2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Locating Occupants in Preschool Classrooms Using a Multiple RGB-D Sensor System

Nicholas Walczak [*§], Joshua Fasching [*], William D. Toczyski [*],
Vassilios Morellas [*], Guillermo Sapiro [††] and Nikolaos Papanikolopoulos [*§]

***Abstract*— Presented are results demonstrating that, in developing a system with its first objective being the sustained detection of adults and young children as they move and interact in a normal preschool setting, the direct application of the straightforward RGB-D innovations presented here significantly outperforms even far more algorithmically advanced methods relying solely on images. The use of multiple RGB-D sensors by this project for depth-aware object localization economically resolves numerous issues regularly frustrating earlier vision-only detection and human surveillance methods, issues such as occlusions, illumination changes, unexpected postures, atypical morphologies, erratic or unanticipated motions, reflections, and misleading textures and colorations.**

This multiple RGB-D installation forms the front-end for a multi-step pipeline, the first portion of which seeks to isolate, *in situ*, 3D renderings of classroom occupants sufficient for a later analysis of their behaviors and interactions. Towards this end, a voxel-based approach to foreground/background separation and an effective adaptation of supervoxel clustering for 3D were developed, and 3D and image-only methods were tested and compared. The project's setting is highly challenging, but then so are its longer term goals: the automated detection of early childhood precursors, ofttimes very subtle, to a number of increasingly common developmental disorders.

I. Introduction

Mental health disorders—including developmental forms with symptoms recognizable even in very early childhood—account for fully one-fourth of all years of productive life lost due to disability and premature mortality [1]. Autism lies at the core of one such group of developmental disorders. Signs for autism spectrum disorders (ASD) appear early (before age three), their consequences can be devastating, but early diagnosis and intervention can also markedly mitigate the worst effects. Recognizing both the value of early treatment and autism's rising incidence, the United States National Institute of Mental Health, as part of its strategic plan, is charting mental illness trajectories to develop policies on proper screening and intervention [2]. Many mental illness symptoms emerging in childhood and early adolescence are now known to be later stages of much earlier processes. Hence, psychiatric research is keenly interested in identifying and detecting such *risk markers* prior to the onset of actual symptoms that are reliable precursors for indicating elevated risks for specific mental illnesses. Such risk markers can consist of genetic, neural, behavioral, and/or social deviations.

[*] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, U.S.A.
[††] Department of Electrical and Computer Engineering, Duke University, Durham, NC, U.S.A.
[§] corresponding authors: walczak@cs.umn.edu (article)
npapas@cs.umn.edu (project)

To facilitate the screening of known risk markers and the discovery of new ones, a long-term project at the University of Minnesota is developing a system for preschool classrooms for automatically monitoring and analyzing child behaviors. Figure 1 shows a classroom at the Shirley G. Moore Laboratory School, where a system is currently installed. The implementation consists of multiple RGB-D sensors to track children for short intervals during their normal daily activities and compute proximity-based relationships.

Fig. 1. A pre-kindergarten classroom used in collecting system RGB-D data.

Figure 2 depicts an outline of the major steps in our processing pipeline. The system works by capturing RGB and depth images from multiple angles, and then generating pointclouds which are transformed into a unified frame of reference (*cf.* §III-A). Background points from this global pointcloud are removed (*cf.* §III-B) and the remaining points are clustered into the objects they represent (*cf.* §III-C). The resulting object detections can then be tracked over time and the resulting tracking information can be used to analyze behavior of occupants of the class room such as computing a measure of social relationship based upon proximity or looking at an average level of activity.

An overview of the whole system was presented along with initial experiments in [3]. Building on this, the focus here is on the first fundamental block of the processing pipeline, with the goal of providing accurate detections of occupants in the classroom. For this work, ground truth was created to permit quantitative evaluations, and our tests confirm the efficacy of the proposed approach (*cf.* §V). Updates to our system that fully leverage the 3D information to remove background points are compared against using solely image-based detection, as well as our previous image-based background subtraction method. Our results (*cf.* §V) show the purely image-based methods to be inferior. The 3D technique presented here (*cf.* §III) marks a significant improvement on our previous system, [3].

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

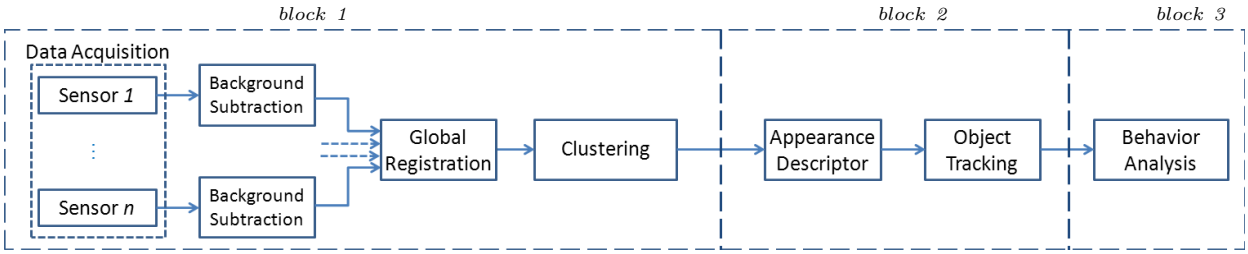


Fig. 2. The three-block, data processing pipeline for monitoring children in a preschool setting.

II. Related Work

There are now numerous working examples of multi-sensor tracking systems, though they generally employ only image-based technologies. An early such example is the multi-sensor "smart room" by Krumm, *et al.* [4], where image-based computer monitoring could control select features in a simulated living room (for example, in controlling the starting or pausing of a movie when viewers sat down or left the room). Using two calibrated stereo cameras to gather both color images and stereo-derived depth data across a temporal window of the thirty latest frames, the smart room was modeled and updated using the means and standard deviations for calculated pixel depths along with these same statistics for the individual *R*, *G* and *B* color channels. Foreground blobs became pixels differing from the room model and were fused into larger objects under the assumption that only people appeared in the foreground. While interesting, this early "smart room" approach was not extensible to cluttered and far more active, unstructured environments.

In [5], a multi-camera system utilizes an array of sixteen wide-baseline stereo cameras. It builds explicit models of people based on color and on a probability of "presence"; using the intersection of epipolar lines across multiple cameras to determine a person's three dimensional position. Its tracker uses these positions projected on to the ground plane, along with smoothness assumptions compatible with a Kalman filtering of trajectories. Under its rather rigid scene assumptions, [5] presents impressive results, but its assumptions of people as always upright, only moving smoothly, not jumping, *etc.* are invalid with respect to our preschoolers.

A more recent multi-camera system for tracking people, [6], likewise incorporates information from a large number of RGB cameras. Like similar systems, a computed ground plane homography maps image locations to a 2D ground plane XY location, with relevant foreground blobs computed from a proprietary image-based background subtraction method. These are then combined with a generative model to complete a quantized *occupancy grid* on the ground plane. Grid locations with sufficiently high degrees of occupancy by foreground blobs rank as positive person detections and get tracked across time. This is however also a method relying on purely *image*-based background subtraction for detecting people, an aspect that is performing poorly in our environment. Additionally, as in [5], key but still inappropriate scene assumptions (*e.g.* upright, walking adults) are made.

III. Methodology

A. RGB-D *Data Acquisition*

As even indirectly inferred depth information can augment a scene's description, many systems (just like those above) have extracted depth information from visual/stereo cues. Today however Microsoft Kinects™ are also available that can provide depths *directly*, at real-time rates of up to thirty frames per second, and while also automatically furnishing the corresponding RGB 640×480 images.

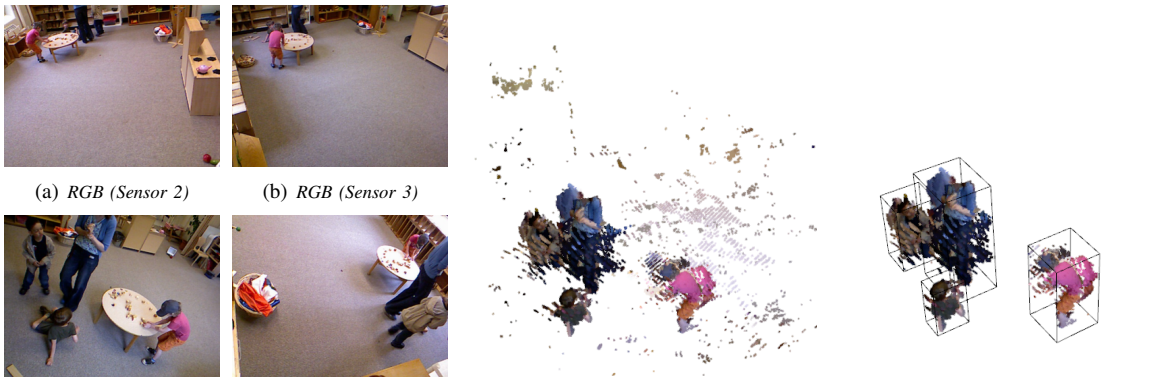
These inexpensive RGB-D devices provide for easy depth acquisition but they too have their limitations, including a restricted operational range of approximately 0.5-4.0m, and no hardwired synchronization capacities for coordinating multiple units. As a *structured light* based technology, distance is found by projecting infra-red patterns to examine their reflected deformations. As such, a projection from one sensor can interfere with an overlapping second sensor's projection, degrading the performance of both. Research in reducing this interference is progressing, *e.g.* [7], but for this application sensor placement keeps such overlap minimal. Moreover, areas with highest overlap, such as floors and tabletops, generally are background points which are ignored and later removed. Finally, given our sensor placements, our objects of interest (OOI), *i.e.* a room's occupants, tend to shield one sensor's structured light pattern from those of oppositely placed sensors, thereby further reducing any interference.

Each sensor's depth data together with its RGB frames can next create multi-featured *point clouds* using the intrinsic parameters of the Kinect. Points are projected into the cloud using

$$\mathbf{x}_c \;=\; \mathbf{K}^{-1}\boldsymbol{x}_c * d\,,$$

where $\mathbf{x}_c$ is a simple 3D point, $\mathbf{K}$ is the 3×3 camera matrix (intrinsic parameters), $\boldsymbol{x}_c$ is a homogeneous 2D camera point, and d is the scalar distance from the sensor.

Each sensor's resulting point cloud is relative to that sensor and so must be merged into a global frame of reference with all point clouds from all sensors aligned. This requires individual rotations and translations of point clouds, corresponding to the extrinsic parameters *R* (rotation) and *C* (camera centered translation) for each unit. The result is a single, unified global point cloud. Calibration, of both intrinsic and extrinsic parameters, employs a variant of the Gold Standard algorithm, [8]. For further details on our system calibration, see our fuller description in [9].



(a) *RGB (Sensor 2)* (b) *RGB (Sensor 3)* (c) *RGB (Sensor 4)* (d) *RGB (Sensor 5)* (e) *Noisy 3D reconstruction of scene* (f) *Detected persons*

Fig. 3. (a)-(d) Sample simultaneous RGB frames from four of five sensors. (e) The fused RGB and range data forming the (noisy) 3D point cloud rendering the scene. (f) The four persons detected and enclosed in four bounding boxes. [Compare (c), (e) and (f).]

B. 3D Background Subtraction

Each Kinect sensor generating up to thirty 640×480 depth maps per second implies raw point clouds totaling over 1.5 million points per time step, *i.e.* from each of the five sensors, 307,200 points per time step. Higher-order processing on such data volumes is not a practical possibility in this domain, and so a vast majority of these points must be removed. Fortunately, only a small fraction constitute points of interest, and identifying these as foreground is the focus of this subsection.

Image-based background subtraction is a thoroughly researched area and many methods exist (see [10] and [11] for recent surveys), but generally among the statistical approaches, each pixel's foreground probability is by some measure determined, and when above some (fixed or dynamic) threshold, promoted from the background. Optionally, an image-based foreground *mask* can also result. As applied to an RGB-D device's registered IMAGE+DEPTH frame pair, this mask can then dictate the corresponding depth pixels eligible for projection into a time step's global foreground point cloud; the end result being greatly reduced subsets of points, and clouds comprised mostly of points of subsequent higher-level interest (plus some noise).

A state of the art image-based background subtraction as described in [12] was tested here. Using local descriptions of texture surrounding a pixel (LBPs - local binary patterns) together with a photometrically invariant color measure a statistical model for each pixel is built as follows: Given a camera sequence of images at N successive time steps, $\{I^t\}_{t=1,\ldots,N}$, for each image pixel $\boldsymbol{x}$, its model M^t is defined as

$$M^t(\boldsymbol{x}) \;=\; \left\{\, \mathrm{K}^t(\boldsymbol{x})\,,\;\; \{m_k^t(\boldsymbol{x})\}_{k=1,\ldots,\mathrm{K}^t(\boldsymbol{x})}\,,\;\; \mathrm{B}^t(\boldsymbol{x}) \,\right\},$$

where $\mathrm{K}^t(\boldsymbol{x})$ is a scalar equaling the total number of $m_k^t(\boldsymbol{x})$ modes observed for pixel $\boldsymbol{x}$ as of time step t, and with the first $\mathrm{B}^t(\boldsymbol{x})$ modes identified as stable background modes.

Each mode for $\boldsymbol{x}$ constitutes a separate history of the pixel through the t^{th} time step and is defined as

$$m_k \;=\; \{\, I_k,\; \hat{I}_k,\; \check{I}_k,\; \mathit{LBP}_k,\; w_k,\; \hat{w}_k,\; L_k\},$$

where, for pixel $\boldsymbol{x}$ at time t, I_k $(\equiv I_k^t(\boldsymbol{x}), \mathit{etc.})$ is that mode's average RGB vector, $\hat{I}_k$ and $\check{I}_k$ are the estimated maximal and minimal RGB vectors thus far, LBP_k is the average local binary pattern, w_k denotes the probability of that mode being background, and $\hat{w}_k$ is the maximal value of w_k up to step k. L_k is the background layer number to which the mode belongs. $L_k = 0$ when this mode belongs to *no* stable background layer, and $\mathrm{B}^t(\boldsymbol{x}) = 0$ when the only background modes found are unstable. (And only background is multi-layered. Foreground is effectively a single layer.) Finally, the unified background model for an entire image I^t then becomes

$$\mathcal{M}^t = \{M^t(\boldsymbol{x})\}_{\forall \boldsymbol{x} \in I^t}\,.$$

Once the model is built, then after each update a background distance map, which is the complement of a foreground probability map, is created comparing the model and the current image. The map's elements are each pixel's distance to the "closest" background mode for that image pixel, or if all $L_k = 0$ and no stable backgrounds are present, distance is set above a foreground threshold. The distance equation used between a pixel and the modes occupying its map location, as well as the modal/model update algorithms, remain identical to those found in [12]. This method outperformed the other image-based background subtraction methods tested, so only its results will be discussed in Section IV. Poorer results, such as found using a Robust PCA-based method [13], were omitted due to space limitations.

There are several drawbacks to purely image-based approaches: those robust for complex scenes can rely on deep statistical modeling and extensive image analysis, which compromise real-time considerations. Techniques to speed processing—for example downsampling—compromise resolutions (and thus a key ancillary consideration, *i.e.* the close monitoring of child behavior). Large numbers of sensors still easily overwhelm overly computationally intensive methods. Cameras, of course, relay no explicit data on ***where*** the background lies, merely images, and so are profoundly affected by occlusions, appearance or lighting changes, reflections, *etc.*

In contrast, since our RGB-D system generates 3D point

clouds, background models that incorporate actual 3D locations are immediately available. A simple *computationally economical* approach first discretizes the 3D space into a regular lattice of 3D volumes $V_s = \{v_1, v_2, ..., v_n\}$ (commonly called voxels). To initialize our 3D classroom model with a room's contents, 3D data points are first recorded before either children or teachers arrive. Any 3D points found within a voxel then mark that voxel as occupied. For these, additional point features are then used to characterize their voxel-level features. Our classroom model then becomes this occupied subset of the voxel lattice.

Essentially, the process will partition the observed scene V_s into two disjoint spaces, the background/obstacle space V_b and the foreground/free space V_f, where $V_s = V_b \cup V_f$ and $\varnothing = V_b \cap V_f$. Later, when creating foreground point clouds during the school day, any points that lie in V_b can be efficiently removed using, for example, octrees [14].

In some situations points created by foreground objects may still occur within V_b. In this case, the appearance model for V_b is utilized. Each $v_i \in V_b$ maintains the corresponding ordered pair $\{\bar{x}_i, \mathbf{\Sigma}\}$ where a vector $\bar{x}_i$ holds the mean RGB values observed for the points in the initialized voxel and $\mathbf{\Sigma}$ holds the corresponding covariances. To determine if an input point q is background, its containing voxel is examined. If v_i encloses q, and if $v_i \in V_b$, then the input point color x_q is compared to the voxel color distribution using the Mahalanobis distance,

$$\text{dist}(x_q, \bar{x}_i) = \sqrt{(x_q - \bar{x}_i)^{\mathsf{T}} \mathbf{\Sigma}^{-1} (x_q - \bar{x}_i)}\,.$$

If the distance exceeds a certain threshold, then that input point is accepted as a foreground element. A resulting 3D point cloud of only 3D foreground points can then form, and then be grouped into higher structures for higher-level processing.

Once this global foreground cloud is isolated, its 3D points must be divided into meaningful sub-clouds. Earlier in the system's development, this clustering used a simple Euclidean metric, *cf.* [9] and [15]. Points less than an ϵ distance apart qualified as connected points and were grouped together as objects. This was straightforward, but it unrealistically presumed clear physical separations between the objects, and was clearly insufficient.

C. Point Cloud Clustering

Clustering is a very active research area. Even seemingly well-established approaches can be of quite recent origin and there are regularly new innovations. Some new, universally best approach however, seems unlikely to emerge—each will have its deficiencies. For instance, k-means (described in [16]), popular as a simple, efficient and well understood choice, requires the number of cluster centers to be assigned in advance, and so is ill-suited to where the cluster count (here, the number of people) is unknown and fluctuating.

Clustering can also be performed by using graph methods and selectively removing edges to create graph partitions where, with high weights indicating large differences, the weights of the final cuts between objects (*i.e.* between subgraphs) are maximized. This forms the so-called *max-cut problem*. Its approximate solution using an agglomorative approach [17] forms the basis of our method for clustering point cloud data. After first applying multiple filters to reduce noise (as detailed in [9]), an initial graph $G = (V, E)$ is defined by placing a connecting edge between each 3D point in the global point cloud and its η nearest neighbors. In practice, η was set to eight, a value found empirically. Vertices in the graph represent 3D feature points, and edge weights are computed as

$$w(v_i, v_j) \;=\; \|v_i - v_j\|_2 \qquad \forall\, e_{i,j} \in E$$

with

$$v_i \;=\; [\,x,\; y,\; z,\;\; r,\; g,\; b\,]^{\mathsf{T}}.$$

The graph-based segmentation then proceeds with a first pass using the efficient graph-based method proposed in [17]. The results of this segmentation are small clouds, on the order of ten to a few hundred points, referred to as supervoxels. Max-cut effectively minimizes intra-class while maximizing inter-class variances so these supervoxels are groupings in space having very similar local feature values, here color distributions (*cf.* Figure 4). Supervoxels and max-cut provided good point clusterings with no *a priori* assumptions such as expected cluster counts or pre-assigned blob sizes. It created small homogeneous clouds which can then be further clustered.

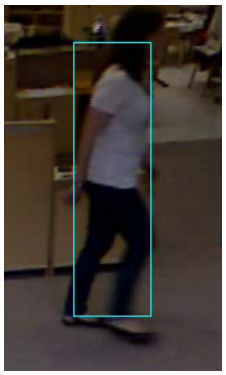
A second noise filtering process is then performed at the supervoxel level to remove isolated supervoxels under a minimal size, currently set to ten 3D feature points. Additional filtering using optical flow was tested (Bruhn's method, [18]), but detection of active children was not greatly improved while inactive children were sometimes missed. The surviving supervoxels are then re-segmented. This time, the graph-based method used edge weights between two supervoxels S_1 and S_2 defined—very conservatively, and using an all-to-all point comparison—as

$$\begin{aligned} v_i' \;&=\; [\,x,\; y,\; z\,]^{\mathsf{T}} \\ w(S_1, S_2) \;&=\; \max \|v_i' - v_j'\|_2, \quad \forall v_i' \in S_1,\; \forall v_j' \in S_2. \end{aligned}$$

Still viable resulting clusters then undergo a third and final round of filtering, again based on object point cloud dimensions and on assumed sizes for children and adults; a method first described for this system in [3]. The final foreground clusters are then considered as true objects of interest (OOI) for later stages in the pipeline.

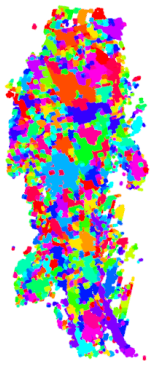
IV. Evaluation

Performance in the detection of objects by the above methods is next evaluated in two ways. The first is to project the OOIs' 3D points back on to image planes so as to create their 2D masks. Masked pixels indicate those pixels assigned as belonging to detected individuals. A comparison of these masks to hand-labeled ground truth masks can then broadly follow the performance detection measures of the PASCAL Visual Object Challenge (VOC) [19], where a bounding rectangle is computed for a detected image and



(a) *RGB (Sensor 1)*

(b) *Noisy 3D reconstruction*



(c) *Computed Supervoxels*

Fig. 4. Images illustrating supervoxel generation. For reference, (a) shows the detection bounding box from one of the five views, (b) is one view of the merged global 3D point cloud for this person, and (c) shows the computed 3D supervoxels visible from the same viewpoint. A connected color blob denotes a single supervoxel. (Colors may get reused however for displaying supervoxels that are fully separated and disconnected.)

then compared to a ground truth bounding rectangle. The overlapping area, a_o, of these two bounding boxes is then calculated to be

$$a_o \;=\; \frac{\text{area}\,(B_p \cap B_{\text{GT}})}{\text{area}\,(B_p \cup B_{\text{GT}})}\,,$$

where B_p is the predicted bounding rectangle and B_{GT} is the ground truth rectangle. If this a_o value exceeds 0.50, the detection is deemed a true positive (TP). The method estimates a detection rate for a single camera, but approximating both the mask and ground truth using only their enveloping projected rectangles is easily improved, and this measure was not intended for assessing multiple coordinated, multi-modal sensors.

A second evaluation recognizes that even a person missed by one sensor might be detected easily by others. In this evaluation, a 3D centroid for the detected person is derived from multiple inputs and the ground plane XY location of that centroid is compared to a ground truth XY location. A computed distance between the detected location and ground truth in the projected XY plane that is within a threshold constitutes a detection. This threshold can be varied to obtain a degree of how well the system localizes a person.

Both of the above metrics can be used to compute true positives, TPs (computed values matched to ground truth values), false positives, FPs (computed values with no matching ground truth values) and false negatives, FNs (ground truth values not matched with computed values). True negatives, TNs, were ignored, since for this data the *vast* number of true negatives would wholly subvert the other measurements (typically, with 100,000s of TNs per TP). To evaluate performance, values for precision, recall, and F_1 (the harmonic mean of precision and recall) are computed as

$$\begin{aligned}\text{precision} &= \mathsf{TP}\,/\,(\mathsf{TP}+\mathsf{FP}),\\ \text{recall} &= \mathsf{TP}\,/\,(\mathsf{TP}+\mathsf{FN}),\\ \mathsf{F}_1 &= 2\mathsf{TP}\,/\,(2\mathsf{TP}+\mathsf{FN}+\mathsf{FP}).\end{aligned}$$

In addition to testing how well the proposed method works, it can also be compared to image-based methods by using a state-of-the-art image-based person detection method. The bounding-rectangle detections from this method can be evaluated with the previously described image-plane evaluation method and the metrics compared with the proposed method.

To perform this analysis a deformable parts-based model for object detection was used. Following [20], a class model is defined by a coarse *root filter* along with several, higher resolution *part filters*, along with a spatial model that weights the part locations. These filters are defined on multi-scale feature maps of histograms of oriented gradient (HOG) features. Speedups were achieved by using PCA dimensionality reduction on the feature maps. An object class—in our case people—is trained on input images using a latent support vector machine (LSVM). The LSVM problem is defined as

$$\min_{\boldsymbol{\beta}}\; \tfrac{1}{2}\,\|\boldsymbol{\beta}\|_2^2 \;+\; c\sum_{i=1}^{n}\max[0,\,1-y_i f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)]$$

where

$$f_{\boldsymbol{\beta}}(\boldsymbol{x}) \;=\; \max_{z\in\boldsymbol{Z}(\boldsymbol{x})}\,\boldsymbol{\beta}\cdot\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{z})\,,$$

and $\boldsymbol{x}_i$ is the feature vector, $\boldsymbol{\beta}$ is the model parameters, y_i is the corresponding label and $c > 0$ is the tuning parameter moderating the trade-off between regularization and the hinge-loss function. The function $f_{\boldsymbol{\beta}}$ selects the latent parameters $\boldsymbol{z}$ from those possible for $\boldsymbol{x}$, referred to as set $\boldsymbol{Z}(\boldsymbol{x})$, which maximizes the linear function. The LSVM problem is convex for the negative examples where $y_i = -1$ however this is not the case for the positive examples. They overcome this by alternating optimization over $\boldsymbol{z}$ for the positive examples and $\boldsymbol{\beta}$ with the latent variables of the positive examples fixed.

The training set for learning the model comes from the VOC 2010 dataset. While lacking any images of children similar to those in our data, the model learned from this training set was still deemed appropriate as as the goal of the learned detector was to be generic to environment and subject.

V. Results

The evaluation methods described in Section IV were applied to roughly five minutes of data recorded during a normal class at the preschool. Three children and two adults were observed during this test. One adult begins present in the room and quickly leaves. Then a boy enters from the playground and sits at a table. After a short time at play, two more children enter and briefly interact with him. Later

Performance Per Sensor				
Sensor	**Method**	F_1	Precision	Recall
1	IMPED	0.4490	0.5000	0.4074
	IMBGS	0.6278	0.6324	0.6232
	VXBGS	**0.6896**	0.5714	**0.8696**
	VXBGSC	**0.6936**	0.5769	**0.8696**
2	IMPED	0.2646	0.5231	0.1771
	IMBGS	0.3182	0.3500	0.2917
	VXBGS	**0.7170**	0.6552	**0.7917**
	VXBGSC	**0.7018**	0.6061	**0.8333**
3	IMPED	0.3260	0.8409	0.2022
	IMBGS	0.3636	0.6667	0.2500
	VXBGS	**0.6000**	0.4615	**0.8571**
	VXBGSC	**0.7000**	0.5385	**1.0000**
4	IMPED	0.1201	0.3500	0.0725
	IMBGS	0.5515	0.8065	0.4190
	VXBGS	**0.8845**	0.8920	**0.8771**
	VXBGSC	**0.7762**	0.7874	**0.7654**
5	IMPED	0.0352	0.2353	0.0190
	IMBGS	0.5905	0.6242	0.5602
	VXBGS	**0.8496**	0.8229	**0.8780**
	VXBGSC	**0.8153**	0.7853	**0.8476**

TABLE I

Summary of results for bounding-box detections on each sensor. IMPED denotes results using image-based person detection, IMBGS denotes results using the image-based background subtraction, VXBGS denotes results using the voxel-based background subtraction with a voxel size of 1cm^3, and VXBGSC denotes results using the voxel-based background subtraction with a color model and a voxel size of 2cm^3.

still, a second adult enters and sits to join the first child, still at the table.

For validation and testing, the multiply recorded scene was hand-labeled using an iterative GrabCut [21] image segmentation program. Personnel other than the authors inspected each frame, marking some object and some background pixels. GrabCut then accurately completed the segmentation and produced the ground truth image masks. Since they align with the corresponding Kinect depth maps, these masks also indicate the depth pixels in the corresponding depth maps that are unrelated to a labeled object. Fortunately, as manually identifying the masks is very tedious, doing so only at sixty frame intervals—about every two seconds for each sensor—has sufficed thus far.

The first comparisons considered are for the per-sensor, 2D segmentation masks. Three different approaches to foreground localization were tested: image-based background subtraction (IMBGS), our voxel-based background subtraction (VXBGS) and the voxel-based method with an added color model (VXBGSC). Additionally, results using the object detection method in [20], now trained for people (IMPED) are presented. For IMBGS, the foreground probability threshold was set to 0.12. This was found empirically but the results were not very sensitive to the value selected.

In general, there is a trade-off in the resolution of voxels for the background model: larger model voxels will remove more background but more foreground points will be lost, however small model voxels will result in not enough background being removed. The size of these model voxels should be characterized based upon the noise in the system. For the VXBGS method, the voxel size was therefore set to 17.5cm^3, a reasonable balance between trimming background without over trimming foreground and with reasonable processing requirements. For the VXBGSC method, a lower resolution of 22.5cm^3 was used. Foreground points now also can be distinguished by their color distributions so positional acuity can be dropped to offset the increased color processing. A threshold of 85 for the Mahalanobis distance was also found to perform the most accurately. It should be noted that the size of the model voxels does not affect the resolution of the point cloud in the foreground voxels.

Table I shows the F_1, Precision and Recall for all five sensors across the different methods. Different segmentation masks produced for the same image are shown in Figure 6. Recall is our most crucial measure, indicating what portion of those present were successfully detected. From the table, it is clear that for every sensor Recall for *both* voxel-based methods consistently outperformed the two image-only methods. Recall percentages ranged from 76.5% to 100% for the voxel methods, and in no case was an individual undetected by all five sensors simultaneously (not shown). Results for Precision were more mixed, but should improve when the system incorporates explicit adult and child models and false positives are reduced.

Table I clearly suggests that even very sophisticated image-based person detection and background modeling methods can be markedly improved with straightforward, conceptually less sophisticated 3D enhancements. The two

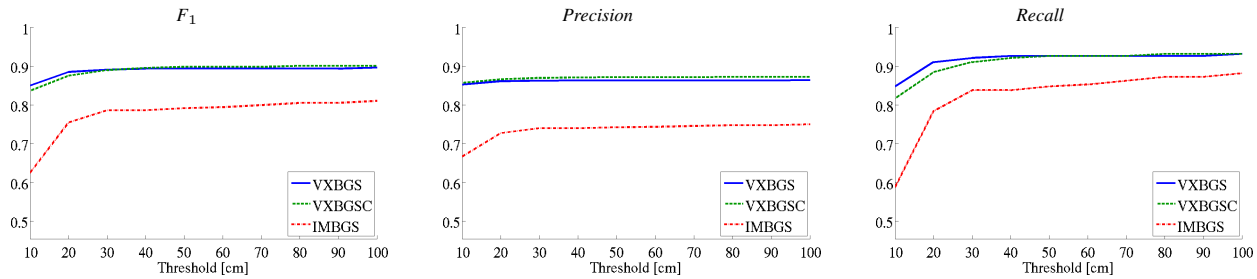


Fig. 5. Summary of results for XY location detections. IMBGS (DASH-DOT RED) denotes results using image-based background subtraction, VXBGS (SOLID BLUE) denotes results using voxel-based background subtraction with a voxel size of 1cm^3, and VXBGSC (DASHED GREEN) denotes results using voxel-based background subtraction with a voxel size of 2cm^3 and combined with a color model.

(a) IMBGS

(b) VXBGS

(c) VXBGSC

Fig. 6. Differing segmentation masks derived from the classroom which depict object points projected back to the image plane, color coded based upon the object ID. From sensor 4, frame # 3,900. (To be viewed in color.)

voxel-based methods achieved the best two results for detecting those in the classroom. Why incorporating color information into VXBGSC did not *consistently* improve performance over VXBGS remains unclear, but their different resolutions may be a factor.

Figure 5 displays evaluations for the XY locations of detected people. Here, the threshold for a correct detection is varied from displacements of ten to one hundred centimeters from ground truth and plotted against the F_1, Precision, and Recall statistics. Improvements generally plateau after a threshold of about 35cm, which corresponds to a reasonable approximation of the space a person occupies. This result shows that the system reliably localizes people in the scene. In Figure 5, IMPED does not appear. Since its Recall and Precision rates proved so exceedingly poor, XY localization using IMPED was not performed.

VI. Conclusions & Future Work

This paper expands on recent enhancements to a system being developed for the automated monitoring and fully 3D analysis of the behavior of preschoolers. The special challenges of this environment made it very problematic for exclusively image-based methods, however our use of multiple RGB-D sensors was able to detect and localize people in the classroom with rates of recall from 76.5% to 100%. These results strongly support the long-term tracking and behavioral analysis now in development, and demonstrate our methods as well matched to this environment.

In addition to extending our research further into our point cloud pipeline, enhancements continue to the portions presented here, including an on-line updating of the voxel-based background model. To further improve our precision, and hence F1, we are incorporating explicit discriminative models to better restrict our detections to only objects of interest.

Acknowledgments

We gratefully acknowledge the efforts of Benjamin Bosch, Walker Krepps and Rachel Redmond in creating the ground truth data, and the assistance from collaborators in the behavioral science field, especially Dr. Kathryn Cullen of our Department of Psychiatry, and Barbara Murphy of our Institute of Child Development. This material is based upon work supported by the National Science Foundation through grants #IIP-0443945, #CNS-0821474, #IIP-0934327, #CNS-1039741, and #SMA-1028076. Additional support for GS from NGA, ONR, ARO, and NSSEFF (AFOSR) is also gratefully acknowledged.

References

[1] 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," *Geneva: WHO*, 2004.

[2] T. Insel. (2012, Feb.) National institute of mental health strategic plan. http://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/index.shtml.

[3] J. Fasching, N. Walczak, R. Sivalingam, K. Cullen, B. Murphy, G. Sapiro, V. Morellas, and N. Papanikolopoulos, "Detecting risk-markers in children in a preschool classroom," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 1010–1016.

[4] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *Proceedings of the Third IEEE International Workshop on Visual Surveillance*. IEEE, 2000, pp. 3–10.

[5] A. Mittal and L. S. Davis, "M_2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[6] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[7] A. Maimone and H. Fuchs, "Reducing interference between multiple structured light depth sensors using motion," in *Virtual Reality Workshops (VR)*. IEEE, 2012, pp. 51–54.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge Univ Press, 2000.

[9] N. Walczak, J. Fasching, W. Toczyski, R. Sivalingam, N. Bird, K. Cullen, V. Morellas, B. Murphy, G. Sapiro, and N. Papanikolopoulos, "A nonintrusive system for behavioral analysis of children using multiple RGB+Depth sensors," in *IEEE Workshop on Applications of Computer Vision*, Jan. 2012.

[10] T. Bouwmans and F. El Baf, *B. Statistical background modeling for foreground detection: A, Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing, Jan. 2010, vol. 4.

[11] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, pp. 1937–1944.

[12] J. Yao and J. Odobez, "Multi-layer background subtraction based on color and texture," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2007, pp. 1–8.

[13] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC, Technical Report UILU-ENG-09-2214, Oct. 2010.

[14] J. Elseberg, D. Borrmann, and A. Nüchter, "Efficient processing of large 3d point clouds," in *International Symposium on Information, Communication and Automation Technologies (ICAT)*. IEEE, 2011, pp. 1–7.

[15] R. Sivalingam, A. Cherian, J. Fasching, N. Walczak, N. Bird, V. Morellas, B. Murphy, K. Cullen, K. Lim, G. Sapiro, and N. Papanikolopoulos, "A multi-sensor visual tracking system for behavior monitoring of at-risk children," in *IEEE International Conference on Robotics and Automation*, May 2012, pp. 1345–1350.

[16] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics, 2001, vol. 1.

[17] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[18] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnörr, "Real-time optic flow computation with variational methods," in *Computer Analysis of Images and Patterns*. Springer, 2003, pp. 222–229.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results," http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[21] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.