operation to disambiguate the similarities in actions. Also, 3D CNNs are not view invariant.

Along the same lines, our focus in the paper is analyzing the stereotype behavior of the ASD children from videos in an uncontrolled environment by action recognition approaches. A vision-based methodology is proposed in [20] where a standard bag-of-words pipeline was utilized for evaluation. Within the context of their project, the Self-Stimulatory Behavior Dataset (SSBD) was collected, in which videos were captured by parents using mobile devices from the daily lives of ASD children and shared in social media. The authors have also proposed an algorithm to detect self-stimulatory behaviors using poselet bounding boxes and global descriptors based on dominant motion patterns [21]. Unlike these studies, we collected a new public video dataset Expanded Stereotype Behavior Dataset (ESBD) from public social media channels consisting of four different behaviors: spinning, arm-clapping, head-shaking, and finger movement [48]. Our dataset includes one more behavior (hand action) and the number of video samples is two times compared to the previous benchmark dataset [20,21]. In addition, the videos are labeled under the guidance of experts of the Child and Adolescent Psychiatrist center at Ataturk University. Other than a proposed bag-of-visual-words method that utilizes a combination of feature histograms for action recognition, we develop different frameworks and their variations to produce extensive baseline results for comparison.

3. Method

In this section, we explain a series of developed methods for evaluating the performance of various action recognition techniques on our collected dataset. These methods leverage different modalities in order to produce baseline results for further research in the future. Based on the feature representation modalities, these methods can be categorized as: **Local descriptor models**, **Skeleton based**, and **deep learning** methods. The entire blocks of the scene understanding pipelines are developed or fine-tuned to adjust the task in hand (person detection, action recognition, etc.). Our focus is on the action recognition task to analyze behaviors of the subjects prone to ASD. Actions can be described as a sequence of articulated poses, representations from extracted local descriptors, or learned representations. The developed frameworks acquire these features to train the parameters of action models. It is assumed that only one action instance is performed in each video. Therefore, even if there are multiple persons in the captured video, only one subject performs the target action.

3.1. Local descriptor action models

Fig. 1 depicts the proposed action recognition framework based on local descriptors. The framework incorporates two phases, which follows a Bag-of-Visual-Words configuration. In the first phase, the visual codeword vocabulary is constructed by extracting the most informative visual features that can mutually discriminate the actions. Random samples from the training set are used for generating the vocabulary. In the second phase, the whole dataset is represented as a series of codewords using the learned vocabulary in the first phase. The videos in the dataset are not recorded in a controlled environment and similar to daily living scenarios, the background is cluttered. Such backgrounds do not help discrimination of the actions and even add extra noise to the representations. Three preprocessing steps are therefore applied to reduce the effects of a cluttered background.

3.1.1. Preprocessing

Person detection: There is no human–human or human-object interaction in the videos. Here, the aim is to detect the subject who performs the actions. To this purpose, three different person detection algorithms are fine-tuned and utilized. **Haar-Cascade** [49] is an object detection method applied to detect people at each frame. The **HOG** [50] descriptor, a gradient-based detector, is also used for detection of the subjects. The gradients are extracted by 8 by 8 pixel grids where the feature histograms are calculated accordingly. After normalization, these histograms are used for detection in the scene. Additionally, **YOLOv3** [51], an end-to-end target object detection network (with 3×3 and 1×1 convolutional and residual layers), is used for person detection. It divides the input image into a grid of $S \times S$ where B anchor boxes for each grid cells are predicted. YOLOv3 predicts bounding boxes at 3 scales. The output tensor contains bounding box coordinates, objectness and class predictions. A pretrained model with COCO [52] dataset with 80 classes is used. Therefore, by predicting 3 boxes at each scale, the predicted tensor is $S \times S \times [3 \times (4 + 1 + 80)]$ for 4 bounding box offsets, 1 objectness prediction and 80 class labels. The bounding boxes labeled as "person" are preserved if they pass filtering imposed by threshold and non-maximum suppression. In our experiments, the input for the network is a 416×416 tensor and outputs can be in form of tensors in three scales: $13 \times 13 \times 255, 26 \times 26 \times 255$, and $52 \times 52 \times 255$. As videos are

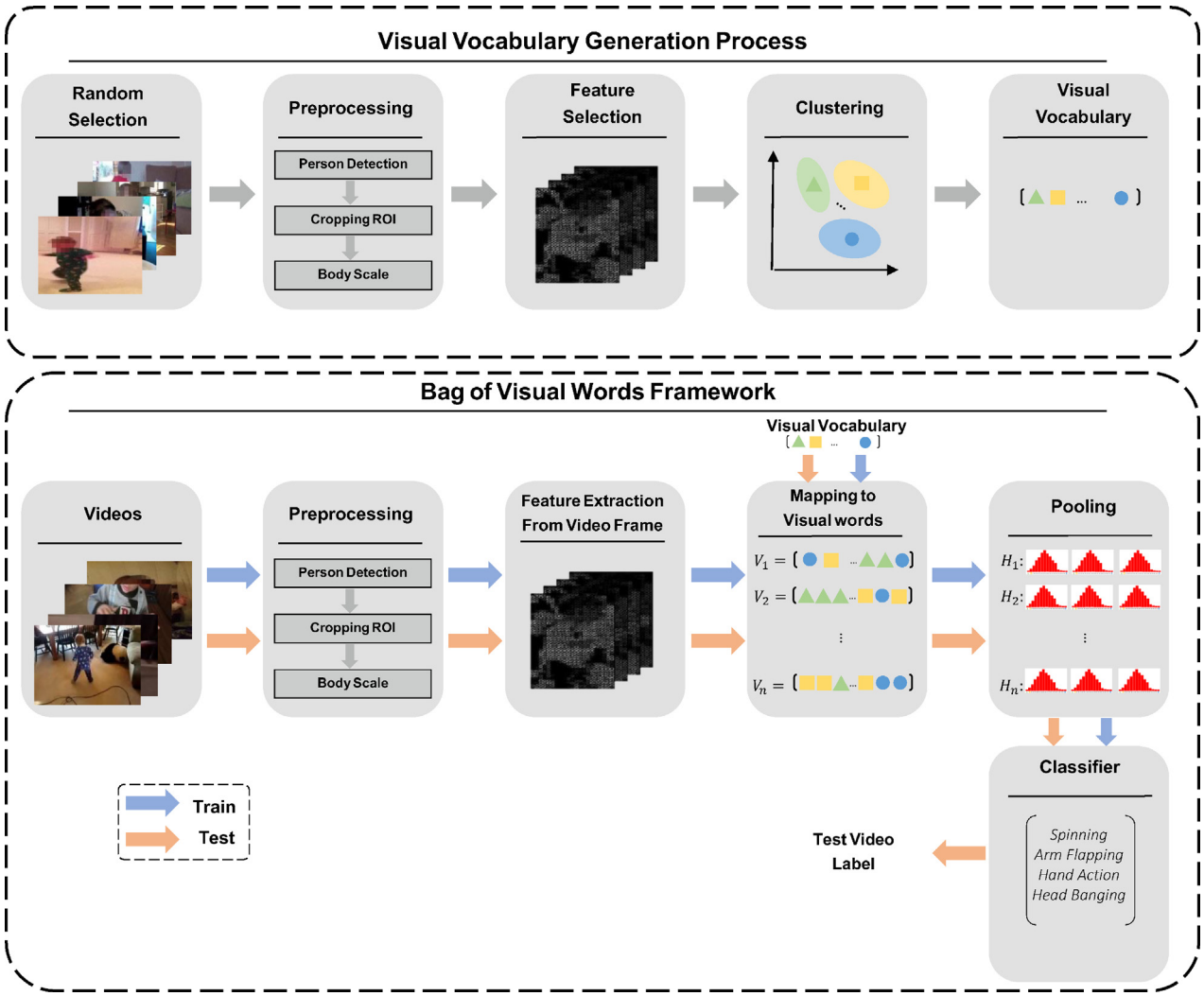

Fig. 1. Action recognition pipeline in a Bag-of-Visual-Words framework. The dashed box on top is the construction of the visual vocabulary and the box below is the histogram representation and classification phase of the framework.

captured in real-world situations, people except for the target subject occasionally appear in the scene, which can make the feature extraction process inconsistent. To overcome this problem, detection is followed by a tracking operation. SORT tracking algorithm [53] based on Kalman filters is employed to perform tracking. SORT receives detected bounding boxes from YOLOv3 and assigns indices to each preventing a potential mismatch among different subjects in the scene.

Detection of Region Of Interest (ROI): The evaluated actions in our dataset do not include interactive actions. Therefore, to recognize an action, information of the subject who performs that action is adequate. Hence, prior to the feature extraction steps, the region including the subject at each frame is cropped using detected bounding boxes.

Body ratio normalization: Subjects are from different age ranges in the videos. The body ratio of a 3 years old child could be significantly different from a 6 years old child. To keep the evaluations independent from the body ratios, the clipped subject bounding boxes are resized to have a fixed size by keeping height to width ratios unchanged.

3.1.2. Feature extraction

Feature extraction is required to describe the motion and appearance of the acquired ROIs. Considering their fast extraction and easy implementation, in this work, we use Histogram of Oriented Gradient (HOG) [50], Histogram of Optical Flow (HOF) [54], HOG-HOF combination, the Scale Invariant Feature Transform (SIFT) [55], and the Speed Up Robust Features (SURF) [56] descriptors. **HOG** is an appearance-based descriptor that relies on the density of local gradients. For each image grid, HOG extracts a $[1, 3780]$ feature vector. **HOF** exploits optical flow information to describe local motion originated from object or camera movements in consecutive frames. It produces a $[1, 18954]$ sized feature vector. The **HOG-HOF** descriptor is a $[1, 22734]$ length vector that is constructed by concatenation of HOG and HOF feature vectors. Furthermore, SIFT ($[1, 128]$ dimension) and SURF ($[1, 64]$ dimension) image-based descriptors are extracted. SIFT depends on image features that stay invariant to illumination, rotation, and scale changes on images. The latter works in a similar way as SIFT, however, instead of heavy Gaussian filter calculations, it employs cost-effective box filters.

3.1.3. Visual codeword calculation

Every extracted feature from an image grid is a visual word. Some of these visual words contain irrelevant, unnecessary information (e.g. noise or background), while some contain important representative information that can be useful to recognize target actions. To get rid of those outliers, and to acquire valuable core descriptors, K-means [57] algorithm is applied. The obtained cluster centers are codewords of the calculated visual vocabulary.

Therefore, the training set with n videos each having variable number of frames will have the frame set $X = [x_1, x_2, x_3, \ldots, x_n]$. Feature vectors of randomly sampled 10 K frames out of set X are computed. After performing iterative K-means process for clustering the descriptors, the visual vocabulary (V) containing representative codewords can be represented as: $V = [P_1, P_2, \ldots, P_K] \in R^{D\times K}$ where D is the dimension of feature vectors and K is the number of codewords in the vocabulary. The feature vector of a video frame is computed and then, the codeword representation of it is constructed by quantization. The quantization is done by calculating the closest Euclidean distance of a given frame descriptor to the codewords of the visual vocabulary. This way, the video is converted into a sequence of codewords: $F = [f_1, f_2, \ldots, f_i, \ldots, f_N] \in R^{D\times N}$ where N is the number of frame in the given video. In order to produce equal-sized descriptors, the codeword sequences are converted into histogram representations. Given F as a sequence of visual codewords for a video, k^{th} bin of the histogram representation is calculated as:

$$Hist_k(F) = card(\{f_i | LB(f_i) = k\}) \tag{1}$$

where $LB(f_i)$ is the codeword label of f_i visual word and $card()$ operator calculates the cardinality of a given set. By codeword representation the temporal information is missed in the calculated feature vectors [58]. Similar to temporal bag-of-words approach [59], to compensate for the loss of temporal information, visual codeword sequence is divided into two segments with an equal number of codewords, where each segment is represented by an individual histogram. The final representation consists of a combination of three histograms that one is calculated from the entire sequence and the other two belong to each one of the segments.

3.1.4. Classification

Three efficient classifiers are applied for classification: Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB) classifiers. MLP is a feedforward neural network that learns to adapt the weights of the neurons at each iteration. It converts inputs to the anticipated responses (labels) based on prior information. At each layer, an activation function (f_h) multiplies weights (w_i) with input values (x_i), adds a bias term (θ), and finally returns layer's activation (y):

$$y = f_h\left[\left(\sum_{i=0}^{N-1} w_i x_i\right) + \theta\right] \tag{2}$$

We choose this classifier because no assumptions about the distribution of the target dataset are required for training. Besides, it allows calculating nonlinear decision boundaries.

Moreover, we use the SVM classifier that finds the optimal hyperplane among data points using statistical learning. It transforms the data points, which are not usually linearly separable, from the input space ($X = [X_1, X_2, \ldots X_m]$ labeled as $Y_1, Y_2, \ldots Y_m$ s. t. $Y_i \in \{-1, 1\}$) to a higher dimension. Therefore, it can find the optimal hyperplane ($w.X + b = 0$) which is capable of separating their feature spaces by maximizing the margin. The optimization is done by resolving:

$$\min_{w,b} \frac{1}{2}||w||^2 \tag{3}$$

s.t. $y_i(w.X_i + b) \geqslant 1$ where $i = 1 \ldots m$. SVM performs binary classification, however, by combining several classifiers, multi-class classification can be carried out. We use one-versus-one strategy for multi-class classification [60].

Finally, we use GNB, which is the most popular extension of Bayesian classifiers. Gaussian Naive Bayes classifiers assume that the feature spaces of classes are independent and follow a Gaussian distribution. Given an input feature vector, class label with higher probability can be calculated by Probability Density Function (PDF) using mean and standard deviation obtained from training data:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\pi\sigma_y^2}\right) \tag{4}$$

where y is the class label, and $X = [x_1, x_2, \ldots, x_i, \ldots, x_N] \in R^{D\times N}$ is a N dimensional input feature vector.

3.2. Skeleton based method

The available depth sensors supply the users with 3D coordinates of 25 or less skeleton joints through their middleware. The videos were captured in uncontrolled environments and were uploaded to YouTube. Therefore, there is no recorded depth information and extracted joint coordinates for each frame. In order to produce baselines with various modalities, we utilize AlphaPose [61], a body part detector algorithm to produce skeleton joints from RGB images. AlphaPose is a CNN-based multi-person 2D pose estimation framework that precisely predicts 17 human skeleton joints from images. The examined actions in the dataset contain only a specific subset of joints. The actions comprise 9 joints: *right and left shoulders*, *right and left elbows*, *right and left hands*, *right and left hips*, and *head* joints. It should be noticed that in some frames, no skeleton is estimated due to occlusion, range from the camera, low image resolution, etc. The framework dismisses such frames and constructs the feature vector, only based on frames with high confidence valued skeletons. To make the skeleton models invariant to the limb size of the subjects, the selected skeletons are normalized to the average limb sizes of the subjects in the entire training dataset. As shown in Fig. 2, pairwise angle descriptors are extracted from the estimated poses at each frame. If $S = [J_1, J_2, \ldots, J_n], n = 9$ represents a set of joints of a skeleton, and $\overrightarrow{J_iJ_j}$ and $\overrightarrow{J_jJ_k}$ are the two vectors created from 3 consecutive joints (J_i, J_j, and J_k), the dot product of the two vectors is calculated as: $|\overrightarrow{J_iJ_j}|.|\overrightarrow{J_jJ_k}|.\cos\theta$ and thereby, the pairwise angle descriptor can be simply calculated as:

$$\theta = \cos^{-1}\left(\frac{\left(\overrightarrow{J_iJ_j}.\overrightarrow{J_jJ_k}\right)}{\left(|\overrightarrow{J_iJ_j}|.|\overrightarrow{J_jJ_k}|\right)}\right) \tag{5}$$

θ is the angle between the two vectors created from three adjacent joints. For example using three shoulder, elbow and hand joints we create two vectors one by shoulder and elbow joints and the other by elbow and hand joints. Then, the θ is calculated as the angle between these two vectors. Since the videos comprise various frame number, the extracted features are padded to create equal-sized descriptors. For padding, the zero-padding method on the basis of the video with the maximum number of frames is utilized. For training action patterns from the extracted features, a two-layered LSTM network is trained. In the first layer, it consists of 100 LSTM modules, whereas in the second layer, there are 100 bi-directional LSTM modules. The LSTMs are trained by Backpropagation Through Time (BPTT) algorithm [62], where it minimizes a Kullback–Leibler divergence loss function. The optimization is carried out by the Adam optimizer [63]. The training takes 50 epochs with batch size set at 5 and a learning ratio of 0.001. As another variation of the skeleton-based method, instead of LSTM modules for training, the calculated features are employed as input to the Bag-of-Visual-Words framework. The same preprocessing steps are taken to produce the skeleton features except at the end, which instead of training the LSTMs, histograms are computed and used for training SVM, MLP, and GNB classifiers. Similarly, the feature

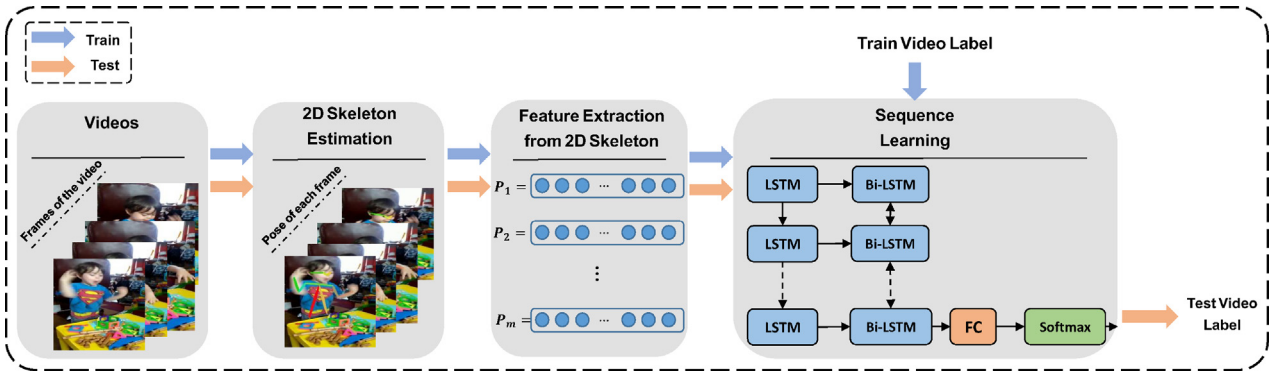

Fig. 2. Skeleton based action recognition pipeline.

histograms of each video are constructed from the concatenation of three histograms (calculated from both the entire video and the video split into two halves).

3.3. Action recognition with deep learning

To compare evaluations on our dataset with deep learning methods, we produce baseline results with complete deep learning pipelines. To do this, we fine-tune two off-the-shelf networks: 3DCNN [64], and ConvLSTM [65]. In order to keep temporal dependency information, instead of 2D spatial convolutions, 3DCNN undertakes 3D spatiotemporal convolutions. On the other hand, ConvLSTM uses a convolutional network (ResNet-152) to extract spatial convolutional feature vectors from video frames, bi-directional LSTMs to learn their temporal dependencies, and Softmax regressors for classification.

4. Dataset, experiments, and evaluations

4.1. Dataset

We collected a new dataset named Expanded Stereotype Behavior Dataset (ESBD) from YouTube videos consisting of children prone to ASD performing four different indicative Stereotype actions. The videos are closely reviewed and their authenticity is approved by professional clinicians. There is no ground-truth information about the condition of the subjects in the videos. Therefore, there is no healthy vs. pathological annotation in the dataset. The clinicians only annotate the repetitive actions that are a potential indication of ASD disorder. The dataset and annotations will be publicly available. The actions in the dataset are divided to four categories (141 videos): *Spinning*, *Arm flapping*, *Hand action*, and *Head banging* actions (Fig. 3). The videos are captured by parents in daily living settings. They are long-term and the target actions are manually clipped. Our dataset includes 141 videos and it is almost two times the size of the previous benchmark dataset SSBD [20] (3 action classes, 68 videos). There is no shared video between the two datasets. The characteristics of the cameras capturing the videos are unknown, therefore, the entire dataset is converted into 25 frames per second rate. With this conversion, in total, there are 46,213 frames in the dataset. The dataset is not subject-oriented and is collected from different individuals. However, there exist videos of different actions from the same subject. In total, there are 108 subjects from which 76 are males and 32 are females. More detailed information about the dataset is shown in Table 1.

4.2. Experiments and evaluations

In this study, we made an effort towards non-invasive recognition of Stereotype actions, which are closely associated with ASD, by means of our novel dataset. For each implemented framework, the evaluations are carried out in 3 different variations. First, with raw data and without any pre-processing; Second, by applying pre-processing steps to inputs; Finally, by calculating the combination of histograms as feature vectors. We use *K-fold cross validation* protocol for both training and evaluation by setting $K = 5$. Given *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)*, and *False Negative (FN)*, accuracy of each evaluation is measured through *Accuracy* $= \frac{TP+TN}{TP+TN+FP+FN}$, and *Unweighted Average Recall (UAR)* metrics. Considering the *Recall* metric as *Recall* $= \frac{TP}{TP+FN}$, UAR is calculated as unweighted average of all *Recall* values of action classes in the dataset. These metrics are useful especially when the dataset is unbalanced (like ours), and there are not enough class instances (e.g. *Head banging* class).

Person detection and tracking in early steps of the pipeline play an essential role in achieving superior recognition performances later. Fig. 4 shows results of applied person detectors. For HAAR-Cascade and HOG descriptors, pre-trained person detection models in the OpenCV library are utilized. These models are trained for detecting the pose of a person in a standing position facing the camera. They fail to detect the subject when there is an occlusion, clutter, or the subject is too close to the camera. The best results are obtained using the YOLOv3 detector, which is used for detection and tracking processes along with the SORT algorithm. In descriptor based framework, for every mode of evaluation, 5 different descriptors, and 3 different classifiers are tried. For training, 10,000 frames from the training set are randomly sampled. The parameters of descriptor extraction and classifiers are fine-tuned to obtain the best results. For the HOG descriptor, block size is set to 16 pixels ($N \times N = (16, 16)$). The best results for the descriptor-based framework are obtained by the HOF descriptor. The best performance is obtained by fixing the parameters *orientation*, *pixel per cell*, *cell per block* to 9, (8, 8), and (3, 3), respectively. SIFT descriptor is extracted by putting *number of octave layers* to 3, and edge threshold to 10. Similarly, for the SURF descriptor, *number of octave layers* is 3, and the threshold is 100. For creating the visual vocabulary, the K-means algorithm is utilized by setting the number of clusters to 200, 400, 600, 800, and 1000. For the initialization of clustering, the K-means++ method is selected where each cluster center is initialized for 10 times. The maximum iteration to reach the best clustering model is 300. Several experiments are carried out by exploiting MLP, GNB, and SVM classifiers, and results are reported in Table 2. The MLP classifier consistently outperforms the other classifiers when it is used with the HOF descriptor. To achieve such results, we tested MLP classifiers with 5 parameters, namely solver (for weight optimization), regularization parameter (α), size of the hidden layer, the maximum number of iteration until convergence, and learning rate. We varied these parameters and obtained optimal performance when we chose *LBFGS* for optimization and L2 for regularization. The network consists of 2 hidden layers where they con-

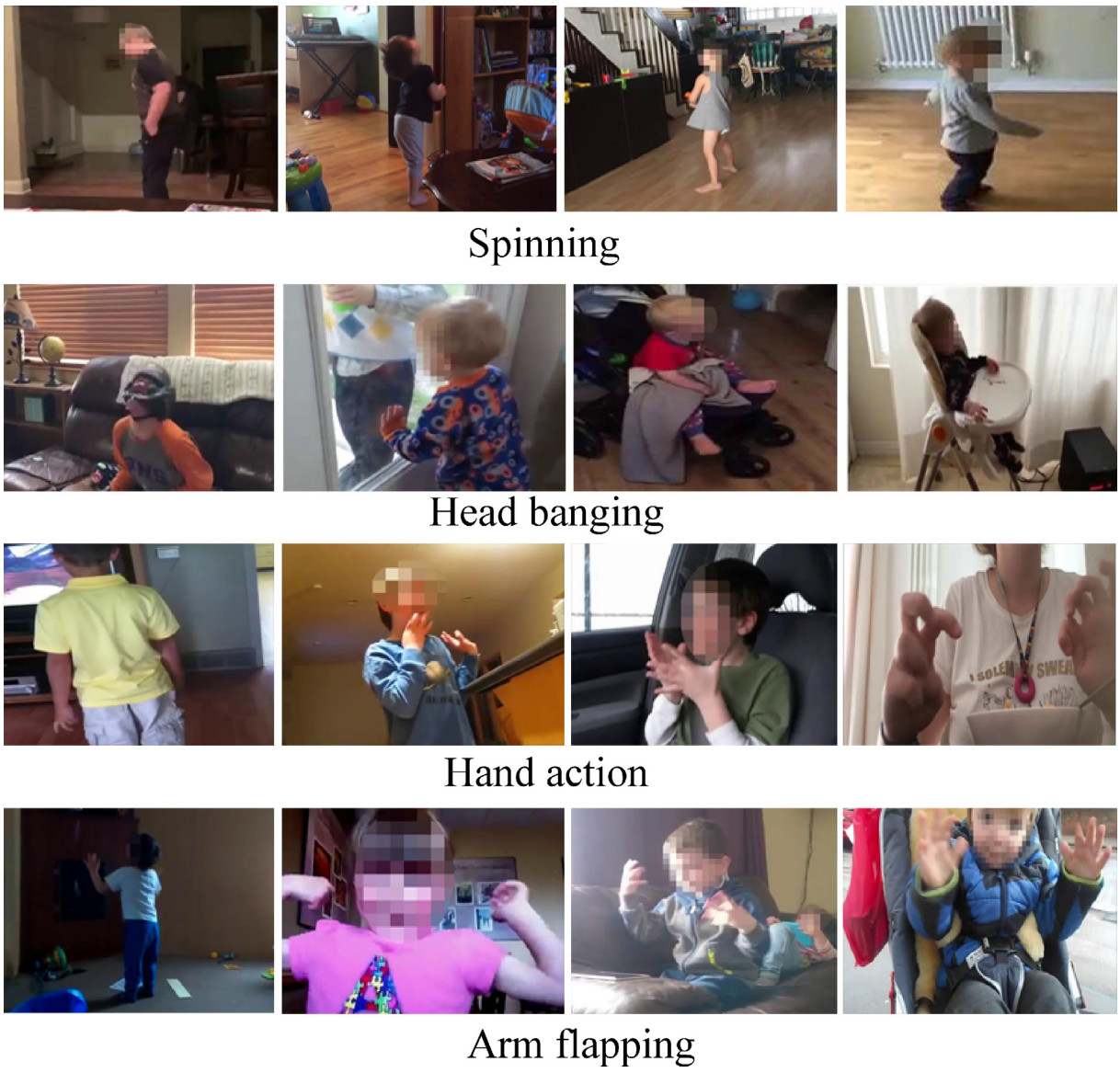

Fig. 3. Instance frames of the videos in the collected dataset. From top row to down: ‘‘Spinning’’, ‘‘Head banging’’, ‘‘Hand action’’, and ‘‘Arm flapping’’.

Table 1
Information about frame number of actions and gender of the subjects in the ESBD dataset.

	Arm flapping	Hand action	Head banging	Spinning
Number of videos	43	31	27	40
Min/Avg/Max number of frames	45/313/138	30/365/3828	30/258/1679	90/548/2545
Total number of frames	5938	11342	6982	21951
Male/Female	24/10	6/7	17/7	29/8

tains 200 and 50 neurons respectively. The adaptive learning rate is selected to schedule weight updates. These blocks of the pipeline are implemented by the Scikit-learn [66] library. The experiments are conducted on a system running Microsoft Windows 10 operating system with an Intel Core-i5 CPU with two 2.7 GHz cores and RAM memory of 16 GB. To run the deep networks a NVidia RTX 2080 GPU is utilized.

The Table 2 represents the effects of preprocessing steps and histogram concatenation. The two lowest performances belong to SIFT and SURF descriptors. The reason may be owing to the lower number of detected features and their less descriptive power compared to the other descriptors. With 56.77%, HOF descriptors achieve the highest performance on the raw dataset (while using 200 visual codewords) which is consistent with relevant studies in the literature [67]. The number of HOF descriptors is higher compared to others showing its capability in capturing dynamic information of the actions. As shown in the middle columns of the Table 2, most of the descriptors take benefit of the preprocessing steps to achieve better performances. In total, there is a 3% boost in performance. Preprocessing improves the best result (achieved by HOF using MLP classifier) by 12% (from 56.77% to 68.83%. This shows the effect of removing clutter which is exposed in the background and in most of the frames. Two left columns in Table 2 indicate that the most accurate recognition is achieved when histogram aggregation is utilized. This big improvement (almost 23%) is obtained with the same feature extraction and the same classification parameters,

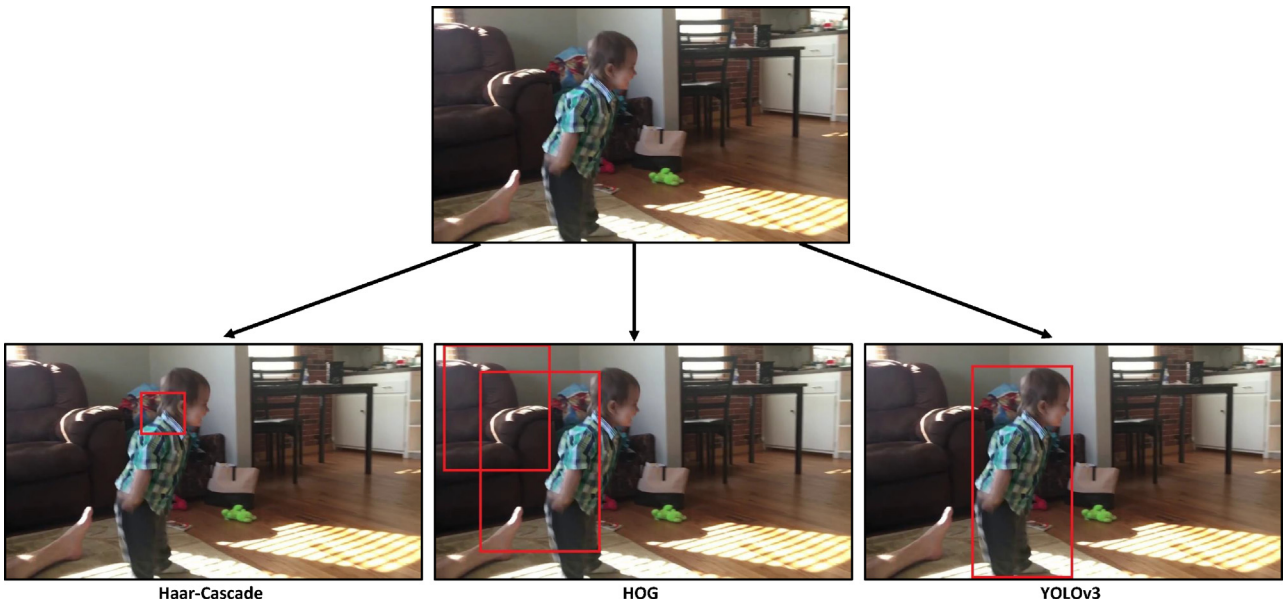

Fig. 4. Examples of applied person detection algorithms: using Haar-Cascade (Left), HOG (Middle), and YOLOv3 (Right) detectors.

Table 2
Results of the descriptor-based framework on the collected dataset when the dataset is not pre-processed (left), pre-processed (center), and is evaluated with concatenated histograms (right).

		Without pre-processing		With pre-processin		With histogram concatenation	
		Accuracy (%)	UAR (%)	Accuracy (%)	UAR (%)	Accuracy (%)	UAR (%)
HOG	*GNB*	46.89	46.44	53.88	53.68	53.89	48.19
	SVM	41.86	35.49	44.67	38.36	41.82	35.41
	MLP	49.82	47.50	53.20	53.05	55.44	55.23
HOF	*GNB*	52.51	50.60	56.79	52.80	60.34	58.03
	SVM	47.51	40.76	48.22	41.93	50.71	42.87
	MLP	56.77	55.05	68.83	68.46	**79.28**	**77.68**
HOG/HOF	*GNB*	54.65	53.79	59.67	57.18	58.94	57.23
	SVM	46.79	40.01	49.65	43.39	48.94	42.55
	MLP	53.19	50.68	69.50	68.88	65.24	63.29
SIFT	*GNB*	51.08	48.94	44.01	39.93	43.30	39.13
	SVM	41.81	36.35	41.87	35.50	43.30	37.38
	MLP	47.50	46.57	54.00	50.36	49.09	47.17
SURF	*GNB*	41.12	40.44	37.51	35.20	31.97	29.55
	SVM	42.59	36.04	39.03	32.71	35.61	34.04
	MLP	45.36	44.32	42.58	39.70	43.25	39.43

and only by concatenation of histograms. This underlines the importance of temporal dependencies in the action recognition task. In addition, the results indicate the superiority of motion-based descriptors over appearance-based descriptors.

Additionally, we compare the produced baselines of the developed frameworks and fine-tuned frameworks altogether (Table 3). We compare the obtained results with a variation of the Bag-of-Visual-Words framework. This framework accepts geometrical skeleton features as input (Indicated as Skeleton-BOVW in the table) instead of local descriptors. We also evaluate the results of the LSTM-based method that receives skeleton features as input (Skeleton-LSTM). Other than these developed frameworks, we evaluate the dataset by fine-tuning the two CNN-based networks (3DCNN and ConvLSTM). As discussed in Section 2, our selected models for producing baselines are representing methods which have shown strong capability in modeling spatiotemporal information (RNN based and 3D CNN based). To train the 3DCNN network, the optimal parameter values are empirically determined through the grid search. The batch size and the number of epoch parameters are set to 8 and 100, respectively. The categorical cross-entropy loss function is optimized via Stochastic Gradient Descent optimizer. The Keras implementation [68] is utilized to generate the results. The best results are obtained by fine-tuning the parameters of the ConvLSTM network when Batch size is set to 16, sequence length to 10, and the number of epoch to 100. Cross-entropy is used for calculating the loss and Adam technique for optimization. To produce the results, Pytorch [69] implementation is utilized. Table 3 contains the results of all developed and fine-tuned frameworks. The table is divided into three categories namely: local descriptor based, pose-based, and CNN-based methods. All the provided results use pre-processed images by YOLO except skeleton-based methods that work directly with joints extracted by AlphaPose from images. In all classes, the descriptor-based method significantly outperforms the two pose-based methods and deep 3DCNN. It is worth mentioning that the descriptor-based method has higher performance compared to the ConvLSTM method, as LSTMs are assumed to store temporal information better. Referring to Table 2, this gain is directly associated with the histogram concatenation technique. Without applying histogram combination, ConvLSTM obtains higher accuracy. In general, in our dataset, the histogram combination technique is more effective in maintaining temporal information. This is also

Table 3
Shows the results of all evaluated frameworks divided into three categories: Local descriptor based, Articulated pose-based, and CNN-based methods.

	Arm flapping		Hand action		Head banging		Spinning		Total	
	Acc. (%)	UAR (%)	Acc. (%)	UAR (%)	Acc. (%)	UAR (%)	Acc. (%)	UAR (%)	Acc. (%)	UAR (%)
Local Descriptor RGB-based										
HOF-BOVW	0.78	0.71	0.87	0.89	0.68	0.70	0.82	0.86	**0.79**	**0.77**
Articulated Pose-based methods										
Skeleton-BOVW	0.69	0.55	0.56	0.81	0.35	0.43	0.73	0.75	0.61	0.64
Skeleton-LSTM	0.59	0.62	0.50	0.61	0.40	0.40	0.77	0.65	0.59	0.57
CNN-based deep learning										
3DCNN	0.36	0.30	0.30	0.75	0.50	0.50	0.55	0.40	0.42	0.49
ConvLSTM	0.72	0.81	0.74	0.76	0.52	0.60	0.93	0.72	0.74	0.73

evident when comparing the two skeleton-based methods. Using the same set of features, BOVW methods achieve relatively higher accuracy (2%) compared to its LSTM counterpart. 3DCNN is failed to achieve higher performance (even less than skeleton-based methods). This can be attributed to the nature of activities in the collected dataset. Unlike action datasets, the convolutional filters encounter difficulties in extracting useful information from the background to recognize the activities. Fig. 5 illustrates the confusion matrices of the top performances in each one of the categories. All frameworks achieve encouraging performances in classifying the "Spinning" class. It can be related to the unique dynamic pattern of this action which is easier to discriminate by classifiers. It can be seen that the "Arm flapping" class is the most confused class. In all frameworks, this class is mostly confused with the "Headbanging" class. The confusion is major in the skeleton-based framework. This can be related to the extracted 2D skeletons. The motion of the head is not reflected in the skeleton joints when the subject is facing the camera. It is surprising that the "Hand action" class is relatively classified accurately (87% using HOF descriptor), although it involves a very intricate configuration of the fingers and complicated motion patterns.

The proposed methods still have some limitations. The limitations are mostly associated with the extraction of robust pose/appearance descriptors, modeling spatiotemporal information and, data. In general, it was expected that pose-based methods achieve better results than the appearance-based descriptors. But in our videos, skeletons extracted from RGB images are not reliable enough to make robust predictions (this could be due to fast repetitive movements of the subject). The estimated joints are not of high confidence and resolving this challenge is still an active research topic. The recognition would be even more accurate with 3D pose information. 3D Pose information is highly informative and robust to illumination and viewpoint changes. But the current 3D pose estimator are not accurate. Having reliable 3D joint information may result in better recognition (such as in depth based sensors). In the lack of robust pose information, motion could be the most characteristic feature of a repetitive action. HOF descriptor performs better in encoding kinetic information hence, results in better recognition. Deep convolutional features also focus more on appearance-based features while the key characteristics of actions come from motion. 3D convolutions could remedy this issue into some extend and achieve competitive results with their temporal stream. However, as discussed previously, these methods suffer from rigid spatiotemporal kernel limitation. The baseline results reaffirm the usefulness of action recognition in unconstrained scenarios such as ASD. However, to completely benefit from such a framework and to carry out the diagnosis of such medical disorders, the limitations of action recognition from videos should be addressed. The developed methods are an endeavor to tackle some of those limitations such as the temporal dependency of actions to establish a ground for future research. Data is an indisputable requirement to meet this objective. The proposed dataset is so far the most comprehensive dataset of repetitive actions of ASD. Based on involved medical experts' opinion, obtained baseline results are promising and encouraging for further studies in the future.

5. Conclusion

A timely diagnosis of neurobiological disorders is very important for providing the best care and effective treatments. Patients diagnosed with ASD can therefore benefit a lot by going through a methodological treatment process based on the level of the disorder. Nonetheless, this requires monitoring and examination of the self-stimulating behaviors by interacting with children for a long time that makes it difficult for early diagnosis of ASD. In this paper, we proposed a comprehensive dataset collected from the daily lives of children prone to ASD. We developed several action-recognition-based frameworks to undergo the recognition of these characteristic behaviors with the aid of computer vision. The frameworks received video recordings and described each action by extracting their discriminative information. Each framework is capable of using various modalities as input and is designed with different architectures from bag-of-visual-words to RNNs and CNNs. We observed from the results that the HOF descriptor outperforms the other methods when used with BOVW architecture. However, deep neural network architectures (especially CNN along with LSTM) achieves competitive performances. In order to put such systems into practice, they

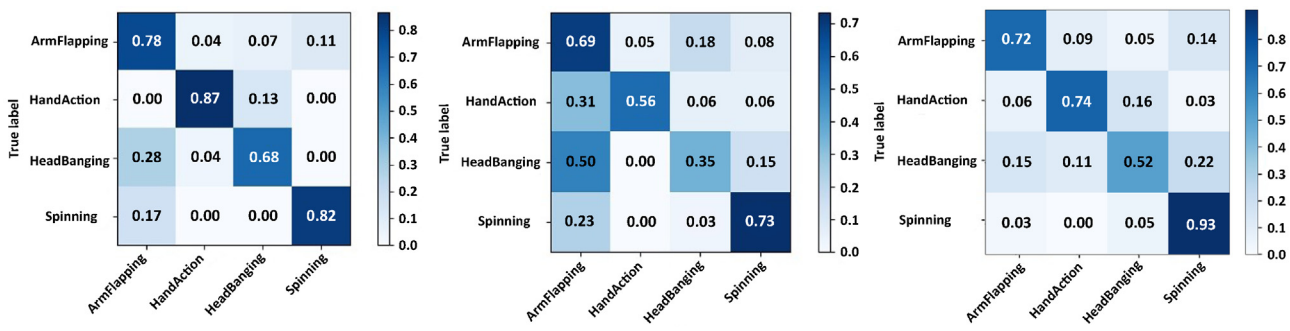

Fig. 5. Confusion matrices of: descriptor-based (Left), skeleton with Bag-of-Visual-Words (middle), and ConvLSTM with preprocessing frameworks.

have to be certified by the domain experts. In addition, it is desirable to expand the dataset with more videos (suitable for data-demanding deep architectures) and particularly, to build a subject-oriented dataset, where there are plenty of long-term recordings of each subject. This will introduce extra challenges such as trimming the videos and isolating the actions throughout the videos. We will address these challenges in upcoming studies to confirm the practicality of computer-aided applications in healthcare and for medical diagnosis.

CRediT authorship contribution statement

Farhood Negin: Conceptualization, Methodology, Writing - original draft. **Baris Ozyer:** Writing - review & editing, Supervision. **Saeid Agahian:** Software, Methodology, Conceptualization, Validation. **Sibel Kacdioglu:** Software, Investigation, Data curation. **Gulsah Tumuklu Ozyer:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank Assoc. Prof. Dr. Ibrahim Selcuk Esin and Assist. Prof. Dr. Hicran Dogru from the Child and Adolescent Psychiatry Department at Ataturk University for their valuable support in creating the dataset. Each video was reviewed and labeled under their guidance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.neucom.2021.03.004.

References

[1] B.J. O'Roak, M.W. State, Autism genetics: strategies, challenges, and opportunities, Autism Research 1 (1) (2008) 4–17.
[2] M. Elsabbagh, G. Divan, Y.-J. Koh, Y.S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C.S. Paula, C. Wang, et al., Global prevalence of autism and other pervasive developmental disorders, Autism Research 5 (3) (2012) 160–179.
[3] A. Knopf, Autism prevalence increases from 1 in 60 to 1 in 54: Cdc, The Brown University Child and Adolescent Behavior Letter 36 (6) (2020) 4.
[4] L. Kanner et al., Autistic disturbances of affective contact, Nervous Child 2 (3) (1943) 217–250.
[5] J.W. Bodfish, F.J. Symons, D.E. Parker, M.H. Lewis, Varieties of repetitive behavior in autism: Comparisons to mental retardation, Journal of Autism and Developmental Disorders 30 (3) (2000) 237–243.
[6] M.H. Lewis, J.W. Bodfish, Repetitive behavior disorders in autism, Mental Retardation and Developmental Disabilities Research Reviews 4 (2) (1998) 80–89.
[7] J. Barbaro, C. Dissanayake, Autism spectrum disorders in infancy and toddlerhood: a review of the evidence on early signs, early identification tools, and early diagnosis, Journal of Developmental & Behavioral Pediatrics 30 (5) (2009) 447–459.
[8] J.L. Matson, J.A. Boisjoli, M.L. Gonzalez, K.R. Smith, J. Wilkins, Norms and cut off scores for the autism spectrum disorders diagnosis for adults (asd-da) with intellectual disability, Research in Autism Spectrum Disorders 1 (4) (2007) 330–338.
[9] R. Landa, Early communication development and intervention for children with autism, Mental Retardation and Developmental Disabilities Research Reviews 13 (1) (2007) 16–25.
[10] B. Reichow, Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders, Journal of Autism and Developmental Disorders 42 (4) (2012) 512–520.
[11] A.P. Association, et al., Diagnostic and statistical manual of mental disorders (DSM-5), American Psychiatric Pub, 2013.
[12] K.S. Lam, M.G. Aman, The repetitive behavior scale-revised: independent validation in individuals with autism spectrum disorders, Journal of Autism and Developmental Disorders 37 (5) (2007) 855–866.
[13] F. Negin, S. Cogar, F. Bremond, M. Koperski, Generating unsupervised models for online long-term daily living activity recognition, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 186–190.
[14] G.S. Young, J.N. Constantino, S. Dvorak, A. Belding, D. Gangi, A. Hill, M. Hill, M. Miller, C. Parikh, A. Schwichtenberg, et al., A video-based measure to identify autism risk in infancy, Journal of Child Psychology and Psychiatry 61 (1) (2020) 88–94.
[15] M. Huerta, C. Lord, Diagnostic evaluation of autism spectrum disorders, Pediatric Clinics of North America 59 (1) (2012) 103.
[16] S. Ozonoff, A.-M. Iosif, G.S. Young, S. Hepburn, M. Thompson, C. Colombi, I.C. Cook, E. Werner, S. Goldring, F. Baguio, et al., Onset patterns in autism: correspondence between home video and parent report, Journal of the American Academy of Child & Adolescent Psychiatry 50 (8) (2011) 796–806.
[17] R.L. Young, N. Brewer, C. Pattison, Parental identification of early behavioural abnormalities in children with autistic disorder, Autism 7 (2) (2003) 125–143.
[18] S. Bereznak, K.M. Ayres, L.C. Mechling, J.L. Alexander, Video self-prompting and mobile technology to increase daily living and vocational independence for students with autism spectrum disorders, Journal of Developmental and Physical Disabilities 24 (3) (2012) 269–285.
[19] S. Agahian, F. Negin, C. Köse, Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition, The Visual Computer 35 (4) (2019) 591–607.
[20] S. Sundar Rajagopalan, A. Dhall, R. Goecke, Self-stimulatory behaviours in the wild for autism diagnosis, in, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, 2013.
[21] S.S. Rajagopalan, R. Goecke, Detecting self-stimulatory behaviours for autism diagnosis, IEEE International Conference on Image Processing (ICIP) 2014 (2014) 1470–1474.
[22] Kaggle, Autistic Children Data Set, (2020, July 16). url: https://www.kaggle.com/gpiosenka/autistic-children-data-set-traintestvalidate.
[23] S. Chen, Q. Zhao, Attention-based autism spectrum disorder screening with privileged modality, IEEE/CVF International Conference on Computer Vision (ICCV) 2019 (2019) 1181–1190.
[24] J. Li, Y. Zhong, G. Ouyang, Identification of asd children based on video data, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 367–372.
[25] B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, L. Shapiro, A facial affect analysis system for autism spectrum disorder, IEEE International Conference on Image Processing (ICIP) 2019 (2019) 4549–4553.
[26] J. Hashemi, T.V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, G. Sapiro, A computer vision approach for the assessment of autism-related behavioral markers, IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL) 2012 (2012) 1–7.
[27] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, V. Murino, Video gesture analysis for autism spectrum disorder detection, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3421–3426.
[28] J. Li, Y. Zhong, J. Han, G. Ouyang, X. Li, H. Liu, Classifying asd children with lstm based on raw videos, Neurocomputing 390 (2020) 226–238.
[29] M. Del Coco, M. Leo, P. Carcagni, P. Spagnolo, P.L. Mazzeo, M. Bernava, F. Marino, G. Pioggia, C. Distante, A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders, IEEE International Conference on Computer Vision Workshops (ICCVW) 2017 (2017) 1401–1407.
[30] W. Liu, T. Zhou, C. Zhang, X. Zou, M. Li, Response to name: A dataset and a multimodal machine learning framework towards autism study, Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) 2017 (2017) 178–183.
[31] J.M. Rehg, G.D. Abowd, A. Rozga, M. Romero, M.A. Clements, S. Sclaroff, I. Essa, O.Y. Ousley, Y. Li, C. Kim, H. Rao, J.C. Kim, L.L. Presti, J. Zhang, D. Lantsman, J. Bidwell, Z. Ye, Decoding children's social behavior, IEEE Conference on Computer Vision and Pattern Recognition 2013 (2013) 3414–3421.
[32] K. Vyas, R. Ma, B. Rezaei, S. Liu, M. Neubauer, T. Ploetz, R. Oberleitner, S. Ostadabbas, Recognition of atypical behavior in autism diagnosis from video using pose estimation over time, in: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), 2019, pp. 1–6.
[33] Behavior Imaging-Health & Education Assessment Technology, (2020, July 20). url: https://behaviorimaging.com/.
[34] E. Marinoiu, M. Zanfir, V. Olaru, C. Sminchisescu, 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[35] O. Rihawi, D. Merad, J. Damoiseaux, 3d-ad: 3d-autism dataset for repetitive behaviours with kinect sensor, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6.
[36] C.-H. Min, Y. Kim, A. Tewfik, A. Kelly, Detection of Self-Stimulatory Behaviors of Children with Autism Using Wearable and Environmental Sensors, Journal of Medical Devices 3 (2), 027506. arXiv:https://asmedigitalcollection.asme.org/medicaldevices/article-pdf/3/2/027506/5570469/027506_1.pdf, url: doi: 10.1115/1.3134931.
[37] T. Westeyn, K. Vadas, X. Bian, T. Starner, G.D. Abowd, Recognizing mimicked autistic self-stimulatory behaviors using hmms, in: Ninth IEEE International Symposium on Wearable Computers (ISWC'05), 2005, pp. 164–167.
[38] C. Min, Automatic detection and labeling of self-stimulatory behavioral patterns in children with autism spectrum disorder, in: 2017 39th Annual

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 279–282.
[39] S. Bansode, T. Shinde, S. Garapati, I. Abdel-Qader, Stereotypic repetitive hand flapping movement detector for children with autism, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–5.
[40] Y. Tian, X. Min, G. Zhai, Z. Gao, Video-based early asd detection via temporal pyramid networks, IEEE International Conference on Multimedia and Expo (ICME) 2019 (2019) 272–277.
[41] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information processing systems, 2014, pp. 568–576.
[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[43] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[44] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[46] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4507–4515.
[47] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[48] S. Kaçdıoğlu, B. Özyer, G.T. Özyer, Otizm spektrum bozukluklarında kendini uyarıcı davranışları tanıma, IEEE, 2020, pp. 1–5.
[49] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, IEEE, 2001, pp. I-I.
[50] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer vision and pattern recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 886–893.
[51] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767.
[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[53] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3464–3468, IEEE.
[54] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8, IEEE.
[55] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[56] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Computer Vision and Image Understanding 110 (3) (2008) 346–359.
[57] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Oakland, CA, USA, 1967, pp. 281–297.
[58] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, 2009.
[59] P. Shukla, K.K. Biswas, P.K. Kalra, Action recognition using temporal bag-of-words from depth maps, MVA (2013) 41–44.
[60] M. Pal, Multiclass approaches for support vector machine based land cover classification, arXiv preprint arXiv:0802.2411.
[61] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: Regional multi-person pose estimation, in: ICCV, 2017.
[62] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proceedings of the IEEE 78 (10) (1990) 1550–1560.
[63] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[64] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, M. Bennamoun, Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 3120–3128.
[65] Y. Jiang, X. Qianqian, X. Cao, Outfit recommendation with deep sequence learning, 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), IEEE (2018) 1–5.
[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Machine learning in python, Scikit-Learn: the Journal of Machine Learning Research 12 (2011) 2825–2830.
[67] I. Dave, K. Carter, M. Shah, "kallis" crcv vipriors challenge submission.
[68] F. Chollet, et al., Keras (2015). url: https://github.com/fchollet/keras.
[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in neural information processing systems, 2019, pp. 8026–8037.

Farhood Negin is working as a research engineer in Computer Vision at INRIA where he obtained his Ph.D. He has also obtained his M.Sc. from Sabanci University in Istanbul, Turkey. Previously, he worked in a reputable industry connected European research project in computer vision, artificial intelligence and assistive technologies such as VIPSAFE, Dem@Care and SAFEE projects. He is a member of Cognition Behaviour Technology (Cobtek) team and also The European Network on Integrating Vision and Language (iV&L Net). He is currently working in the Spatio-Temporal Activity Recognition Systems (STARS) team at INRIA in order to develop next generation technologies in computer vision and human-computer interaction with a focus on activity and gesture recognition.

Baris Ozyer received the B.Sc. degree from the Electronics Engineering Department, Erciyes University, Turkey, in 2002, and the integrated Ph.D. degree from the Electrical and Electronics Engineering Department, Middle East Technical University, Turkey, in 2012. He worked as an Intern Researcher with CNS Computational Neuroscience Laboratories in ATR, Japan, from 2007 to 2008. He is currently an Assistant Professor in Computer Engineering at Ataturk University, Turkey. His current research focus is mainly on Computer Vision and Machine Learning applications in the field of robotics and understanding human behaviors.

Saeid Agahian is an Assistant Professor in Computer Engineering Department at Erzurum Technical University, Turkey. He received his M.Sc. and Ph.D. degrees from the Computer Engineering Department, Karadeniz Technical University, Trabzon, Turkey, in 2012 and 2018, respectively. His research interests include Computer Vision, Machine Learning and Combinatorial Optimization.

Sibel Kacdioglu obtained a bachelor degree in Computer Engineering from Anadolu University in 2016. She received her M.Sc. in 2020 in Computer Engineering Department at the Ataturk University, Erzurum. Currently she pursues her Ph.D. in Computer Engineer at Computer Engineering Department of Ataturk University, Erzurum. Along with her Ph.D. she works as a research assistant in Erzurum Technical University. Her areas of interest are Machine Learning, Computer Vision and Image Processing.

Gulsah Tumuklu Ozyer received the B.Sc. degree from Erciyes University, Turkey, in 2002 and the integrated Ph.D. degree from Computer Engineering Department of Middle East Technical University, Turkey, in 2012. She was a visiting researcher in James Z. Wang Research Group in Penn State University, USA during her Ph.D. studies. She is currently an Assistant Professor in Computer Engineering Department at Ataturk University, Turkey. Her research interest includes Visual Attention, Pattern Recognition and Computer Vision.