Machine Vision and Applications (2018) 29:703–717
https://doi.org/10.1007/s00138-018-0926-y

ORIGINAL PAPER

# Toward identifying behavioral risk markers for mental health disorders: an assistive system for monitoring children's movements in a preschool classroom

Nicholas Walczak[1] · Joshua Fasching[1] · Kathryn Cullen[2] · Vassilios Morellas[1] · Nikolaos Papanikolopoulos[1]

Received: 23 December 2016 / Revised: 22 January 2018 / Accepted: 25 March 2018 / Published online: 18 April 2018

**Abstract**

Mental health disorders are a leading cause of disability in North America. An important aspect in treating mental disorders is early intervention, which dramatically increases the probability of positive outcomes; however, early intervention hinges upon knowledge and detection of risk markers for particular disorders. Ideally, the screening of these risk markers should occur in a community setting, but this is time-consuming and resource-intensive. Assistive systems could greatly aid in the detection of risk markers in a hectic environment like a preschool classroom. This paper presents a multi-sensor system consisting of 5 RGB-D sensors that detects and tracks the location of occupants in a preschool classroom and computes a measure of activity level and proximity between individuals, an index of social functioning. This assistive system operates in near real-time and is able to track occupants and deal with difficult situations both with occupants (children sitting and laying on the ground, hugging, playing dress-up, etc) and their environment (i.e., changing light levels from artificial and natural sources). The system is installed at, and validated on recordings taken from, the Shirley G. Moore Lab School, a research preschool classroom at the University of Minnesota. The work described herein provides the initial groundwork for monitoring basic elements of child behavior; future efforts will be geared toward identifying and tracking more sophisticated behavioral signatures relevant to mental health.

## 1 Introduction

Early intervention can dramatically improve an individual's quality of life for many psychiatric disorders. Accounting for 25% of all years of life lost to disability and premature mortality, mental health disorders are the leading cause of disability in the USA and Canada according to the World Health Organization [36]. Symptoms of mental illness which emerge in childhood and early adolescence are actually the later stages of a process which began years earlier. Hence, psychiatric research is immensely interested in identification and detection of risk markers (genetic, neural, behavioral, and/or social deviations) that indicate elevated risk for development of specific mental illnesses prior to the onset of symptoms.

The task of defining these risk markers is challenging since it requires a lengthy process of collecting, manually annotating, and analyzing data. An example of examining behavioral risk markers through video analysis lies in the study of [8], where retrospective home videos of patients with manually annotated skeletonizations were used. The researchers analyzed the symmetry in the body and were able to show that it could be used as a predictor for the development of autism. Large amounts of effort must be expended in collecting, processing, and analyzing such data. Automated systems can assist both in the process of analyzing the data and also in allowing the collection of data that would not be feasible through manual observations.

Once relevant risk markers have been identified, the next challenge is to detect them reliably in practice. Often, this is performed by a trained clinician who must spend hours either in an exam room with a patient or analyzing recorded videos. The time-consuming nature of this process severely limits the number of patients these professionals can attend to. This process of identification and detection can be greatly aided by assistive systems.

✉ Nikolaos Papanikolopoulos
npapas@cs.umn.edu

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA

[2] Department of Psychiatry, University of Minnesota, Minneapolis, USA

Considering this motivation, this paper presents a system that is able to detect and track the occupants in a preschool classroom. The over-arching goal of this system is to identify disease-specific neurobehavioral predictors in a system that could be widely disseminated and used in settings such as schools to facilitate early intervention. At this stage in development, the system is focused on the capability to capture basic, observable characteristics of children behavior that are relevant to mental health. This is realized by computing a measure of activity level for the tracked individuals and also computing a measure of social relationship based on proximity, which can then be used to summarize relationships via a graph. This designed system has been installed in and validated on data recorded at the the Shirley G. Moore Lab School, a research preschool classroom at the University of Minnesota. An example of the classroom observed can be seen in Fig. 1.

Ecological validity is a concern for psychiatric assessments [2], so preservation of the classroom's natural environment is paramount. This means that the system must be minimally invasive and cannot burden the teachers with additional considerations. A preschool classroom represents a challenging environment, where children be moving in sporadically and far less predictably than adults. This can involve running, jumping, and rolling on the floor, or it could involve sitting at a table and calmly playing with blocks. The children can don costumes as they play dress-up and utilize a number of different props in the classroom. Environmental changes are also a concern, such as varying illumination levels.

The system consists of multiple RGB-D sensors that allow us to develop a 3D understanding of the environment. The environment creates a challenging situation characterized by cluttered scenes in which many people act spontaneously within a small geographical area. We must also deal with drastic illumination changes such as the sun shining through the windows and the lights being turned on and off. This multi-sensor setup helps us overcome the problem of numerous occlusions among people as well as allowing us to acquire multiple views of subjects for robust identification. The depth information combined with sensor calibration allows us to register objects from multiple sensors and to segment out objects of interest. We then create descriptors for the 3D objects and track them over time. This tracking enables us to generate statistics and perform higher-level analysis on the activity of children in the classroom. All of our data have been collected with the approval of the Institutional Review Board (IRB) at the University of Minnesota.

This paper is the first work that combines all of the different aspects of this system to describe an end-to-end pipeline that works on long sequences of real-world data along with ground truth based-validation. The main contributions of this paper consists of updates to our system for detecting, tracking, and analyzing the behavior of people in a challenging real-world setting along with validation of system performance. Both the detection and tracking phases of our processing pipeline have undergone changes that both increase performance and reduce runtime. The performance of the system is validated on over 22,000 frames from several recordings taken from a preschool classroom. Validation includes the detection metrics seen in previous work but also includes validation of the tracking system, which has not been previously published. Our system leverages 3D information to remove the background in a challenging setting where traditional RGB-based methods perform poorly. We are able to detect people in a much broader range of poses compared to similar works, where upright people are frequently taken for granted. We also present a novel method for dealing with merged detections, which allows us to accurately deal with situations where two people are in very close proximity (i.e., hugging). The system also performs in near real time, with

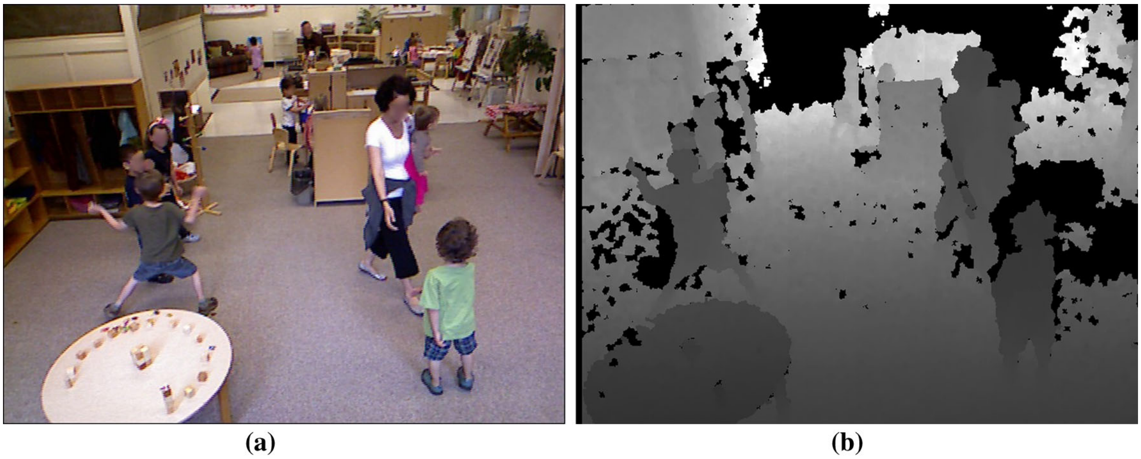


**Fig. 1** One RGB-D frame taken at the Shirley G. Moore Lab School, yielding an **a** RGB room image and **b** depth data. Depth is contrast normalized for visualization. (Faces blurred to preserve anonymity. Taken with IRB approval.) **a** RGB example **b** Depth example

an average frame processing time of approximately half a second.

## 2 Related work

The interaction between young children and caregivers is important yet complex, and in fact is central to the care and education of children under 3 years [7]. Works such as [4] indicate that the presence of a caregiver yields a positive impact, providing children with a "secure base" for interacting with other children. In addition, the acquisition of complex communication skills in children can be aided by the presence of adult caregivers. Other work however, such as [17], indicates that adult caregivers have an inhibiting effect on peer interactions. The work in [23] claims that these theories are not antagonistic, rather they are contextually dependent. This was shown through a painstaking study which required large amounts of video analysis and manual video coding schemes. Their work involved approximating the proximity between room occupants by assigning a grid of cells to the floor and counting the number of cells between individuals. Another example can be found in [33], which studies the role geography (location) plays in micro-social interactions of young children. The study's data are collected by manual annotation from observers watching the children. While these kinds of studies are important to the field of psychiatry, the amount of painstaking labor involved can limit the amount of work. Systems that can automate parts of the coding process, such as directly computing proximity, would have a large impact on this kind of work.

Computer vision researchers have begun to explore applying their techniques to neurodevelopmental disorder risk marker detection. More specifically on using vision in recognizing symptoms related to autism. A professional and child subject undergoing the Rapid ABC exam for ASD are monitored in the work of [30]. The work examines methods for predicting the subject's engagement in an activity using audio and visual information. Focus on engagement of the subject being examined in a protocol similar to the Rapid ABC is further explored in [18]. Their method focuses on modeling the appearance of body part landmarks on the subject's face to measure the exact orientation of the face relative to an object. Gait asymmetry in a subject, another risk marker associated with autism, is assessed using body pose tracking in this work as well.

Characterizing motor stereotypical behaviors associated with autism have been another area of exploration. Using the spatiotemporal interest point detector of [22], three motor stereotypic behaviors associated with autism were examined in [29] for classification. Their dataset is comprised of YouTube videos of children at various ages performing motor stereotypic behaviors. Classification of motor stereotypies was also explored in [11], where a new dataset was collected from classroom data and comparisons were also made to the previous YouTube dataset.

To our knowledge there is not another computer vision-based study that monitors children in an unconstrained classroom environment. Multi-sensor systems for object detection and tracking are a widely researched topic. An early example of a multi-sensor system for use in a "smart room" is presented in [20]. This "smart room" system uses computer vision-based monitoring to control certain aspects of a simulated living room, such as playback control of a movie when a person gets up or sits down from a couch. The system operates by using two calibrated stereo cameras, which yield depth and color images that can be used to create foreground blobs. These blobs are created by computing a mean and standard deviation for depth, and each individual color channel in the RGB images across 30 frames and comparing pixels to these models. Blob merging techniques are used to merge foreground blobs, and the assumption is made that only people will appear in the foreground. Another example of a multi-camera tracking system for "smart room" applications can be found in [15].

Another multi-camera system is presented in [24], which utilizes an array of 16 wide baseline stereo cameras. In that system, explicit models of people are built up based on color and "presence" probabilities. Intersections of epipolar lines across multiple cameras are used to determine a 3D location of tracked people on the ground plane. A Kalman filter is then used to track these locations over time.

A more recent work, [14], presents another multi-camera system for tracking people that incorporates information from a large number of RGB cameras. This work, like many other systems, uses a ground-plane homography to match image-plane locations to a 2d ground-plane location. Proprietary background subtraction is used to compute foreground blobs in the image. These foreground blobs are then combined with a generative model to compute occupancy in a quantized occupancy grid on the ground plane. Grid locations with a high enough degree of occupancy are considered positive person detections, and these are tracked across time.

In [25], Munaro et al. describe a method for detecting and tracking people using RGB-D that shares many similarities with our method. Their detection pipeline works by downsampling the point cloud created from the RGB-D input with a voxel grid filter, removing a ground plane detected with RANSAC, then clustering the remaining points based on Euclidean distance, and then finally running a histogram of oriented gradient (HOG)-based person detection algorithm on the RGB image projections of the clustered point clouds. This approach faces a similar problem to our approach in that multiple people standing close to each other will get merged into one detection. Munaro et al. solve this with their sub-clustering step which relies upon detecting heads. This, along with some of their other decisions for removing detections such as detections too close to the ground or too high off the ground, makes this approach unsuitable for our application. Frequently children will be sitting, laying, or rolling on the ground, and the vast array of different poses the children can assume invalidate the assumptions that make the approach in [25].

## 3 System

This section details the technical aspects of the system, highlighting the latest improvements where appropriate. The processing pipeline from the RGB-D sensors to the final output is summarized in Fig. 2.

### 3.1 Point cloud creation

The presented system consists of 5 Microsoft Kinect RGB-D sensors. These sensors are setup with a wide baseline, with each sensor sitting in the four corners of the room, and the fifth sensor in the middle of a long wall. Figure 6 depicts the areas of activity in the classroom, but also gives a visual depiction of sensor placement. While the Kinect sensors operate by projecting an infrared speckled dot pattern which can lead to interference from multiple sensors, this interference only causes points to be lost in the areas were patterns overlap. Since the sensors are placed with a wide baseline, the amount of overlap on objects of interest is minimized, so the vast majority of points lost are on areas of little interest (such as the floor). Additionally, the wide baseline aids in problems with occlusions; if the view of a subject is occluded in one sensor there is likely to be another sensor that has a better view of the same subject.

Each RGB-D sensor generates a color image and depth map which can be used to project a point cloud where each point consists of RGB color and 3D *XYZ* location. The extrinsic and intrinsic parameters for each sensor are extracted by calibration with a rigid calibration rig, which allows the point clouds from each sensor to be transformed into a global frame of reference. The calibration rig defines an origin in the world coordinate system and by computing the pose of the sensor an affine transformation can be applied to each sensor pointcloud, transforming it into the global frame of reference. Figure 3a–d depicts the color images from 4 sensors as part of an example of point cloud generation.

Each sensor can generate up to 307,200 points, which leads to over a million points for 5 sensors. Background subtraction is employed to reduce these points for later steps of processing. A 3D model of the background is learned by

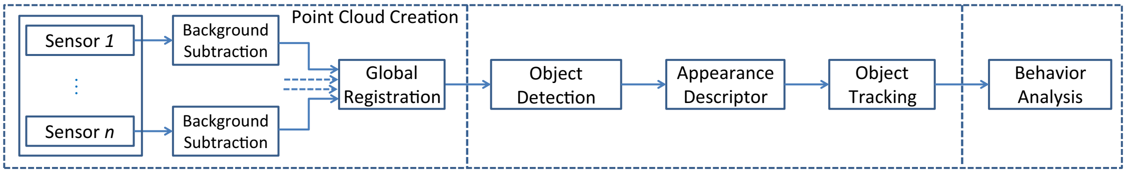


**Fig. 2** A flowchart of the processing pipeline

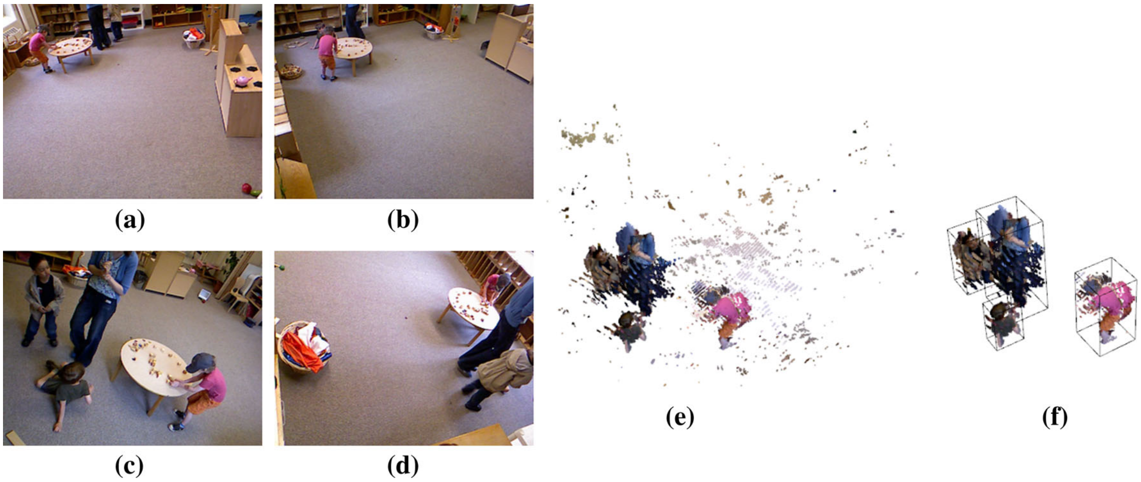


**Fig. 3** **a**–**d** Simultaneous RGB frames from four sensors. **e** The fused RGB and range data forming the 3D point cloud rendering the scene. **f** The four persons detected and enclosed in four bounding boxes. [Compare (**c**), (**e**) and (**f**)]. **a** RGB (Sensor 2), **b** RGB (Sensor 3), **c** RGB (Sensor 4), **d** RGB (Sensor 5), **e** noisy 3D reconstruction, and **f** detected objects

voxelizing the space of the classroom and learning a model of occupied voxels when the classroom is empty. Figure 3e depicts an example where data from each sensor are used to create the background-removed cloud.

The resulting background-removed global point cloud is then used as the input to the detection phase of the processing pipeline. Further details on the setup and point cloud generation of this system (including important issues such as temporal synchronization and dealing with variable frame rates in the sensors) can be found in previous works: [12,35].

### 3.2 Detection

In a track-by-detection approach, the first step is generation of detections. The same multi-step hierarchical clustering scheme described in our previous work is employed [12], referred to later in this work as HGB for hierarchical graph based. The detection process works by first collecting groups of homogeneous points into super voxels using an efficient graph-based segmentation algorithm [13]. The supervoxels are filtered based on size, as small supervoxels are more likely to be generated from noise. The remaining supervoxels are then clustered into detections and again filtered based upon size.

The graph-based supervoxel computation was handled by solving a max-cut problem, where a graph was created between each point in the point cloud with edges weighted based on color and spatial Euclidean distance. An example of supervoxels computed with the graph-based approach can be see in Fig. 4. This graph-based approach ignores some important underlying geometric aspects that arise from the nature of the data. Ideally, supervoxels should not cross object boundaries (i.e., all points in a single supervoxel should belong to only one object) and should tend toward continuity.

In light of this, we have employed a new supervoxel computation method, Voxel Cloud Connectivity Segmentation (VCCS) [27], which has more stringent geometric constraints and better meets the previous criteria. The detection approach that utilizes VCCS is later referred to as HVB, for hierarchical voxel based. VCCS operates by creating an adjacency grid using a voxelized version of the point cloud. Initial supervoxels are seeded, and then clustered based on the following feature vector:

$$\mathbf{F} = [x, y, z, L, a, b, \mathrm{FPFH}_{1..33}], \tag{1}$$

where $x, y, z$ are the spatial coordinates, $L, a, b$ are color in CIELab space, and $\mathrm{FPFH}_{1..33}$ are the 33 elements of Fast Point Feature Histograms (FPFH), a local geometrical feature proposed by [32]. Since VCCS operates on a voxelized version of the point cloud, runtime complexity scales with the size of the space and not with the number of points, which makes this method faster than the previous graph-based approach. An example of supervoxels computed with VCCS can be seen in Fig. 4.

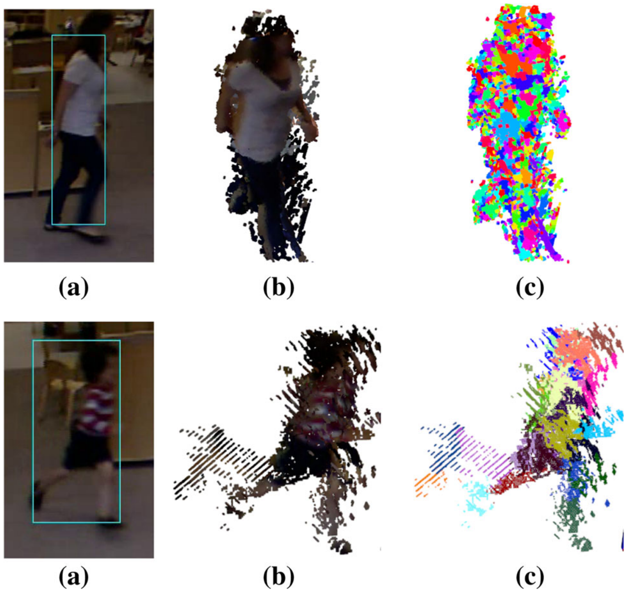


**Fig. 4** Images illustrating supervoxel generation, contrasting between graph-based supervoxels and VCCS supervoxels. **a** The detection bounding box from one view, **b** is a 3D view from the global 3D point cloud, and **c** shows the computed 3D supervoxels. A color blob denotes a single supervoxel, and colors may be reused for distinct supervoxels

Once the supervoxels are created and filtered, the remaining supervoxels are again clustered to create detections. Our previous work used a second pass of graph-based clustering with edge weights based on the maximum distance between two points between two supervoxels. This relied on the assumption that a standing person would have a dense cluster of points when projected onto the $X$–$Y$ plane, however this caused issues with situations like adults leaning over children. Instead, we now employ the density-based clustering method DBSCAN [9]. Distance between supervoxels is computed as the $L_2$ spatial distance between supervoxel centroids. This allows us to cluster supervoxels into detections without requiring the number of clusters beforehand; however, it can lead to issues if individuals are close enough that there is no clear density boundary.

For supervoxel filtering, supervoxels with fewer than 150 points are removed. For detection filtering, detections with fewer than 2300 points were removed. These numbers were found experimentally considering the point density of the returns on the sensors. In general, background points that remain after background subtraction are sparse, and when supervoxels are formed background supervoxels will have a low point count. When it comes to detections, there needs to be enough points to ensure an accurate appearance descriptor so even if a sparse detection is correct, it may still be worth discarding.

### 3.3 Appearance descriptors

Once objects (individuals) are detected, additional information is required to help track and identify the object. Each object detection consists of a number of points in the global point cloud and each point has associated with it an $X, Y, Z$ position, an $R, G, B$ color, and information about which pixel in which sensor generated that point. The object's point cloud provides position and shape information, and we can also leverage the color information provided by the RGB-D sensor. All of this heterogeneous information can be efficiently and compactly encapsulated within a region covariance descriptor (RCD). These descriptors were first introduced by [34] as an appearance descriptor for textures, and they were originally applied to tracking in [28]. Region covariance descriptors are computed by defining a feature vector which is computed for each point in a region of interest and then computing a covariance for that feature vector across all points. Our feature vector is defined as

$$\begin{aligned} f_s(p) = [&X,\ Y,\ Z,\ R(p),\ G(p),\ B(p), \\ &\partial_x I(p),\ \partial_y I(p),\ \partial_{xx} I(p),\ \partial_{yy} I(p),\ \partial_{xy} I(p), \\ &||\partial I(p)||,\ atan(\partial_y I(p),\ \partial_x I(p))], \end{aligned} \tag{2}$$

where $p$ represents a point from an object cloud situated at $\{X, Y, Z\}$ in the global reference frame. This point is projected back to the image plane for sensor $s$, where the color and intensity image derivative information is collected. In our previous work, depth information was incorporated as well; however, this was shown to have a negative effect in performance. This feature vector will yield a $13 \times 13$ covariance descriptor for each detected object, which incorporates information from each sensor that observed the object.

Non-singular covariance matrices are symmetric positive definite matrices that reside in the space $\mathrm{Sym}_d^+$ whose distance between two matrices is not accurately defined by the Euclidean distance. The affine-invariant Riemannian (AIRM) metric defined as

$$d_{\mathrm{R}}(P, Q) = \left\| \log\left(P^{-1/2} Q P^{-1/2}\right) \right\|_F, \tag{3}$$

captures the curvature and geodesic distance along the Riemannian manifold of $\mathrm{Sym}_d^+$. However, this metric is expensive to compute with matrix inversions, square roots, and logarithms. One efficient approximation of distance between covariance matrices is the Log-Euclidean Riemannian metric (LERM) [1],

$$d_{\mathrm{LE}}(P, Q) = \| \log(P) - \log(Q) \|_F, \tag{4}$$

where $P, Q \in \mathrm{Sym}_d^+$. LERM is a lower bound to AIRM [6]. This approximation was chosen for computational efficiency as $\log(X \in \mathrm{Sym}_d^+)$ can be precomputed for each object and used for several comparisons unlike AIRM.

### 3.4 Tracking

Tracking of detections in the scene was performed using a Kalman filter with a constant velocity model [3]. The state consists of

$$\mathbf{x} = \left[\mu_x, \mu_y, \dot{\mu}_x, \dot{\mu}_y, \sigma_x^2, \sigma_y^2, \sigma_z^2, \sigma_{xy}, \sigma_{xz}, \sigma_{yz}\right]^T, \tag{5}$$

where $\mu$ and $\dot{\mu}$ are 2D position and velocity and latter 6 values are the independent parameters in the spatial covariance.

The process noise covariance $Q$ was assumed to be uncorrelated with a differing parameter for position, velocity and shape. Likewise, the measurement noise covariance $R$ was also assumed to be uncorrelated with differing parameters for position and shape. The process noise parameters were empirically set to $\sigma_p^Q = 0.5$, and $\sigma_v^Q = 0.75$ and $\sigma_s^Q = 1$ for position, velocity and shape, respectively. The measurement noise parameters were empirically set to $\sigma_p^R = 0.55$, and $\sigma_s^R = 0.1$ for position, and shape, respectively. The sensor characteristics and accuracies were used to determine these parameters, and they are not expected to change significantly unless different sensors are used.

Two fundamental issues with tracking by detection arise in track/observation association and track management. To deal with the the association problem, the appearance descriptors are leveraged. In each frame, a decision has to be made about which observation to match to which tracker. This is solved by computing a dissimilarity score between each track and observation. The dissimilarity score used is

$$D(tr, obs) = D_p(tr, obs) + \alpha D_s(tr, obs) + \beta D_a(tr, obs) \tag{6}$$

$$D_p(tr, obs) = \sqrt{(obs_p - tr_p)^T \Sigma_p (obs_p - tr_p)} \tag{7}$$

$$D_s(tr, obs) = \sqrt{(obs_s - tr_s)^T \Sigma_s (obs_s - tr_s)} \tag{8}$$

$$D_a(tr, obs) = || \log tr_a - \log obs_a ||_F \tag{9}$$

This can be summarized as a weighted combination of the appearance descriptor dissimilarity and the Mahalanobis distances between the 2D position and shape parameters of the tracker and observation. The covariances for the Mahalanobis distances, $\Sigma_{pos}$ and $\Sigma_{shape}$, come from the covariance in the Kalman filter. The parameters $\alpha = 0.5$ and $\beta = 0.02$ were set empirically and used to shift the scale of each term so that none of the terms overpowered the others. Given the classroom space, the spatial distances tend to be in the range of 0–1 m, at least for matches that have any semblance of being a match. The $\alpha$ and $\beta$ parameters are set to bring the shape and appearance distances into this scale, based on the distribution of values observed in practice. Dissimilarity scores that exceed a threshold of $T_d = 2$ are removed from consideration, which was determined empirically to be a good cutoff value for never matching.

The associations between tracks and observations are then made in a greedy fashion by iteratively matching the lowest scoring track/observation pair until there are no more free tracks or observations to match. The greedy approach is not guaranteed to yield an optimal association; experiments were performed using the Munkres algorithm [26], although in practice the greedy method was sufficient.

Track management is handled by computing a track score based on a log-likelihood ratio of whether a track is a true target or a false alarm [3]. This track score is used to determine when tentative tracks (tracks created from unmatched observations) become confirmed tracks (those tracks which are considered to correspond to a true target), and when tracks should be deleted. The score is computed by recursively summing log-likelihood contributions from the process model and sensor model, or a reduction if a track is not matched on a particular timestep. The initial track score for a track is a function of position: if tracks are formed near the entrances/exits to the room, then the track has a far higher initial track score than if the track appears elsewhere.

Given the nature of the classroom, occupants are frequently in very close proximity, or even touching, each other. This can lead to one detection consisting of several merged detections, which can starve tracks of observations causing them to be eliminated. This is overcome by a mechanism to split apart these merged detections. Following the intuition that an established track will not vanish, several criteria were created to determine if a single detection consists of several merged objects:

$$D(tr_m, obs) - D(tr_u, obs) < T_s \tag{10}$$

$$D(tr_u, obs) < T_d, \tag{11}$$

where $tr_u$ is an unmatched tracker (a tracker without a matched observation), and $tr_m$ is the tracker that is matched with $obs$ which is the observation with the smallest distance to $tr_u$. That is, an observation $obs$ is considered to consist of multiple merged detections if it is the closest match to an unmatched tracker and its distance is less than a threshold, and the distance between $obs$ and its matched tracker is less than a threshold. For the experiments presented here, $T_s = 0.35$ and $T_d = 2$. The $T_d$ parameter matches the previous cutoff for not considering a match, while the $T_s$ parameter is set such that unmatched tracks close to observations are allowed to split the observation, but not so large as to cause many false splits. The exact values were determined by empirically studying the data.

These criteria are also augmented with bounding volume information. For a tracker without a matched observation, the bounding volume from the previous matched observation is updated using the tracker's velocity information. This bounding volume is compared with the bounding volumes of the observation with the smallest distance. If this bounding volume overlap is over a threshold of 0.7, and the distance is below the $T_d$ distance threshold, the observation is considered to consist of merged detections. This allows an unmatched track that significantly overlaps an observation to split that observation (if the track's bounding volume overlaps by 70% or more, it is strong evidence of a merged detection).

If any observations are marked as consisting of multiple detections, they are then split and a new association is computed. Detections are split using $K$-means [16] on spatial distance. The number of clusters for $K$-means, $K$, is determined based upon the number of tracks that want to match to the observation in question. That is, one for the original track match plus one for each unmatched track that meets the split criteria.

## 4 Behavior analysis

The overall motivating factors for this work are twofold. First, the creation of a system that can aid mental health providers in screening for early signs of mental illnesses. Second, the creation of a platform for gathering data to aid in the discovery of new risk markers for the development of mental illness. A fundamental issue to consider is that the broad scope of this work raises the challenge of identifying which behaviors to measure that would be relevant to many different types of mental illness and not just one disease area. Another fundamental challenge is that the design of the system can limit the kind of data that can be generated; for instance, there is a need to balance the desire to capture micro-motor movements at high resolution versus the desire to capture broader, whole-area information at lower resolution. In the present work, the design favors a system that could optimally capture larger movements and interactions within the broader system, as a starting point for generating basic behavioral information that would be useful to mental health professionals.

Using the position and velocity information from the previous steps, three types of information regarding observed subjects are extracted:

1. an adult/child classification,
2. behavior related to individual activity levels, and
3. quantification of social interaction.

An important aspect for further information analysis is determining if a track belongs to an adult or a child. Given the developmental differences between adults and children, the height of a track can be a strong indicator. At preschool age, the average height of a child is about 1 m, depending upon exact age and gender [21]. For child/adult classification purposes, height is determined by using the Z component of the bounding box around a detection. For each matched detection in a track, the height is computed and is thresholded against 1.3 m. If the number of matched observations that are 1.3 m or taller is above 25%, then the track is determined to be an adult, otherwise a child. This approach was very effective in the data collected from the Shirley G. Moore Lab School.

Quantifying activity level is a straightforward process. Since the Kalman filter itself estimates velocities, that information can be used directly. The estimated velocities over all observations can be used to form a set of velocities, and taking the median velocity provides a robust estimate of the activity level of a child. This information represents a measure that our system can generate that would be all but impossible to generate for traditional non-intrusive observational studies.

Since the positions of detected occupants is known, it is possible to define a measure to estimate social relationship based on proximity. Previous psychiatric studies such as [23,33] used similar measures; however, proximity was approximated due to manually coded observational studies. Based on the proximity metrics, a social graph can be computed which can provide valuable information about the relationships between occupants in the classroom. The approach employed here is similar to [5] in that a connectivity score is assigned between two people at a given time through a function, $f_{\text{group}}$, which is a measure that incorporates relative position between two subjects $i$ and $j$ as well as the headings of each subject:

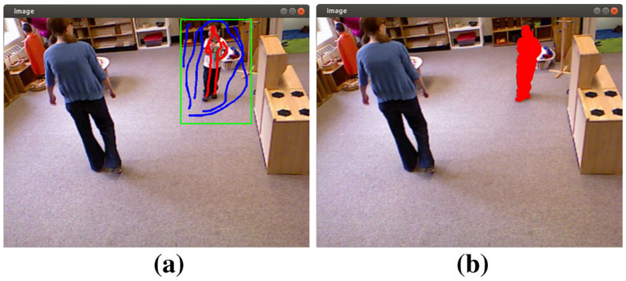
$$f_{\text{group}}(i, j) = \exp\left(-\frac{1}{2}\mu_d^T (C_i^T C_j)^{-\frac{1}{2}} \mu_d\right), \tag{12}$$

where $\mu_d$ is the difference between the centroid locations of the Kalman tracker for person $i$ and person $j$ and $C_i$ is defined in Eq. (13).

To incorporate heading information, a 2D oriented Gaussian kernel is used. The covariance of the Gaussian kernel is defined as

$$C_i = \begin{bmatrix} \frac{\cos\theta_i^2}{\sigma_x^2} + \frac{\sin\theta_i^2}{\sigma_y^2} & \frac{-\sin 2\theta_i}{2\sigma_x^2} + \frac{\sin 2\theta_i}{2\sigma_y^2} \\ \frac{-\sin 2\theta_i}{2\sigma_x^2} + \frac{\sin 2\theta_i}{2\sigma_y^2} & \frac{\sin\theta_i^2}{\sigma_x^2} + \frac{\cos\theta_i^2}{\sigma_y^2} \end{bmatrix}. \tag{13}$$

The parameters $\sigma_x$ and $\sigma_y$ were set such that they incorporate the assumption that people in a group tend to walk side by side. A corresponding kernel w.r.t. $\theta_j$ was also computed, with a covariance $C_j$ as defined by Eq. (13). These covariances were blended together to form the kernel shown in Eq. (12).



(a) (b)

**Fig. 5** Depiction of the ground truth generation process. The program allows the user to draw a box around the intended subject, then mark the foreground and background pixels. The interactive segmentation algorithm can then generate a segmentation mask which can be further refined. **a** Hand labeling. **b** Ground truth mask

The instantaneous similarity score between any two individuals in the observed scene is accumulated over time to yield a pairwise affinity score between them. This yields an $n \times n$ affinity matrix $W$ representing the interactions between all the observed people. The affinity matrix is then normalized such that each row sums to 1, yielding an asymmetric affinity $W_{\text{asym}}$. The $i$th row in $W_{\text{asym}}$ represents the distribution of interaction individual $i$ has with everyone else in the scene (including themselves). The symmetrized version of this matrix, $W_{\text{sym}}$, is used to generate a graph embedding, where each individual is a node and the edges represent the interactivity between them.

## 5 Results

Data were collected over the course of several days at the Shirley G. Moore Lab School, and broken down into recorded segments, each 4–5 min in length. The results from three such consecutive recordings are reported here: the first recording contains 7853 frames, the second contains 8093 frames, and the third contains 6673. More recordings were taken and have been processed, but the results presented here are only on the sequences with corresponding ground truth data.

A labeling program was created that allows a user to select a region of interest in an RGB image and interactively segment an individual, creating a segmentation mask that gives a pixel-level identity association. This labeling process is depicted in Fig. 5, and the implementation is based upon the OpenCV[1] implementation of GrabCut [31]. Since the RGB and depth images are registered, the segmentation masks in the color images can also be applied to the depth maps. By masking out the depth maps, a point cloud can be created where every point belongs to an individual and that individual's identity is associated with the point (a ground truth pointcloud). Despite having an interactive segmenta-

[1] https://opencv.org/.

**Table 1** Summary of performance for current methods

| | | | |
|---|---|---|---|
| (a) Recording 1 | | | |
| Performance per sensor | | | |
| Sensor | $F_1$ (%) | Precision (%) | Recall (%) |
| 1 | 83.1 | 73.9 | 95.1 |
| 2 | 91.2 | 83.8 | 100 |
| 3 | 86.9 | 86.7 | 87.1 |
| 4 | 82.5 | 93.0 | 74.1 |
| 5 | 83.7 | 76.6 | 92.1 |
| Tracking performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 5.63 cm | 90.0 | 1.51 | 1.81 |
| (b) Recording 2 | | | |
| Performance per sensor | | | |
| Sensor | $F_1$ (%) | Precision (%) | Recall (%) |
| 1 | 83.7 | 77.4 | 91.1 |
| 2 | 89.1 | 84.0 | 94.7 |
| 3 | 68.8 | 73.3 | 64.7 |
| 4 | 85.9 | 84.9 | 86.9 |
| 5 | 89.9 | 84.8 | 95.7 |
| Tracking performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 4.60 cm | 96.5 | 0.5 | 0.99 |
| (c) Recording 3 | | | |
| Performance per sensor | | | |
| Sensor | $F_1$ (%) | Precision (%) | Recall (%) |
| 1 | 68.5 | 55.4 | 89.9 |
| 2 | 69.2 | 66.7 | 72.0 |
| 3 | 94.0 | 91.9 | 96.1 |
| 4 | 66.7 | 58.3 | 77.8 |
| 5 | 97.8 | 97.8 | 97.8 |
| Tracking performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 4.44 cm | 98.0 | 0 | 1.0 |

tion method for labeling, this process is still time and labor intensive so only every 60 frames were labeled and currently only 3 recorded segments are labeled.

The detection portion of the pipeline can be validated by reprojecting 3D points belonging to a person back to the image plane, creating a computed segmentation mask. The computed and ground truth segmentation masks can then compared by computing their bounding boxes and examining the area of overlap. This is inline with the PASCAL Visual Object Classes (VOC) Challenge segmentation task [10], where the area of overlap must exceed 50% for a true positive (TP). False positives (FP) occur when computed masks have no matching ground truth masks and false negatives (FN) when ground truth masks are not matched with computed masks. These can then be used to compute the standard values of precision, recall, and $F_1$ (the harmonic mean of precision, and recall).

To evaluate tracking performance, the CLEAR MOT metrics MOTA and MOTP are used [19]. MOTP, multiple object tracking precision, measures the total error in estimated position averaged by the number of matches, and indicates the ability of the tracker to estimate precise objection positions. MOTA, multiple object tracking accuracy, accounts for all object configuration errors made by the tracker over all frames, and gives an intuitive measure of the tracker's performance. Additionally, miss rate and false positive rate are computed.

The results are summarized in Table 1 and Figs. 6, 7, 8, and 9. One thing to note in the performance table is that, despite the per sensor recall rates, the miss rate for the tracking in each recording is quite low. This illustrates the advantage of multiple sensors, as a detection may be missed in several sensors but it is likely at least one sensor will have a positive detection. This is an important aspect of this system, since low recall can cast doubt on the ability of the system to accurately monitor behavior.

In additional to performance on the previously described system, results from previous iterations of this system are presented to provide a point of comparison. For these comparisons, the methods for generating detections were adjusted. In the first method, referred to as Euclidean, detections are formed from the background subtracted global point cloud by performing clustering on Euclidean distance between points: any points that are within an epsilon Euclidean distance are considered part of the same object. Additionally, to test the idea of hierarchically computing supervoxels and then clustering supervoxels into detections, we present results using the hierarchical graph-based (HGB) approach described in previous work [12]. The results of these comparisons are presented in Table 2. Performance results were computed across the same three recordings presented in Table 1; however, the results were averaged together to provide a single result for each method.

Figure 6 shows contour plots that depict the placement of the sensors with respect to the activity of the classroom where red denotes areas of high activity and blue represents areas of low activity. The areas of high activity in recordings 1 and 3 end up occurring near the table seen in Figs. 3a–d; the table ends up being a focus of activity for many of the classroom occupants. In Fig. 6d, the two contour regions branching out from the table area are entrances/exits to the classroom area. The branch off to the left side of the diagram leads to another area of the classroom while the branch that

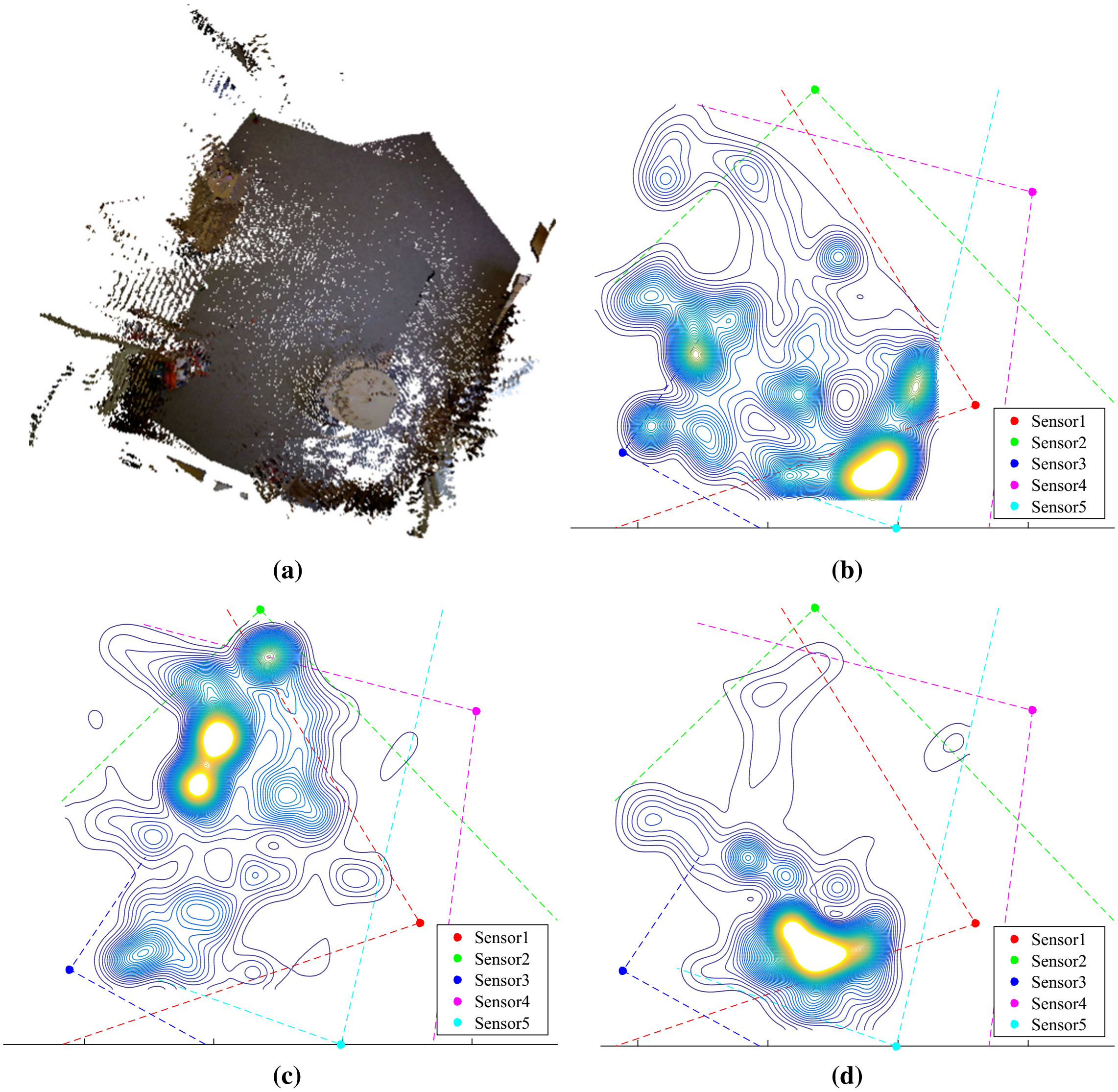


**Fig. 6** Sensor placement and activity levels. **a** gives a layout of the ground of the classroom, which corresponds approximately to the activity level contours. The contours denote regions of activity with blue corresponding to the lowest activity and red the highest. The large dots represent sensor locations and the dashed lines show their fields of view. **a** Ground plane projection, **b** recording 1, **c** recording 2, and **d** recording 3

heads to sensor 2 ends in a door to the playground. The first recording contains a number of children playing further back in the classroom who then venture into the area of recording; this recording contains nine distinct children and three adults. The second recording contains nine distinct children and four adults. The majority of the third sequence consists of several children entering the room and heading over to the table to play, eventually being joined by an adult who sits at the table. The third recording has four distinct children and two adults (although one child and one adult are only visible for a few seconds as they pass through the room).

Example trajectories of ten occupants are shown in Fig. 9, with the trajectories being color coded such that blue represents where the track begins and red where the track ends. The trajectories belong to individuals who are deemed to have a high relationship in Fig. 7. Figure 9i depicts the trajectory of a child who enters from the playground and then sits at the table to play, while Figure 9h shows the trajectory of an adult who enters the room and sits at the table to play with Child 1. Figure 9b belongs to a child who enters with an orange cloak from a dress-up area, and he plays in the area around the table. Adult 5 from recording 1, depicted in Fig. 9a, enters the room and interacts with Child 11, eventually removing the cloak. The tracking system is able to maintain the identity of Child 11 through this de-cloaking. Figure 9e, f also depicts another interesting interaction, where Adult 2 enters from further into

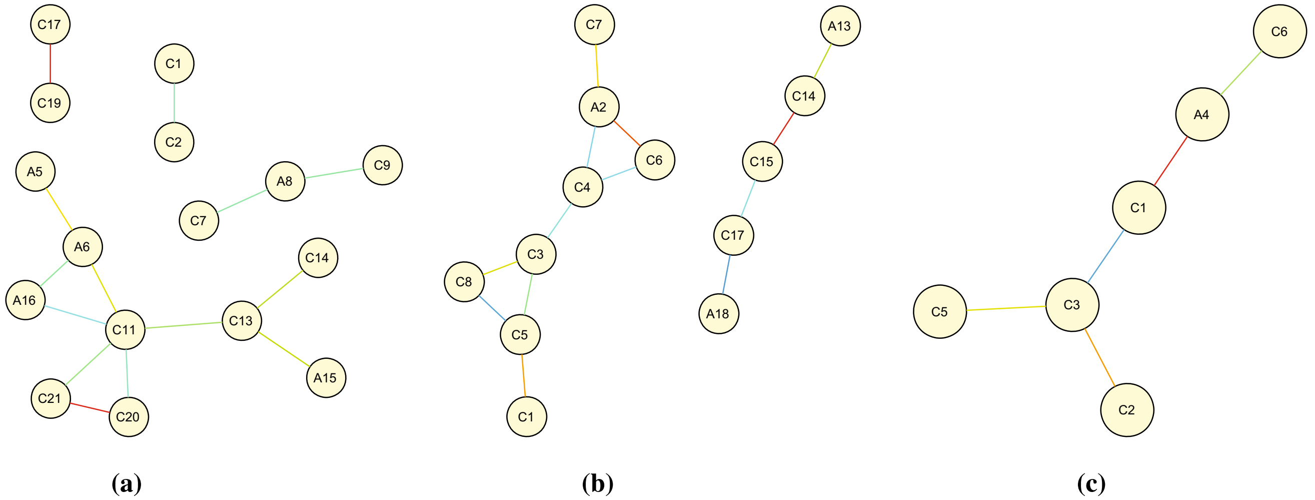


**Fig. 7** Graph embeddings from the symmetric affinity matrix $W_{\mathrm{sym}}$ depicting the interaction between different individuals. Prefix C indicates child, A indicates adult). **a** Recording 1, **b** Recording 2, **c** Recording 3

the classroom and walks up to Child 6 then kneels down and hugs the child. The identify of both occupants is successfully maintained throughout this interaction.

Figure 7 depicts the social proximity measures discussed in Sect. 4 for all three recordings. For the graph embedding, nodes without edges have been omitted and additionally some edges have been filtered to remove weak social relationships. Of note, for recording 2, there is a strong link between Child 1, who enters the room and sits down at the table to play and Adult 1 who enters the room and sits at the table to play with Child 1. Figure 8 depicts the median velocity of the tracked children. Due to the number of children in some recordings, the median velocity graphs were limited to the seven most observed children. Since the recordings were taken during a free play session, the children in the classroom are playing in an unstructured environment, and their level of activity can vary drastically, frequently leading to large fluctuations in velocity. In particular, note that Child 1 in recording 3 has a small mean absolute deviation, since the child spends much of the observed time sitting at the table.

The quantitative evaluation of system performance is summarized in Table 1. Per-sensor performance tends to vary between sensors, which is generally related to the sensors position with respect to the activity in the room. Frequently, activity may be occurring either at the edge of a sensor, or far away from it. In general, the recall rates for each sensor are good, which is important as the system should prioritize not missing observations in order to provide accurate behavioral trends. Additionally, the benefit of combining information from multiple sensors is shown in the tracking performance, where all metrics are quite good. The MOTA is brought down a bit because of misidentification errors, which generally occur when an occupant leaves the room and returns.

Comparing the different detection method performances in Table 2, it is clear that the current method performs the best. The two previous detection methods frequently suffer from a large number of false positives, which bring the MOTA down. In additional to better performance, the average frame processing time is significantly lower for the current method. The Euclidean method takes an average of 2.3 seconds across all processed frames, while the HGB method takes 0.84 seconds. The current method is the fastest at 0.54s per frame.

## 6 Conclusions and future work

This paper presented a multi-stage, multi-component system for tracking children in a challenging classroom setting. This processing pipeline enables us to quantify child-child interactions, child-caregiver interactions, and child activity levels in an automated fashion. The processing pipeline has been designed such that each component is modular and can be updated as better methods become available. This system has been validated on data collected from an unstructured real-world environment, and the results have been promising.

The validation of this system indicates that our methodology for detecting and tracking room occupants works well in the challenging environment. The labeled data contain segments where the room lighting changes drastically, and also difficult scenarios where children add or remove clothing during a dress-up game. There are also situations where two individuals come into contact (hug) and then separate, which the system is able to successfully track and maintain identity. The system is able to deal with children who are sitting, standing, walking, running, or even laying on the ground.

While the work here has yet to achieve our ultimate goal of automatically identifying disease-specific neurobehavioral

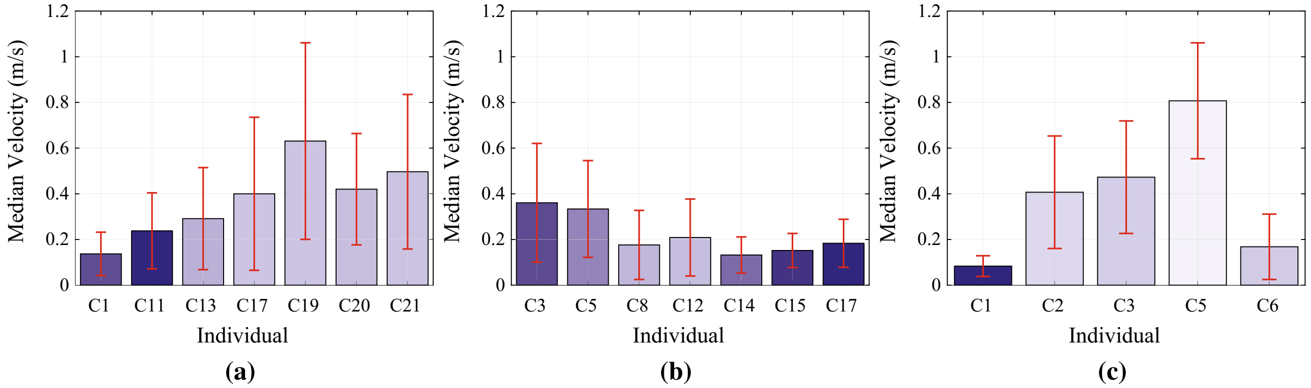


**Fig. 8** Median velocities tracked children. Transparency indicates relative amount of time each child is observed; darker bars were observed for longer. The error-bars show the median absolute deviation. For recordings with a larger number of children, only the top 7 longest observed children are shown. **a** Recording 1, **b** recording 2, and **c** recording 3

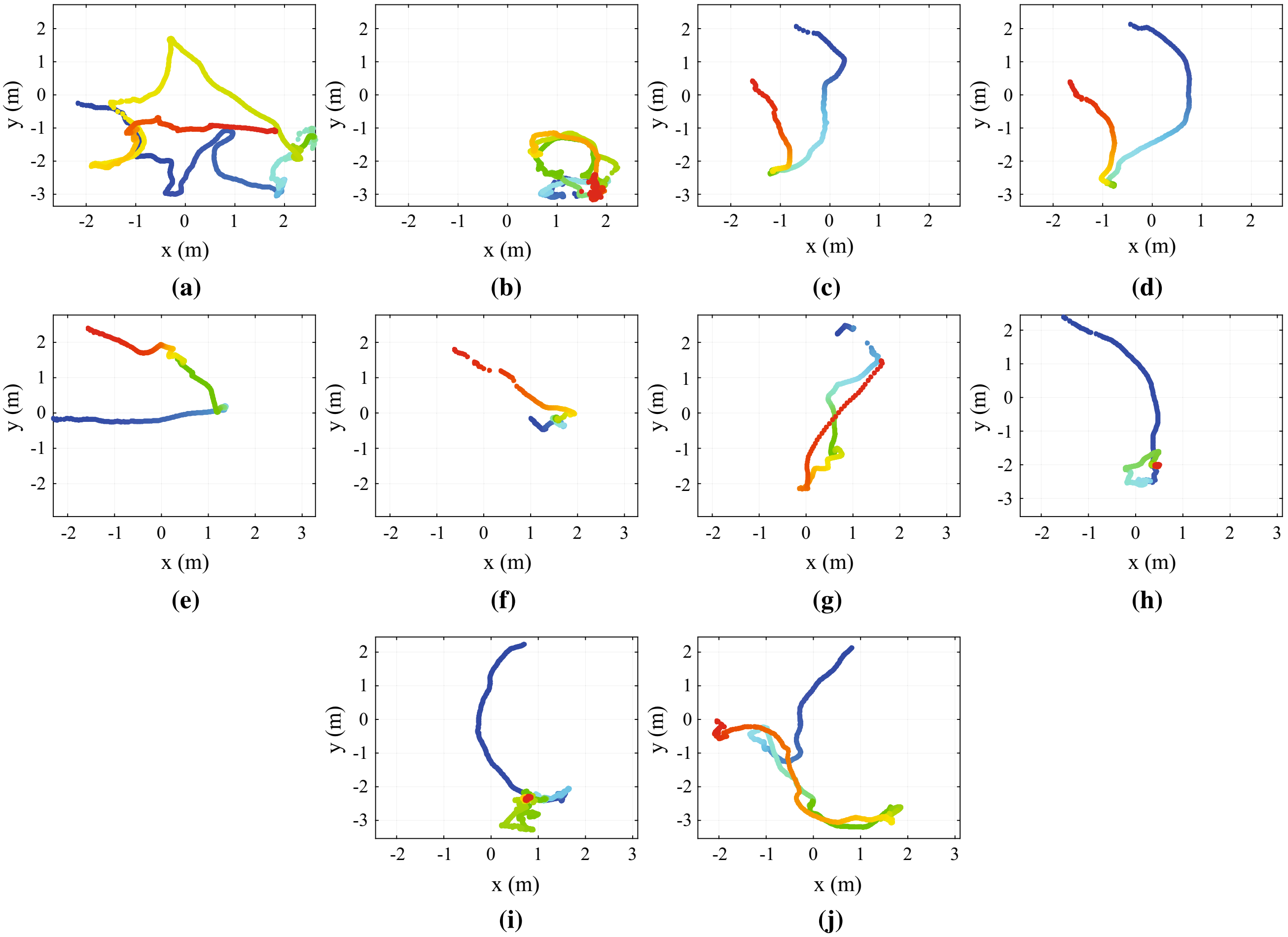


**Fig. 9** Top-down view of tracking results on several individuals. Tracks begin as blue dots and transition to red by the end. Each track is displayed for only one individual at a time, to disambiguate the individuals trajectory from all of the other activity going on in the room. **a** Recording 1 Adult 5, **b** Recording 1 Child 11, **c** Recording 1 Child 17, **d** Recording 1 Child 19, **e** Recording 2 Adult 2, **f** Recording 2 Child 6, **g** Recording 2 Child 4, **h** Recording 3 Adult 4, **i** Recording 3 Child 1, **j** Recording 3 Child 3

**Table 2** Summary of performance across different detection methods

| (a) Euclidean | | | |
|---|---|---|---|
| Performance per sensor | | | |
| Sensor | $F_1$ (%) | Precision (%) | Recall (%) |
| 1 | 69.8 | 56.6 | 91.6 |
| 2 | 81.3 | 74.4 | 90.0 |
| 3 | 77.5 | 71.9 | 84.3 |
| 4 | 81.0 | 83.4 | 79.6 |
| 5 | 65.2 | 50.6 | 95.6 |
| Tracking Performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 7.55cm | 41.5 | 1.4 | 50.2 |
| (b) HGB | | | |
| Performance per sensor | | | |
| Sensor | $F_1$ (%) | Precision (%) | Recall (%) |
| 1 | 62.3 | 47.9 | 90.1 |
| 2 | 67.9 | 55.3 | 88.9 |
| 3 | 70.3 | 64.4 | 78.7 |
| 4 | 58.5 | 49.4 | 73.5 |
| 5 | 71.4 | 59.9 | 89.0 |
| Tracking performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 8.54 cm | 13.1 | 7.9 | 54.5 |
| (c) HVB | | | |
| Performance per sensor | | | |
| Sensor | $F_1$ | Precision | Recall |
| 1 | 78.4 | 68.9 | 92.0 |
| 2 | 83.2 | 78.2 | 88.9 |
| 3 | 90.5 | 86.4 | 95.2 |
| 4 | 83.2 | 84.0 | 82.6 |
| 5 | 78.4 | 78.7 | 79.6 |
| Tracking performance | | | |
| MOTP | MOTA (%) | Miss Rate (%) | FP rate (%) |
| 4.89 cm | 94.8 | 0.7 | 1.3 |

predictors for early intervention, the system is able to capture basic, observable characteristics of children behavior that are relevant to mental health. There remains a number of important tasks to accomplish in order to reach the our goal. The work presented here treats each recording independently, so tracks belonging to the same individual from different recordings are not identified in any way. This re-identification task is an important step in the long term processing of this system, and represents the next important advancement. With the ability to re-identify previously tracked occupants, this would allow the system to combine data across several days or weeks. One solution to this would be perform manual re-identification; with a small number of recordings (around 10 or 20), the number of tracks is small enough that the amount of effort to perform manual re-identification is reasonable. However, an automated approach to this is desired. The appearance descriptors presented for tracking could be useful in re-identifying occupants, however appearance descriptors may not work across days as clothing changes.

Once data have been accumulated over a long period of time, a better means of representing the data to the user is also necessary. As part of this, work has begun on creating an interactive visualization tool that display a summary of social relationship to the user via a social graph. The user can interact with this social graph by adjusting the display and layout, but can also click on the nodes (representing individuals) and edges (representing a summary of the relationship between two individuals) in order to plot information. Plots are broken out over each recording the individuals are present in and the social relationship is graphed for each frame. Similar plots can be shown for velocity. Each of these plots also allows the user to click and define a time window and then watch the video at that time from different perspectives.

Additionally, in the future the output of this system could be used to as input to other behavioral analysis modules, such as activity recognition. The objects detected in 3D could be used to generate a bounding box on the 2D image, which could be used to constrain the video to a particular region for activity recognition. Additionally, information about social relationships could be used to determine when to detect for certain activities.

**Acknowledgements** This material is based upon work supported by the National Science Foundation through Grants IIP-0443945, CNS-0821474, IIP-0934327, CNS-1039741, SMA-1028076, and CNS-1338042.

## References

1. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magn. Reson. Med. **56**(2), 411–421 (2006)
2. Barkley, R.A.: The ecological validity of laboratory and analogue assessment methods of adhd symptoms. J. Abnorm. Child Psychol. **19**(2), 149–178 (1991)
3. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House, Norwood (1999)
4. Bowlby, J.: A Secure Base: Clinical Applications of Attachment Theory. Taylor & Francis, New York (2005)
5. Chang, M.C., Krahnstoever, N., Ge, W.: Probabilistic group-level motion analysis and scenario recognition. In: IEEE International Conference on Computer Vision, 2011, pp. 747–754 (2011)
6. Cherian, A., Sra, S., Banerjee, A., Papanikolopoulos, N.: Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices. IEEE Trans. Pattern Anal. Mach. Intell. **35**(9), 2161–2174 (2013)

7. Elicker, J., Ruprecht, K.M., Anderson, T.: Observing infants and toddlers relationships and interactions in group care. In: Lived Spaces of Infant-Toddler Education and Care, pp. 131–145. Springer (2014)
8. Esposito, G., Venuti, P., Apicella, F., Muratori, F.: Analysis of unsupported gait in toddlers with autism. Brain Dev **33**(5), 367–373 (2011)
9. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining, vol. 96, pp. 226–231. AAAI (1996)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
11. Fasching, J., Walczak, N., Morellas, V., Papanikolopoulos, N.: Classification of motor stereotypies in video. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4894–4900. IEEE (2015)
12. Fasching, J., Walczak, N., Sivalingam, R., Cullen, K., Murphy, B., Sapiro, G., Morellas, V., Papanikolopoulos, N.: Detecting risk-markers in children in a preschool classroom. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1010–1016. IEEE (2012)
13. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. Int. J. Comput. Vis. (IJCV) **59**(2), 167–181 (2004)
14. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 267–282 (2008)
15. Focken, D., Stiefelhagen, R.: Towards vision-based 3-D people tracking in a smart room. In: IEEE International Conference on Multimodal Interfaces, pp. 400–405. IEEE (2002)
16. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer Series in Statistics, vol. 1. Springer, Berlin (2001)
17. Garvey, C.: Play. Harvard University Press, Cambridge (1990)
18. Hashemi, J., Tepper, M., Spina, T.V., Esler, A., Morellas, V., Papanikolopoulos, N., Egger, H., Dawson, G., Sapiro, G.: Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. Autism Research and Treatment (2014)
19. Keni, B., Rainer, S.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J. Image Video Process (2008)
20. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S.: Multi-camera multi-person tracking for easyliving. In: IEEE International Workshop on Visual Surveillance, pp. 3–10. IEEE (2000)
21. Kuczmarski, R.J., Ogden, C.L., Guo, S.S., Grummer-Strawn, L.M., Flegal, K.M., Mei, Z., Wei, R., Curtin, L.R., Roche, A.F., Johnson, C.L.: 2000 CDC growth charts for the united states: methods and development. Vital and Health Statistics. Series 11, Data from the National Health Survey pp. 1–190 (2002)
22. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. (IJCV) **64**(2–3), 107–123 (2005)
23. Legendre, A., Munchenbach, D.: Two-to-three-year-old children's interactions with peers in child-care centres: effects of spatial distance to caregivers. Infant Behav. Dev. **34**, 111–125 (2011)
24. Mittal, A., Davis, L.S.: $M_2$Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. Int. J. Comput. Vis. (IJCV) **51**(3), 189–203 (2003)
25. Munaro, M., Menegatti, E.: Fast RGB-D people tracking for service robots. Auton. Robots **37**(3), 227–242 (2014)
26. Munkres, J.: Algorithms for the assignment and transportation problems. J. Soc. Ind. Appl. Math. **5**(1), 32–38 (1957)
27. Papon, J., Abramov, A., Schoeler, M., Worgotter, F.: Voxel cloud connectivity segmentation-supervoxels for point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2027–2034. IEEE (2013)
28. Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 728–735 (2006)
29. Rajagopalan, S.S., Dhall, A., Goecke, R.: Self-stimulatory behaviours in the wild for autism diagnosis. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 755–761. IEEE (2013)
30. Rehg, J.M., Abowd, G.D., Rozga, A., Romero, M., Clements, M.A., Sclaroff, S., Essa, I., Ousley, O.Y., Li, Y., Kim, C., et al.: Decoding children's social behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3414–3421. IEEE (2013)
31. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive foreground extraction using iterated graph cuts. In: ACM Transactions on Graphics (TOG), vol. 23, pp. 309–314. ACM (2004)
32. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (FPFH) for 3D registration. In: IEEE International Conference on Robotics and Automation, pp. 3212–3217. IEEE (2009)
33. Torrens, P.M., Griffin, W.A.: Exploring the micro-social geography of childrens interactions in preschool: a long-term observational study and analysis using geographic information technologies. Environ. Behav. **45**(5), 584–614 (2013)
34. Tuzel, O., Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: Lecture Notes in Computer Science, vol. 3952, pp. 589–600. Springer (2006)
35. Walczak, N., Fasching, J., Toczyski, W.D., Morellas, V., Sapiro, G., Papanikolopoulos, N.: Locating occupants in preschool classrooms using a multiple RGB-D sensor system. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2166–2172. IEEE (2013)
36. World Health Organization: The World Health Report 2004: Changing History, Annex Table 3: Burden of Disease in DALYs by Cause, Sex, and Mortality Stratum in WHO Regions, Estimates for 2002. WHO, Geneva (2004)

**Nicholas Walczak** is a Research Scientist at Raytheon BBN Technologies. He received his PhD from the Department of Computer Science and Engineering at the University of Minnesota in 2017. The focus of Dr. Walczak's research is computer vision, particularly working with point cloud data. His past work has involved computer vision for behavior analysis, both using point cloud data and also purely image-based approaches.

**Joshua Fasching** received his bachelor of science degree in software engineering from the Milwaukee School of Engineering in 2009. He received the PhD degree in computer science from the University of Minnesota in 2017. He is a research scientist at Raytheon BBN Technologies in Massachusetts. His research interests include activity recognition, machine learning and virtual reality.

**Kathryn Cullen** is an Associate Professor and Director of the Child and Adolescent Psychiatry at the University of Minnesota Medical School in Minneapolis. She received a bachelor of arts degree from the University of Chicago. She completed medical school, psychiatry and child and adolescent psychiatry clinical training, and postdoctoral research training at the University of Minnesota. Dr. Cullen's research, which focuses on adolescent depression and related problems such as self-injury, is funded by the National Institute of Health among other sources.

**Vassilios Morellas** received his diploma degree in Mechanical Engineering from the National Technical University of Athens, Greece, his MSME degree from Columbia University, NY, and his PhD degree from the department of Mechanical Engineering at the University of Minnesota. He is Program Director in the department of Computer Science & Engineering and Executive Director of the NSF Center for Robots and Sensors for the Human Well-Being. His research interests are in the area of geometric image processing, machine learning, robotics and sensor integration.

**Nikolaos Papanikolopoulos** (IEEE Fellow) received his Diploma of Engineering in Electrical and Computer Engineering, from the National Technical University of Athens in 1987. He received his M.S. in 1988 and PhD in 1992 in Electrical and Computer Engineering from Carnegie Mellon University. His research interests include robotics, computer vision, sensors for transportation and precision agriculture applications, and control systems. He is the McKnight Presidential Endowed Professor at the University of Minnesota and has received numerous awards including the 2016 IEEE RAS George Saridis Leadership Award in Robotics and Automation.