Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1808–1816
https://doi.org/10.54941/ahfe1008409

Room-Scale Digital Twin Sensing for Quantifying Young Children's Gaze-Based Interest and Movement Behavior

Yuya Kawabe[1], Shihomi Kuribayashi[1], Koji Kitamura[2], Tatsuhiro Yamanaka[3], Yoshifumi Nishida[1], and Mikiko Oono[2]

[1]Department of Mechanical Engineering, Institute of Science Tokyo, Tokyo, Japan
[2]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[3]Safe Kids Japan, Tokyo, Japan

ABSTRACT

This study proposes a room-scale digital twin sensing system for jointly quantifying young children's gaze-based interest and movement behavior within the same three-dimensional environment. The system integrates skeletal data acquired from multiple Azure Kinect sensors, a three-dimensional point cloud model of the room, and stimulus-state data within a unified coordinate system. Gaze-based interest is estimated from head orientation and visualized as an interest heatmap, while movement behavior is quantified using cosine similarity between the child's movement direction and each target direction. As a proof-of-concept study, the system was validated in a stimulus-controlled room where the brightness and color of visual targets could be electronically controlled. Experiments with two four-year-old children showed that the activated LED target achieved the highest interest score and movement-direction score under all tested conditions. These results demonstrate the feasibility of integrating gaze-based interest estimation and movement-direction analysis within a common room-scale digital twin, providing a technical basis for quantifying how children visually attend to and move within everyday environments.

Keywords: Digital twin sensing, Young children, Gaze-based interest, Movement behavior, Stimulus-controlled room, Interest heatmap, Child safety

INTRODUCTION

Domestic accidents involving infants and young children remain a significant social issue, with falls, burns, and poisoning among the leading causes of injury (Peden et al., 2008). Accident risk increases rapidly during infancy as motor and cognitive development expand children's opportunities for exploration and interaction with their surroundings (Safe Kids Worldwide, 2009; Ma et al., 2022). At the same time, infants and young children have a limited ability to recognize hazards despite their strong curiosity and exploratory behavior (Safe Kids Worldwide, 2009; Bowman (James) Associates, 1998).

Previous studies have investigated children's gaze, attention, and exploratory behavior in relation to cognitive development, joint attention, and developmental assessment (Gibson, 1979; Adolph and Robinson, 2015;

Received September 25, 2026; Revised October 25, 2026; Accepted November 2, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Franchak et al., 2011; Mundy and Newell, 2007; Charman, 2003; Jones and Klin, 2013). Visual attention in natural behavior is closely linked to ongoing and upcoming actions, suggesting that gaze can provide useful cues for understanding movement intentions and exploratory behavior (Hayhoe and Ballard, 2005). Wearable eye tracking records what a child looks at (Franchak et al., 2011) but is difficult to apply in natural living environments, whereas room-scale skeletal tracking captures movement but not what attracted the child (Coppens and Maquil, 2025). Consequently, few systems quantify how objects that attract a child's gaze-based interest are related to subsequent movement behavior in the same three-dimensional living space.

Making gaze and movement comparable requires that skeletal data, the room geometry, and the state of each object be handled within a single coordinate system and on a common time base.

Therefore, this study aims to develop a room-scale digital twin sensing system for jointly quantifying young children's gaze-based interest and movement behavior. The proposed system integrates skeletal tracking, a three-dimensional point cloud model of the room, and stimulus-state data within a unified digital twin environment, enabling head-orientation-based gaze information to be visualized as an interest heatmap and movement behavior to be evaluated using a cosine-similarity-based metric. To validate its feasibility, a proof-of-concept experiment was conducted in a stimulus-controlled room where the brightness and color of visual targets could be electronically controlled, allowing reproducible interest induction and quantitative analysis of gaze-based interest and movement behavior in the same three-dimensional space.

The intended application is domestic accident prevention. Injuries to young children often occur when a child approaches a hazard, and that approach is usually preceded by the child noticing it. Measuring interest and the movement that follows in the room itself would therefore allow the hazards of a particular home to be identified from how children actually behave, and could in principle alert a caregiver while the child is still approaching.

The main contribution of this study is the integration of gaze-based interest estimation and movement-direction analysis within the same room-scale digital twin environment, rather than treating gaze and movement as separate measurements. This integration requires coordinate alignment and temporal synchronization among multiple depth sensors, together with estimation of visual attention from head orientation, since direct eye tracking is impractical at this age.

ROOM-SCALE DIGITAL TWIN SENSING SYSTEM

Figure 1 shows an overview of the proposed room-scale digital twin sensing system. Skeletal data of young children were acquired using multiple Azure Kinect sensors, and body tracking was performed using the Azure Kinect Body Tracking SDK (Shotton et al., 2013). A LiDAR-equipped smartphone and Scaniverse were used to generate a three-dimensional point cloud model of the room, while the states of the LED stimuli were recorded. By integrating skeletal data, the point cloud model, and stimulus-state data in a common coordinate system, the system represents the child's position, posture, head

direction, and surrounding environment within a digital twin environment. Coordinate alignment and temporal synchronization among the Kinect sensors were implemented with reference to a previous multi-RGB-D sensor integration method (Coppens and Maquil, 2025), and the skeletal data with the highest confidence were selected for each frame. Figure 2 shows an example of the resulting digital twin environment.

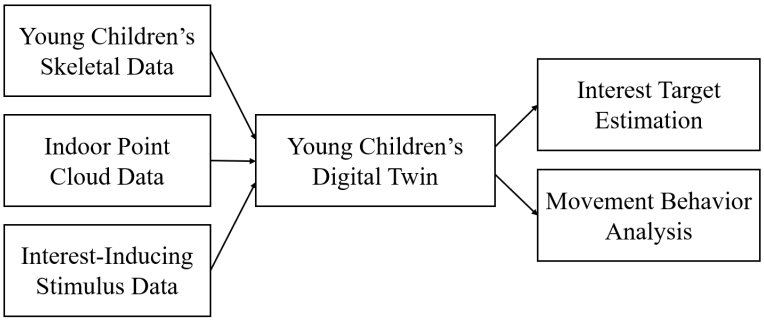


Figure 1: Overview of the proposed room-scale digital twin sensing system, in which a digital twin of the child is constructed from skeletal, point cloud, and stimulus-state data.

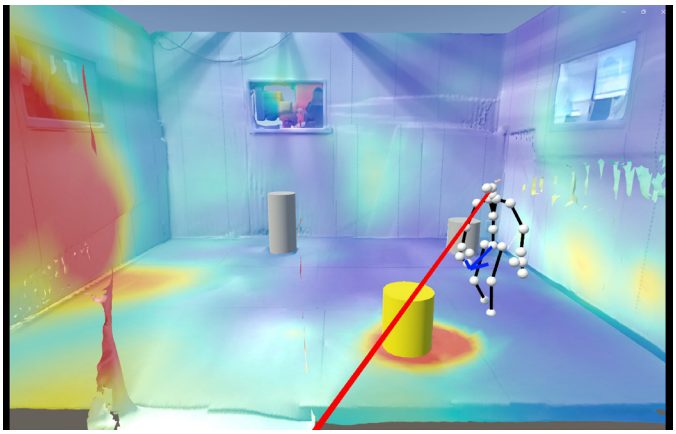


Figure 2: Example of the generated digital twin environment: the child's skeleton and the estimated gaze direction (red line) superimposed on the three-dimensional point cloud model of the room, with the interest heatmap rendered on the point cloud, where warmer colors indicate positions that entered the interest region more often.

GAZE-BASED INTEREST AND MOVEMENT BEHAVIOR ANALYSIS METHOD

Interest Estimation

In this study, the vector from the Head joint to the Nose joint was used as an approximation of the child's gaze direction. Head orientation provides a practical approximation of visual attention when direct eye tracking is

unavailable, although it does not always coincide with the true gaze direction (Murphy-Chutorian and Trivedi, 2009). Head-mounted eye trackers, moreover, are difficult to keep in place on four-year-old children during free movement. Because the actual viewing direction is systematically offset downward in pitch from this vector, a calibration experiment was conducted using five markers placed at known positions. Figure 3 shows the calibration setup and the resulting errors. The average pitch offset observed at the center marker was 24.4° downward, and the head-direction vector was corrected by rotating it downward by this amount. After correction, the combined head-direction estimation error across markers A–D was 10.14°, which was used to define the half-angle of the conical interest region.

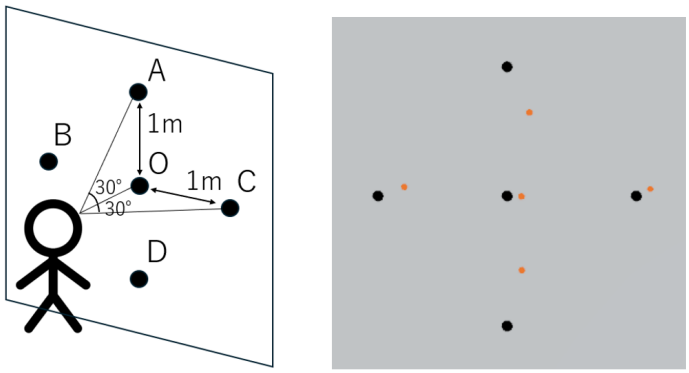


Figure 3: Experimental setup and results of head direction estimation error evaluation (black dots: true directions, orange dots: corrected estimated directions).

Using the corrected head-direction vector, a conical region whose half-angle equals the measured estimation error of 10.14° was defined as the interest region. A point cloud position p_i was considered to be inside the interest region when it satisfied (1).

$$\theta_i = \cos^{-1}\left(\frac{\mathbf{g}\cdot\mathbf{v}_i}{|\mathbf{g}|\,|\mathbf{v}_i|}\right) \leq 10.14^{\circ}, \tag{1}$$

where $\mathbf{g}$ denotes the corrected gaze-direction vector, $\mathbf{v}_i$ denotes the vector from the head position to $\boldsymbol{p}_i$, and θ_i represents the angle between the two vectors.

For each frame, an object was regarded as attended when at least one point belonging to the object was included within the interest region. In such cases, the interest score was incremented by one. The interest score I_k for object k is defined by (2).

$$I_k = \sum_{t=1}^{T} B_k(t), \tag{2}$$

$$B_k(t) = \begin{cases} 1 & \text{if object } k \text{ is inside the viewing cone} \\ 0 & \text{otherwise,} \end{cases}$$

where T denotes the total number of frames. To visualize the spatial distribution of interest, an interest heatmap was generated. Quantitative evaluation was performed using the object-wise interest scores, while the heatmap was used for visualization.

The interest score was adopted because it makes no assumption about how long a single look lasts and remains defined while the child is moving, which makes it robust to brief occlusions. On the other hand, an object counts as attended when any one of its points falls inside the cone, so the score is affected by the apparent size of the object. In the present experiment the three LED targets are taken to be of essentially the same size, so this effect is disregarded.

Behavior Analysis

To evaluate whether movement was directed toward objects that attracted gaze-based interest, the consistency between movement direction and object direction was calculated. For each frame, a movement vector was obtained from the displacement of the Pelvis joint position, and the cosine similarity between the movement vector and the direction vector toward each object was computed using (3).

$$C_k(t) = \frac{\mathbf{m}(t) \cdot \mathbf{o}_k(t)}{|\mathbf{m}(t)|\,|\mathbf{o}_k(t)|}, \tag{3}$$

where $\mathbf{m}(t)$ represents the movement vector at time t and $\mathbf{o}_k(t)$ represents the direction vector toward object k. The cosine similarity ranges from –1 to 1, with larger values indicating a stronger tendency to move toward the object. The average cosine similarity over all evaluation frames was used as the movement-direction metric.

$$\overline{C}_k = \frac{1}{N}\sum_{t=1}^{N} C_k(t), \tag{4}$$

where N denotes the number of evaluation frames. A higher average cosine similarity indicates a stronger tendency for the young children to approach the target object.

Cosine similarity was adopted in order to judge, from the direction of movement alone, whether the child approached each object. Frames in which the child is almost stationary yield a direction dominated by tracking noise and are currently included in the average. The metric also weights every frame of a trial equally, so it reflects the overall tendency of movement during the trial rather than the movement that immediately follows a look.

PROOF-OF-CONCEPT EXPERIMENT

Stimulus-Controlled Room and Procedure

A proof-of-concept experiment was conducted in a stimulus-controlled room measuring 3.6 m × 2.7 m, shown with the sensor layout in Figure 4. Three LED targets were placed at separate locations, and their brightness

and color could be electronically controlled to present reproducible visual stimuli. During the experiment, one LED target was randomly activated at a time to elicit the child's interest. Two four-year-old children participated in the experiment. Skeletal data, estimated gaze direction, and LED states were recorded to generate interest heatmaps and movement trajectories, and the consistency among the activated target, estimated interest target, and movement direction was evaluated. This study was conducted with the approval of the Ethics Committee of the Institute of Science Tokyo (Approval No.: 2025244).

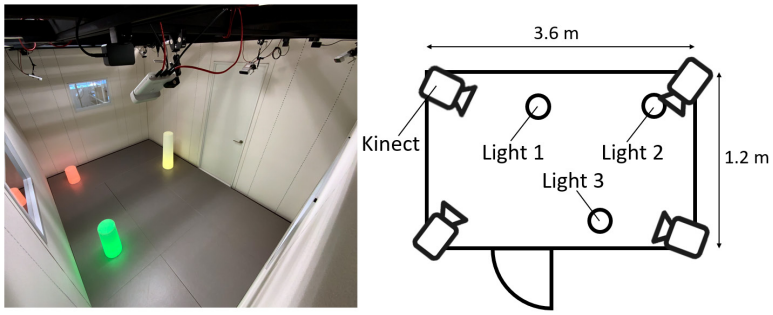


Figure 4: Experimental environment and sensor layout.

RESULTS AND DISCUSSION

Interest Estimation Results

Table 1 summarizes the interest scores obtained under each condition. As shown in Figure 5, heat values were concentrated around Light1, Light2, and Light3 under the Light1 ON, Light2 ON, and Light3 ON conditions, respectively. In addition, the illuminated LED achieved the highest interest score under all conditions for both participants. The Total Interest column gives the number of frames in which any point in the room point cloud, including the wall and floor surfaces as well as the three LED targets, fell inside the interest region; it therefore exceeds the sum of the three target scores and indicates the total amount of attended space rather than attention to the targets alone.

Table 1: Estimated interest scores under different LED conditions.

Participant	Condition	Light 1	Light 2	Light 3	Total Interest
A	Light 1 ON	233	74	99	683
A	Light 2 ON	100	220	144	686
A	Light 3 ON	55	0	380	628
B	Light 1 ON	75	18	45	268
B	Light 2 ON	10	64	45	245
B	Light 3 ON	14	6	154	275

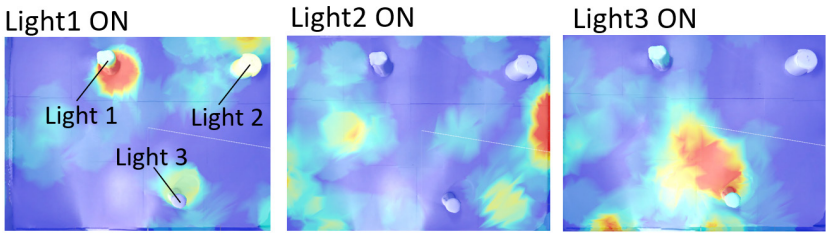


Figure 5: Interest heat maps of Participant A under different LED conditions.

Movement Analysis Results

Table 2 summarizes the movement-direction evaluation results obtained with (4), and Figure 6 shows the corresponding movement vectors for Participant A. For both participants, the highest average cosine similarity was observed in the direction of the illuminated LED under all conditions. In contrast, the average cosine similarity for the non-illuminated LEDs was generally lower and was negative in many conditions.

Table 2: Average cosine similarity for each target under different LED conditions.

Participant	LED Condition	Light 1	Light 2	Light 3
A	Light 1 ON	0.294	−0.057	−0.295
A	Light 2 ON	−0.119	0.125	−0.138
A	Light 3 ON	−0.262	−0.217	0.341
B	Light 1 ON	0.172	−0.131	−0.132
B	Light 2 ON	−0.051	0.217	−0.233
B	Light 3 ON	−0.183	−0.053	0.207

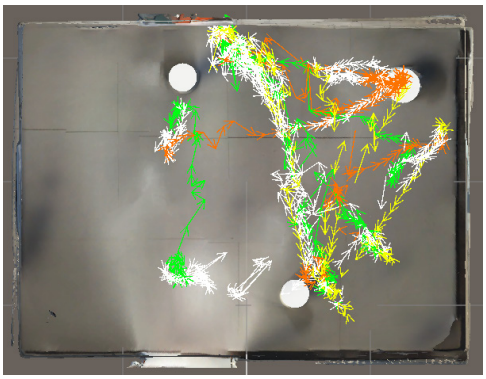


Figure 6: Movement vectors of Participant A under different LED conditions (green: Light 1 ON, orange: Light 2 ON, yellow: Light 3 ON).

DISCUSSION

In this experiment, the object with the highest interest score also showed the highest movement-direction score in every condition and for both participants, indicating that the two measures ordered the three targets

consistently and that the system can estimate an interest target and quantify movement direction within the same environment. Because both measures were aggregated over a whole trial, the present results demonstrate consistency between the two measures in this controlled proof-of-concept setting rather than a population-level or causal gaze–movement relationship.

Because only one target was lit at a time, the interest score accumulates on the lit target while the child's head is turned toward it and the cosine similarity becomes positive while the child displaces toward it, so both quantities respond to the same behavior of orienting to and approaching an attractive object. For domestic accident prevention, the objects that matter are those high in both the interest score and the movement-direction score, that is, the objects a child actually approaches. A hazard that repeatedly attracts a child's attention and is accompanied by approach behavior may represent a particularly relevant target for preventive intervention. Because position, posture, head direction, and the state of each object are held in a single coordinate system, the system can be developed into one that detects a child's approach to a hazard and notifies a caregiver. Realizing and validating this is a subject for future work.

CONCLUSION

This study developed a room-scale digital twin sensing system for quantifying young children's gaze-based interest and movement behavior. A proof-of-concept experiment in a stimulus-controlled room showed that the activated target achieved the highest interest score and the highest movement-direction score under all tested conditions, demonstrating the technical feasibility of jointly measuring gaze-based interest and movement behavior within the same digital twin environment. This integrated representation provides a basis for extending the system from aggregate measures to event-level analysis of when a child attends to and begins moving toward an object, with the longer-term aim of sensing-based prediction and prevention of domestic accidents involving young children.

ACKNOWLEDGMENT

The authors would like to thank the children and their families who participated in this study. This paper is partially based on results obtained from a project, JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

Adolph, K. E. and Robinson, S. R. (2015). "Motor Development", in: Handbook of Child Psychology and Developmental Science, Vol. 2: Cognitive Processes, 7th ed., Liben, L. S., Mueller, U., and Lerner, R. M. (Eds.). Hoboken, NJ, USA: John Wiley & Sons, pp. 113–157. doi: 10.1002/9781118963418.childpsy204

Bowman (James) Associates. (1998). Safety First: Preventing & Managing Childhood Injuries: Training Guides for the Head Start Learning Community. Washington, DC, USA: U.S. Department of Health and Human Services, Administration for Children and Families, Head Start Bureau.

Charman, T. (2003). "Why is Joint Attention a Pivotal Skill in Autism?", Philosophical Transactions of the Royal Society B: Biological Sciences, Vol. 358, No. 1430, pp. 315–324. doi: 10.1098/rstb.2002.1199

Coppens, A. and Maquil, V. (2025). "Skeletal Data Matching and Merging from Multiple RGB-D Sensors for Room-Scale Distant Interaction with Multiple Surfaces", Electronics, Vol. 14, No. 4, Art. no. 790. doi: 10.3390/electronics14040790

Franchak, J. M., Kretch, K. S., Soska, K. C., and Adolph, K. E. (2011). "Head-Mounted Eye Tracking: A New Method to Describe Infant Looking", Child Development, Vol. 82, No. 6, pp. 1738–1750. doi: 10.1111/j.1467-8624.2011.01670.x

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, MA, USA: Houghton Mifflin.

Hayhoe, M. and Ballard, D. (2005). "Eye Movements in Natural Behavior", Trends in Cognitive Sciences, Vol. 9, No. 4, pp. 188–194. doi: 10.1016/j.tics.2005.02.009

Jones, W. and Klin, A. (2013). "Attention to Eyes Is Present But in Decline in 2–6-Month-Old Infants Later Diagnosed with Autism", Nature, Vol. 504, No. 7480, pp. 427–431. doi: 10.1038/nature12715

Ma, Y., Song, J., Hu, M., Yang, R., and Yang, P. (2022). "Risk factors of in-home unintentional injuries among 0–6-year-old children in Changsha city of China: a cross-sectional survey based on Bronfenbrenner's ecological system theory", BMC Pediatrics, Vol. 22, Article 598.

Mundy, P. and Newell, L. (2007). "Attention, Joint Attention, and Social Cognition", Current Directions in Psychological Science, Vol. 16, No. 5, pp. 269–274. doi: 10.1111/j.1467-8721.2007.00518.x

Murphy-Chutorian, E. and Trivedi, M. M. (2009). "Head Pose Estimation in Computer Vision: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 4, pp. 607–626. doi: 10.1109/TPAMI.2008.106

Peden, M., Oyegbite, K., Ozanne-Smith, J., Hyder, A. A., Branche, C., Rahman, A. F., Rivara, F., and Bartolomeos, K. (Eds.). (2008). World Report on Child Injury Prevention. Geneva, Switzerland: World Health Organization and UNICEF.

Safe Kids Worldwide. (2009). Raising Safe Kids: One Stage at a Time: A Study of Child Development and Unintentional Injury. Washington, DC, USA.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). "Real-Time Human Pose Recognition in Parts from Single Depth Images", Communications of the ACM, Vol. 56, No. 1, pp. 116–124. doi: 10.1145/2398356.2398381