Classifying Children with 3D Depth Cameras for Enabling Children's Safety Applications

Can Basaran[1,2], Hee Jung Yoon[1], Ho Kyung Ra[1], Sang Hyuk Son[1], Taejoon Park[1], and JeongGil Ko[3]
[1]Daegu Gyeongbuk Institute of Science and Technology (DGIST), [2]METU Northern Cyprus Campus
[3]Electronics and Telecommunications Research Institute (ETRI)

ABSTRACT
In this work, we present *ChildSafe*, a classification system which exploits human skeletal features collected using a 3D depth camera to classify visual characteristics between children and adults. ChildSafe analyzes the histograms of training samples and implements a bin-boundary-based classifier. We train and evaluate ChildSafe using a large dataset of visual samples collected from 150 elementary school children and 43 adults, ranging in the ages of 7 and 50. Our results suggest that ChildSafe successfully detects children with a proper classification rate of up to 97%, a false negative rate of as low as 1.82%, and a low false positive rate of 1.46%. We envision this work as an effective sub-system for designing various child protection applications.

Author Keywords
Child Classification, Kinect-based Applications

ACM Classification Keywords
I.5.2 Computing methodologies: Design Methodology

INTRODUCTION
The need to provide safe living environments, coupled with the advancements in sensing technologies have introduced a diverse set of human identification and classification systems for various surveillance applications [3, 4, 12, 13, 14, 15]. Nevertheless, among such work, it is surprising that there is only a limited number of systems that target child safety applications, which represent one of the weakest population sets [5, 11]. An essential feature that these systems require is the capability to detect the presence of children in a target environment. By doing so, the application system will be able to issue alert messages or change the envrionmental parameters to ensure a safe environment when children are present.

In this work, we try achieving this goal of effectively classifying children using 3D depth cameras. By designing such a classification system, we aim at opening the possibilities for developing novel applications that contribute to preserving the safety of children. While the characteristics of each person is 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 [1, 8].

UbiComp'14, September 13 - 17, 2014, Seattle, WA, USA

http://dx.doi.org/10.1145/2632048.2636074

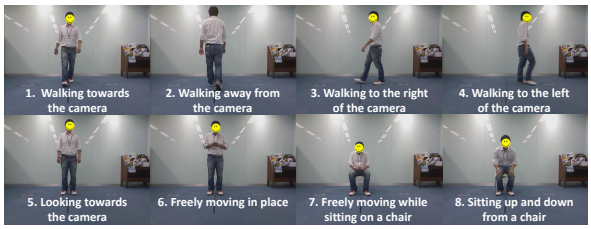


Figure 1. Eight different actions performed by each participant for our visual data collection phase.

Age Group	Label	Age	# M	# F	# Total
child	Child-1	7	27	13	39
	Child-2	8	10	17	27
	Child-3	9	9	22	31
	Child-4	10	6	9	15
	Child-5	11	12	7	19
	Child-6	12	12	7	19
adult	Adult-1	13-19	5	6	11
	Adult-2	20's	14	6	20
	Adult-3	30's	5	2	7
	Adult-4	40's	4	0	4
	Adult-5	50's	0	1	1

Table 1. Distribution of the 193 subjects across age groups and genders. M: Males, F: Females

Using a Microsoft Kinect camera sensor, we start this work by collecting joint position data from 193 people, spanning a diverse set of age groups that can be classified as children, i.e., elementary school students, and adults, i.e., post-teenage to senior. These data samples are arranged so that we can easily extract various parameters on the body joints and facial points. We propose the ChildSafe classification system that is based on the histogram analysis of the training data that accurately classifies the observed human as a child or an adult. Evaluations using our data set show that with large enough training data, ChildSafe can correctly classify children with an accuracy of up to ∼97%. Performance comparisons against a system based on C-Support Vector Classifier Support Vector Machines (C-SVC SVMs) show that ChildSafe achieves higher classification accuracy while maintaining low classification errors of only 1.46% false positives and 1.82% false negatives.

DATA COLLECTION PROCESS
Our work starts with a data collection phase where we use a Microsoft Kinect camera sensor to collect facial and body joint information from a number of subjects. Specifically, we collected time-stamped body and facial

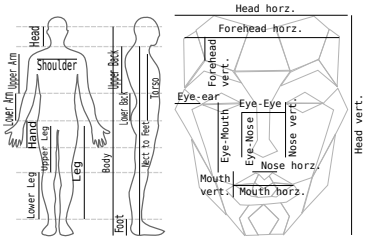


Figure 2. Body joint and facial metrics used in this study.

data from 150 children and 43 adults, which include recordings of eight different actions as shown in Figure 1.

In this work, as in many studies, we define *children* as a human between the stages of birth and puberty [9]. Since detecting infants or kindergarten children is relatively simple (due to noticeable differences in height and posture), we focus on the proper detection of elementary school students. Therefore, we collect a children data set of 150 elementary school students. The adult data set, consisting of 43 samples, was collected from individuals between the ages of 13 to 50 [9]. We show the distribution of participants in Table 1.

CHARACTERIZING HUMAN JOINT INFORMATION

Using the collected body and facial data, we built a knowledge base, which consists of positions of the body joints and face points. As mentioned, each subject, $p \in P$, followed a predetermined script and executed a sequence of actions, $a \in A$. 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, $g \in G$. Based on this knowledge base, we identify useful features using the following steps.

• **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., $s_n = \{j_{n1}, j_{n2}, \dots\}$ 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 Figure 2. The length of each body part, $L(t, s_n, a, p)$ and facial length, $L(t, f_m, a, p)$, is computed by adding the distance between directly connected joints and facial points:

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

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

• **Step 2:** In this step, we compute the average value for each body and facial metric separately for each individual across different actions (Eqn. 3). This process essentially cleans the data while also reducing the computation complexity in the steps to follow. By including various actions, we address the fact that depth cameras report slightly different measurements based on the body posture. Furthermore, diversifying the training set, 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.

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

• **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 head length, 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 ratio of a body part's length s_i, to another body part s_j is computed using Equation 4.

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

We refer to these ratios as *body features* and *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 $R(\exists, p)$ such that $\exists$ represents a feature as the ratio of an arbitrary pair of facial or body lengths. The main reason behind this design choice is that methods based on absolute measurements, e.g., height, are more vulnerable to measurement errors compared to relative metrics.

A BIN-BASED CLASSIFIER FOR IDENTIFYING CHILDREN

ChildSafe is a bin-based classifier that uses 3D depth cameras to detect the presence of children in a geographical region. We use bin-based classifiers ChildSafe since they provide an empirical occurance-based classification method, rather than "best fit"-based approachs that require the establishment of a learning model. When interacting with application systems, ChildSafe acts as a core sub-system with the role of properly detecting the presence of children to assist the application system in making proper actuation decisions. Specifically, using the extracted features, ChildSafe performs the following steps to achieve robust classification.

• **Bin Creation:** ChildSafe starts by creating 25 *bins* for each feature such that each sample $R(\exists, p)$ from a feature can be mapped to a single bin, $\beta(R(\exists, p)) \rightarrow b$, i.e., a discrete value within the closed range [0,24] according to Equation 5. We use 25 bins based on empirical performance validation using our data set.

$$\beta(R(\exists, p)) = \lfloor \frac{R(\exists, p) - V_{min}}{(V_{max} - V_{min})/25} \rfloor \quad (5)$$

Here, V_{max} takes the maximum of $R(\beth, \forall p_o \in P)$ and V_{min} takes the minimum. Note that bin-boundaries are global across age groups due to $\forall p_o \in P$ in Equation 5; therefore, computing 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, an age group, and a bin number b. Hence, a frequency tuple $Freq(\beth, g, b)$ refers to the number of samples from an age group falling into b, normalized to the total samples of the age group (Eqn. 6); thus, represents the likeliness of a feature in a given age group falling into that bin.

$$Freq(\beth, g, b) = \frac{|\beta(R(\beth, p_o \in g)) \in b|}{|\beta(R(\beth, p_o \in g))|} \tag{6}$$

• **Weight Computation:** Lastly, 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 (Eqn. 7).

$$W(\beth, g, b) = \frac{Freq(\beth, g, b)}{\sum_{g_o \in G} Freq(\beth, g_o, b)} \tag{7}$$

These weights are used for mapping crisp features into membership degrees of age groups by first computing the b, for a sample, and then calculating the weight, i.e., the certainty of membership, via Equation 7.

Classification Process Operations

From the three phases, ChildSafe is designed so that it **(1)** takes as input the body and facial features of a human, **(2)** identifies the bins that each feature of the target falls into, and **(3)** makes an estimate on the age group through the classification process. All features from an individual are separately processed to obtain per-feature membership degrees to an age group. Since the *minimum* is the standard intersection operator in fuzzy set theory [10, 16], per-feature estimations are aggregated by taking the minimum of all membership degrees for each age group.

ChildSafe finally classifies the subject into the age group with a stronger membership degree. Hence, given the membership degrees to each age group, the subject is classified as a child if $\mu_{child} \geq \mu_{adult}$, else, as an adult.

EVALUATION

Before evaluating ChildSafe we first start by constructing a fair and competitive comparison scheme. Specifically, we use a Support Vector Machine (SVM)-based method, a widely used machine learning approach in various applications. We chose SVMs given that recent studies show that the SVM outperforms other machine learning algorithms when classifying coordinate data sets [7]. For implementing an SVM-based classifier, we use LibSVM [2] and configure the parameters as default, unless specified. In detail, we use the C-Support Vector Classification (C-SVC) SVM [2], which is suitable for binary classification. When utilizing C-SVC SVM, one of the most critical parameters affecting the performance is the *cost value*. When large, the SVM chooses a smaller-margin hyperplane as long as it well-classifies the training samples. On the other hand, a small cost value leads to larger-margins despite some mis-classified samples. Logically, the cost value also determines the complexity of C-SVC SVM; thus, it is good to keep a low cost value to minimize the complexity while maintaining it high enough to ensure accuracy. We empirically determine the cost value for our target data set to 10^7. Furthermore, we note that by comparing the performance of linear, polynomial, and radial SVM kernels, we observed a ∼10% higher accuracy when using the linear SVM. Therefore, we utilize the linear SVM kernel.

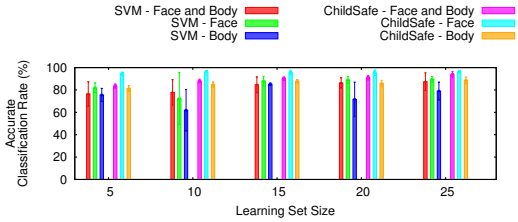
Body Features	Facial Features
arm / neck to feet	nose vert / fhead vert
hand / body	nose vert / eye eye
lower arm / shoulder	eye nose / nose hrz
lower arm / neck to feet	eye ear / eye eye
upper back / upper leg	head vert / eye nose
neck to feet / head	fhead hrz / eye nose

Table 2. List of features used in our experiments.

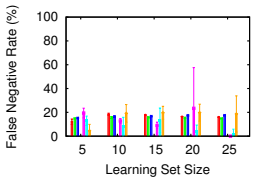
Using this, we now compare the performance of ChildSafe under different configurations (e.g., varying feature sets and learning set sizes). All experiments were executed five times with randomly selected training sets, and the rest were used for testing. The decision of using ratios, results in an explosion in the possible features. To this end, we generated all possible features by considering all pairs of lengths and eliminated those that looked similar, or otherwise insignificant, across age groups. After eliminating these features, we calculated the information gain per feature and selected features providing the highest information gain [6]. The resulting features used for ChildSafe and C-SVC SVM are organized in Table 2.

Figure 3 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, the classification accuracy increases as the learning data set grows, since more learning data will well-train the systems. Nevertheless, compared to the C-SVC SVM, ChildSafe shows higher accuracy. Quantitatively, this is an improvement of ∼37% in the most extreme case. We note that in comparisons with the K-Nearest Neighbors classifier, ChildSafe showed ∼43% higher performance. By analyzing the mis-classified cases, we noticed that the SVM was weak to the measurement noise embodied in the collected data. ChildSafe, on the other hand, is more robust to measurement noise since the bins smooth the data and using frequencies reduces the contribution of outliers.

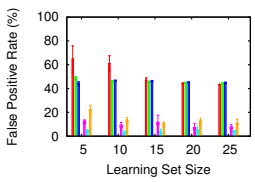
While accurately classifying children is of utmost importance, high false detection rates cause faults at the larger application. For this, we plot the false negative detec-



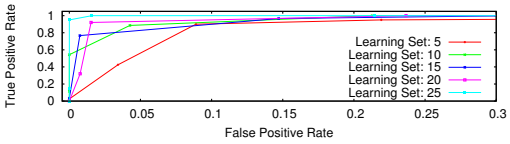
(a) Detection Accuracy



(b) False Negative Rate



(c) False Positive Rate



(d) ROC Curve for ChildSafe

Figure 3. Detection accuracy, false negative and false positive rates for ChildSafe with different learning set sizes compared to a SVM-based estimation.

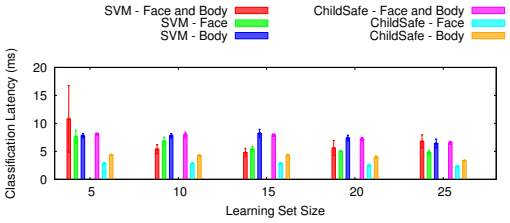


Figure 4. Comparison of detection latency for the SVM-based child detection approach against ChildSafe.

tion rates, i.e., detecting a child's data as an adult, and the false positive detection rates, i.e., detecting an adult as a child, in Figures 3(b) and 3(c), respectively. These results indicate that ChildSafe, with a learning set size $\geq$20, successfully reduces detection errors compared to the C-SVC SVM. Lastly, in Figure 3(d) we present the ROC curve for ChildSafe, from which we can notice that, with a learning sample size of 25, ChildSafe shows a reliable classification performance. Based on these results, we conclude that ChildSafe minimizes the incorrect classification of children and also minimizes the number of false alarms which can potentially affect the user experience of a child protection system.

On a practical perspective, for ChildSafe to be used in a variety of applications, it should be able to operate with minimal latency. For this, we examine the detection latency of our schemes of interest. This experiment was done on a PC with an Intel i7, 1.90GHz Dual-core processor and 4GBs of RAM. As Figure 4 shows, the latency of ChildSafe is relatively lower than that of the SVM. This result provides evidence that ChildSafe can be easily used as a sub-system component within an application for low-latency demanding applications.

Considerations for Practical Deployments

We now outline systematic issues that affect ChildSafe's performance and discuss how we overcome them.

• **Body orientation:** Given that ChildSafe uses a 3D depth camera, the angle and position of the camera can affect the physical space skeletal joint orientation. To address this, we utilize the "ratios" of the body parts rather than the measured length. Using simple calibration, we can simplify the process of accounting for different camera angles when extracting the features for ChildSafe.

• **Variations in joint positions:** Throughout this work we utilize data collected from people performing eight different actions. It is true that the accuracy of the ChildSafe system can be affected by the quality of the subject's image. Nevertheless, since ChildSafe utilizes only the features' ratios, even a single snapshot of the target is enough to extract required joint positions. Furthermore, considering a person's walking speed, we envision that the Kinect can gather multiple images of the objects to select the best quality image.

• **Detecting multiple people:** We currently evaluate ChildSafe under the assumption that a single person is present in the environment. Nevertheless, when deployed practically, the system should process data for multiple individuals. Based on our experiments, the processing overhead is not much of an issue, but rather face hardware limitations since the current Kinect can detect up to only two individuals' skeleton data. Nevertheless, the Kinect 2 detects as many as six people simultaneously. With further hardware improvements and dense camera deployments, we envision that ChildSafe can be easily applied to a variety of applications.

DISCUSSIONS AND SUMMARY

This work presents our efforts to systematically detect the presence of children in a target environment using 3D depth cameras. Specifically, we exploit a Microsoft Kinect camera for collecting facial and body joint features from people of various age groups, and with this data, we design a bin-based classification scheme for distinguishing children from adults, *ChildSafe*. The high accuracy of ChildSafe, as our evaluations show, suggests that it can be applied to larger application systems to provide feedback on the presence of children in the target environment, to ensure users that children in a target environment are safe.

Acknowledgements

The authors would like to thank the reviewers for their comments and Hyenpung Elementary School in helping with the data collection. This work was supported by the DGIST R&D Program of MSIP of Korea (CPS Global Center) and by the IT R&D Program of MSIP/KEIT project #10035570.

REFERENCES

1. E.C. Burns, J.M. Tanner, M.A. Preece, and N. Cameron. Final Height and Pubertal Development in 55 Children with Idiopathic Growth Hormone Deficiency, Treated for Between 2 and 15 Years with Human Growth Hormone. *European Journal of Pediatrics*, 137(2), 1981.
2. C.C. Chang and C.J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM TIST*, 2(3), May 2011.
3. W. Dai and J. Ge. Research on Vision-based Intelligent Vehicle Safety Inspection and Visual Surveillance. In *CIS*. IEEE, 2012.
4. Y. Gu, M. Kim, Y. Cui, H. Lee, O. Choi, M. Pyeon, and J. Kim. Design and Implementation of UPnP-Based Surveillance Camera System for Home Security. In *ICISA*, 2013.
5. R.K. Lee, C.H. Yu, M.S. Liang, and M.W. Feng. An Approach to Children Surveillance with Sensor-Based Signals Using Complex Event Processing. In *ICEBE*, pages 596–601. IEEE, 2009.
6. Wenke Lee and Dong Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143. IEEE, 2001.
7. M. Luštrek and B. Kaluža. Fall Detection and Activity Recognition with Machine Learning. *Informatica*, 33(2), 2008.
8. W.V. Mechelen, J.W.R. Twisk, G.B. Post, J. Snel, and H.C.G. Kemper. Physical Activity of Young People: The Amsterdam Longitudinal Growth and Health Study. *MSSE*, 32(9), 1981.
9. G.C. Patton and R. Viner. Pubertal Transitions in Health. *The Lancet*, 369(9567), 2007.
10. W. Pedrycz. *Fuzzy Control and Fuzzy Systems*. Research Studies Press Ltd., 1993.
11. J. Rajamaki, P. Rathod, A. Ahlgren, J. Aho, M. Takari, and S. Ahlgren. Resilience of Cyber-Physical System: A Case Study of Safe School Environment. In *EISIC*, 2012.
12. T.D. Raty. Survey on Contemporary Remote Surveillance Systems for Public Safety. *SMCS*, 40(5), 2010.
13. C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. *TCSVT*, 21(5), 2011.
14. C.C. Tao. A Two-Stage Safety Analysis Model for Railway Level Crossing Surveillance Systems. In *ICCA 2009*, 2009.
15. W.K. Wong, H.L. Lim, C.K. Loo, and W.S. Lim. Home Alone Faint Detection Surveillance System Using Thermal Camera. In *ICCRD*, 2010.
16. L.A. Zadeh. Fuzzy Sets. *Information and Control*, 8(3), 1965.