Fuzzy Bin-Based Classification for Detecting Children’s Presence with 3D Depth Cameras

HEE JUNG YOON and HO-KYEONG RA, DGIST
CAN BASARAN, HERE
SANG HYUK SON, DGIST
TAEOJOON PARK, Hanyang University
JEONGGIL KO, Ajou University

With the advancement of technology in various domains, many efforts have been made to design advanced classification engines that aid the protection of civilians and their properties in different settings. In this work, we focus on a set of the population which is probably the most vulnerable: children. Specifically, we present ChildSafe, a classification system that exploits ratios of skeletal features extracted from children and adults using a 3D depth camera to classify visual characteristics between the two age groups. Specifically, we combine the ratio information into one bag-of-words feature for each sample, where each word is a histogram of the ratios. ChildSafe analyzes the words that are normalized within and between the two age groups and implements a fuzzy bin-based classification method that represents bin-boundaries using fuzzy sets. We train and evaluate ChildSafe using a large dataset of visual samples collected from 150 elementary school children and 150 adults, ranging in age from 7 to 50. Our results suggest that ChildSafe successfully detects children with a proper classification rate of up to 94%, a false-negative rate as low as 1.82%, and a low false-positive rate of 5.14%. We envision this work as a first step, an effective subsystem for designing child safety applications.

CCS Concepts: • Computing methodologies → Scene anomaly detection; Biometrics; Supervised learning;

Additional Key Words and Phrases: Child classification, child safety, fuzzy logic, kinect-based applications

ACM Reference format:
https://doi.org/10.1145/3079764

This work was partially supported by the Ministry of Trade, Industry and Energy (Grant #N0002312) as part of the Industrial Infrastructure Program for Fundamental Technologies, DGIST Research and Development Program (CPS Global Center) funded by the Ministry of Science, ICT & Future Planning for the project “Identifying Unmet Requirements for Future Wearable Devices in Designing Autonomous Clinical Event Detection Applications”, and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. B0101-15-0557, Resilient Cyber-Physical Systems Research).

Authors’ addresses: H. J. Yoon, H.-K. Ra, and S. H. Son, Information & Communication Engineering Department, DGIST, 333 Techno Jungang-daero, Hyeongpung-myeon, Dalseong-gun, Daegu, Republic of Korea, 42988; emails: [heejung8, hk, sonj}@dgist.ac.kr; C. Basaran, HERE, Invalidenstrasse 116, Mitte, 10115, Berlin, Germany; email: basaran.can@gmail.com; T. Park, Robotics Engineering Department, Hanyang University, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do, Republic of Korea, 15588; email: taeejoon@hanyang.ac.kr; J. Ko, Software and Computer Engineering Department, Ajou University, 206 Worldcup-Ro, Yeongdong-Gu, Suwon, Gyunggi-Do, Republic of Korea, 16499; email: jgko@ajou.ac.kr.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2017 ACM 1550-4859/2017/08-ART21 $15.00
https://doi.org/10.1145/3079764
1 INTRODUCTION

A modernized environment takes on various efforts to protect the people that reside within its boundaries. Among these efforts, one of the utmost goals is to provide a safe environment for the socially weak, and this goal is, in many cases, used as an index of civilization. The need to provide such safe environments, coupled with advances in sensing and communication technologies, has introduced a diverse set of classification engine-based systems for various applications [11, 18, 37, 38, 44]. Nevertheless, among such work, it is surprising that there is only a limited number of systems within the domain of child safety applications, which target one of the weakest population sets [1, 3–5, 23, 36]. Even within these studies, most of the efforts aim to provide safety for infants inside an enclosed space such as a vehicle. To catalyze the development of applications that focus on children beyond the infancy stage and their safety in an open environment, one of the inevitable steps is to design a baseline system that effectively distinguishes children from adults.

In this work, we try to address this goal of effectively classifying children in a target geographical region using body and facial ratios extracted as raw features from 3D depth camera sensors. By designing such a system, we aim at opening the possibilities for developing novel applications that contribute to preserving the safety of children. For this purpose, we define children as people who are between the stages of infancy and puberty: specifically, boys and girls between the ages of 3 and 12 [34]. While the characteristics of each person are unique, many studies in the field of anthropology, human engineering, and kinesiology show that we can potentially find common characteristics among people in similar age groups [6, 28].

Using a Microsoft Kinect camera sensor, we start this work by collecting joint position data from 300 people, spanning diverse age groups that can be classified as children (i.e., elementary school students) and adults (i.e., teenagers to seniors). These data samples are arranged so that we can easily extract various parameters on body joints and major facial points. We propose the ChildSafe classification system, which uses the histogram analysis of bag-of-words features to accurately classify the observed human as a child or an adult. To take our studies a step further, we provide an organized guideline of body proportions of children and adults, considering the inaccuracies of Kinect’s tracking capabilities. We also investigate eliminating the use of facial information to work with only body information since there are circumstances when a person’s face may be covered by accessories such as a hat or sunglasses. To ensure high detection/classification accuracy, even with only the body joint information, we introduce fuzzified bin-boundaries to the ChildSafe system. Evaluations using the collected dataset show that, with large enough training data, ChildSafe can correctly classify children with an accuracy of up to 94%. Performance comparisons of ChildSafe against a system based on C-Support Vector Classifier Support Vector Machines (C-SVC SVMs) show that ChildSafe achieves ~9% higher classification accuracy with a ~31% lower false-positive rate of 5.14%. Furthermore, we record a false-negative rate as low as 1.82%, which is ~13% lower than that of the SVM-based approach.

We summarize the contributions of this work as follows:

- We introduce the importance, potential applications, and technical challenges that relate to detecting the presence of children beyond the stages of infancy in a targeted open environment. This step of our study opens the possibilities for new applications to be developed within the research community.
- We collect a large set of Kinect sensor-based anthropometric data from various age groups to not only perform our own study, but also to catalyze active research on child detection-related applications. Our dataset consists of time-stamped body joint and facial data of eight different actions recorded with a Microsoft Kinect sensor from 150 children and 150 adults,
Fuzzy Bin-Based Classification for Detecting Children’s Presence with 3D Depth Cameras

resulting in 2,400 unique actions. We believe that this is the largest publicly available age-labeled dataset containing children and adult skeletal data.

- We present a guideline based on deep analysis of the prominent body ratios of children and adults. To the best of our knowledge, this is the first work to provide detailed information on children and adults’ body proportions. Rather than using realistic, practical body ratios, some previous studies merely reference drawing books to work with ideal body proportions [40].
- Using the knowledge from our dataset, we design ChildSafe, a lightweight system that targets effectively classifying children against older age groups by using the face and (or only) body joint information collected using a 3D depth camera sensor.
- We intensively evaluate the performance of ChildSafe with real human data and compare its performance with well-known machine learning techniques to showcase its high classification accuracy and low processing overhead under various learning data set configurations.

The remainder of this article is structured as follows. In Section 2, we introduce several potential applications where systems that effectively identify children can be applied to protect their safety, and we discuss some of the technical challenges that lie on the path to realizing them. We start our technical discussions by characterizing the collected visual data in Section 3. Using this data and our observations, the design of ChildSafe is presented in Section 4, and we show performance evaluations in Section 5. Finally, we position our work among others in Section 6 and conclude the paper in Section 7.

2 POTENTIAL APPLICATIONS AND TECHNICAL CHALLENGES

Designing a system that detects the presence of children in a targeted environment can be the basis of a variety of social safety and minor-protection applications. In this section, we introduce some of the applications that can be enabled by such a technology, along with their possible impacts on society, and we discuss some of the technical challenges that remain to be addressed.

2.1 Applications

In-household Children Safety Protection: While a home environment is considered to be relatively safe, children are often exposed to materials that they should have restricted access to, such as sharp kitchenware, home improvement tools, drugs, or inappropriate web/multimedia contents. Keeping and securing such items from the reach of children would be an ideal manual solution, but there are cases where this is not viable. For example, children can be injured when there are changes to the usual routine due to poor housing and overcrowded conditions, being in a hurry, or being in unfamiliar surroundings (e.g., when visiting friends’ or relatives’ homes) [9]. According to the Royal Society for the Prevention of Accidents (ROSPA), more than 1 million children under 15 years of age are taken to the emergency department each year due to domestic injuries [15]. A system that automatically detects the presence of children and warns their guardians against hazards can substantially reduce the number of children suffering from domestic injuries and/or emotional damage.

School and Social Safety Protection: Unfortunately, while school environments are considered relatively safe, environments with a high population of children present the biggest threats to the safety of children. For example, in the United States and in many other parts of the world, we frequently see news articles that report incidents where shooters and sexual offenders intrude into the school environment and harm the emotional states or even the lives of the children. In the United States, for example, there has been a rapid increase in school shootings. In 2012, there were 10 school shootings that left 41 people dead and 13 wounded [41]. Since many cities have
already deployed a number of advanced cameras on the streets and public buildings to record and detect crimes, a system that detects the presence of children using these cameras, or perhaps more advanced ones, can enforce the environment to be more alert when children are present.

2.2 Technical Challenges
We now detail a few technical challenges in realizing a system for efficiently detecting children in a targeted environment drawn from our experiences and existing literature.

Lack of Anthropometric Data: Based on our experiences, the first challenge in realizing the design of a system that detects the presence of children is the lack of child-related anthropometric data to be used for ground truth learning. While this is not a technical challenge in itself, collecting a sufficiently large amount of data is essential in the design and evaluation stages of a classification system.

Minimal Detection Latency: Most underage protection applications require low latency detection when children are present in the environment, and, hence, the value of the system output degrades with increased response time. Therefore, the detection system should be applicable to incoming data streams and should be able to detect and report the presence of children with minimal response times.

Differences in Aging Patterns: While automatically recognizing different characteristics of a person, such as identity, emotion, and gender, has been extensively studied, automatic age group estimation has not been as much explored, despite the fact that it is an interesting problem in its own right. One reason behind this is that aging works in a personalized, uncontrollable way, making it challenging to classify: Different people physically age in different ways. Studies show that different parts of human bodies age faster than others [43]. Moreover, people can look younger or older than their chronological age due to external factors, such as health, diet, exercise, and lifestyle. Therefore, considering different aging patterns is essential in designing an age classification system.

Adaptiveness to Diverse Population Characteristics: When monitoring people in a physical environment, the system should take into account the fact that people can have unique characteristics from fashion to body, facial, and vocal expressions. It is most likely that when an automated monitoring system is deployed, new characteristics will be exhibited by the person being monitored. Hence, the system should be flexible enough and use only the information that it is capable of collecting in making a proper detection.

Inaccuracy in Skeletal Tracking: Skeletal tracking in many commercial 3D depth cameras is not fully reliable. While supporting features such as joint tracking, its reliability can be easily affected by noise in practical settings where occlusions are present. Therefore, this inaccuracy should be considered when designing and testing a 3D camera-based system.

Preserving Privacy During Processing: The main purpose of surveillance cameras is usually to monitor behavior, activities, or other varying activities. Observing people while simultaneously protecting their privacy can be challenging. Although it may not be possible to provide complete privacy protection, the system should try its best to minimize tensions that some people may have with privacy concerns.

False Alarms: In any safety system, false alarms are a common yet important issue to address. We believe that there must be a balance between fully leaving the judgment of classification with the machine and burdening the operator with correcting the system. This is especially important.
for our study, where readings obtained from camera sensors in a practical deployment can be extremely noisy. Moreover, in a real-world scenario, the person being monitored is not always facing the camera, but is engaged in daily activities. Properly compensating the noise, relative distance, and body orientation between the person and the camera is mandatory to reduce false alarms while accurately classifying the input feature data.

3 DATA COLLECTION PROCESS

To design a system for detecting the presence of children, this work utilizes a 3D depth camera sensor with a novel classification method. Specifically, we utilize the Microsoft Kinect camera sensor to collect facial and body joint information from a number of subjects in order to extract and evaluate the features to be used for age-group classification. For these reasons, we start by collecting a large number of data samples from both children and adults. We perform our data collection process in two separate phases: the first phase where we collect physical samples for facial and body joint data from various volunteers, and the second phase where we extract the features that we need in order to design and evaluate our target classification algorithms.

3.1 Data Collection Platform: Microsoft Kinect

Before we introduce the two steps of our data collection process, we first introduce the hardware platform that we chose to perform our study with. The system we propose in this work, ChildSafe, uses streaming skeletal data from the Kinect for the detection and classification processes. This device includes an infrared (IR) camera and an RGB camera. The depth image generated by the IR camera has a maximum resolution of $640 \times 480$ and has a recording rate of 30 frames per second. The Microsoft Kinect SDK is used for tracking 20 different skeletal joints, where each of these joints has x, y, z coordinates on the three-dimensional space. The skeletal joint information is effective for our work as it allows us to measure the length between joints, which we use in order to calculate limb lengths. Under this configuration, we point out that no specific pose or calibration action need be taken for a person to be tracked. We also use the Kinect SDK for tracking human faces, where it tracks 87 two-dimensional points that indicate the features of the face and additional 13 points that denote the center of the eye, the corner of the mouth, the center of the nose, and the bounding box around the head.

3.2 Collecting Visual Data from Volunteers

As introduced earlier, one of the main challenges in designing a practically working child detection/classification system is the process of collecting real samples for learning and acceptable evaluations. We address this issue by conducting a data collection phase for visual data (i.e., Kinect sensing data) from a large population distributed over various age groups. Specifically, we collect time-stamped body and facial data from 150 children and 150 adults, which include recordings from eight different actions resulting in 2,400 actions. Our dataset is especially large because it consists of more than just joint data. We also collected RGB video streams for ground truth and depth images that can be used for any possible future studies.

In this work, as in many studies, we define the children that we target for detection as humans between the stages of birth and puberty. Since detecting infants or children younger than kindergarten age is relatively simple (due to noticeable differences in height and posture), we focus on the proper detection of elementary school students. Therefore, we collected a child dataset of 150 students from a nearby elementary school for 39, 27, 31, 15, 19, and 19 children from the 1st, 2nd, 3rd, 4th, 5th, and 6th grades, respectively (Table 1). To the best of our knowledge, this is the largest dataset that contains different age groups from elementary school children. The dataset is publicly available and can be downloaded from http://aeis.ajou.ac.kr:8000/jgko/childsaf.
Table 1. Distribution of the 300 Subjects Across Age Groups and Genders. M: Males, F: Females

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Label</th>
<th>Age</th>
<th># M</th>
<th># F</th>
<th># Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>Elem-1</td>
<td>7</td>
<td>27</td>
<td>13</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Elem-2</td>
<td>8</td>
<td>10</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Elem-3</td>
<td>9</td>
<td>9</td>
<td>22</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Elem-4</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Elem-5</td>
<td>11</td>
<td>12</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Elem-6</td>
<td>12</td>
<td>12</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>Adult</td>
<td>Adult-1</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Adult-2</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adult-3</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Adult-4</td>
<td>17</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Adult-5</td>
<td>18</td>
<td>23</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Adult-6</td>
<td>19</td>
<td>11</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Adult-7</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Adult-8</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Adult-9</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adult-10</td>
<td>23</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Adult-11</td>
<td>24</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Adult-12</td>
<td>25</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Adult-13</td>
<td>26</td>
<td>15</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Adult-14</td>
<td>27</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Adult-15</td>
<td>28</td>
<td>9</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Adult-16</td>
<td>29</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Adult-17</td>
<td>30</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Adult-18</td>
<td>31</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adult-19</td>
<td>32</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adult-20</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-21</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-22</td>
<td>38</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adult-23</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-24</td>
<td>41</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-25</td>
<td>42</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-26</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Adult-27</td>
<td>50</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

It is important to note that we collected data under the approval of school administration, faculty, and parents of individual students for elementary school children. We provided information on why the research was being done, why they were being asked to participate, how long the participation was to last, what would happen during the research, and how personal information would be protected in the future. The adult dataset, consisting of 300 samples, was also collected after a verbal statement of the purpose of our study from various post-puberty individuals between the ages of 14 and 50 [34]. For teenagers who may or may not have passed puberty, we try to classify them as adults since the applications that ChildSafe tries to address are mostly related to less mature people (e.g., elementary school and younger.

Each participant of the study performed eight different actions, as illustrated in Figure 1. These actions include walking toward the camera, walking away from the camera, walking to the right...
of the camera, walking to the left of the camera, looking toward the camera (standing still), freely moving in place, freely moving while sitting on a chair, and repeating the action of getting up and sitting down in the chair. We use three Kinect sensors for this process: one for collecting data, a second to run simple games which allowed each participant to interact with the camera while collecting data, and a third to display the RGB information so that each participant could see how their joints are detected in real-time. Originally the purpose of these different actions was to utilize movement-based data in our classification scheme. However, as the following sections will show, we designed a lightweight and robust system that only utilizes the face and (or only) body joint ratios to classify the two different age groups.

3.3 Extracting and Characterizing Human Joint Information

Using the data from our visual data collection phase, we built a knowledge base which consists of positions of the body joints and face points. Each subject, \( p \in P \), followed a predetermined script and executed a sequence of actions, \( a \in A \) (see Figure 1). We use a tuple notation and refer to the position of a joint \( j \in J \) at time \( t \) by \( D(t, j, a, p) = (x, y, z) \), while the position of a face point, \( f \in F \), is referred to as \( D(t, f, a, p) = (x, y, z) \). Furthermore, each subject \( p \) belongs to one of the age groups, \( g \in G \), listed in Table 1, where we present the distribution of subjects over age groups and genders. Based on this knowledge base, we start the second phase of our data collection process, which consists of identifying useful features from the data. Specifically, \( D(t, j, a, p) \) and \( D(t, f, a, p) \) are processed through four discrete steps to build the system knowledge base as follows.

- **Step 1**: The first step is to calculate the length of each body part and face metric. For this purpose, we define a body part, \( s_n \in C_S \), as a chain of joints, i.e., \( s_n = \{j_{n1}, j_{n2}, \ldots \} \) and a face metric, \( f_m \in C_F \), as a pair of face points, i.e., \( f_m = \{f_{m1}, f_{m2}\} \). While all face metrics are defined using two points, a body part chain can be as long as six joints. The body parts and facial lengths we define are illustrated in Figures 2 and 3, respectively.

The length of each body part, \( L(t, s_n, a, p) \) is computed by adding the distance between directly connected joints:

$$L(t, s_n, a, p) = \sum_{i < |s_n|} |D(t, j_{ni}, a, p) - D(t, j_{ni+1}, a, p)|. \quad (1)$$

Similarly, a facial length is computed as the distance between two facial points:

$$L(t, f_m, a, p) = |D(t, f_{m1}, a, p) - D(t, f_{m2}, a, p)|. \quad (2)$$

Fig. 1. Eight different actions performed by each participant for our visual data collection phase.
Fig. 2. (a) Skeletal joints of the body that are extracted from Kinect and (b) body metrics that are used in this study.

Fig. 3. Facial metrics used in this study.
We point out that the equations used for analyzing the facial and body data are similar; hence, we omit the computational details for the facial features and describe the detailed process of feature extraction using body joint data as the example throughout the rest of this section.

- **Step 2:** In this step, we compute the average value for each body and facial metric separately for each individual across different actions (Equation 3). This process essentially cleans the data while also reducing computation complexity in the steps to follow. By including various actions, on average, we address the fact that depth cameras report slightly different measurements based on body posture. Furthermore, by training with a diverse set of activities, we can design the system to be robust enough to tolerate some level of noise and detect, in reality, the presence of children with only a limited amount of “natural” actions:

$$\text{Avg}(s_n, p) = \frac{\sum_{a \in A} \sum_{t} L(t, s_n, a, p)}{|t| \cdot |A|}. \quad (3)$$

- **Step 3:** This step focuses on computing the ratio of the body and facial metrics. For example, we compute the ratio of eye-to-eye distance to the vertical length of the head or the ratio of the lower arm to the length of the leg. While doing this, we handle face and body data separately and do not compute the ratio of a facial metric to a skeletal metric. The main reason for this separation is that we believe the face and body data have different application domains, as we will discuss in Section 4. The ratio of a body part’s length $s_i$ to another body part $s_j$ is computed using Equation 3:

$$R(s_i, s_j, p) = \frac{\text{Avg}(s_i, p)}{\text{Avg}(s_j, p)}, s_i \in C_S, s_j \in C_S, s_i \neq s_j. \quad (4)$$

We refer to these computed ratios of body skeletal length as the *body features* and the ratios of facial lengths as the *facial features*, and we use the term *features* to collectively refer to both body and facial ratios. Following this step, we only work on the ratio data and can discard the initial position data; therefore, we will relabel $R(s_i, s_j, p)$ as $R(\ell, p)$ such that $\ell$ represents a feature as the ratio of an arbitrary pair of facial or body lengths. The main reason behind this design choice is our observation that methods based on absolute measurements are significantly more vulnerable to measurement errors compared to those based on relative metrics. For the same reason, we do not consider the height of individuals.

Using these various body and facial features extracted from a person from our data collection phase, the following sections will discuss in detail ChildSafe, our proposed system to classify and detect the presence of children in a targeted environment.

### 3.4 Guidelines for Body and Facial Proportions for Children and Adults

We now analyze the modeled features from two complementary perspectives and provide guidelines that compare the body proportions of children and adults. For the first part of our analysis, we consider the imperfect accuracy of Kinect’s skeleton tracking. Specifically, since Kinects skeleton tracking is not fully accurate [33], we try to overcome this by first selecting the features that show the most prominent differences in length between children and adults. Furthermore, we use these specific features to identify what proportions of the body are prominently different among children and adults.

#### 3.4.1 Best Body Features

There are a number of issues to consider when utilizing human skeleton data. Due to real-world issues such the distance from the camera to the person being monitored and the imperfect accuracy of depth cameras tracking capability, we categorize features that work best when using skeleton data that are extracted from the Kinect camera sensor. In doing so, we measure and compare the average length of each body part. As we show in Figure 4, the body, neck
to foot, leg, upper leg, and arm show the largest differences in length between children and adults. We use these specific features for computing the ratios and eliminate the use of other features since they do not show a major difference and could thus later complicate the overall system’s operations. We can say that either the actual lengths of these particular body features are similar between children and adults or that they also can be an artifact of the depth camera not being able to make detailed measurements.

### 3.4.2 Ratios

Using the absolute length of a particular body part can give varying results from person to person with respect to how a camera is positioned. For example, we can clearly see that adults are typically taller than children. However, children can look taller than adults if they are closer to the camera. Additionally, not all adults are taller than children nor are children always shorter than adults. Therefore, we specifically use ratio information to work with body proportions. Using the best features selected in Section 3.4.1, we compute all possible ratios between these features. Figure 5 summarizes these ratios. We see that there is a clear difference in the ratio between the body features of children and adults. We go into further detail and analyze the differences by children’s school grade. As shown in Figure 6, we see that few of the ratio pairs do not show much difference between 6th graders and adults, such as arm and neck-to-foot pair, upper leg and leg pair, and neck-to-foot and body pair. We eliminate these pairs and use the remaining seven ratio pairs (upper leg and body; arm and body; upper leg and neck-to-foot; leg and body; arm and leg; upper leg and arm; and neck-to-foot and leg) that show a clear difference for classification.
4 CHILDSAFE: BIN-BASED CLASSIFIER FOR IDENTIFYING CHILDREN

ChildSafe is a bin-based classifier system that uses 3D depth sensors to detect the presence of children in a geographical region. To present the big picture of how ChildSafe can be used in real-world applications, we start by illustrating a sample application in which a warning is generated when an individual of a restricted age group interacts with a set of tools to which she is not allowed access (see Figure 7). ChildSafe acts as a core subsystem in such an application and takes on the role of properly detecting the presence of children to allow the larger application system to make proper actuations. We now begin our detailed discussions on ChildSafe by introducing how the system performs its classification.

4.1 Constructing the Bin-Based Classifier

Using the features extracted through the processes described in Section 3.3, we design a system that classifies a human as a child or an adult using facial and/or body features. Specifically, ChildSafe performs the following three steps to achieve robust classification:

- **Bin Creation:** ChildSafe starts by creating 25 bins for each feature such that each sample $R(\mathcal{F}_p)$ from a feature can be mapped to a single bin, $\beta(R(\mathcal{F}_p)) \rightarrow b$, i.e., a discrete value within the closed range $[0,24]$ according to Equation 5. The number of bins, 25, is extracted experimentally by evaluating the performance of various bin numbers. Our observations...
indicate that there is a lower bound for this number to achieve acceptable performance; however, the system is less sensitive to a wide range of values higher than this lower bound. Later, in Section 4.5, we describe a process that further reduces the system’s sensitivity to this parameter.

\[
V_{\text{max}} = \max(R(\mathcal{I}, \forall p_o \in P)) \\
V_{\text{min}} = \min(R(\mathcal{I}, \forall p_o \in P)) \\
\text{step} = (V_{\text{max}} - V_{\text{min}}) / 25 \\
\beta(R(\mathcal{I}, p)) = \left\lfloor \frac{R(\mathcal{I}, p) - V_{\text{min}}}{\text{step}} \right\rfloor.
\]  

Note that the bin-boundaries for features are global across age groups due to the terms \( \forall p_o \in P \) in Equation 5, which results in the calculation of the minimum and maximum values without considering the age groups of individuals.

- **Frequency Detection**: We now utilize the bins created in the previous step to generate the notion of frequencies. Frequencies are tuples defined by a feature, age group, and a bin number \( b \). Hence, a frequency tuple \( \text{Freq}(\mathcal{I}, g, b) \) refers to the number of samples from an age group falling into a specific bin, normalized to the total samples from that age group (Equation 6); thus, this represents the likeness of a feature from a person in a given age group falling into that bin:

\[
\text{Freq}(\mathcal{I}, g, b) = \frac{|\beta(R(\mathcal{I}, p_o \in g)) \in b|}{|\beta(R(\mathcal{I}, p_o \in g))|}.
\]  

- **Weight Computation**: Last, we compute the weight of each age group for a specific bin and feature as the ratio of the age-local frequency to the sum of frequencies across all age groups (Equation 7):

\[
W(\mathcal{I}, g, b) = \frac{\text{Freq}(\mathcal{I}, g, b)}{\sum_{g_o \in G} \text{Freq}(\mathcal{I}, g_o, b)}.
\]  

These weights are used for mapping crisp feature samples into membership degrees of age groups by first computing the bin, \( b \), for a sample and then calculating an age group’s weight (i.e., the certainty of membership) via Equation 7.

### 4.2 Classification Process Operations

From the three preceding phases, we design ChildSafe so that it (1) takes as input the raw body and facial features of a human currently located in the target environment; (2) computes the ratios of the body and facial metrics; (3) combines the features into one bag-of-words, where each word is a histogram of the ratios; (4) normalizes the bag-of-words features within and between the two age groups; and (5) makes an estimate of the age group of the observed human through the classification process. Note that a histogram contains the distribution of ratio data sampled from each activity. With eight activities that we use for training, which makes eight histograms (words) and seven different age groups, we get a total of 56 words. Moreover, all features from an individual are separately processed in order to obtain per-feature membership degrees into an age group. Since the minimum is the standard intersection operator in fuzzy set theory \([35, 51]\), these estimations are aggregated into a final decision by taking the minimum of all membership degrees for each age group (Equation 8):

\[
\mu_g(\mathcal{I}) = \min(W(\mathcal{I}, g, \beta(R_S(\mathcal{I}, p)))) \quad \forall \mathcal{I}.
\]
Table 2. Accuracy of ChildSafe With and Without Fuzzy Bin Boundaries Using Different Types of Features

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature Type</th>
<th>Classify Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin-based</td>
<td>Body &amp; Face</td>
<td>84.64%</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>84.33%</td>
</tr>
<tr>
<td></td>
<td>Body</td>
<td>78.26%</td>
</tr>
<tr>
<td></td>
<td>Height</td>
<td>49.51%</td>
</tr>
<tr>
<td>Fuzzy</td>
<td>Body &amp; Face</td>
<td>88.13%</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>88.04%</td>
</tr>
<tr>
<td></td>
<td>Body</td>
<td>85.09%</td>
</tr>
<tr>
<td></td>
<td>Body (only child data)</td>
<td>82.41%</td>
</tr>
<tr>
<td></td>
<td>Height</td>
<td>58.86%</td>
</tr>
</tbody>
</table>

ChildSafe, at this point, classifies the subject as being in the age group that has a stronger degree of membership. Hence, given the membership degrees of the age groups of interest, the subject is assigned an age group according to Equation 9. Note that for classifying face and body combined features, a late fusion method is used by adding up the resulting membership degrees for both face and body and then identifying the age group:

\[
\text{Est} = \begin{cases} 
\text{kid} & \mu_{\text{child}} \geq \mu_{\text{adult}} \\
\text{adult} & \text{otherwise} 
\end{cases}
\] (9)

In the remainder of this section, we discuss in greater detail how utilizing different feature sets can affect the performance of ChildSafe. Based on these results, we explain how we can further improve the performance of the system under different feature usage scenarios.

4.3 Utilizing Both Facial and Body Features

We first start applying this bin-based classification scheme using combinations of the features in Figure 6 for the body and random combinations of features for the face. Using a subset of individuals (specifically, a subset of each age group), the bins for each feature are determined, and age group weights for each feature and bin pair are calculated via the process described in Section 4.1. This part of the processing is called the training process. With these per-bin, per-age-group frequencies, we use the data from the remaining individuals to evaluate our proposed bin-based classification scheme.

We present the results of our experiment using both body and facial features, only facial features, and only body features in Table 2 when the training set contains 20 individuals from each age group. According to these results, the system is most accurate in detecting children using a combination of body and facial features with \(\sim 84\%\) classification accuracy. Using only face features yields a slightly inferior accuracy, whereas the system can attain as little as \(\sim 78\%\) accuracy when using only the body features. With this, we conclude that the bin-based ChildSafe system does not provide satisfactory performance using only body features.

4.4 Eliminating the Use of Facial Data

In many cases, children and adults can easily hide their facial features by means of accessories such as hats or sunglasses. This can be a severe challenge for school intrusion detection systems, in which case the intruder will not be collaborating with the classifier system. Therefore, we argue that a practical system should minimize its dependency on facial features and utilize only body features, which are more difficult to explicitly hide, for the detection process. Unfortunately,
the results in Table 2 suggest that body features alone do not yield satisfactory detection rates. The reduced classification accuracy is at a level only slightly higher than a random guess. In the following section, we introduce an advanced method that addresses this issue of utilizing only the body joint features while sustaining high classification accuracy.

4.5 Enhancing ChildSafe for Body Joint Data with Fuzzy Bin-Boundaries

One of the key observations we made during our experiments with the bin-based ChildSafe system was that the discrete bin-boundaries from Equation 5 create sudden “jumps” in the membership degrees across neighboring bins. Even though a sample is very close to a bin with a high number of samples from an age group, if it falls out of the bin, the nearby mass has no effect on the classification process. This issue, however, cannot be trivially eliminated by increasing or decreasing the number of bins, mostly because each feature requires a different bin layout that fits its distribution. We exemplify the case with the lower arm-to-shoulder ratio in Figure 8, which only plots the bins with non-zero memberships to any one of the targeted age groups. We note that a sample from a child falling into the 5th bin gets 0 degree of membership to the child age group despite the fact that the neighboring 6th bin assigns a degree of 1. Furthermore, there is a sharp change in $\mu_{\text{child}}$ from the 16th bin to the 17th bin regardless of how close the sample is to the boundary.

Based on these observations, we now describe a method for representing the bins using Fuzzy Logic [21, 51] instead of making sharp transitions. The proposed method aims at generating gradually varying age group membership degrees for nearby bins, similar to SIFT [25]. The system makes this possible by allowing feature samples to fall into, at most, two bins such that a sample belongs to a bin, $b$, and a neighboring bin, $b_{i-1}$ or $b_{i+1}$ with some certainties.

In operation, the bin-boundaries of Equation 5 fall into the bin with 0.5 certainty, while the center of the bin is assigned a certainty of 1. Hence, given a feature sample, we first determine the bins that a sample belongs to along with the respective membership degrees and use the centroid method to compute the final age group membership degree [35]. As a quantitative example, let

Fig. 8. Discrete bins and corresponding membership degrees for lower-arm to shoulder ratio.
us assume that a given feature sample is assigned two bins, $b_1$ and $b_2$, with fuzzy membership degrees $\mu_{b_1} = 0.25$ and $\mu_{b_2} = 0.75$. Furthermore, let’s assume that the age group membership for the child group is 0.8 and 0.6 for these bins, respectively. The final age group membership degree is calculated as $(0.25 \times 0.8) + (0.75 \times 0.6) = 0.65$. This method relaxes the strict bounds on bins and results in a very robust system because it incorporates the notion of distance into the computations.

We test our fuzzy bin-based method using the body and facial features, only facial features, only body features, and only child body features with a training set of 20 individuals from each age group, and we present the results in Table 2. Notice that using only body features with the fuzzy bin margin method achieves higher accuracy than strict bin margins at $\sim 85\%$ prediction accuracy.

### 4.6 ChildSafe with Biased Training Data

Certain application scenarios may require biased learning (i.e., data for one of the categories to be detected may not be available). This can be the case if one of the categories is hard to observe experimentally, such as detecting the body gestures of an injured person. In this section, we modify our algorithm with fuzzy bin-boundaries in order to evaluate its performance when only one of the categories has training data. Specifically, we ignore adult data and only use the data collected from children in the training process.

The lack of data from one of the age groups necessitates a modification to Equation 7. Rather than computing the weight of an age group within a bin, we compute the weight of a bin within the samples from a feature. For example, if half of the samples collected from children fall into a bin for some feature, that bin is assigned 50% weight (Equation 10):

$$ W(\perp, \text{child}, b) = \frac{\text{Freq}(\perp, \text{child}, b)}{\sum_{b=0}^{24} \text{Freq}(\perp, \text{child}, b)}. \quad (10) $$

The system still allows a sample to fall into at most two bins, as described in Section 4.5 and calculates the membership to the child group according to Equation 8; however, the final estimation simply becomes a result of using a threshold of 0.5 (Equation 11):

$$ \text{Est} = \begin{cases} \text{kid} & \mu_{\text{child}} \geq 0.5 \\ \text{adult} & \text{otherwise} \end{cases}. \quad (11) $$

We present the detection accuracy when only 20 child samples are used in the training phase. As we present in Table 2, the overall accuracy of the biased learning system drops to 82.41% with an approximately 2.68% decrease when compared to the unbiased system that utilizes both child and adult data.

### 4.7 Absolute Length Data vs. Relative Datasets

It is easy to think that height, which is an absolute length data that can be captured from a human, can be an accurate measure of a person being a child or not. To verify the effectiveness of using height data (or any other absolute length data of the body) as the classification metric in ChildSafe, we measure the height of the human object collected from our dataset based on the locations of visual anchors. We present the classification results for the height dataset using ChildSafe both with and without the fuzzy bin-boundary extensions in Table 2. The results show that the classification performance with height data is significantly poorer compared to the results using body point (and/or facial length) ratios. The main reason for this poor performance is the fact that the installation angle of the Kinect camera affects absolute length-related metrics. On the other hand, since the metrics that we use in ChildSafe are all relative metrics (e.g., body joint ratios), our proposed system is less impacted by such measurement noise-like factors.
5 EVALUATION

We now perform extensive evaluations on the performance of ChildSafe in greater detail while varying different parameters related to the learning dataset. For performance metrics, we select the accurate classification rate along with the false-negative and false-positive rates to measure the detection accuracy of the system. Furthermore, we take a look at the computation overhead (e.g., processing latency) to examine the feasibility of adopting ChildSafe to various application scenarios that may require short response times.

5.1 Constructing an SVM-Based Classifier

Before evaluating the performance of ChildSafe, we first start by constructing a fair and competitive comparison scheme for classifying children and adults. Specifically, for comparisons, we use an SVM-based method, a widely used machine learning approach in various applications. For implementing the SVM-based classification scheme, we use the LibSVM library [8] and implement the target scheme in Java. The parameter $C$ is configured to 1,000. In detail, we utilize the C-SVC SVM [8], which is most suitable for binary classification applications (e.g., deciding whether someone is a child or not).

We performed preliminary performance evaluations using the extracted body and facial features with the SVM operating under different types of SVM kernels. Specifically, in Figure 9, we show the detection accuracy for linear, polynomial, and radial SVM kernels and also present the performance for the K-nearest neighbors–based classification scheme. The training dataset consisted of 25 individuals’ data, and the evaluation set consisted of 100 individuals’ data. Notice here that the SVM-based approaches perform better than the K-nearest neighbors–based classification scheme, which agrees with the findings from previous literature [26]. Furthermore, the results suggest that, among the three types of SVM kernels, linear SVM shows the best classification performance. Based on the findings of Figure 9, we use the linear kernel for the C-SVC SVM in this work.

Fig. 9. Preliminary detection accuracy plots for three types of SVM kernels and a K-Nearest Neighbors-based classification scheme.
Fig. 10. Detection accuracy for ChildSafe with and without fuzzy bin margins with different sizes of learning sets compared with the detection ratio of an SVM-based estimation approach.

5.2 Comparisons with C-SVC SVM

We now compare the performance of ChildSafe in different configurations (i.e., with and without the fuzzy bin-boundaries) against the aforementioned C-SVC SVM-based approach using the features that we extracted in Section 3.4. While varying the size of the training dataset (e.g., from 20 to 100) in steps of 20 for both children and adults, we also test the two systems with different datasets as well. Note here that bag-of-words features are used for both SVM and ChildSafe to run a fair comparison. All experiments were executed 30 times with randomly selected individuals for training, and the rest were used for testing. We report the mean and standard deviations of each performance metric.

Figure 10 reports our first experimental result, where we present the accurate classification rate as the ratio of correctly classified persons over the total number of individuals. Notice that, as expected, the classification accuracy of the test cases increases with increasing learning dataset size. This is an expected result since more learning data would basically mean that the system is well-trained for various samples with different characteristics. Nevertheless, we note that, compared to the C-SVC SVM approach, ChildSafe shows a more satisfactory performance both with and without the fuzzy bin-boundaries. Especially in the case of utilizing only body feature data with fuzzy bins, the mean classification rate is 94%. In other words, this is an improvement of ~9% when compared to the C-SVC SVM-based approach. We can explain this difference in performance from the fact that our learning data naturally contains measurement noise, and SVMs are known to perform poorly when the input data contain noise [24]. ChildSafe, on the other hand, is more robust to measurement noise since the use of bins smoothes out the data and the use of frequencies reduces the contribution of the outliers. Note that of most importance are the results with the dataset configuration that only includes the body features since this configuration is most likely to be used in reality. The results for this case suggest that ChildSafe, especially when utilizing the
fuzzy bin-boundaries, is an effective system for classifying children from adults, both relatively and absolutely.

Correct classification of the age group for children is of utmost importance to ChildSafe since classifying a child as an adult can be a significant fault that can block ChildSafe from being used in practical child safety applications. False-positive rates, on the other hand (i.e., an adult classified as a child) can issue false alarms in child-detection applications and can be considered significant in applications such as school intruder detection systems. For this reason, using the same training and testing data as the previous evaluation for SVM, bin-based, and fuzzy method, we now examine the false-negative detection rates (i.e., detecting a child’s data as an adult) and the false-positive detection rates, and we present the results in Figures 11 and 12, respectively. Notice that compared to the false-positive rates in Figure 12, both systems succeed in bringing down the false-negative rates in Figure 11. These results also indicate that ChildSafe, especially with the fuzzy boundaries and a large enough learning set size (e.g., ≤100) successfully reduces the “error rates” by ~13.4% for false negatives and ~31% for false positives when compared to the C-SVC SVM approach with both systems utilizing only body features.

According to the results, we see there are more false positives than false negatives. The reason behind this is that there is a larger concentration of data that we collected from children than adults. Specifically, we collected 150 instances of data from children between the ages of 7 and 12, and 150 instances of data from adults between the ages of 14 and 50. Clearly, the adult data are more distributed over different ages and, therefore, we found more misclassified cases in this age group. This also links to the reason behind why ChildSafe has a noticeably significant lower false-positive rate when compared to the performance of C-SVC SVM. Our approach utilizes bins which consider a wide range of feature data within the same class to reduce the false-positive rates despite a certain age group having more distributed data. Ideally, we agree that a balanced (and larger) learning set can help both ChildSafe and the SVM approach show improved performance.
Fig. 12. False-positive detection ratios (detecting an adult data as a child) for ChildSafe with and without fuzzy bin margins with different sizes of learning sets compared with the false-positive rate of an SVM-based estimation approach.

Overall, we can conclude from the results that ChildSafe minimizes incorrect child classification and also minimizes the number of false alarms, which can potentially affect user experience with a child protection system.

Furthermore, in Figure 13, we plot the ROC curve of ChildSafe (with fuzzy bin boundaries) and the C-SVC SVM approach for varying learning set sizes. By observing the relationship between the true-positive and false-positive detection rates, we note that the overall detection accuracy of ChildSafe naturally improves with increasing learning set sizes and also note that the detection performance of ChildSafe is much more accurate compared to the C-SVC SVM approach with our collected dataset.

From a practical perspective, for ChildSafe to be utilized in a variety of potential applications, it should be able to determine and classify a child with minimal latency. Therefore, we now examine the detection latency of ChildSafe under different conditions and compare it with the latency of the C-SVC SVM-based classification. This experiment was done on a PC-scale machine with an Intel i7, 1.90GHz dual-core processor, 4GB of RAM, using Java. As we see in Figure 14, we note that the latency for ChildSafe is relatively lower than that of the SVM-based approach. Even in the worst case (i.e., the fuzzy bin-boundary approach), the latency is measured at <5 msec. These results provide strong evidence that ChildSafe can easily be used as a subsystem component within an application that targets the detection of children in dangerous situations or environments that require short response times.

Last, we perform an extra experiment to examine the performance of ChildSafe under biased training data (i.e., adult data are neglected in the training phase). In other words, we emulate a configuration where only child data are used as the learning data of the system. We use both child and adult data, however, for performance evaluation purposes. The design of ChildSafe makes this possible since bin frequencies can be computed even with the lack of adult data, using small
modifications to the original algorithm as discussed in Section 4.6. Moreover, since we are implementing a binary classification system with two target age groups (e.g., child or not), if the observed features do not fall into the child age group, then we can simply classify the sample as an adult.

In Figure 15, we present the accurate classification, false-positive, and false-negative rates with different learning set sizes for ChildSafe with the fuzzy bin-boundaries using only the body joint data. The SVM-based approach requires data from all output categories during the training phase; hence, we can only evaluate ChildSafe for this experiment. Furthermore, unlike the experiments where we included adult data as learning sets, since we have a sufficient number of child data (c.f., Table 1), we configure the learning set size to be up to 50 samples in this experiment. We note in Figure 15 that while the accuracy is slightly lower compared to the case where ChildSafe learns both adult and children data, ChildSafe is flexible enough to achieve satisfactory performance both in terms of classification accuracy and error rates. Such results suggest that ChildSafe is a robust system that can account for various types of inputs.

5.3 Comparison with Existing Age-Group Detection Studies

We now compare the performance of ChildSafe in context with other age-group detection-related studies. Table 3 evaluates the performance of ChildSafe with existing behavioral, kinematic, and Kinect-based works. The details of these works can be found in their respective papers and also in Section 6. Note that the values that we compare in Table 3 are reported from the respective papers [12, 20, 39, 52]. From this, we see that ChildSafe is tested with the largest dataset, more than 5 to 20 times the amount of data than that of previous studies. Despite this significant difference in the amount of data, which translate directly to the diversity and complexity of data, ChildSafe manages to provide a similar (if not better) accuracy. Additionally, we note that none of these works evaluates their capability to classify the target samples in real-time. Finally, we point out
Fig. 14. Comparison of detection latency for the SVM-based child detection approach against ChildSafe with and without the fuzzy bin margins.

Fig. 15. Classification ratio and false-positive and false-negative rates for ChildSafe with fuzzy bin margins when learning sample consists of only child body joint data (no adult learning data).
Table 3. Comparison with Existing Age-Group Detection Studies

<table>
<thead>
<tr>
<th>Work</th>
<th>Feature(s)</th>
<th>Data type</th>
<th>Ages groups</th>
<th># of subjects</th>
<th>Accuracy</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>Locomotion patterns</td>
<td>Visual features</td>
<td>3 to 5 (child), 30-52 (adult)</td>
<td>15</td>
<td>93-95%</td>
<td>N/A</td>
</tr>
<tr>
<td>[20]</td>
<td>Human motion</td>
<td>Silhouette</td>
<td>20 to 37 (young), 56 to 80 (elderly)</td>
<td>53</td>
<td>94.30%</td>
<td>N/A</td>
</tr>
<tr>
<td>[52]</td>
<td>Gait</td>
<td>Contour</td>
<td>25 to 30 (young), 60 to 65 (elderly)</td>
<td>14</td>
<td>83.33%</td>
<td>N/A</td>
</tr>
<tr>
<td>[39]</td>
<td>3D body motion</td>
<td>3D skeletal joints</td>
<td>7 to 10 (1st group), 11 to 16 (2nd group)</td>
<td>28</td>
<td>95.20%</td>
<td>N/A</td>
</tr>
<tr>
<td>ChildSafe</td>
<td>Body ratio</td>
<td>3D skeletal joints</td>
<td>7 to 12 (child), 14 to 50 (adult)</td>
<td>300</td>
<td>94%</td>
<td>&lt;5 msec</td>
</tr>
</tbody>
</table>

Fig. 16. Experiment run in two different real-world environments: (a) a hallway with an open space and (b) a living room that has occlusions and darker lighting.

that all of these previous efforts focus on analyzing patterns of body movement, while ChildSafe examines anthropometric information of the subject being monitored.

5.4 Evaluation in a Real-World Environment

To evaluate the performance of ChildSafe in a practical environment, we deploy our system in two different settings. As Figure 16(a) shows, the first environment is a school hallway that has minimal occluding objects in its surroundings. Figure 16(b) shows our second testing environment, which is a living room setting with furniture and lower lighting. We purposely ran our experiment in these two noticeably different settings so that we can see a clear difference in accuracy between open space and closed space areas. Note that a closed space reveals more noisy data since there are occlusions in the surroundings that can interfere with the subject being monitored. The experiment is performed using eight adults of different ages and height. Each of the adults performs 20 rounds and is asked to either walk, stand, move, or sit in each of the two locations. They are not constrained to direction, speed, or distance from Kinect. The purpose of this experiment is to evaluate and
compare ChildSafe in a natural environment with natural movements of individuals. Note that we train ChildSafe using 100 learning sets.

We compare the overall accuracy of ChildSafe and the SVM-based approach by averaging the results from all 160 rounds from each of the two locations. As we show in Figure 17, ChildSafe performs with an accuracy of 93.75% in the hallway. Furthermore, despite the presence of many occluding objects in the living room, ChildSafe successfully achieves a high accuracy of 87.08%. Using the SVM approach, we achieve an accuracy of 82.32% in the hallway and 75.13% in the living room environment. From this, we note that the SVM method is affected by noisy data slightly more than is ChildSafe.

One of few reasons that there is not a significant accuracy drop from a bright open space and dark closed space for both ChildSafe and the SVM-based approach is that using a depth camera allows a person to be detected even in areas of low lighting. Additionally, when using depth information, we using skeletal information that is less sensitive to occluding objects than other tracking methods, such as using a bounding box [27]. In our previous works, we identified that bounding boxes can interfere with objects in the background more easily and lead to conflicts with surrounding objects [50]. However, one direction of our future work is to find better ways to maintain a higher accuracy such as integrating multiple cameras for large-scale and occlusion-free monitoring, as demonstrated in state-of-the-art approaches [14, 17, 45].

5.5 Considerations for Practical Deployments

Based on the satisfactory results just discussed, we now outline some of the systematic elements that can affect the performance of ChildSafe and discuss how we overcome these challenges.

- **Body orientation:** Given that ChildSafe is a 3D depth camera-based system, the angle of the camera and how it takes pictures of the human object can be important since this can affect the physical space of skeletal joint orientation. Nevertheless, by only utilizing the ratios of the body parts derived from the joint positions in the system, we reduce the issue of the camera positioning. Using simple calibration techniques, we can simplify the process of accounting for different camera angles and extract the proper ratios (e.g., features) used in ChildSafe.
• **Variations in joint positions:** Throughout this work, we are utilizing data collected from people performing eight different actions as discussed in Section 3.2. It is true that the accuracy of the ChildSafe system, or any other classification method, can be significantly affected by the quality of the data received from the target person. Nevertheless, as discussed, ChildSafe only utilizes ratios as features. Therefore, even a single snapshot of the targeted human can be enough to extract the information of interest to our system (e.g., joint positions). The only issue here is that the calibration process to minimize any outlying measurements can take as long as 4–5 seconds. We envision that once ChildSafe is deployed to detect a large geographic region, the camera sensor can properly collect the information required to classify the human object that it observes as the human object moves into (or within) the target region.

• **Detecting multiple people:** We currently evaluate our system under the assumption that there is a single person within the target environment. This is a valid assumption given that the cameras can be deployed at the entrances to regions that can potentially be of danger (e.g., entrance to basement or kitchen). Nevertheless, when deployed to wider targeted regions, the system must be able to process data for multiple individuals. Based on our experiments, the processing overhead is not much of an issue, but we are restricted due to hardware configurations since the current Kinect sensor can detect up to only two individuals’ skeletal data. However, the Kinect 2 detects as many as six people’s skeletal data simultaneously. With further improvements in the hardware, along with dense deployments of sensors, each monitoring disjoint regions, we envision that ChildSafe can be easily incorporated in a variety of applications.

### 6 RELATED WORK

Computer-based automatic age estimation has become a particularly prevalent topic because of emerging real-world applications, such as security control and surveillance, forensic art, electronic customer relationship management, biometrics, and entertainment [16]. A number of approaches have been proposed using various physiological and behavioral characteristics to train models of different age groups.

**Physiological features:** Numerous studies have been proposed using facial information to estimate age [2, 19, 22, 42, 46–48, 53]. Although appealing, as we introduced earlier in Section 4.4, the downside of using facial data is that they can easily be hidden.

The system proposed by Mirhassani et al. estimates age groups using acoustic analysis of the speech by dividing the data into six groups based on the vowel classes [49]. Similarly, Meinedo and Trancoso present an age and gender classification system using the fusion of four and six individual subsystems trained with short- and long-term acoustic and prosodic features, different classification paradigms, and different speech corpora [29]. The weakness of these approaches is that they require an active engagement with the person while collecting detection data, which is inconvenient and reduces the system’s usability. Compared to these techniques, ChildSafe is a completely noninvasive monitoring system.

Holding the similar motivation of supporting the safety of children in school, Rajamaki et al. propose that the use of identification cards, wristbands, and necklaces that are linked with video surveillance can provide the solution for identifying only authorized people as having access to building facilities [36]. Lee et al. present an approach toward child surveillance using complex event processing with sensor-based signals and deploy their system in a real-world school environment [23]. The major drawback of these work is that they do not provide any in-depth evaluation results to prove their effectiveness. We, in contrast, present details on the data collection process and discuss quantitative results to demonstrate a reliable system for child-present environments.
Behavioral and kinematic features: Studies on walking patterns are performed to understand their association with age. Davis presents an approach for visual discrimination of children from adults using stride-based properties of their walking style [12]. Handri et al. propose an age and gender classification method using computer vision and machine learning to extract shape-motion by discrete wavelet and fast Fourier transform based on spatiotemporal information [20]. AdaBoost algorithms were used afterward to analyze age and gender discrimination based on extracted features. In addition, Zhang et al. propose a framework for age classification based on human gait using hidden Markov models [52]. The downside of these techniques is that they only use two-dimensional RGB images and do not provide skeletal-level information to accurately classify different age groups. Additionally, compared to video images, the skeletal information from Kinect is far less privacy intrusive. A study by Demiris et al. [13] explores older adults’ privacy considerations for technology. The study shows that older adults are more amenable to vision-based systems if they use anonymized imaging systems. Participants in this study expressed no privacy concerns with anonymized images and emphasized their appreciation for not being recognizable in the video sequences.

Kinect-based systems: Although the general topic of automatic age estimation received considerable interest, only a small number of researchers carried out research on this topic using depth information. As an example, Clavin et al. use a three-dimensional human skeletal model to measure three different body joint ratios: (1) head width to shoulder, (2) arm length to body height, and (3) body height to head height ratio. Using these metrics the authors estimate a person’s age to provide parental control settings for media applications [10]. However, the system only utilizes a subset of features compared to ChildSafe, and the work does not provide any specific details on how the estimations are executed, nor do they present any detection-related evaluations. Our work presents a more empirical and comprehensive approach in proposing a system that detects children’s presence for designing safety applications for children. On the other hand, a study by Sandygulova proposes a neural network–based learning architecture to estimate child age and gender based on 3D body motion information extracted from Kinect [39]. Though this study can classify with high accuracy, the range of ages that the system can detect is small and only concentrates on detecting between two groups of children. Apart from other previous work, our own previous work presents an initial version of the ChildSafe system using bin-based classification [5]. In addition to detailed explanation of the scheme and application scenarios of how ChildSafe can be used in real applications, this work improves our previous work technically in three different ways. First, we introduce the concept of fuzzy bin-boundaries to ChildSafe, which helps smooth the discrete decisions that “fine-cut” bins can cause. Second, we test for cases where the facial information is not used since there are situations when a person’s face is covered. Finally, third, we present a more thorough set of evaluations to showcase the effectiveness of ChildSafe’s performance.

Mining fuzzy association rules: Madden et al. propose a solution to the sharp boundary problem using Fuzzy Set theory [7]. The proposed method is similar to the fuzzy boundary solution we propose in this work; however, we eliminate the problem of overemphasized rules, which stems from the sum of bin memberships being greater than 1, by always having a total certainty of 1 across memberships. Moreover, unlike the process described in [7], the system we propose in this article does not require user input in order to generate fuzzy association rules.

7 DISCUSSIONS AND SUMMARY
This work presents one of the first efforts to systematically detect the presence of children in a targeted environment using 3D depth camera sensors. Specifically, we utilize a Microsoft Kinect camera sensor to collect facial and body joint features from people of various age groups, and, with
this data, we design a bin-based learning scheme for classifying children from adults, ChildSafe. We further optimize the system so that ChildSafe achieves high classification ratios even with the lack of facial features by including a method that fuzzifies the bin-boundaries of the initial ChildSafe design. Our evaluations show that ChildSafe classifies children properly with an accuracy of up to 94%, and our performance comparisons with a C-SVC SVM shows that this classification ratio is ∼9% higher, with a false-positive rate that is ∼31% lower and a false-negative rate that is ∼13% lower.

Despite ChildSafe showing very high detection accuracy with the dataset that we collected, we acknowledge the fact that there are more than 60 billion people across a diverse set of ethnic groups on Earth, all with unique characteristics. Therefore, ChildSafe will definitely need further improvement to suit the entire population. Nevertheless, as our evaluations suggest, the flexible design of ChildSafe allows the system to be adaptive to various learning datasets. We envision that such a system can act as a catalyst to finally realize useful applications that can protect our children from potentially dangerous environments.

REFERENCES


Received March 2016; revised April 2017; accepted April 2017