Towards Skeleton Biometric Identification Using the Microsoft Kinect Sensor

CONFERENCE PAPER · JANUARY 2013

DOI: 10.1145/2480362.2480369

3 AUTHORS, INCLUDING:

Ricardo Matsumura Araujo
Universidade Federal de Pelotas
39 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Gustavo Graña
Universidade Federal de Pelotas
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Towards Skeleton Biometric Identification Using the Microsoft Kinect Sensor

Ricardo M. Araujo
PPGC - CDTec
Federal University of Pelotas
Pelotas, RS, Brazil
ricardo@inf.ufpel.edu.br

Gustavo Graña
CDTec
Federal University of Pelotas
Pelotas, RS, Brazil
gustavoggs@gmail.com

Virginia Andersson
PPGC - CDTec
Federal University of Pelotas
Pelotas, RS, Brazil
virginia.andersson@gmail.com

ABSTRACT

In this paper, we consider the viability of using Microsoft Kinect sensor to extract skeleton points from walking subjects and use these points for biometric identification. We do so by capturing several subjects using the sensor, calculating the length of several body parts inferred from the extracted points and training a model for later classification using these lengths and labels identifying the subjects as training examples. We consider the cases where one wants to discriminate each subject individually and where only recognizing a single subject is enough, showing that in both cases a Nearest Neighbor algorithm is able to achieve high accuracy when considering a relatively small group of subjects. However, our approach requires a moderately large number of training examples and we discuss the impact of such caveat in certain scenarios. Finally, we consider the contribution of different combinations of body parts to the identification process.

Categories and Subject Descriptors

I.2.1 [**Computing Methodologies**]: Artificial Intelligence: Applications and Expert Systems; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection: Authentication

General Terms

Experimentation,Performance

Keywords

Biometrics, Machine Learning, Anthropometry

1. INTRODUCTION

The automatic identification of individual people is a task required in several applications, most notably in security scenarios where access to resources are only allowed to prespecified individuals, or where a multi-user system is customized to the current user [16, 17]. Biometric identification is a convenient way to accomplish this task, since they do not require individuals to carry authentication tokens (e.g. keys, cards) or remember user names and passwords. Additionally, when using biometric identification the subject is ideally required to be physically present at the point of identification.

Biometric identification systems can be understood as pattern recognition systems [16] that stores information on physical attributes of subjects of interest that allows it to subsequently match a subject to one of the stored identities. They can be divided into two approaches. An active biometric system requires the subject to interact in some way with an interface. This is the case of fingerprint recognition, arguably the most widely used biometric identification system. In passive (or non-cooperative) biometrics systems, on the other hand, the subject is not required to directly interact with the system, or even be aware that any identification is taking place. Facial and voice recognition fall into this category and are becoming increasingly common in daily tasks (for instance, mobile phones running Android can employ face recognition to unlock the operating system).

Another approach to passive identification is to use measurements of several different parts of the body, such height or arms and legs length. The assumption behind this approach is that the combination of these measurements are shared by very few individuals. One possible advantage of this approach is the ability to use coarser measurements that may be obtained from relatively low-resolution photos and videos, in contrast to facial identification that requires a detailed mapping of a face, which is a much smaller target than a whole body.

There is a number of scenarios where such passive identification may be required or beneficial. Two scenarios are of our interest. In user recognition tasks, a system is required to recognize (possibly even authenticate) a user and either allow access or customize some interface to the recognized user [16, 17]. In this case, there can be a reasonable long training stage, where the system may collect information from the user. Moreover, only a limited number of subjects are required to be recognized.

For another scenario, we consider the task of person re-identification [2], where the goal is to recognize when a subject was seen before by the system (e.g. to estimate the number of unique people in some area). There is no proper training stage in this case, since subjects may not be aware of the system, and typically subjects are moving (e.g. walking down on a corridor or aisle). Furthermore, a potential large number of subjects have to be identified. While we

SAC'13 March 18-22, 2013, Coimbra, Portugal.

focus on the former scenario, we make considerations on the usefulness of our approach to the latter.

In this paper, we tackle the problem of performing individual identification using passively collected information about physical attributes from subjects. For this purpose, we use data provided by the Kinect device, introduced by Microsoft as a companion to its X-Box video game console. The Kinect contains sensors that allows it to recognize the position in space of several body parts from a person standing or moving in front of it, thus enabling the person to control the console using body movements and gestures.

We captured data from several subjects in motion, calculating their members' lengths in several video frames. We then proceed to train models using machine learning algorithms to classify individual subjects, testing different models and different attributes sets. We report good success in the task, with an overall accuracy close to 99% in the best case. However, we show that relatively large training sets are required to reach such levels of accuracy, limiting its usefulness for person re-identification.

This paper is organized as follows. Section 2 discusses related work and states how the present work differentiates from them. Section 3 defines our goals and methodology used throughout the paper. Section 4 presents our main results and Section 5 concludes the paper and details future research lines.

2. RELATED WORK

While originally intended as a video game controller, since its launch in 2010 the Kinect sensor has seen several applications outside gaming. Examples of such applications include tracking people across rooms [6], creating interfaces with robots [5], robot control [15] and detecting human poses [4]. For this purpose, Microsoft provides an SDK (Software Development Kit), that allows the Kinect to communicate with a PC [1].

Biometric identification is built-in in X-Box consoles using the Kinect. However, the only anthropometric information used is the users' height; it mostly relies on facial recognition complemented by clothes' colors [10].

Biometric identification based on body measurements was moderately common in the late 1800's in criminal investigations [1]. However, taking measures of several body parts was very time consuming and prone to human errors and this type of system was supplanted by fingerprint identification. With the popularization of photographic and video cameras, along with more precise measurement systems [11], anthropometric identification has seen a rise in interest, in particular in areas such as person re-identification (i.e. recognizing whether a person was seen before) [2,14]. A related line of work is the attempt to use body movements (such as walking) as a source for biometric identification [7,13].

To the best of our knowledge, there are not prior work that try and analyze anthropometric identification using Kinect data, especially with moving subjects.

[1]http://www.microsoft.com/en-us/kinectforwindows/

3. GOALS AND METHODOLOGY

The main goal of this paper is to assess the viability of using Kinect's data to create a biometric identification system. In particular, we are interested in using static anthropometric measurements (i.e. without considering motion) for this purpose, but without requiring subjects to stand still in front of the sensors (so that the system is fully passive).

Specific goals include: (i) apply and test different machine learning algorithms to train classifiers; (ii) compare the difficulty of identifying multiple subjects (multi-class classification) versus identifying a single subject (binary classification) and (iii) identify relevant attributes to perform an accurate identification.

3.1 Data Capture and Pre-Processing

Our approach starts by capturing data from different volunteers. In this paper, we report results from 8 volunteers (2 female and 6 male), all from the same age range (25-35 years-old) and without (noticeable) physical problems. The Kinect provides *frames* containing information from its sensors, at a rate of 30 frames per second. We only used data regarding the captured skeleton, which is composed of three spatial coordinates for the points depicted in Figure 1. This data was captured while subjects walked in front of the sensor.

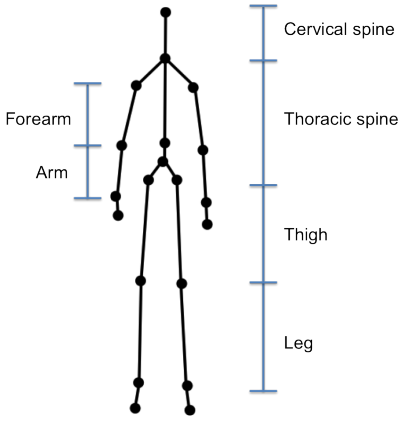

Figure 1: Skeleton points returned by Kinect's API (circles) and defined body parts.

Each subject performed ten walks in different directions relative to the sensor (five perpendicular and five parallel), in order to obtain a good diversity of scenarios. On average, each walk took about 4 seconds and each subject generated a total of 1400 frames across all walks. For each frame, we calculated the length of different parts of the body based on the coordinates provided by the sensors. The used parts are depicted in Figure 1 and listed below[2]:

- Left and right arm
- Left and right forearm
- Left and right leg
- Left and right thigh

[2]The names are not anatomically accurate, but rather rough representatives of the general area captured

- Thoracic spine
- Cervical spine
- Height

Height was calculated by summing the Cervical Spine, Thoracic Spine, the mean between left and right legs and the mean between left and right thighs. It must be noted that clearly these attributes are not independent from each other - indeed, they are expected to be highly correlated. We explore the relative importance of each part in a later section.

At this point, each frame of each subject is composed of a label (the subject's identification) and an attribute list containing the lengths of the 11 different parts. We observed that the data contained many noises, evidenced by anatomical impossibilities of the positioning of members. In order to reduce this noise, we filtered the data by calculating the averages and standard deviations of each attribute, detecting any measurements that were beyond two standard deviations from the average. Whenever we detected an outlier in at least one attribute we removed the entire frame containing this attribute. Close to 15% of frames were removed during this pre-processing.

By the end of the pre-process stage we had 11200 rows of data, each row representing a frame with 11 numerical attributes and labeled with the subjects' identification. From this set, we randomly selected 1000 frames for each individual, so that all individuals had the same number of frames associated to them.

3.2 Training Models

We propose training a model using the data set in order to be able to identify the targeted subjects. This task can be seen as a machine learning task. We tested three models and four learning algorithms on our data: Multi-layer Perceptrons (using Backpropagation, with 10 hidden units [3]) [9], Decision Trees (using C4.5 [12] and Random Forests [3]) and K-Nearest Neighbors ($K = 1$) [12]. We used the implementations provided by Weka, version 3.6.7 [8]. As we'll show, K-Nearest Neighbors (KNN) outperforms the other methods and is used for all other experiments.

In order to accomplish our goals, we considered two cases. In the first, we trained our model using all labels, leading to a multi-class classification problem. Therefore, a successful model must correctly discriminate all individual subjects. This is the case when we want to track and recognize several subjects at once. In the second experiment we trained our model using frames from a single individual as positive examples, while all other individuals' frames were labeled as negative examples. This binary classification problem covers the scenario where one desires to authenticate or recognize a single user (e.g. personal devices, such as smartphones or tablets, typically employ a single-user authentication scheme).

Except where stated, presented results were calculated by using a 10-fold cross-validation method. We observed the overall average accuracy of the trained models as well as their specificity and sensitivity. Whenever a claim of statistical significance is made, we used a t-test with 95% confidence interval.

[3]We tested several different configurations and found that a single layer with around 10 hidden units provided the best average accuracy.

Table 1: Accuracy and total number of hits and misses for each learning method. Accuracy is given by the fraction of correctly classified instances of a 10-fold cross-validation.

Method	Accuracy	Hits	Misses
Multi-Layer Perceptron	83.0%	6642	1358
C4.5	92.4%	7392	608
Random Forests	97.5%	7800	200
Nearest Neighbor	99.6%	7975	25

Table 2: Confusion matrix for a 10-fold cross-validation training of the Nearest Neighbor algorithm.

Subject	1	2	3	4	5	6	7	8
1	993	1	0	2	0	4	0	0
2	0	998	0	0	0	1	0	1
3	0	0	998	0	0	0	0	2
4	2	1	0	995	1	0	1	0
5	0	0	0	0	1000	0	0	0
6	2	0	0	1	0	996	1	0
7	0	1	0	1	0	0	998	0
8	1	0	0	0	0	2	0	997

4. RESULTS

4.1 Comparing Learning Algorithms

For our first experiment we trained each model to identify individual subjects. Table 1 shows the results of a 10-fold cross-validation training for all tested learning algorithms. All results are statistically significant. KNN greatly outperform the other methods, achieving a 99.6% accuracy. This method also has as an advantage that training is incremental and very fast, since it only stores the observed examples. However, Random Forests performed well and, since it builds an explicit model, it can be much faster to provide a classification after training.

While these results are promising, the problem of identifying a subject using a single frame can be too strict. In a real-world scenario, frames are not independent - i.e. consecutive frames are likely to contain information from the same subject. If we allow for the classifier to take a majority vote on the classification of N consecutive frames, accuracy can be improved. Indeed, initial results shows that for $N = 3$, KNN was able to make no mistakes at all in our data set. Nonetheless, we focus the remaining of this paper on single-frame identification, as it provides a basic unit for the task. Moreover, we also focus the remaining of the paper on the KNN method, as it was the best-performing among the tested methods.

Table 2 shows the confusion matrix when using KNN. The results are very satisfactory, with good and consistent recall rate across all subjects. Even among subjects that visually have similar height and physical attributes (e.g. 5 and 6), there are no consistent misclassification between them.

4.2 Attributes

While using as many attributes as possible is likely to im-

Table 3: Accuracy for different combination of attributes, ranked by accuracy.

Combination	Accuracy
All attributes	99.60%
Spines + Left Arms and Legs	97.13%
Legs and arms	97.09%
Arms only	81.32%
Left legs and arms	79.55%
Legs only	66.51%
Spines	41.80%
Height	32.34%

prove accuracy, one could ask whether a smaller set could explain most of the shown results. In particular, the X-Box console uses height as the only anthropometric signal (relying mostly on facial recognition) [10]. To answer this question, we run the nearest neighbor algorithm using different subsets of attributes.

Table 3 shows the results of 8 different combinations. The first thing to notice is that using only one side of the body (left or right arms and legs only) decrease performance considerably. This is likely to be an artifact of our data, since not all side walks were performed on both sides (i.e. some subjects walked more from right to left than the other way around). This forces the sensor to infer the positions of body parts that are not visible to it, hence reducing information available for the learning algorithm.

Using information only from both legs and arms showed a reasonable performance, explaining most of the accuracy from the full attribute set. Adding spine information increase accuracy by 1.8%. Arms were more useful than legs and using only information from the spines led to a poor performance. Interestingly, using only height information had the worse performance.

It must be noted that we only use a limited number of subjects, hence these results are biased towards the stature and physical attributes of this particular group. Nonetheless, the results are useful to better understand the interaction between attributes. It is clear that no single attribute is solely responsible for the high accuracy observed when using the full attribute set. Using a combination of attributes is important for an accurate classification.

4.3 Pose Dependency

The previous results used the full data set, which consisted of data captured with subjects walking orthogonally to the sensors and in parallel to the sensors. Since we used cross-validation to assess the model's accuracy, both training and test sets contained both types of walks. An important issue is whether it is sufficient to capture data in only one pose - i.e. if Kinect data is pose independent.

Table 4 shows the multi-class results when a model is trained using data captured from only one type of walk and tested on the other type of walk. It is possible to observe that the results are much worse than when training using both types of walks, even though they are still much better than random guess.

The observed difference (16.8% and 23.8%) was not found to be statistically significant, an evidence that no particular pose provides more information when compared to the other. We also applied the other learning methods using this methodology, resulting in similar or worse results.

Table 4: Average accuracy when training using only one type of walk and testing on the other.

Train on	Test on Front Walks	Test on Side Walks
Front Walks	98.5%	16.5%
Side Walks	23.8%	99.6%

In our data, the sensors provided different measurements for the same attributes when the subject is under different poses (front and side). An inspection of the results shows that there are statistically significant differences. The average difference across all attributes and all subjects was of 4.6% but for some attributes it was as high as 15% (thoracic spine) across all subjects and as high as 20% for a single subject.

This pose-dependent measurement is able to explain the observed result. An immediate conclusion is that one must be careful to include a good diversity of poses in the training data. However, we only tested two poses; one could ask whether other poses could provide more accurate information than these two. We plan on pursuing an answer to this question in future work.

4.4 Number of Frames

An important question is how many frames are necessary to have a good classification accuracy. To answer this question, we trained our models using a limited number of frames per subject. We made sure the training set contained the same number of frames representing side and front walks, but otherwise chose randomly which frames would be included. We repeated the process 10 times and averaged the results.

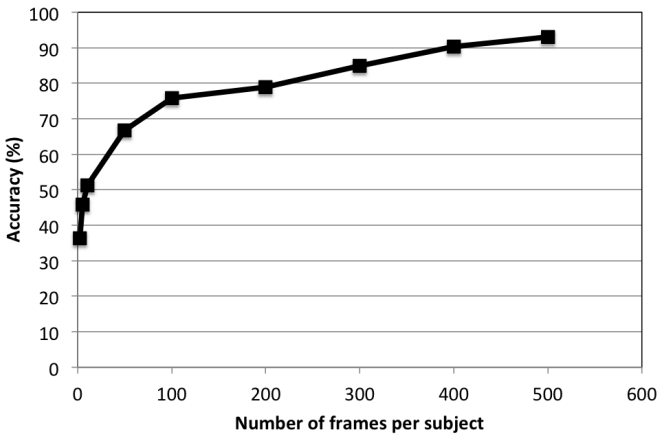

Figure 2: Average accuracy after training on a limited number of frames per subject. The fewest number of frames tested was two.

Figure 2 shows the results of applying the trained models on 500 frames that were not used in the training stage. For a low number of frames available for tranining, accuracy quickly improves with increases in the number of frames, but a clear logarithm behavior is observed, typical of learning tasks.

To reach over 90% accuracy about 500 frames are needed, which translate to about 17 seconds of pre-processed frames

Table 5: Accuracy on individual subjects.

Subject	Accuracy	False Positive	False Negative
1	99.86%	0.12%	0.10%
2	99.82%	0.08%	0.89%
3	99.93%	0.03%	0.20%
4	99.95%	0.20%	0.25%
5	99.87%	0.04%	0.66%
6	99.89%	0.05%	0.42%
7	99.72%	0.07%	0.22%
8	99.84%	0.07%	0.82%

or about 20 seconds of unfiltered frames. This is a reasonable amount of time if subjects are aware of a training stage or if subjects are not walking in front of the sensors. However, moving subjects in front of a static sensors take a little under 3 seconds to cross the sensors' field of view, which is too fast to capture enough frames for an accurate classification after training.

4.5 Binary Classification

In this type of scenario, instead of requiring the classifier to discriminate each subject we could ask it to identify a single subject, arguably an easier task. To accomplish this, we built one data set for each subject, where frames of the target subject were labeled as positive examples and the rest of the frames were labeled as negative examples.

In Table 5, we can see that this is indeed an easier problem. All subjects are recognized with great accuracy, with very few false positives or false negatives.

It is also possible to learn to recognize a single subject faster. Figure 3 shows the true positive and true negative rates when we vary the number of positive frames, while keeping a constant number of negative frames pre-stored (500 frames). In the binary classification problem, we need less than half the number of frames to recognize the subject with an over 90% accuracy. Nonetheless, 200 frames translates to about 7 seconds, which is still a long period.

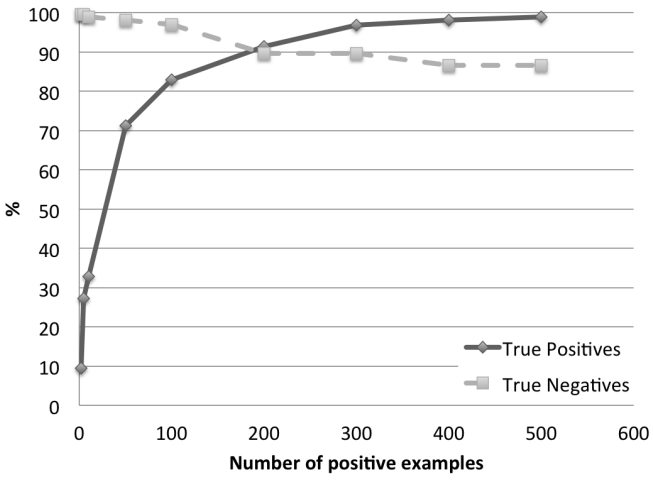

Figure 3: True positive and true negative rates when varying the number of positive examples. The number of negative examples is kept at 500. The fewest number of frames tested was two.

5. CONCLUSIONS

In this paper we considered the problem of performing subject identification using skeleton data extracted from Microsoft Kinect sensors. By capturing several subjects walking in front of the sensor and storing skeleton information provided by the sensor, we were able to calculate the relative lengths of several body parts of each subject. This information was pre-processed to reduce noise and used to train machine learning algorithms and their effectiveness in classifying each subject was tested.

We considered multi-layer perceptrons, decision trees and K-nearest neighbors learning algorithms. Our results showed that a nearest neighbor approach provided the best results. For the 8 subjects in our test set, KNN achieved a 99.6% overall accuracy, with low rates of false positives and false negatives, followed closely by a decision tree trained using the Random Forests algorithm. Random Forests were good enough and provide advantages over KNN - in particular, a decision tree is much faster to evaluate. However, KNN has the advantage of allowing incremental learning, making it straightforward to learn new subjects.

When transforming the multi-class classification problem of identifying multiple individual subjects to a binary classification problem, where only a single subject must be identified, we observed an increase in accuracy across all subjects. Such improvement was relatively small due to the already strong performance displayed in the multi-class case. However, we showed that learning can be made faster in this case, requiring far less positive examples to achieve high levels of accuracy.

When considering which attributes (i.e. body parts) to consider, our results showed that information about arms and legs were more useful than information about the spines or total height. Nonetheless, the best accuracy was obtained using all attributes. Using only height, in particular, was the worse combination of attributes. These results are evidence that it is the proportionality between different lengths that allows for a proper and unique identification.

Pose dependency was shown to be a major issue when using data from the Kinect sensors. When training using data from subjects walking in parallel to the sensor, performance was poor on walks performed perpendicular to the sensor, as well as the other way around. This shows an inconsistency in how Kinect infers the position of different body parts that is highly detrimental to anthropometric identification. One workaround to this issue is to make sure training data is composed of subjects in different poses.

Finally, we considered the number of frames needed to correctly classify a subject. While in the binary classification task this number was smaller (200 frames or 7 seconds of data capture) than for the multi-class cases (500 frames or about 20 seconds of data capture), we considered both cases to be excessively long for tasks such as moving person re-identification. This is because a walking subject completely crosses the sensor field-of-view in about 3 seconds. Different approaches seems to be needed for this task, including different lenses, moving sensors or improved processing methods.

The main contribution of this paper was to provide a first attempt at using Kinect data in anthropometric identification. The results were very promising and some caveats were identified that may help building a working system based on the ideas presented here. In general, the observed accuracy was better than those reported in e.g. [14] (where detailed

3D models are used), but it must be considered that our data set is still much smaller.

We are currently working towards expanding the set of captured subjects, including more people and different poses and situations. This expanded data set will allow for an improved analysis of the results presented in this paper. Furthermore, we plan on looking into body dynamics (as in e.g. [13]) to improve identification accuracy and, most importantly, reduce the size of the training set required.

Acknowledgments

This work is supported by CNPq (Brazilian National Research Council) through grant number 477937/2012-8.

6. REFERENCES

[1] *Alphonse Bertillon: Father of Scientific Detection*. George G. Harrap, 1956.

[2] S. Bak, E. Corve, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2010.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] J. Charles. Learning shape models for monocular human pose estimation from the microsoft xbox kinect. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1202–1208, 2011.

[5] M. V. den Bergh. Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In *IEEE International Symposium on Robot and Human Interactive Communication*, 2011.

[6] A. Dubois, A. Dib, and F. Charpillet. Using hmms for discriminating mobile from static objects in a 3d occupancy grid. In *IEEE International Conference on Tools with Artificial Intelligence*, 2011.

[7] S. Fazli, H. Askarifar, and M. J. Tavassoli. Gait recognition using svm and lda. In *Proc. of Int. Conf. on Advances in Computing, Control, and Telecommunication Technologies*, 2011.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[9] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.

[10] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo. Kinect identity: Technology and experience. *Computer*, 44(4):94–96, 2011.

[11] J.-M. Lu and M.-J. Wang. The evaluation of scan-derived anthropometric measurements. *IEEE Transactions on Instrumentation and Measurement*, 59(8):2048– 2054, 2010.

[12] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[13] G. V. Narasimhulu and D. S. A. K. Jilani. Gait recognition : A survey. *International Journal of Electronics Communication and Computer Engineering*, 3(1), 2012.

[14] D. B. Ober, S. P. Neugebauer, and P. A. Sallee. Training and feature-reduction techniques for human identification using anthropometry. In *4th IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010.

[15] J. Stowers. Altitude control of a quadrotor helicopter using depth map from microsoft kinect sensor. In *IEEE International Conference on Mechatronics (ICM)*, 2011.

[16] M. N. Uddi, S. Sharmin, A. H. S. Ahmed, E. Hasan, S. Hossain, and Muniruzzaman. A survey of biometrics security system. *International Journal of Computer Science and Network Security*, 11(10), 2011.

[17] L. Wang. Some issues of biometrics: technology intelligence, progress and challenges. *International Journal of Information Technology and Management*, 11(1):72–82, 2012.