

# Self-Stimulatory Behaviours in the Wild for Autism Diagnosis

Shyam Sundar Rajagopalan<sup>1</sup>

Abhinav Dhall<sup>2</sup>

Roland Goecke<sup>1,2</sup>

<sup>1</sup>Vision & Sensing Group, HCC Lab, ESTeM, University of Canberra, Australia

<sup>2</sup>IHCC Group, RSCS, Australian National University, Australia

shyam.rajabopalan@canberra.edu.au, abhinav.dhall@anu.edu.au, roland.goecke@ieee.org

## Abstract

*Autism Spectrum Disorders (ASD), often referred to as autism, are neurological disorders characterised by deficits in cognitive skills, social and communicative behaviours. A common way of diagnosing ASD is by studying behavioural cues expressed by the children. We introduce a new publicly-available dataset of children videos exhibiting self-stimulatory (stimming) behaviours commonly used for autism diagnosis. These videos, posted by parents/caregivers in public domain websites, are collected and annotated for the stimming behaviours. These videos are extremely challenging for automatic behaviour analysis as they are recorded in uncontrolled natural settings. The dataset contains 75 videos with an average duration of 90 seconds per video, grouped under three categories of stimming behaviours: arm flapping, head banging and spinning. We also provide baseline results of tests conducted on this dataset using a standard bag of words approach for human action recognition. To the best of our knowledge, this is the first attempt in publicly making available a Self-Stimulatory Behaviour Dataset (SSBD) of children videos recorded in natural settings.*

## 1. Introduction

The area of computational behaviour modelling deals with the study of machine analysis and understanding of human behaviour. An important application area is in assisting clinicians in diagnosing ASD, which is a condition affecting children at their early developmental ages and is more pronounced in boys than girls. Unfortunately, it is growing at a fast rate worldwide and currently the number of children having autism is 1 in 110 [14]. The genetic basis for ASD is still unknown and one of the common ways of diagnosing is by using behavioural cues of the children [14]. These behavioural cues are identified by exercising instruments such as the Autism Diagnostic Observation Schedule (ADOS) [12], the Autism Diagnostic Interview [13], the Autism Observation Scale for Infants (AOSI) [16], and the Diagnostic

and Statistical Manual of Mental Disorders (DSM-5) [5]. *Self-stimulatory behaviours* are one type of such atypical behaviour cues assessed in these instruments. The diagnosis involves clinicians interacting with children in multiple long sessions to identify the behaviour cues, risking a delay in diagnosis. ASD is typically diagnosed at the age of 5yr, while a diagnosis by the age of 2yr can lead to an early intervention [6].

Self-stimulatory behaviours refer to stereotyped, repetitive movements of body parts or objects. These behaviours can be studied from videos captured in an uncontrolled (natural) or controlled environment. In an uncontrolled setting, when the children are performing their regular day-to-day activities, some of the self-stimulatory behaviours, such as arm flapping, head banging, or spinning behaviours can be studied by automatically analysing the captured videos. This provides an important clue to parents/caregivers and also to clinicians for early intervention. In a controlled setting, such as in a dyadic conversational setting, a therapist will be exercising a defined protocol of play action with the child to elicit higher level behaviours [15]. Studying the children behaviours in both controlled and uncontrolled scenarios is equally important for early intervention and diagnosis.

An important first step for modelling the children behaviour towards autism diagnosis is the availability of standard public datasets. Recently, a Multimodal Dyadic Behaviour Dataset (MMDDB), is released for public to study the behaviours in a dyadic scenario [15]. In a similar way, here we propose a new publicly available dataset, Self-Stimulatory Behaviour Dataset (SSBD), for studying the behaviours from videos recorded in an *uncontrolled natural settings*. The proposed dataset's videos are also dyadic in nature as it involved parent prompting or playing with the children. These videos do not intend to convey the presence or absence of autism in a child, rather they meant to be used for analysing stimming behaviours to provide inputs to clinicians for further diagnosis.

The remainder of this paper is organised as follows. The motivation for introducing this dataset is discussed in Sec-

tion 1.1. In Section 2, some of the key studies in the literature related to the applicability of computer vision and affective sensing approaches for autism diagnosis and the availability of related datasets are discussed. The details on the proposed new dataset, the annotations and statistics about the videos are provided in Section 3. In Section 4, the complexities involved in the analysis of such datasets are discussed. The application of a common action recognition algorithms on this new dataset and the experiment results are presented in Section 5. Section 6 concludes this paper.

## 1.1. Motivation

Children with ASD exhibit certain atypical behavioural cues during their regular day-to-day activities. When these atypical behaviours are captured on video over a period of time, they can be analysed to identify “red-flags” in the childrens behaviour [17] and used by clinicians in their diagnosis. Technology in the form of a diagnostic aid can help in reducing the time needed for ASD diagnosis. Moreover, it can provide alerts for caregivers for early intervention. The advancements in affective sensing technology in the diagnosis of neurological problems can also assist in understanding the behaviour. Body expressions can provide information about the affective state of a person [9] and are, for example, used in depression analysis [8]. In a similar way, body expressions can be analysed for identifying atypical behavioural cues from children suffering from ASD.

Stimming behaviours are more common in autistic children and it is usually observed during children’s regular daily activities. These atypical behaviours, when observed early, can lead to an early intervention and diagnosis. However, it is impractical for a clinician or carer to observe the children at all times during the day. Instead, the videos recorded in a natural setting during children’s regular activities can be analysed for these atypical behaviours by developing robust behaviour analysis algorithms. The first step towards this goal is to build a baseline dataset of children’s behaviour recorded in an uncontrolled natural settings. Children’s videos posted by parents / caregivers on public domain websites, such as YouTube, Vimeo, DailyMotion, etc., can be used for analysing children behaviours. Towards this goal, we have collected and annotated a rich set of videos for sharing it with the academic community for research purposes.

From the perspective of affective computing and computer vision, the proposed dataset can be categorised as an ‘in the wild’ corpus. In the problem of face recognition, human action recognition and facial expression analysis much advancement has been made due to the availability of ‘in the wild’ datasets (such as Labeled Faces In The Wild [7], Hollywood dataset [11] and Acted Facial Expressions In The Wild [4]) which represent the real-world scenarios. The proposed dataset SSBD has been recorded

in different real-world scenarios which presents challenges from the perspective of behaviour modelling. This opens up the opportunity for stimulating behaviour analysis ‘in the wild’.

## 2. Related Work and Datasets

There has been a growing interest to study child behaviours using computer vision and affective sensing approaches in recent years. Visual tracking, the level of attention, sharing interest and motor patterns are behaviours that are studied in [6]. Vision techniques are used for head pose estimation, facial feature tracking and arm symmetry measurement, from which atypical behaviours are identified. Experiments are performed on a small dataset in a structured environment. In [14], the author outlines the possible research directions for behaviour imaging that captures and analyses social and communicative behaviours. ASD diagnosis is discussed as one of the potential application areas of behaviour imaging. The author discusses a case study of applying state of the art vision algorithms for extracting and analysing social behaviours from unstructured videos. In [15], the authors have introduced a new problem domain in *activity recognition*, i.e. to extract social and communicative behaviour cues of children in a dyadic interaction between an adult and child. They discussed the applicability of existing algorithms to determine fine-grained behaviour information about the conversation, e.g. the engagement level of the child during the interaction.

The dataset requirement for behaviour analysis will depend on the type of behaviours. In an individualised setting, it involves behaviours observed during the child’s regular day-to-day activities. The behaviours can occur in short bursts at different times, on random days and in different places. Some behaviours can be seen only during conversational settings, when the child is interacting with clinicians. These situations make the problem of creating a new standard dataset a complex process. In addition, ethics approval also plays a vital role. The Multimodal Dyadic Behaviour Dataset (MMDB), a publicly available dataset used for eliciting behaviour cues in a dyadic settings, has recently been released to the academic community. It contains 160 sessions of interactions between an adult and child, each lasting 3-5min, in a very controlled environment. The ASD Video Glossary [1] is a web based tool for parents / caregivers to learn about ASD and to identify “red flags” for early intervention. Autism Generic Resource Exchange (AGRE) [2] is yet another source for behaviour dataset of autistic children. In addition, it is possible that certain types of behaviours can be modelled as higher level actions, activities and gestures. When this is possible, a subset of videos from human action recognition

datasets (KTH<sup>1</sup>, Weizmann<sup>2</sup>, UCF101<sup>3</sup>), facial expression datasets (Cohn-Kanade AU-Coded Expression Database<sup>4</sup>, Acted Facial Expressions in the Wild<sup>5</sup>), and gesture recognition datasets can be used for analysing behaviours.

In order to analyse the behaviours and provide alerts to parents / caregivers based on the day-to-day regular activities of children, it is necessary to have the algorithms developed and tested on real-world datasets taken in an uncontrolled settings. We present such a dataset collected from videos posted by parents / caregivers on public domain websites.

### 3. Details on Self-Stimulatory Behaviour Dataset

Stimming behaviour videos of children available on public domain websites and video portals, such as Youtube, Vimeo, Dailymotion, etc., are searched and collected for automatic behaviour analysis. There is no post-processing done on these videos, thereby, preserving the original natural setting recordings. We have closely watched every video and grouped them into three stimming categories: arm flapping, head banging and spinning. Spinning includes head spinning, walking in circles and body rotation behaviours. These three categories of behaviours belong to atypical motor behaviours [16]. A snapshot of the videos is shown in Figure 1 and the duration of each video in Figure 2. The snapshots are blurred to hide the children’s identity. However, the other information such as postures, illumination, places, clutter backgrounds and multiple objects can be seen indicating real world scenarios. The captions maps to a original video URL.

A reference to a list of public domain URLs of this dataset and their corresponding annotations are released as part of this work <sup>6</sup>. In addition, the source code used for baseline experiments and the computed STIP for every video are also provided. The dataset contains 75 videos grouped in three categories each containing 25 videos. The mean duration of a video is 90s. The resolution of the videos varies but are greater than 320x240 pixels. The exact resolution is provided as part of the annotations. The videos are annotated with a set of representative attributes of the behaviour. The attributes and their descriptions are shown in Table 1 and a sample XML schema is provided in Figure 3.

<sup>1</sup><http://www.nada.kth.se/cvap/actions/>  
<sup>2</sup><http://www.wisdom.weizmann.ac.il/vision/SpaceTimeActions.html>  
<sup>3</sup><http://crcv.ucf.edu/data/UCF101.php>  
<sup>4</sup><http://www.pitt.edu/emotion/ck-spread.htm>  
<sup>5</sup><http://cs.anu.edu.au/few>  
<sup>6</sup><http://staff.estem-uc.edu.au/roland/research/datasets/>

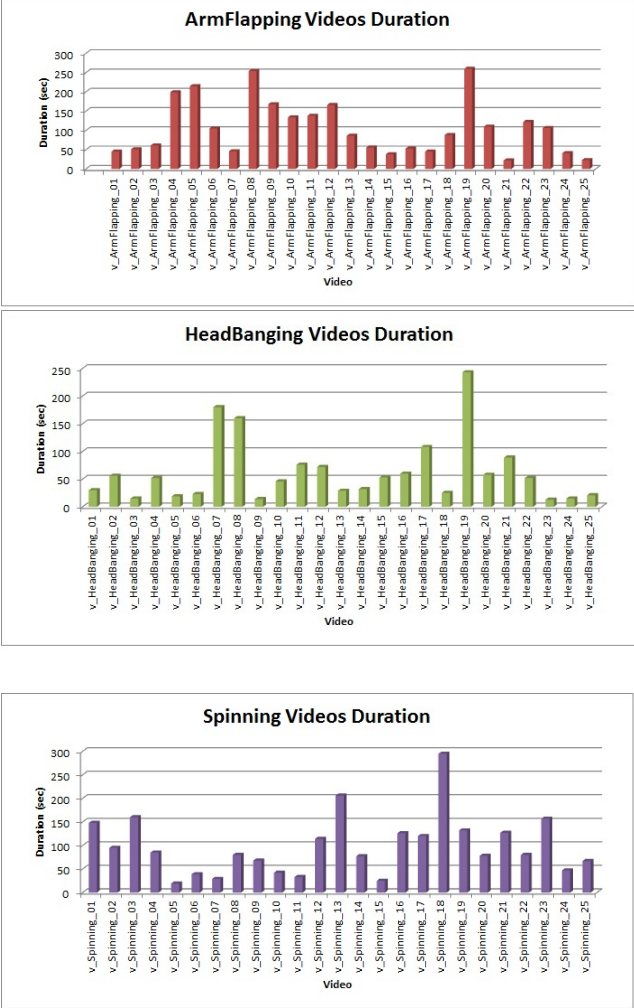


Figure 2. Duration of individual videos in all three categories

Attributes	Description
URL	Reference website URL to the video
Persons	Number of persons in the video
Behaviours	Number of stimming behaviours in the video
Time	The behaviour start:end time instant in a video
BodyPart	Hand, Head, Face, Eye, Full
Category	Stimming behaviour category
Intensity	Behaviour Intensity - Low, Medium and High
Modality	Dominant behaviour modality like audio, video. For future use

Table 1. Attributes used for video annotation



### Arm Flapping Videos



### Head Banging Videos



### Spinning Videos

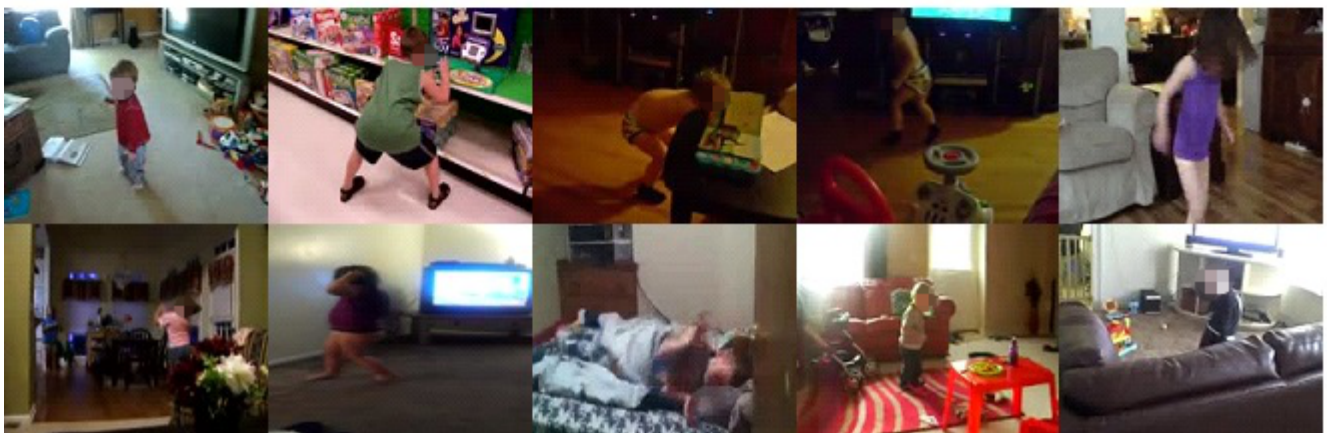


Figure 1. Snapshots(blurred to preserve the anonymity of identity) of videos in all three stimming categories. The children exhibit different postures and were in different places. Moreover, seen are varying backgrounds, clutter and multiple objects.

```

<?xml version="1.0"?>
<Stimming count="75">
  <ArmFlapping count="25">
    <video id="v_ArmFlapping_01">
      <url>http://www.youtube.com/watch?v=I7fdv1q9-m8</url>
      <persons>1</persons>
      <duration>46 sec</duration>
      <behaviours id="b_Set_01" count = "2">
        <behaviour id="b_01">
          <time>18:24</time>
          <bodypart>hand</bodypart>
          <category>armflapping</category>
          <intensity>high</intensity>
          <modality>video</modality>
        </behaviour>
        <behaviour id="b_02">
          <time>27:32</time>
          <bodypart>full</bodypart>
          <category>headbanging</category>
          <intensity>medium</intensity>
          <modality>video</modality>
        </behaviour>
      </behaviours>
    </video>
  </ArmFlapping>
</Stimming>

```

Figure 3. Sample XML schema for video annotation

#### 4. Challenges in Self-Stimulatory Behaviour Dataset

The analysis of the videos recorded in an uncontrolled environment has provided us more domain specific insights, which are extremely challenging to deal with for automatic behaviour analysis. In addition to the common computer vision challenges, such as camera motion, pose, illumination, cluttered background, video quality, occlusion, etc., the following domain specific challenges have to be carefully studied.

1. **Subtle behaviours** - The children videos may not contain the stimming behaviour for the full duration. The behaviours will be seen in spurts and there may be other dominant activities. Hence, it is important for the algorithm to pick up the subtle behaviour precisely and use that to characterize the behaviour.
2. **Intensity and Continuity** - The intensity level of the behaviours will vary and there is no specific pattern that is followed by the child. Furthermore, the behaviour will not happen continuously in one go, instead, it will happen multiple times with varying levels of intensity. This makes the analysis difficult and the time series history has to be learnt. This is quite in contrast to the actions that are observed in standard action recognition datasets, wherein the specific action is dominant, continuous and with similar intensity.
3. **Spatial Variance** - It is quite common for the child to start exhibiting the behaviour in one location in the house or play area and move to different location while continuing on the behaviour. This implies that there will be many other objects and persons in the trajectory

of the child's movement. Potentially, the motions of objects not of interest will start to influence the scene more than the child's behaviour.

4. **Social Cues** - It is usually the case where the person recording the video, be it a parent or caregiver, prompts the child to perform some action or look at the camera. It is in these circumstances that the child may choose to completely ignore and/or gets excited to perform atypical behaviours. Both these situations yield an important cue for the analysis. In the former case, a lack of response or interest behaviour can be elicited, while the regular stimming behaviour can be studied in the latter case. Hence, the algorithm for automatic behaviour analysis ideally has to be multimodal, taking cues from speech, video, face, eye and body expressions.
5. **Mixing Behaviours** - It is usual for the child to start with a behaviour and transition into a completely new behaviour within a video. It is also possible that the child combines both the behaviours in one swift action. For example, the child might start with arm flapping, then slowly transitions to rotating in circles with and without arm flapping. These types of mixing behaviours have to be dealt with while designing new algorithms.
6. **Object Influence** - During the child's stimming behaviour, the associated objects might also follow the same behaviour. The most challenging fact is that the object alone continues with the behaviour after the child has stopped performing the behaviour. For example, the child might sit on a chair and keep rotating in circles and after some time, the child will stop stimming but make the chair to rotate constantly. Hence, the influence of associated objects has to be carefully separated during the analysis
7. **Person Anxiety** - It is very common for the parent or caregiver to get anxious when the child starts performing the stimming behaviour with high vigour. In these circumstances, there will be significant camera motions reducing the video quality and in turn adding complexity to the analysis
8. **Context Stimming** - When the child is just observing an activity, say, a TV program and if there are stimming behaviours in the TV programs, they are not to be confused with the child's behaviour. A mechanism to delineate these 'false' stimming events is important

#### 5. Experiments and Results

The standard action recognition pipeline of interest point detection, feature extraction, generation of feature descrip-

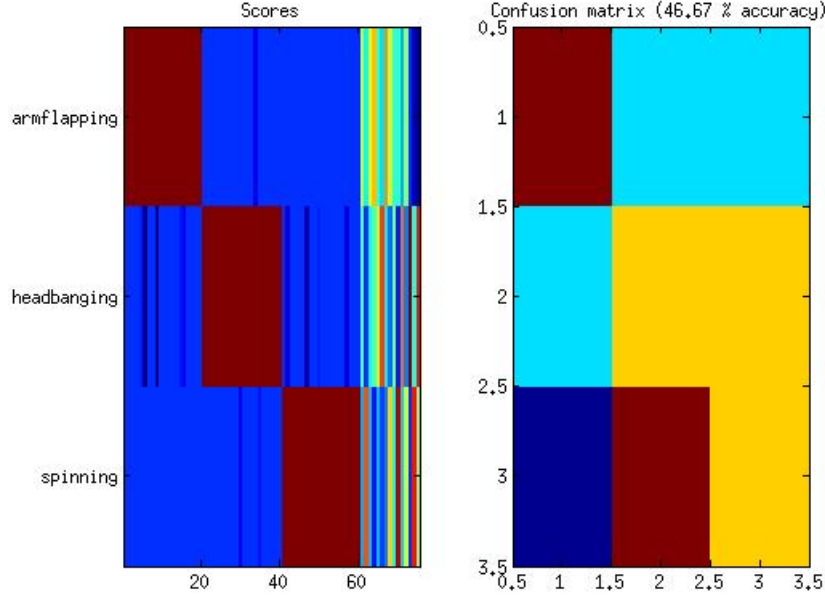


Figure 4. Confusion matrix and Classification Accuracy for three stimming categories. This example is a 5-fold validation results for a run in a “Leave-One-Group-Out” scenario

tors, model training and using it for recognition for a test video has been followed on this new dataset. The well known Space Time Interest Points (STIP) [10] are employed with Harris3D detectors in a Bag Of Words (BOW) framework to train a 3-class classifier. The three classes are arm flapping, head banging and spinning corresponding to the videos in the dataset. The experiments are conducted using multiple codebook sizes with n-fold validation in a “Leave-One-Group-Out” scenario. A group of 5 videos from every class is left out for testing, while the classifier is trained on the remaining videos. The details on the codebook sizes and the mean accuracy values corresponding to different folds are provided in Table 2. We report on mean accuracy result of 47.1% for 5-fold validation across three codebook sizes of 500,1000 and 1500. The confusion matrix for one of the runs is shown in Figure 4. The experi-

ments are conducted by modifying the image classifier code provided as part of VLFeat library [3]. The features used in this baseline experiment did not capture the challenges discussed in Section 4 and, hence, contribute to the low overall accuracy. In particular, many videos in this dataset had stimming behaviours for a short period of time and other activities for the remaining duration. The current baseline algorithm has not modelled this aspect of delineating the interested behaviour from other activities. There are different types of spinning behaviours observed in this dataset such as rotating in circles and rolling on the bed. The inherent periodic motion of these behaviours have not been learnt in the current baseline algorithm resulting in low accuracy. Hence, a robust model of behaviour analysis needs to be developed for analysing self-stimulatory behaviour videos taken in natural settings.

Codebook Size	5-fold	10-fold
<b>500</b>	<b>44.0%</b> ( $\sigma$ :5.3)	<b>42.0%</b> ( $\sigma$ :6.6)
<b>1000</b>	<b>50.7%</b> ( $\sigma$ :3.2)	<b>47.3%</b> ( $\sigma$ :5.5)
<b>1500</b>	<b>46.6%</b> ( $\sigma$ :12.6)	<b>44.6%</b> ( $\sigma$ :11.1)

Table 2. Classification accuracy results on proposed dataset. The mean accuracy and standard deviation across the folds are computed corresponding to different codebook sizes.

## 6. Conclusions

We have collected and annotated a rich dataset of children’s self-stimulatory behaviour videos recorded in an uncontrolled environment. We share this with the research community along with per-video level annotations. The baseline experiments on this dataset yield low class recognition accuracy and, hence, future work is needed to improve the performance against this new benchmark. Also, we plan to increase the size of the dataset to include more videos and categories. We hope this dataset can act as a good benchmark for studying the children’s self-stimulatory behaviour



in a regular day-to-day activities and ultimately developing technology that will be useful to parents / caregivers and clinicians for early intervention and diagnosis.

## References

- [1] Asd video glossary - <http://www.autismspeaks.org/what-autism/video-glossary>. 2
- [2] Autism generic resource exchange (agre) - <http://research.agre.org/agrecatalog/agrecatalog.cfm>. 2
- [3] V. Andrea. Image classifier on caltech-101 data - <http://www.vlfeat.org/applications/apps.html>. 6
- [4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. A semi-automatic method for collecting richly labelled large facial expression databases from movies. *IEEE Multimedia*, 2012. 2
- [5] DSM. Diagnostic and statistical manual of mental disorders (dsm-5) - <http://www.dsm5.org/pages/default.aspx>, May 2013. 1
- [6] J. Hashemi, T. V. 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 *ICDL-EPIROB*, 2012. 1, 2
- [7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2
- [8] J. Joshi, R. Goecke, M. Breakspear, and G. Parker. Can body expressions contribute to automatic depression analysis? In *IEEE FG*, 2013. 2
- [9] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013. 2
- [10] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64 number 2/3:107–123, 2005. 6
- [11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, CVPR’08, pages 1–8, 2008. 2
- [12] C. Lord, S. Risi, L. Lambrecht, J. Cook, Edwin H., 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. of Autism and Dev. Dis.*, 24:659685, 2000. 1
- [13] C. Lord, M. Rutter, and A. Le Couteur. The autism diagnostic interview revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. of Autism and Dev. Dis.*, 24:659686, 1994. 1
- [14] J. Rehg. Behavior imaging: Using computer vision to study autism. In *MVA2011 IAPR Conference on Machine Vision Applications*, 2011. 1, 2
- [15] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, and et al. Decoding children’s social behavior. In *IEEE CVPR*, 2013. 1, 2
- [16] B. Susan E., Z. Lonnie, M. Catherine, R. Vicki, and J. Brian. The autism observation scale for infants: Scale development and reliability data. *J Autism Dev Disord*, 38:731–738, 2008. 1, 3
- [17] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, and P. Brian, J. and Szatmari. Behavioral manifestations of autism in the first year of life. *Int. J. Dev. Neuroscience*, 23:143–152, 2005. 2