AFRL-RH-WP-TP-2014-0014

GENDER AND ETHNICITY CLASSIFICATION FROM SMALL SUBSETS OF HUMAN BODY MEASUREMENTS

Huaining Cheng
Dustin Bruening
Darrell Lochtefeld
711th Human Performance Wing
Air Force Research Laboratory
Human Effectivess Directorate

Zhiqing Cheng
IST: a DCS Company
4027 Colonel Glenn Highway
Suite 210
Dayton OH 45431

MARCH 2014
INTERIM REPORT

Distribution A: Approved for public release; distribution is unlimited.

See additional restrictions described on inside pages

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

STINFO Copy

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TP-2014-0014 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
Huaining Cheng
Work Unit Manager
Human Signatures Branch

//signature//
Louise Carter, Ph.D.
Chief, Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YY)***
31-03-14

2. REPORT TYPE
Interim

3. DATES COVERED ***(From - To)***
2 January 2012 – 31 January 2013

4. TITLE AND SUBTITLE
Gender and Ethnicity Classification from Small Subsets of Human Body Measurements

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Huaining Cheng
Dustin Bruening
Darrell Lochtefeld
Zhiqing Cheng*

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H04N (7184A002)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
*IST: a DCS Company
4027 Colonel Glenn Highway/Suite 210
Dayton OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Human-Centered ISR Division
Human Signatures Branch
Wright-Patterson Air Force Base, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHXB

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TP-2014-0014

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88ABW-2014-1247; Cleared 27 March 2014

14. ABSTRACT

This paper investigates and compares machine learning models for classifying gender and ethnicity with human anthropometric measurements as input attributes. Optimal attribute sets are identified through individual measurement ranking and subset selection. These sets are further down-selected by taking into consideration the acquirability of measurements. Using the Civilian American and European Surface Anthropometry Resource (CAESAR) database as training and test datasets, the investigation has achieved a classification rate over 96% for gender (male/female) and 80% for ethnicity (White American/African American), respectively. Furthermore, the effect of random measurement noise on the classification performance is investigated to find a preliminary performance boundary for the classifiers. This study shows that gender can be predicted with high confidence and robustness from a few torso dimensions, while ethnicity can only be estimated roughly from limb dimensions under real-world conditions. The approach developed in this paper can be used with other image analysis software to achieve better understanding of human attributes contained in video imagery and to facilitate automated content analysis and decision making.

15. SUBJECT TERMS
Linear Discriminant Analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	17	Huaining Cheng
U	U	U			**19b. TELEPHONE NUMBER** *(Include Area Code)*

THIS PAGE LEFT BLANK INTENTIONALLY.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE LEFT BLANK INTENTIONALLY.

Gender and Ethnicity Classification from Small Subsets of Human Body Measurements

Huaining Cheng[1], Zhiqing Cheng[2], Dustin Bruening[1], Darrell Lochtefeld[1]

[1] Air Force Research Laboratory, Wright-Patterson AFB, Ohio
[2] Infoscitex Corporation, Dayton, Ohio

1.0 INTRODUCTION

The wide use of computer vision technology often requires recognizing human characteristics (e.g., gender and ethnicity) in a natural environment. For gender recognition, human face has been widely used [1-8]. However, the accurate front-view face images used in these studies are difficult to obtain in a dynamic setting or from a distance. Face images can be easily altered or obscured by head movements, facial hair, and glasses. Therefore, they are useful for controlled environments but less effective for unstructured environments. Other researchers have investigated using gaits for the prediction of gender [9, 10] based on either silhouettes or motion capture data. However, capturing accurate human locomotion in an unstructured environment is a difficult task. Furthermore, these studies are limited by the small size of the datasets; hence variances in human gaits have not been quantified. Recently, human shape has been used for gender recognition also [11, 12]. However, how to accurately acquire human shape data and how to effectively handle high-dimensional shape data still remain as two technical challenges.

With the development of various 2D and 3D imaging and computer vision technologies, it is possible to directly extract basic human body measurements from the data acquired at a distance by 2D/3D cameras, LIDARs, and radars. This potentially allows for the inference of gender and ethnicity from human anthropometric measurements. In fact, for gender recognition, body measurements are good discriminative factors, as shown in a Linear Discriminant Analysis (LDA) study for gender classification [13]. Since body measurements are relatively larger compared to the facial features and more stable compared to the gait features, they are more suitable for unstructured and dynamic environments. Compared to 3D shape data, body measurements are easier and more efficient to implement. They can be extracted or estimated from 2D or 3D imagery via 2D or 3D image analysis such as edge detection, body segmentation, and 3D model reconstruction. These measurements are intuitive to human analysts and can be considered as pose-invariant.

Compared to gender recognition, current ethnicity classification also relies heavily on facial images but is less studied [5-8, 14]. Therefore, this paper investigates the anthropometry-based gender and ethnicity recognition by identifying the optimal anthropometric feature subsets as wells as evaluating the various classification models. The generation of the optimal feature sets is designed as an induction of three different types of feature selection schemes. The classification models explored are logistic regression, Support Vector Machine (SVM) [15], and AdaBoost [16] classification methods. A large publicly available anthropometric survey, CAESAR [17], is used to produce separate training and test datasets. Furthermore, the effects of random measurements noise on the classification performance are studied to provide guidance on classifiers' performance boundaries.

To the best of our knowledge, this is the first paper that studies feature selection, model performance, and noise effect together for anthropometry-based classification of gender and ethnicity. The open source machine learning toolkit WEKA [18] is used in this study to facilitate the analytical process. Since features are also called attributes in WEKA, anthropometric measurements, features, and attributes have the same meaning in this paper and thus are used interchangeably in different context.

2.0 RESEARCH DATASET

In this study, the CAESAR anthropometric survey is selected as the source for both training and test data. An anthropometric survey measures and collects data on individual demographic and physical dimension information from a human subject pool that statistically represents the target population. Therefore, it is ideal for training and testing gender and ethnicity classifiers. CAESAR is a large scale NATO country survey sponsored by the US Air Force and carried out by a group of organizations in North America (US and Canada), the Netherlands – the tallest population in NATO, and Italy – the shortest population in NATO. The dataset used by this paper is restricted to the North American set only, which was collected at multiple regions of North America. After removing any records with missing field values, a dataset is extracted which consists of 2263 subjects in three roughly equal age groups (18-29, 30-44, and 45-65). Among the subjects, there are 1208 females and 1055 males. The ethnicity composition is 243 African Americans and 2020 White Americans. Hispanic and other multi-racial subjects are not included in the dataset because of their small sample numbers in the original CAESAR survey, and also because of our desire for investigating two-label ethnicity classification first, i.e., White vs. African American.

Since the dataset is not balanced in ethnicity, a data balancing process is made against the ethnicity field by randomly removing White American subjects down to 250 subjects. Though the removal of a large number of White American samples may have some negative effects on the classification precision for the White American, it is nevertheless needed in order to control the false positives in the predicted African Americans and to attain valid classification accuracy results.

In the CAESAR survey, there are 83 one-dimensional anthropometric measurements (40 from traditional hand measurements and 59 calculated from scan landmarks). The overall measuring accuracy is within 10 mm at the confidence level of 95%. For each subject, the survey has detailed demographic information and 3D laser body scans that show the hand-placed anatomical landmarks.

The initial training and test datasets include 23 measurements selected from the 40 traditional ones and 2 measurements chosen from scan-derived ones as the input attributes. Gender and ethnicity are the class labels and taken from the demographic records. These input attributes are selected in order to include length, breadth, and circumference of different body segments plus height and weight (Table 1). These measurements capture size and proportion information among different body segments as well as important joint distances. Furthermore, individual subject's measurements, excluding height and weight, are normalized with the subject's height to bring them into a similar proportional scale.

Table 1. Input attributes and corresponding measurements

Attribute Name	Measurement Name	Attribute Name	Measurement Name
HEAD-CIR	Head Circumference	HIP-CIR-MAX-HT	Height at Maximum Hip Circumference
HEAD-LTH	Head Length	HIP-BRTH-SIT*	Hip Breadth, Sitting
HEAD-BRTH*†	Head Breadth	THI-CIR	Thigh Circumference
IPD-SE†	Inter-pupillary Distance Scan Extracted	BUTT-KNEE-LTH†	Buttock to Knee Length
NECK-BASE-CIR	Neck Base Circumference	ANK-CIR-MALL	Ankle Circumference at Malleolus
NECK-HT-SE	Neck Height Scan Extracted	KNEE-HT-SIT†	Knee Height, Sitting
CHE-CIR-SCY	Chest Circumference at Scye	FOOT-LTH†	Foot Length
CHE-CIR	Bust/Chest Circumference	SHDR-WRST-LTH†	Shoulder to Wrist Length
SHDR-BRTH*	Shoulder (Bideltoid) Breadth	SHDR-ELB-LTH	Shoulder to Elbow Length
WST-CIR-PREF	Waist Circumference at Preferred Waist	HAND-LTH†	Hand Length
CRO-LTH-PREF	Crotch Length to Preferred Waist	HT*	Height
CRO-HT	Crotch Height	WT	Weight
HIP-CIR-MAX	Hip Circumference, Maximum		

* Selected for gender classification, see Section 4.1.
† Selected for ethnicity classification, see Section 4.1.

The complete set of measurements listed in Table 1 is only used to obtain the best possible performance benchmarks for gender and ethnicity classification. Experimental classifiers are trained and tested using several subsets that contain a much smaller number of measurements from Table 1, derived through the feature selection process described in Section 3.1.

3.0 APPROACHES

3.1 Input Attribute Selection

Attribute selection is beneficial to a learning task because it usually reduces the dimensionality of the data and hence the hypothesis space. This leads to not only better algorithm performance but also a lesser burden on the sensor systems. The latter is critically important for real-world operability. In this study, a three-scheme induction process is designed for the attribute selection. The three selection schemes consist of the information gain for coarse-grained individual attribute ranking and the subset evaluation using a correlation-based filter and an AdaBoost wrapper. The inducted attributes are the intersection of the returns from the three selection schemes. Our goal of feature selection is to find an input attribute set having the smallest number of measurements but without a significant loss of class discriminative power.

The information gain evaluates the significance of each attribute according to the expected reduction in the information needed for assigning class labels if the attribute value is known. The average required information is quantified by the entropy term $-\sum_i p_i \ln(p_i)$, where p_i is the probability that a training sample belongs to class i. The information gain is an individual attribute ranking scheme which does not take the combining effects of multiple attributes into consideration.

Unlike information gain, subset selection identifies a set of significant attributes through either a filter or wrapper, both of which are employed in this study. The filter approach is represented by the Correlation-based Feature Selection (CFS)[19] which uses a heuristic merit score to select "subsets containing features highly correlated with the class, yet uncorrelated with each other," i.e.,

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{1}$$

where $\overline{r_{cf}}$ is the average feature-class correlation and $\overline{r_{ff}}$ is the average feature-feature inter-correlation. The main advantage of the filter approach is that it is independent of any classification algorithm; hence the returned subset is applicable to various algorithms.

The wrapper approach [20] typically measures the worthiness of a subset by running an actual learning algorithm and cross validation for accuracy. Since there are many possible combinations of feature subsets, the wrapper takes a longer time to conclude and may get a different optimal subset depending on the chosen learning algorithm. In this study, the standard AdaBoost is chosen as the algorithm for the wrapper. This boosting algorithm is a weighted additive model of component hypothesis $h_t(\boldsymbol{x})$ in the form of,

$$H(\boldsymbol{x}) = sign(\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})), \tag{2}$$

where α_t is the weight and in this study $h_t(\boldsymbol{x})$ is a decision stump of the individual attribute. We set the iteration number T to 50 after some experimentation. Because AdaBoost minimizes l_1-norm $\parallel \alpha \parallel_1$, which leads to sparsity, those component classifiers with higher weights can form the significant subset naturally.

Denote $\{m_i\}_I$ as the set containing the top 25% information gain rankings and $\{m_i\}_F$ and $\{m_i\}_W$ as the return subset from the CFS and the AdaBoost wrapper, respectively. The final inducted attribute set $\{m_i\}$ is,

$$\{m_i\} = \{m_i\}_I \cap \{m_i\}_F \cap \{m_i\}_W. \tag{3}$$

If necessary, we could pare down further those measurements in $\{m_i\}$ that are deemed to be difficult to acquire accurately by a real world sensor system, through an evaluation of their effects on the classification performance.

3.2 Classification Models

In this study, three classification algorithms, logistic regression, SVM, and AdaBoost are compared. The overall performance of a classification model is assessed in terms of classification accuracy and the tolerance to random noise. We will use the inducted significant attribute subsets as the input instead of the entire 25 measurements. The benefits of this reduced dimensionality are simpler classifiers and smaller generalization error, especially when a large number of training samples are used.

Logistic regression makes a prediction according to the log-odds (logit) of posterior probabilities of class *C* given the input $\boldsymbol{x}$. For the two-label case in this study, the logit function is given by,

$$log\frac{\Pr(C=1|X=\boldsymbol{x})}{\Pr(C=2|X=\boldsymbol{x})}=\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{x}. \tag{4}$$

Therefore $\Pr(C=k|X=\boldsymbol{x})$ is a logistic function and the weights β_0 and $\boldsymbol{\beta}$ are computed by maximizing the log conditional likelihood of C given the set of training sample $\boldsymbol{x}$. Logistic regression doesn't make any particular assumption on the marginal density Pr(X) so it is relatively robust, compared to LDA. Its potential problem is overfitting when more variables are used with limited training samples, though it is not a particular concern for this study, as explained above.

Instead of estimating posterior probability of class, SVM finds the optimal separating boundary between the classes directly by maximizing the minimum margin M between the training samples and the hyperplane. The optimization problem is stated as,

$$\min_{\boldsymbol{\beta},\beta_0,\xi}\{\frac{1}{2}\parallel\boldsymbol{\beta}\parallel^2+\ C\sum_i\xi_i\}, \tag{5}$$
$$subject\ to\quad y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0)\geq 1-\xi_i\ \ \forall i,\ \xi_i\geq 0,$$

where $\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0=0$ is the hyperplane with margin $M=1/\parallel\boldsymbol{\beta}\parallel$. ξ_i is a slack variable with a non-zero value if sample *i* is on the wrong side of the hyperplane, i.e., if there are some overlaps between the classes. *C* is a parameter controlling the tradeoff between maximizing the margin and minimizing the amount of slack. The problem can be generalized to a nonlinear boundary through kernelization.

The AdaBoost classifier is implemented as (2). For gender and ethnicity classification, the application of SVM requires a normalization or standardization of the input attribute space. This is due to the fact that each measurement is scaled against individual subject's height, except for height itself and weight. This introduces large differences between the height/weight and all other height-normalized measurements. It is known that large margin classifiers are sensitive to the way features are scaled and our preliminary experiments confirm this. Therefore, for all samples in the training and test sets, each attribute's *z* (standard) scores are computed using the corresponding measurement's mean and standard deviation. The *z* scores are then used as the actual input attribute values. Though WEKA has an option for SVM feature standardization, we have to complete the process outside of WEKA because of the planned addition of random noise to the gender and ethnicity test data later.

4.0 EXPERIMENTS

4.1 Optimal Input Measurement Subset

The purpose of identifying an optimal input attribute set is to enhance the classification algorithm's workability in a real world environment. Due to the sensor's viewing angle and occlusion, it is very difficult to obtain a good capture or estimation of circumferences and obscured lengths, such as waist circumference and crotch length, etc. In addition, the extracted measurement values inevitably include some noise. Taking these factors into account, it is ideal if one can find an optimal input subset with small number of input measurements, and preferably straight-line measurements such as breadth and length of the torso and limbs. At the same time, the subset should maintain sufficient discriminative power so that a good classification rate can still be attained by the reduced number of inputs.

Using the aforementioned attribute selection scheme, the top 25% attributes from the information gain ranking and the two subsets from the CFS filter and AdaBoost wrapper are presented in Tables 2 and 3, respectively. The attributes highlighted by the bold font in the two tables are those appearing on the returns from all three selection schemes. According to our selection criteria (3), they are inducted as the final input attributes for classification. There are some divergences in the returns for the gender attributes; only three common attributes (HT, HIP-BRTH-SIT/HT, and SHDR-BRTH/HT) are present in all three returns. However, the removal of other gender attributes may not be a huge loss because they include some hard-to-estimate or occluded measurements such as height at maximum hip circumference and crotch length, etc.

Table 2. Returns for gender attribute selection

InfoGain Ranking (Top 25%)	**Subset by CFS Filter**	**Subset by AdaBoost Wrapper**
HT	IPD-SE/HT	FOOT-LTH/HT
HIP-BRTH-SIT/HT	SHDR-WRST-LTH/HT	**HIP-BRTH-SIT/HT**
CRO-LTH-PREF/HT	HAND-LTH/HT	**SHDR-BRTH/HT**
WT	**HIP-BRTH-SIT/HT**	**HT**
SHDR-BRTH/HT	HIP-CIR-MAX-HT/HT	WST-CIR-PREF/HT
HIP-CIR-MAX/HT	**SHDR-BRTH/HT**	
NECK-BASE-CIR/HT	**HT**	
	CRO-LTH-PREF/HT	
	WT	
*Bold-highlighted attributes are used for gender classification		

Table 3. Returns for ethnicity attribute selection

InfoGain Ranking (Top 25%)	**Subset by CFS Filter**	**Subset by AdaBoost Wrapper**
SHDR-WRST-LTH/HT	**SHDR-WRST-LTH/HT**	**SHDR-WRST-LTH/HT**
BUTT-KNEE-LTH/HT	**BUTT-KNEE-LTH/HT**	**BUTT-KNEE-LTH/HT**
HAND-LTH/HT	**HAND-LTH/HT**	**HAND-LTH/HT**
IPD-SE/HT	**IPD-SE/HT**	**IPD-SE/HT**
FOOT-LTH/HT	**FOOT-LTH/HT**	**FOOT-LTH/HT**
KNEE-HT-SIT/HT	**KNEE-HT-SIT/HT**	**KNEE-HT-SIT/HT**
SHDR-ELB-LTH/HT	CRO-HT/HT	SHDR-ELB-LTH/HT
		CRO-HT/HT
		HIP-CIR-MAX-HT/HT
		HEAD-BRTH/HT
		CHE-CIR/HT
*Bold-highlighted attributes are used for ethnicity classification		

The small size of the gender attribute subset may also indicate that these few measurements have significant discriminative capabilities. Intuitively, the ratio of shoulder breadth to the hip breadth is an easy-to-see differentiator between male and female, with the former having a higher ratio and the latter having a lower one. The corresponding bivariate chart (Figure 1) of the training dataset seems to support this assertion. There is a perceived difference in the limb length between White and African Americans, but it is not as distinctive as the breadth ratio for gender, as evidenced by Figure 2. These results are consistent with human observations.

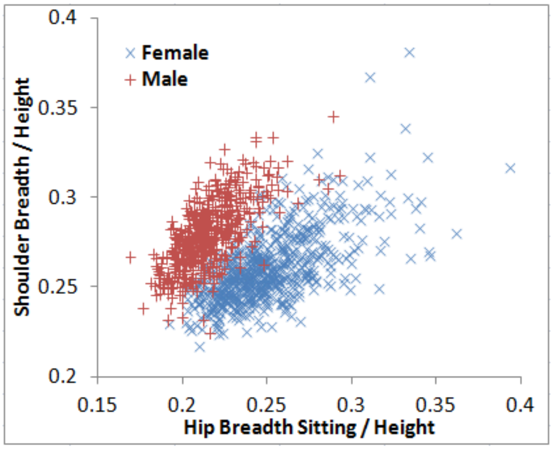

Figure 1. Clustering of two genders against shoulder and hip breadth

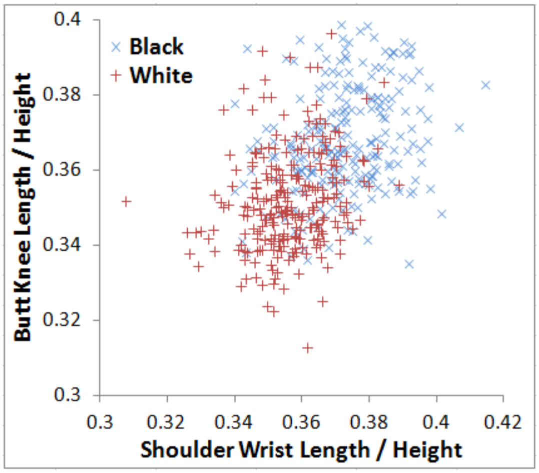

Figure 2. Clustering of two ethnicities against limb lengths

Unlike those of the gender case, the top-ranked measurements for the optimal ethnicity feature set are fairly consistent across the returns of three attribute selection schemes. The difference between the two cases can be explained as follows. Since there are quite a few equally-prominent attributes for gender classification, any two or three of them may be capable to provide sufficient discriminative power. This leads to different top rankings by the three attribute selection schemes. On the other hand, for ethnicity classification, since there are only several better-than-average attributes, achieving satisfactory classification accuracy requires all of them.

4.2 Classification Results Based on Cross-Validation

WEKA has built-in 10-fold cross-validation for its learning process. For gender classification, we first use the complete 25-measurement dataset to run Logistic Regression, SVM, and AdaBoost with 10-fold cross-validation. All three algorithms have achieved a similar classification accuracy of 99%. This value is used as the ideal benchmark to gauge how much degradation of performance will occur when the reduced input attribute set is used instead. A similar ideal accuracy benchmark is obtained for ethnicity classification. It has a value of 88%.

Since the number of available gender samples is large, the entire dataset can be divided into two equal parts, one with 1132 samples as the training set and the other with 1131 samples as the test set. For ethnicity classification, the total 493 samples are also divided into two equal groups with a balanced number of African Americans and White Americans. The separation of the dataset into training and test sets serves two purposes. First, by using only the training set for input attribute selection and learning with cross-validation, one can better test and validate the classification accuracy separately using the test set "uncontaminated" by the attribute selection process. Second, it allows the introduction of noise to the test set for investigating the effect of random noise on the classifier's performance. The performance based on the separate test sets and the noise effects are discussed later in Section 4.3. The classification accuracy results presented in this section are associated with the training sets only.

Table 4 lists the training performance of gender classification for three models using 1132-sample training set with three selected input attributes – {HT, HIP-BRTH-SIT/HT, SHDR-BRTH/HT}. The results demonstrate the effectiveness of these three easy-to-acquire input measurements for separating the two genders. All classifiers have similar performance and are very close to the ideal benchmark (within 3%), though logistic regression performs slightly better than the other two in overall. One can conclude from these results that a high-performance and efficient gender classifier can be built from any one of the three classification models using these three input measurements.

The ridge parameter (regulating the weights) in the logistic regression and tradeoff parameter *c* (regulating the slack amount) in SVM do not have significant effect on the performance due to the nature of the dataset. Both of them are set to 1. It should be pointed out that the SVM results provided in Table 4 are from standardized *z* scores. If normalization or standardization is not performed, the learning task takes much longer time to complete and the classification accuracy degrades to 86.66%. Therefore, normalization or standardization is very important for SVM.

Table 4. Performance of three-attribute gender classifiers

	Logistic Regression	**SVM**	**AdaBoost (50 Iter.)**
Cross-V. Accuracy	96.82%	96.47%	96.11%
True Positive	497	492	494
False Positive	17	22	20
False Negative	19	18	24
True Negative	599	600	594
*positive = male, negative = female			

A similar learning procedure is applied to the ethnicity training dataset of 247 samples with 6 input measurements – {SHDR-WRST-LTH/HT, BUTT-KNEE-LTH/HT, HAND-LTH/HT, FOOT-LTH/HT, IPD-SE/HT, KNEE-HT-SIT/HT}. Table 5 lists the classification performance parameters. All three classifiers perform very close to the ideal benchmark. Overall, the ethnicity classifiers cannot attain the same level of accuracy as that achieved by the gender classifiers, even with more input measurements. However they are still very effective for ethnicity recognition purpose with over 87% classification accuracy.

Table 5. Performance of six-attribute ethnicity classifiers

	Logistic Regression	**SVM**	**AdaBoost (50 Iter.)**
Cross-V. Accuracy	87.45%	87.85%	87.45%
True Positive	98	99	98
False Positive	17	16	17
False Negative	14	14	14
True Negative	118	118	118
*positive = African American, negative = White American			

Within the ethnicity input attribute set, the hand length, foot length, and distance between eye sockets (IPD) may not be easy to obtain accurately through real-world sensors, due to small size of the measurements and occlusion. If these three attributes are removed from the input, the classification accuracy drops to around 80%, as shown in Table 6. This also supports the previous hypothesis that there aren't dominant body size characteristics for the ethnicity classification, in contrast to the case of gender classification.

Table 6. Performance of three-attribute ethnicity classifiers

	Logistic Regression	SVM	AdaBoost (50 Iter.)
Cross-V. Accuracy	80.97%	80.17%	78.54%
True Positive	90	93	88
False Positive	25	22	27
False Negative	22	27	26
True Negative	110	105	106
*positive = African American, negative = White American			

4.3 Effects of Measurement Noise

This section presents the test accuracy of the three classifiers. In addition, the effects of measurement noise are evaluated in order to draw some preliminary performance guidelines. For the case of gender classification, random noise is added to the test data set of 1131 samples. The above three classifiers trained using the original 1132-sample training dataset are then tested against the noisy test dataset with different levels of noise. It is assumed that random noise has a Gaussian distribution of mean 0 and standard deviation σ. Two levels of σ are designed to represent noise level of 30 mm and 50 mm, respectively. Inverse transform sampling is used to generate the random noise which is then incorporated into each individual test sample. The test performance values for different noise levels are listed in Table 7.

Table 7. Test accuracy for three-attribute gender classifiers

Noise Level (mm)	Logistic Regression	SVM	AdaBoost (50 Iter.)
0	96.37%	96.73%	96.20%
30	89.21%	88.77%	88.06%
50	78.07%	77.54%	79.30%

It is shown that all three classification models perform fairly well and robustly under moderate noisy condition in terms of their classification accuracies. Measurement noise within the threshold of 30 mm (which is approximately 1.8 percent of the average height) can be tolerated without significant reduction of classification accuracy. This threshold can also be used to define an acceptable range for measurement errors if they can be considered as unbiased and normally distributed.

A similar test is performed on the ethnicity classification. The results are shown in Table 8 for the three-attribute case. It is apparent that the performance of the ethnicity classifier is further deteriorated to the point that at the 30 mm noise level, the body measurement alone may not be sufficient for a high confidence decision. Other non-anthropometric information may be needed to boost the performance.

Table 8. Test accuracy for three attribute ethnicity classifiers

Noise Level (mm)	Logistic Regression	SVM	AdaBoost (50 Iter.)
0	78.05%	78.86%	80.08%
30	74.39%	72.36%	73.98%
50	59.76%	62.60%	63.42%

We have also compared six- and three-attribute ethnicity classifier under noisy conditions. It is found that the former has a worse performance than the latter, even though it has more input features. For example, at 30 mm noise level, the six-attribute logistic, SVM, and AdaBoost models are only able to achieve an accuracy of 59%, 67%, and 63%, respectively. It is possible that the 30 mm measurement noise on the three smaller body dimensions (hand length, foot length, and distance between eye sockets) could be large enough to negatively affect the classification results. By removing them, the same noise level has a much smaller effect on the remaining larger input attributes. This result favors the use of large body measurements in the inference of human characteristics, besides the inherit benefit of easier data acquisition for large measurements. It also demonstrates the importance of checking models under noise conditions.

5.0 CONCLUSIONS

This paper presents anthropometric measurement-based gender and ethnicity recognition. A new attribute selection method has been designed to identify the optimal input measurement sets for gender and ethnicity classification. Three classification models have been experimented with separate CAESAR training and test sets. The effects of measurement noise on the classification accuracy have been simulated and analyzed. Based on the results of the investigation, the following conclusions are in order.

1. The three classifiers, logistic regression, SVM, and AdaBoost, perform equally well if accurate measurements can be acquired. They achieve very high accuracy for gender classification (up to 96%) and acceptable accuracy for ethnicity classification (up to 80%) with only three input measurements.
2. Different anthropometric measurements have different discriminative power for gender and ethnicity recognition. It is found that height, hip breadth, and should breadth constitute an effect input attribute set for gender recognition, whereas arm and leg length related measurements provide effective classification between White and African Americans.
3. The measurement noise has larger adverse effect on the performance of classifiers having input attributes made of smaller measurements. For the gender classifier, a standard deviation of 30 mm can be reasonably used as the preliminary threshold for the measurement noise. For the ethnicity classifier, additional information other than anthropometric measurements may be required in order to maintain high confidence in classification results under noisy environment.

REFERENCES

[1] Lawrence, D. T, Golomb, B. A., and Sejnowski, T. J., "Sexnet: A neural network identifies sex from human faces," Neural Information Processing Systems, 1991, pp. 572–577.

[2] Moghaddam, B. and Yang, M., "Gender classification with support vector machines," in Proc. IEEE Conf. on Automatic Face and Gesture Recognition, 2000, pp. 306–311.

[3] Xu, X., and Huang, T. S., "SODA-boosting and its application to gender recognition," in Proc. IEEE Workshop on Analysis and Modeling of Faces and Gestures, 2007.

[4] Baluja, S. and Rowley, A.H., "Boosting sex identification performance," Int. J. of Computer Vision, Vol. 71 (1), 2007, pp. 111-119.

[5] Gutta, S., Wechsler, H., and P. J. Phillips, "Gender and ethnic classification of face images," in Proc. IEEE Conf. on Automatic Face and Gesture Recognition, 1998, pp. 194–199.

[6] Lu, X., Chen, H., Jain, A., "Multimodal facial gender and ethnicity identification," Advances in Biometrics, Lecture Notes in Computer Science, Vol. 3832, 2005, pp. 554-561.

[7] Shakhnarovich, G., Viola, P.A., and Moghaddam, B., "A unified learning framework for real time face detection and classification," in Proc. IEEE Conf. on Automatic Face and Gesture Recognition, 2002. pp. 14-21.

[8] Yang, Z. and Ai, H., "Demographic classification with local binary patterns," Advances in Biometrics, Lecture Notes in Computer Science Vol. 4642, 2007, pp. 464-473.

[9] Li, X., Maybank, S., Yan, S., Tao, D., Xu, D., "Gait components and their application to gender recognition," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38, 2008, pp. 145–155.

[10] Davis, J. W. and Gao, H., "An expressive three-mode principal components model for gender recognition," J. Vision 4 (5), 2004, pp. 362–377.

[11] Tang. J., Liu., X., Cheng, H., and Robinette, K., "Gender recognition using 3-D human body shapes," IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 41., no. 6, 2011.

[12] A. Fouts, R. McCoppin, M. Rizki, L. Tamburino, and O. Mendoza-Schrock, "Exploring point-cloud features from partial body views for gender classification," in Proc. SPIE 8402, Evolutionary and Bio-Inspired Computation: Theory and Applications VI, 2012.

[13] Fullenkamp, A., Robinette, K., and Daanen, H., "Gender difference in NATO anthropometry and the implication for protective equipment," in Proc. NATO Research and Technology Organization (RTO) Human Factors and Medicine Panel (HFM) Symposium, Turkey, 2008.

[14] Hosoi, S., Takikawa, E., and Kawade, M., "Ethnicity estimation with facial images," in Proc. IEEE Conf. on Automatic Face and Gesture Recognition, 2004, pp.195-200.

[15] Vapnik, V.N., Statistical learning theory, Springer, New York, 1998.

[16] Freund, Y. and Schapire, R., "A decision-theoretic generalization of online learning and an application to boosting," J. Computer Sys. Sci., vol. 55, 1997, pp. 119–139.

[17] Robinette, K., Blackwell, S., Daanen, H., Fleming, S., Boehmer, M., Brill, T., Hoeferlin, D., and Burnsides, D., Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Volume I: Summary, Technical Report AFRL-HE-WP-TR-2002-0169, National Technical Information Service Accession No. ADA406704, United States Air Force Research Laboratory, 2002.

[18] Available: http://www.cs.waikato.ac.nz/~ml/weka/

[19] Hall, M. A., "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," in Proc. 7th International Conference on Machine Learning, p.359-366, June 29-July 02, 2000

[20] John, G. H., Kohavi, R., and Peger, P., "Irrelevant features and the subset selection problem," in Proc. 11th International Conference on Machine Learning, 1994.