Children Behavior Tracking and Personal Identification By Multiple Kinect Sensors

Bin Zhang[*1] Tomoaki Nakamura[*1] Kasumi Abe[*1] Muhammad Attamimi[*1]
Takayuki Nagai[*1] Takashi Omori[*2] Oka Natsuki[*3] Masahide Kaneko[*1]

[*1]The University of Electro-Communications [*2]Tamagawa University
[*3]Kyoto Institute of Technology

In this paper, we proposed a children behavior tracking and personal identification system based on multiple Kinect sensors. In our system, two Kinect sensors are used, one is set horizontal in front of the children, to get their frontal face information, and the other one is set slanted, to get the whole scene with few occlusions among the children. Face, clothe color and moving information are used for personal identification, and this identification results are integrated to our extended Markov Chain Monte Carlo (MCMC) Particle Filter for robust tracking. The motion trajectory of each particular person can be extracted. The effectiveness of our method is proved through the experiments conducted in a nursery school.

1. INTRODUCTION

In recent years, double-income households keep increasing, and taking care of the children has become a big problem in Japan. However, the number of qualified nursery teachers is not enough at all. Taking care of all the children and recording their behaviors is difficult and time intensive for the nursery teachers. At present, the nursery teachers have to focus on the children while holding the class activities, and record their performances afterwards by memories. Thus, it is necessary to automate the recording work by machines. Developing a child care assisting system for effective utilization of teacher resources is necessary. The aim of our research is to develop a child care assisting system to help the nursery teachers with their work. As the basic technique, tracking the behaviors of the children is essential. However, there are few devices or softwares that can be applied for tracking the children and collecting data. An easy setting monitoring and data collection system is under needed.

In this paper, we proposed a children behavior tracking and personal identification system by utilizing multiple Kinect sensors. We proposed to set 2 Kinect sensors from different views in different height. One is set in the average height of children in front of the class to monitor the children with high qualified frontal face images, and the other is set slanted in a higher height to monitor the children with less occlusions. We integrate the information from these two sensors to track and identify the children by our extended the Markov Chain Monte Carlo (MCMC) particle filter [Choi 13] method. Notice that our system can be easily extended to more sensors if the view is not enough. It is a general multiple sensors children tracking system. Figure 1 shows the scene that we need to track in a nursery school where we conducted the experiments.

Contact: Bin Zhang; The University of Electro-Communications; 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585; 042-443-5216; zhang@radish.ee.uec.ac.jp

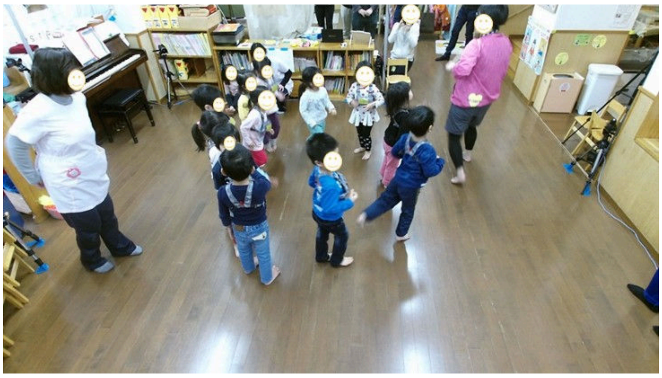


図 1: A tracking scene in a nursery school.

2. ROBUST CHILDREN TRACKING

In order to monitor the children behaviors in an unstructured classroom, we proposed a easy setting multi-sensor children tracking system by only using two Kinect sensors. The sensor positions in the classroom are shown in Fig.2. Kinect 1 is set in the average height of children in front to monitor the children with high qualified frontal face images, and Kinect 2 is set slanted in a higher height to monitor all of the children with less occlusions. The accurate positions and slanted angles are not required in the sensor setting step, which ensures that the sensing equipments are easy to mounting. The accurate coordinates of Kinect sensors will be estimated during the calibration process. The scene in Fig.1 is taken from the slanted Kinect 2.

2.1 Children Position Detection

We detect the positions of the children from the point cloud information gotten from Kinects. As the detection process is familiar with each other, we take the slanted Kinect 2 as example to explanation. As the sensor is set roughly at the beginning for easy-setting and the accurate position and slanted angle of the Kinect $(x, y, h, \phi, \theta, \psi)$ is unknown, we need to estimate accurately the height and

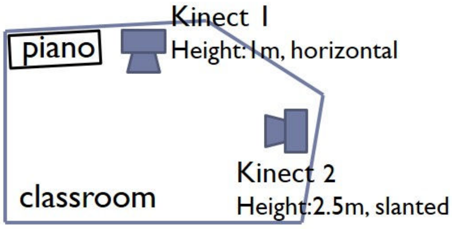


図 2: Sensors placement in the classroom.

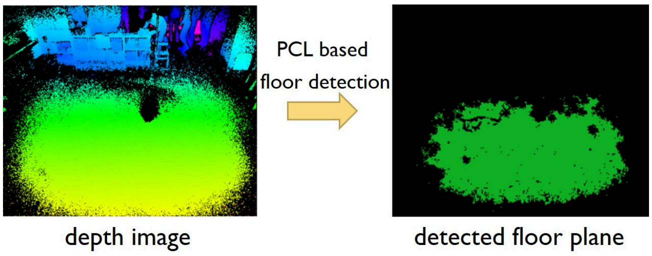


図 3: The depth image and detected floor plane.

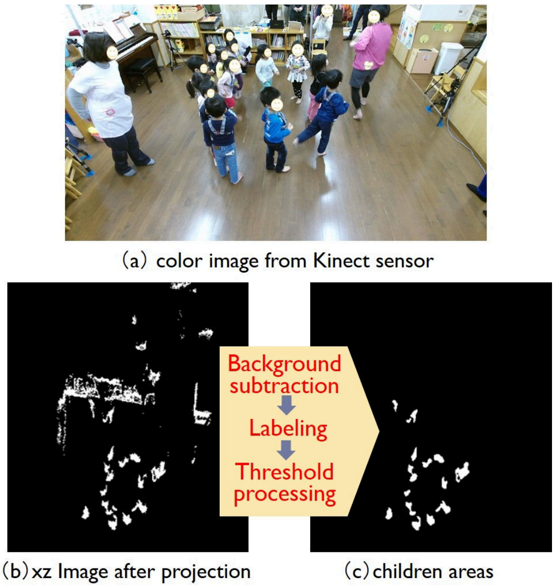


図 4: Children detection process.

slanted angle of the Kinect. We detected out all the planes using Point Cloud Library (PCL) in the empty classroom, and segment out the horizontal plane with the lowest height as the floor plane. During the process, the original rough slanted angles (roll $\tilde{\phi}$, pitch $\tilde{\theta}$) are used to limit the floor plane range with an angle allowance ϵ. That is to say, the detected floor plane (roll ϕ_f, pitch θ_f) needs to meet Eq. (1) in the Kinect coordinate.

$$\tilde{\phi} - \epsilon \leq \phi_f \leq \tilde{\phi} + \epsilon \ ; \ \tilde{\theta} - \epsilon \leq \theta_f \leq \tilde{\theta} + \epsilon \qquad (1)$$

The detected floor plane for the scene in Fig.1 is shown in Fig.3. We can calculate the accurate Kinect slanted angle (roll ϕ, pitch θ) by Eq. (2).

$$\phi = \phi_f ; \theta = \theta_f \qquad (2)$$

Then we transformed the point cloud as if the Kinect runs parallel with floor and detected out the lowest horizontal plane to calculate the height h of Kinect. For each frame during the detection process, we can get the 3D information of the environment from the Kinect sensor, and we transform the point cloud into the horizontal Kinect coordinate system. We detect the children by projecting the transformed point cloud with the child height range (0.5m-1.2m) on the ground, and finding out projected points after deleting the background parts. These areas can be considered as children candidates. Then we use connected-component labeling [17] process to detect out the areas with children size by Eq. (3).

$$l_{min} \leq l_{width} \leq l_{max}$$
$$l_{min} \leq l_{length} \leq l_{max}$$
$$S_{min} \leq S_{area} \leq S_{max} \qquad (3)$$

Here, l_{width}, l_{length} means the length and width of a child area, S_{area} means the size of the area. $l_{min}, l_{max}, S_{min}, S_{max}$ means the thresholds of the candidate area.

The detection process is shown in Fig. 4. For the scene shown in Fig.4 (a), the projected result on the ground plane is shown in Fig.4 (b). The final children position detection result is shown in Fig.4 (c). In this way, we can get the children position information from two Kinect sensors.

2.2 Multiple Sensors Calibration

We address the calibration problem by matching the same corresponding points between the two children position detection results of two Kinect sensors. We change the matching process into matching the corresponding points on XZ maps to decrease the complexity. We get the children position information from two Kinect sensors, and they actually expressed the same position information in two Kinect coordinates. We try to match the points by affine transformation. The affine transformation matrix is calculated from a frame of position information that is easy to be matched, as shown in Fig. 5. We choose 10 corresponding points, and use 3 of them (like (a1)-(a3) in Fig.5) to calculate the affine transformation matrix. The remaining points are used to check the error of residual sum of squares (RRS). We repeat this process until finding the best affine transformation matrix with the least RRS. The corresponding points can be gotten easily by asking a person walking around in the beginning. The single detected person by the 2 Kinect coordinate systems are surely the same person. Finally, we apply this best affine transformation matrix to each frame and get the children position detection result after calibration. From Fig. 5, we can find that the view from Kinect 1 is easy to detect the faces, and the view from Kinect 2 has few occlusions. After transforming into the same coordinate system (e.g. Kinect 1 coordinate system), the person can be detected even if she is totally occluded in Kinect 1 (like (b) in Fig. 5). Once we set the sensors, we can apply this method to do the calibration process. It is especially

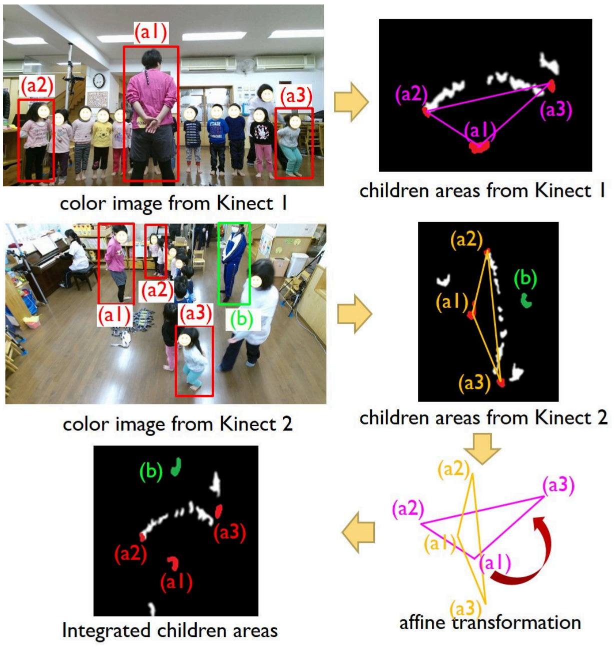

(a1)
(a2)
(a3)
color image from Kinect 1
(a2)
(a1)
(a3)
children areas from Kinect 1
(a1)
(a2)
(a3)
(b)
color image from Kinect 2
(a2)
(b)
(a1)
(a3)
children areas from Kinect 2
(b)
(a3)
(a2)
(a1)
Integrated children areas
(a2)
(a3)
(a2)
(a1)
(a1)
(a3)
affine transformation

図 5: Multi-Kinect calibration result.

useful for the scenes that we need to assemble and disassemble the equipments frequently. Like our experiments in a nursery school, we cannot leave the equipments in the multi-purpose classroom.

2.3 Children Identification And Tracking

We model the particular child identification and tracking problem using a sequential Bayesian framework. A child's state at time t can be expressed as X_t (location, velocity and acceleration in 2D). When the observation information Z_t is gotten from sensors information at time t, we estimate the children states by finding the maximum-a-posteriori (MAP) solution of the joint probability.To find the most probable configuration, we estimate the MAP solution of $P(X_t|Z_t)$ by Eq. (4).

$$P(X_t|Z_t) \propto P(Z_t|X_t) \int P(X_t|X_{t-1})P(X_{t-1}|Z_{t-1})dX_{t-1} \tag{4}$$

Here, $P(Z_t|X_t)$ represents the observation likelihood at time t, X_t, given the sensor input at time t, Z_t. This measure the confidence of a hypothetical configuration. $P(X_t|X_{t-1})$ is the motion model, which shows the smoothness of the trajectory over time. $P(X_{t-1}|Z_{t-1})$ is the posterior probability of time $t-1$. The posterior probability at arbitrary time t can be calculated from the probabilities from time 1 to $t-1$ sequentially if the posterior probability at initial time is given. The best configuration X_t is then the MAP solution.

2.3.1 Motion Model

The motion model $P(X_t|X_{t-1})$ can be modeled by giving the update rules as

$$X_t = X_{t-1} + X'_t d_t \;;\; X'_t = X'_{t-1} + X''_t d_t \;;\; X''_t = X''_{t-1} + \mu \tag{5}$$

Here, μ is a process noise for a child's motion getting from a Gaussian noise.

2.3.2 Observation Likelihood

Given a hypothesized location of a child on the image, the observation likelihood measures the accuracy of the location. In our system, we proposed to use three detectors to evaluate the observation likelihood: a face detector, a color detector and a motion detector. Each single detector has its strength and weakness. The face detector is extremely reliable when a frontal face is shown, but the face information may not be always available as the child may show his back to the sensor. The color detector is always available, but the accuracy is relatively low when different children wear similar clothes. The motion detector can effectively limit the motion range of the child as he/she cannot move a long distance in a single frame time. However, this detector is hard to distinguish the candidates that show up in the motion range. We propose to combine the detectors by using a weighted combination of detection responses as shown in Eq. (6).

$$P(Z_t|X_t) \propto exp(\sum w_j P_j(Z_t|X_t)) \tag{6}$$

Here, w_j is the weight of a detector.

Face detector is used to detect and recognize a particular child's frontal face. We employ the OKAOVISION software in our system. The particular child face detector likelihood is calculated from the maximum recognition confidence score S_{face}.

$$P_{face}(Z_t|X_t) = \alpha(S_{face} - Th) \tag{7}$$

Here, Th is the threshold of face identification confidence. α is the coefficient to adjust the range of the OKAOVISION recognition confidence.

Color detector is used for searching out the child with similar color. As a Kinect sensor is set slanted in a relatively high height to decrease occlusions, the color information is available all the time. We find out the points in the Kinect point cloud that corresponds to the detected areas and match their histograms with the child that we are trying to track. The observation likelihood is calculated from the correlation histogram comparing result S_{color}.

$$P_{color}(Z_t|X_t) = S_{color} \tag{8}$$

Motion detector is a strong indicator of the presence of a person. The areas around the predicted position trend to have higher possibility to be the tracked target. The observation likelihood is calculated from the distance D between the predicted position and detected children areas.

$$P_{motion}(Z_t|X_t) = -\beta D^2 \tag{9}$$

Here, β is the coefficient to adjust the range of the motion likelihood.

2.3.3 Tracking with extended MCMC

We have discussed the motion model and how to evaluate proposed tracking states through observation likelihood.

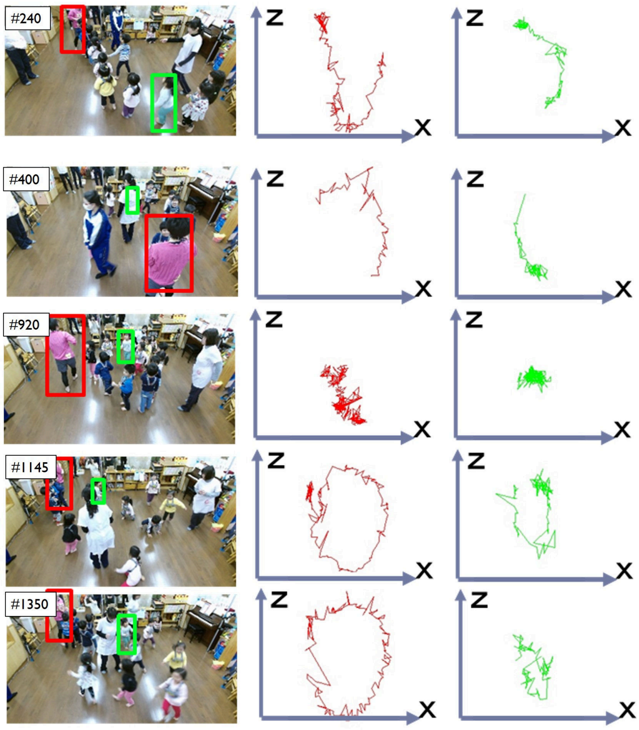
Then we need to explore the space of these hypotheses to find the MAP solution. To effectively explore the configuration space, we extended the MCMC particle filter [Choi 13], referring the reversible jump parts. Different from the work in [Choi 13], we use independent particles to track the particular child and these particles will be never used for other children's tracking. Our goal is to find out the area from the position detection results that takes the highest probability as the particular child. The area that is closest to most of the re-sampled particles will be chosen as the real position of the child. The position of the child is calculated from the center of the most possible area, instead of mean value of all the re-sampled particles. Notice that one area in the children position detection results at arbitrary time t can be recognized as multiple children in our system. When more than one child are close to each other in a frame, the detected child area may fuse with each other, and it turns out to be some "big" areas. The position of these areas should be used for more than one child in practice. Our system can deal with the problem of number changing of detected children in this way. The tracking results are shown as the most possible trajectory for each child. All children can be tracked simultaneously by running multiple sets of these kinds of particles at the same time.

3. EXPERIMENTAL RESULTS

In this section, we present the experimetal tracking results of children during a rhythmics class in a nursery school. Each class is consisted of 10-15 kids in the age between 2-5. Children in the same class have similar ages. It is a challenge to track the children when they move and occlude with each other. We took a scene, in which the teacher leaded the children to do a drum game, as an example to show our tracking results. In the beginning of the game, the teacher leaded the children to walk slowly in the clockwise direction. The speed was synchronous with drum rhythm. As shown in Fig. 6, the teacher stopped for a while (frame 240), and continued to walk on (until frame 400) with time going on. The children followed the motion of the teacher with a time delay. After that, the teacher stopped moving in a circle, and explained the game in detail. We can see from the trajectories (frame 920) that the teacher moved slowly with a few displacement, and the child almost kept static. Then the teacher started to run in the clockwise direction and stopped after one lap. This process is repeated twice with the faster drum rhythm. The trajectories (frame 1145 and 1350) showed the same moving tendency. We can see that the trajectories responded the motion tendency of the tracked persons well. The trajectories also contained the swings of their bodies when moving forward. Our system can track all of the children in different classes and output their trajectories.

4. CONCLUSIONS

In this paper, we proposed a novel application of an easy-setting multiple Kinect sensors system for identification and tracking children behaviors, towards the goal of assisting the nursery teacher with the child care work. We explained in detail the way to set the sensors and the techniques to calibrate them. An extended MCMC particle filter is proposed to track each particular child. All the children can be tracked in the same way. However, the color information of each child can not be repeatedly used as the children change their clothes everyone. More robust personal features need to be proposed for personal identification, especially for children in crowed scenes. Future work will also be focused on building children behavior database, including the changing with time going on. It will be very useful to analyze the psychology development of the children and provide big data for artificial intelligence.



(a) motion of different persons, including a teacher (red) and a particular child (green)

(b) motion trajectories of the teacher (red) and the particular child (green)

図 6: The tracking trajectories of the teacher and a particular child during the whole drum game.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas 26118003. Special thanks to all the members in the nursery school, especially for the cooperative children and their families.

参考文献

[Choi 13] Choi, W., Pantofaru, C., and Savarese, S.: A general framework for tracking multiple people from a moving camera, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 35, No. 7, pp. 1577–1591 (2013)