Detecting self-stimulatory behaviours for autism diagnosis

Article · January 2015

DOI: 10.1109/ICIP.2014.7025294

2 authors:

Shyam Sundar Rajagopalan
University of Canberra
6 PUBLICATIONS 67 CITATIONS

SEE PROFILE

Roland Goecke
University of Canberra
166 PUBLICATIONS 3,686 CITATIONS

SEE PROFILE

DETECTING SELF-STIMULATORY BEHAVIOURS FOR AUTISM DIAGNOSIS

Shyam Sundar Rajagopalan[1]

Roland Goecke[1,2]

[1]Vision & Sensing Group, HCC Lab, ESTeM
University of Canberra, Australia
shyam.rajagopalan@canberra.edu.au

[2]IHCC Group, RSCS, CECS,
Australian National University, Australia
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. An algorithm for detecting three types of self-stimulatory behaviours from publicly available unconstrained videos is proposed here. The child's body is tracked in the video by a careful selection of poselet bounding box predictions using a nearest neighbour algorithm. A global motion descriptor – *Histogram of Dominant Motions (HDM)* – is computed using the dominant motion flow in the detected body regions. The motion model built using this descriptor is used for detecting the self-stimulatory behaviours. Experiments conducted on the recently released unconstrained SSBD video dataset show significant improvement in detection accuracy over the baseline approach. The robustness of the method is validated using benchmark action recognition datasets. The proposed poselet bounding box selection algorithm is validated against the ground truth annotation data provided with the UCF101 dataset.

***Index Terms*—** Computaional behaviour analysis, autism, stimming, poselets, histogram of dominant motions

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 diagnosed with ASD is 1 in 88 in the USA [1]. The genetic basis for ASD is still unknown and a common way of diagnosing is by using behavioural cues of the children [2]. *Self-stimulatory behaviours* are one type of such atypical behavioural cues used for the diagnosis. They refer to stereotyped, repetitive movements of body parts or objects, such as arm flapping, head banging, and spinning. The diagnosis involves clinicians interacting with children in multiple long sessions to identify the behavioural cues, risking a delay in diagnosis and being an expensive process. Moreover, suitably trained clinicians may not be available in some areas, in particular in regional areas. Computational behaviour analysis can provide a way of assisting clinicians and parents. ASD is typically diagnosed at the age of 5yr, while a diagnosis by the age of 2yr would be desirable [3, 4]. A retrospective analysis of reports from parents and analysing children's home videos are commonly used mechanisms in the diagnosis [4].

A new dataset of self-stimulatory behaviours (SSBD) of children videos captured in unconstrained conditions was released recently [5]. These videos are extremely challenging for analysis, due to camera pan / zoom, camera motion / wobble, low resolution, lighting conditions, and also due to the domain specific challenges such as spatial variance, subtle behaviours, context stimming, etc. The main contributions of this paper are

1. A nearest neighbour algorithm for the selection of poselet bounding boxes tracking the child's body in a video.

2. A global dominant motion field based descriptor for capturing motion patterns in self-stimulatory behaviours.

2. MOTIVATION

The time until an ASD diagnosis is made is long due to (a) scarce availability of specialists and (b) the diagnosis involving the study of behaviour cues over a period of time. Technology that assists in the automatic analysis of videos to quantify the child's behaviours and to summarise the videos for the specialists may help to reduce the turnaround time for the diagnosis and provide better access to clinicians remotely. In addition, technology that provides alerts to parents by analysing unconstrained videos, such as videos taken in normal day-to-day play activities of a children, can go a long way towards early intervention. Hence, this work aims to take a step towards this goal. Advancements in computer vision research in understanding human actions and activities naturally lead to the next stage of analysing more subtle child behaviours [2, 6].

3. RELATED WORK

Research on studying self-stimulatory behaviours can broadly be grouped into two categories. The first category involvse using wearable sensors in the child's hands or on the body and tracking the sensor data obtained over a fixed period of time. The tracked body part locations are then used to analyse for repetitive self-stimulatory behaviours. Westyn *et al.* [7] used wrist band worn sensors to track the hand motion of the child and constructed a Hidden Markov Model using 3-axis readings obtained from the sensors to classify the behaviours. Ploetz *et al.* [8] used sensors attached to the limbs to obtain acceleration data. The captured data is segmented to form behaviour episodes that are used to extract features to train models, which are then used for detecting aggression, disruption and self-injury behaviours.

The second category of work involves tracking the child's body part motion in a video and then analysing the tracks to detect the behaviours. Marwa *et al.* [9] proposed a new descriptor to detect and localise the rhythmic body movements using colour and depth images. In [10], visual tracking, the level of attention, sharing interest and motor patterns are the studied behaviours using vision techniques. To the best of our knowledge, there is not yet any work reported on automatically analysing self-stimulatory behaviours in children videos captured in unconstrained conditions.

4. SELECTION OF POSELET BOUNDING BOXES

Poselets are body part detectors trained on various human poses [11] that are used to identify person locations in an image. In this work, poselet detectors are used to obtain bounding box predictions. The frames are marked as detected if the maximum confidence score among all the predictions is greater than a defined threshold δ. This value is empirically chosen as 1.0 in the experiments. The unconstrained children videos contain multiple occlusions, illumination changes, low resolution and significant pose changes. Due to these challenges, poselets fail to detect the child's body region with high confidence in certain frames within the video. One of the contributions of this paper is to obtain the best possible poselet bounding box prediction of a child in such frames.

Assume poselets detected a body region, R_i, with high confidence in frame t. A new cost function f is defined and used to identify the body region in neighbouring frames with high accuracy. The cost function is dependent on the difference area of region R_i (ψ_i) between two frames, the maximum intersection area between R_i and all poselet bounding boxes in frame $t-1$ (ϕ_i), and the distance between R_i and all poselet bounding boxes in frame $t-1$ (δ_i). It is given by

$$f = \delta_i + \Delta\psi_i + 1/\phi_i.$$

The bounding box selection algorithm is run against all bounding boxes in neighbouring frames and the one that minimises the cost function is chosen as the body region bounding box for undetected frames. Algorithm 1 gives detailed steps involved in this procedure and Figure 1 provides a visual illustration.

Algorithm 1: Selection of poselet bounding boxes

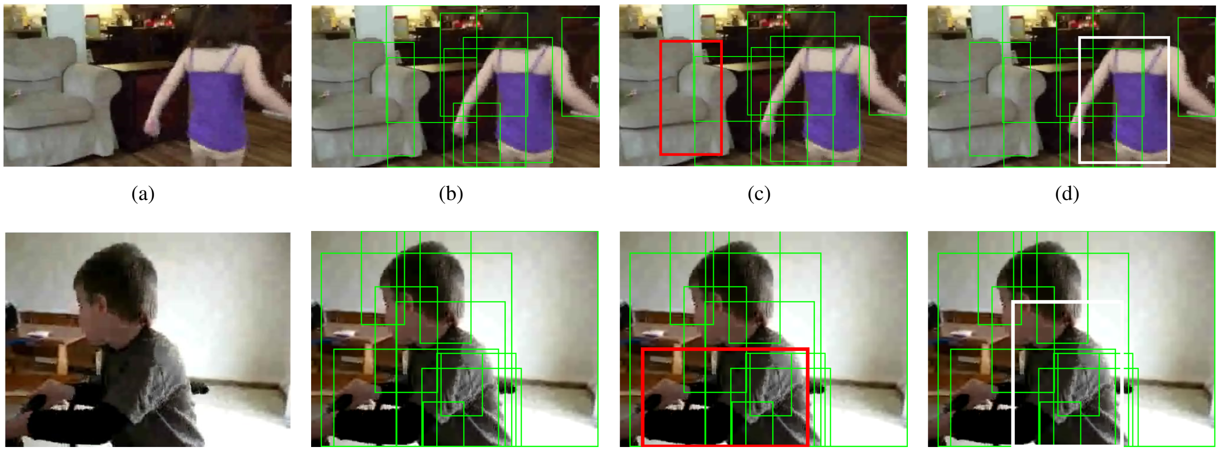
```
  Input: Set $A = \{a_1, a_2, \ldots, a_n\}$ of bbox predictions
  Output: Set $P = \{p_1, p_2, \ldots, p_n\}$ of detected bboxes
for i=1 to n do
$s = score(a_i)$
$[s_{max}, idx] = max(s)$
if $s_{max} \geq \delta$ then
    $sel_i \leftarrow true$
    $p_i.bb = a_i.bb_{idx}$
else
    $sel_i \leftarrow false$
repeat
    for each $i$ of $sel_i$ do
        // compute bbox for left and right neighbours
        $prev.bb = sort(a_{i-1}.bb, score, descend)$
        $prev.bb = TopK(prev.bb)$
        $cost = f(\delta, \phi, \Delta\psi)$
        $[b, idx] = min(cost)$
        $p_{i-1}.bb = prev.bb_{idx}$
        $next.bb = sort(a_{i+1}.bb, score,' descend')$
        $next.bb = TopK(next.bb)$
        $cost = f(\delta, \phi, \Delta\psi)$
        $[b, idx] = min(cost)$
        $p_{i+1}.bb = next.bb_{idx}$
until bboxes for all frames are detected
```

5. HISTOGRAM OF DOMINANT MOTIONS (HDM)

Repetitive translational and rotational motions are the two predominant motion patterns in children's self-stimulatory behaviours. The proposed approach builds on the pixel level optical flow to compute the *Dominant Flow* for the frame, which is a higher level abstraction over the pixel level optical flow. A new global motion descriptor obtained from dominant flow, *Histogram of Dominant Motions (HDM)*, is used for modelling the motion patterns, thereby, the self-stimulatory behaviours.

5.1. Computation of HDM descriptor

DeepFlow [12] is a recently released optical flow algorithm designed to handle larger displacements in unconstrained videos. It is used in this work to compute the optical flow between successive frames. The gradient orientation is obtained at every pixel location using the computed optical flow. To identify a dominant motion, a Histogram of Oriented Optical



(a) (b) (c) (d)

(e) A frame

(f) Poselet bounding box predictions

(g) Red box showing selection based on maximum score

(h) White box showing selection based on proposed algorithm

Fig. 1. Person detection using poselets for a single frame (blurred to preserve the anonymity of identity) in two video sequences

Flow (HOOF) [13] is calculated. HOOF captures the distribution of gradient orientations for a frame along with the magnitude. The dominant motion is obtained from HOOF by finding more common, high scoring, gradient orientations. The per frame dominant motion is aggregated across all frames to form a *Histogram of Dominant Motions (HDM)*. The variance of the horizontal and vertical motions indicates the dominant motion direction in a video. The HDM is fused with the dominant motion direction variance to form a global HDM descriptor. Detailed steps for computing HDM descriptor are given in Algorithm 2.

Algorithm 2: Computation of HDM descriptor

```
Input: Set $O = \{O_1, O_2, \dots, O_n\}$ of optical flows
Output: $HDM$ a $38X1$ Dominant Motion Descriptor
while $i \leq n$ do
    $u \leftarrow O_i.u$
    $v \leftarrow O_i.v$
    $u_{var} \leftarrow variance(u)$
    $v_{var} \leftarrow variance(v)$
    if $u_{var} \leq v_{var}$ then
        $d_v \leftarrow d_v + 1$
    else
        $d_h \leftarrow d_h + 1$
    $HOOF_i \leftarrow ComputeHOOF(u, v)$
    $\theta_i \leftarrow GetDominantAngle(HOOF_i)$ 3
$h \leftarrow ComputeHistogram(\theta)$
$d_v \leftarrow d_v / sum(d_v)$
$d_h \leftarrow horiz / sum(d_h)$
$HDM \leftarrow [d_v, d_h, h]'$
```

Algorithm 3: GetDominantAngle

```
Input: $HOOF$ - Histogram of Oriented Optical Flow
Output: $DomAngle$ - Dominant Angle
$angles \leftarrow TopK(SORT(HOOF, descend))$
// place the angles in one of four quadrants
$quadrants \leftarrow FindQuadrant(angles)$
$q \leftarrow SelMaxQuadrant(quadrants)$
$DomAngle \leftarrow FindAvgQuadAngle(q)$
```

6. EXPERIMENTS AND RESULTS

The effectiveness of the proposed poselet based body region estimator is tested by comparing the bounding box estimations with the ground truth annotations provided in the UCF101 dataset [14]. The HDM descriptors around the detected body regions are used to train a discriminatory model for classifying self-stimulatory behaviours. To the best of our knowledge, the recently released SSBD is currently the only publicly available dataset of videos of children with ASD captured in unconstrained conditions and is thus used here. In addition, to validate the generalisation capability of the proposed method, a subset of videos from the benchmark UCF101 [15] and Weizmann [16] datasets are analysed.

6.1. Selection of Poselet Bounding Boxes

The bounding box annotations of a person are provided for 24 classes in UCF101 and a single video from the first group in all classes is used for the experiments. The intersecting area is computed between the estimated and ground truth bounding boxes of a person and the results are shown in Fig. 2. There is a considerable improvement in the bounding box intersecting

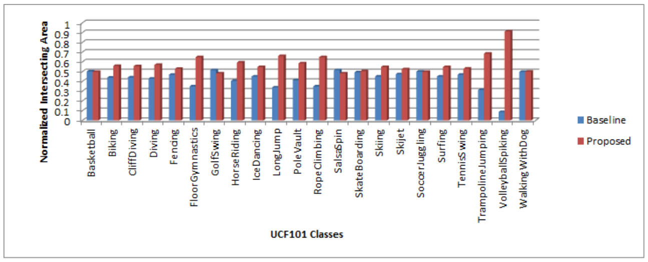


Fig. 2. Selection of poselet bounding boxes for UCF101 classes

area with the proposed approach indicating its effectiveness in detecting the child's body regions. The baseline area is computed by using the poselet detection based on maximum confidence score. In situations, where poselet detections are accurate with high confidence scores for most of the frames, the need for a new method does not exist and hence the contribution from the proposed algorithm is not very significant.

6.2. SSBD Dataset

The videos in SSBD dataset are pre-processed to obtain subclips of self-stimulatory behaviours. The baseline results are obtained by computing HOG-HOF features around space-time interest points[17] of a frame. These features are used in a Bag-of-Words framework to build a model. The experiments are performed using the VLFeat library [18] and the results are reported in Table 1. The 3-Class refers to *ArmFlapping*, *HeadBanging* and *Spinning* classes. *HeadBanging* and *Spinning* behaviours are used for the 2-Class problem and due to clear separation of motion patterns, the detection accuracy of these behaviours are significant. The baseline approach uses appearance and flow information at all interest points to build a model, including non-body regions of the child resulting in degraded performance. In the proposed method, features are computed in the detected body region. The errors introduced due to challenges with unconstrained videos have been minmizsed due to the choice of higher level global dominant features leading to improved performance.

	Test	Baseline	Proposed HDM
2-Class	5-fold	58.6% (σ 15.6)	83.4% (σ 6.9)
	10-fold	69.5% (σ 15.8)	86.6% (σ 13.0)
3-Class	5-fold	54.5%(σ 9.5)	73.6%(σ 8.2)
	10-fold	45.8%(σ 19.1)	76.3%(σ 13.6)

Table 1. k-fold cross validation results for the SSBD

6.3. UCF101 Dataset

To test the proposed method more widely and to show its generalisation capability, a subset of videos containing actions similar to self-stimulatory behaviours are chosen from the benchmark UCF101 dataset. We select a total of 375 videos for the tests from 15 classes with one video from every group in a class ensuring complete representation. The classes are grouped into two categories based on the motion type (rotational or non-rotational). The classes *SalsaSpin*, *IceDancing*, *JumpingJack*, *Swing*, *GolfSwinging*, *CricketBowling*, *HulaHoop*, *Nunchucks* and *YoYo* have actions that have rotational motions and *BenchPress*, *PullUps*, *PushUps*, *JumpRope*, *RopeClimbing* and *Skiing* belong to the non-rotational motion category. The poselet detector is not run for this dataset as the videos contain motion only in the region of interest. Tests on these two categories indicate comparable performance with the baseline approach (see Table 2).

	Test	Baseline	Proposed HDM
2-Class	5-fold	78.8% (σ 5.8)	82.1% (σ 6.1)
	10-fold	77.8% (σ 6.4)	82.7% (σ 5.6)

Table 2. k-fold cross validation results on UCF101

6.4. Weizmann Dataset

The Weizmann dataset of 10 classes is well studied with precise 100% accuracy using state-of-the-art algorithms. While this dataset could therefore be considered as easy, we use it here for a second generalisation experiment. The proposed algorithms also achieves 100% accurary (see Table 3). Poselet detectors are not run for this dataset as the only motion is in the region of interest. The classes, *wave2*, *bend*, *wave1*, *jack* are categorised as rotational motion and the remaining *walk*, *skip*, *side*, *run*, *pjump*, *jump* are placed in the other category.

	Test	Baseline	Proposed HDM
2-Class	5-fold	100.0%	100.0%
	10-fold	91.6%	100.0%

Table 3. k-fold cross validation results on the Weizmann dataset

7. CONCLUSION AND FUTURE WORK

An approach for detecting self-stimulatory behaviours in videos shot in unconstrained conditions is proposed. An algorithm for selecting poselet bounding boxes and a higher level global descriptor based on dominant motion patterns in the video are proposed and used for modelling the self-stimulatory behaviours. Experiments conducted on a challenging real-world dataset show good recognition accuracy, while demonstrating robustness on benchmark action recognition datasets. Future work will involve spatio-temporal localisation of the child and the behaviours in the video.

8. REFERENCES

[1] Centers for Disease Control and Prevention, "Prevalence of autism spectrum disordersautism and developmental disabilities monitoring network, 14 sites, United States, 2008.," in *MMWR Surveillance Summaries*, 2012, vol. 61, pp. 1–19.

[2] James Rehg, "Behavior imaging: Using computer vision to study autism," in *IAPR Conference on Machine Vision Applications (MVA2011)*. IAPR, June 2011, pp. 14–21.

[3] Shattuck PT, Durkin M, Maenner M, Newschaffer C, Mandell DS, Wiggins L, Lee LC, Rice C, Giarelli E, Kirby R, Baio J, Pinto-Martin J, and Cuniff C., "Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study.," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 48, no. 5, pp. 474–483, May 2009.

[4] Lonnie Zwaigenbaum, Susan Bryson, and Nancy Garon, "Early identification of autism spectrum disorders," *Behavioural Brain Research*, vol. 251, pp. 133–146, Aug. 2013.

[5] Shyam Sundar Rajagopalan, Abhinav Dhall, and Roland Goecke, "Self-Stimulatory Behaviours in the Wild for Autism Diagnosis," in *IEEE ICCV Workshop on Decoding Subtle Cues from Social Interactions*, Dec. 2013, pp. 755–761.

[6] James M Rehg, Gregory D Abowd, Agata Rozga, Mario Romero, Mark A Clements, Stan Sclaroff, and et al, "Decoding Children's Social Behavior," in *IEEE CVPR*, 2013, pp. 3414–3421.

[7] Tracy Westeyn, Kristin Vadas, Xuehai Bian, Thad Starner, and Gregory D. Abowd, "Recognizing mimicked autistic self-stimulatory behaviors using hmms," in *In IEEE International Symposium on Wearable Computers*. 2005, pp. 164–169, IEEE Computer Society.

[8] T. Ploetz, N.Y. Hammerla, A. Rozga, A. Reavis, N. Call, and G. D. Abowd, "Automatic assessment of problem behavior in individuals with developmental disabilities," in *Proceedings of the 14th ACM International Conference on Ubiquitous Computing (Ubicomp 2012)*, 2012.

[9] Marwa Mahmoud, Louis-Philippe Morency, and Peter Robinson, "Automatic Multimodal Descriptors of Rhythmic Body Movement," in *International Conference on Multimodal Interaction*, 2013.

[10] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro, "A computer vision approach for the assessment of autism-related behavioral markers," in *ICDL-EPIROB*, May 2012.

[11] Lubomir Bourdev and Jitendra Malik, "Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations," in *International Conference on Computer Vision (ICCV)*, 2009.

[12] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *IEEE Intenational Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013.

[13] Rizwan Chaudhry, Avinash Ravichandran, Gregory D. Hager, and René Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1932–1939.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Bounding box annotations of humans (24 classes)," in *THUMOS-ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," in *CRCV-TR-12-01*, Nov. 2012.

[16] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as Space-Time Shapes," in *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, 2005, pp. 1395–1402.

[17] Ivan Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, pp. 107–123, 2005.

[18] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," 2008.