Optimization of an individual re-identification modeling process using biometric features

Alejandro Heredia-Langner[1], Brett G. Amidan[1], Shari Matzner[1], and Kristin H. Jarman[1]
[1]Pacific Northwest National Laboratory, 902 Battelle Boulevard, PO Box 999 Richland, Washington 99352 USA

Abstract— *We present results from the optimization of a re-identification process using two sets of biometric data obtained from the Civilian American and European Surface Anthropometry Resource Project (CAESAR) database. The datasets contain real measurements of features for 2378 individuals in a standing (43 features) and seated (16 features) position. A genetic algorithm (GA) was used to search a large combinatorial space where different features are available between the probe (seated) and gallery (standing) datasets. Multiple linear regression models are employed to estimate one set of features from the other. Results show that optimized model predictions obtained using less than half of the 43 gallery features and data from roughly 16% of the individuals available produce better re-identification rates than two other approaches that use all 43 gallery set features and information from all 2378 individuals.*
Key Words: Genetic Algorithm, Re-identification, CAESAR.

1. Introduction

Re-identification is the task of accurately recognizing a person that has been previously observed and for whom some information is available in a database. For example, an image obtained from a photograph or video can be employed to estimate measurements of certain body features or other characteristics, and those estimates used to interrogate a database in search of a match. In particular, the database may contain biometric information obtained using a controlled and systematic process that can be reliably used to identify an individual. Subsequently, the same or other measurements may only be obtainable under a different and more challenging set of circumstances.

Numerous research efforts have been conducted recently on person re-identification, including gait recognition [1], clothing appearance [2], and anthropometry [3], [4]. The work in references [2] and [5] provide two excellent surveys on person re-identification. Other research has also focused on finding anthropometric features for clustering individuals along gender [6] and in reducing the number of dimensions needed for clustering [10]. In this work, we present results for person re-identification using two sets of biometric data. The two sets of data were obtained by the Air Force Research Laboratory (AFRL) and form part of the Civilian American and European Surface Anthropometry Resource Project (CAESAR) database [7]. The datasets used in this work are 1D North American anthropometric measurements for 2384 individuals in a standing position and 2380 individuals in a seated position. The standing dataset contains measurements of 43 body features (measurements between two landmarks) and the seated dataset contains measurements of 16 body features, with values for both sets reported in millimeters. The two datasets have 2378 persons in common. Data for five body features are identified with very similar names in the two sets but, because the measures are obtained in standing or seated positions, the numerical values of the features with similar names are not equal for a given person. Some measurements are missing for some individuals in each of the two datasets. Because the standing dataset contains information for a larger number of body measurements and the largest number of individuals, it is used as the gallery set, the set containing sufficient information for unique identification, and the seated set is used as the probe (or secondary) data from which gallery set feature value estimates will be obtained. The names of the features in the standing set are shown in Table 1 and the names of the features in the seated set are shown in Table 2.

Table 1. Names of the 43 Features in the North American 1D Standing Set (Gallery Set)

Feature Name	Feature Name	Feature Name
Acromial Ht Stand Lt	Bitrochant.Brth Stand	Malleolus Med Rt
Acromial Ht Stand Rt	Bustpoint Brth	Neck Ht
Acromion-Radiale Len Lt	Cervicale Ht	Radiale-Stylion Lt
Acromion-Radiale Len Rt	Chest Ht Stand	Radiale-Stylion Rt
Ankle Ht Lt Malleolus,Lat	Elbow Ht Stand Lt	Sellion Supramenton
Ankle Ht Rt Malleolus,Lat	Elbow Ht Stand Rt	Sleeve Outseam Lt
Arm Inseam Lt	Foot Brth Lt	Sleeve Outseam Rt
Arm Inseam Rt	Foot Brth Rt	Sphyrion Ht Lt
Axilla Ht Lt	Infraorbitale Ht Lt	Sphyrion Ht Rt
Axilla Ht Rt	Infraorbitale Ht Rt	Suprasternale Ht
Biacromial Brth	Inter-pupillary Dst	Trochanterion Ht
Bicristale Brth	Interscye Dst Stand	Trochanterion Ht
Bigonial Brth	Knee Ht Stand Lt	Waist Back
Bispinous Brth	Knee Ht Stand Rt	
Bitragion Brth	Malleolus Med Lt	

Table 2. Names of the 16 Features in the North American 1D Seated Set (Probe Set)

Feature Name
Acromial Ht Sit Lt
Acromial Ht Sit Rt
Bi-lateral Femoral Epicondyle Brth Sit
Bi-lateral Humeral Epicondyle Brth Sit
Bitrochanteric Brth Sit
Buttock to Trochanter Lth
Femoral Epicondyle Lat to Malleolus Lat Lt
Femoral Epicondyle Lat to Malleolus Lat Rt
Infraorbitale Ht Sit Lt
Infroarbitale Ht Sit Rt
Trochanter to Femoral Epicondyle Lat Lt
Trochanter to Femoral Epicondyle Lat Rt
Trochanter to Seated Surface Lt
Trochanter to Seated Surface Rt
Elbow Ht Sit Lt
Elbow Ht Sit Rt

2. Materials and Methods

The process of re-identifying an individual by searching a database is fairly simple. A numerical vector with values for some body features from an unknown individual is compared against the corresponding values in a database. Standardized distances between the vector of the unknown individual and every person in the database are calculated and ranked. Standardized distance metrics are often used in re-identification because body feature measurements vary in magnitude. The individual in the database with the smallest distance, ideally zero, to the vector from the unknown person is reported as the closest match. This matching process creates a single, real-valued metric that determines how similar any two subjects are. If a correct match is found as the top-ranked standardized distance, it is said to have a Rank of 1. If the correct match is found, say, in the fifth position of ranked standardized distances, it is said to have a Rank of 5.

Using standardized Euclidean distance between two vectors of body measures as a metric to identify an individual is reasonable because sets of body measurements for a person tend to be unique. In the absence of noise, very few features in the gallery set are needed to unambiguously identify an individual in the database. Most combinations using only two of the 43 gallery features available provide perfect discrimination under conditions of unchanging measurements and the fixed number of individuals in the database. Not surprisingly, using a larger number of features results in better identification power, measured as better separation between all pairs of distinct individuals in the gallery set. Figure 1 shows distributions of the minimum standardized Euclidean distance among all pairs of individuals in the gallery when one thousand samples with 5, 10, 15, 20, 25, 30, and 35 features from the gallery set are selected at random and used to calculate pairwise standardized Euclidean distances.

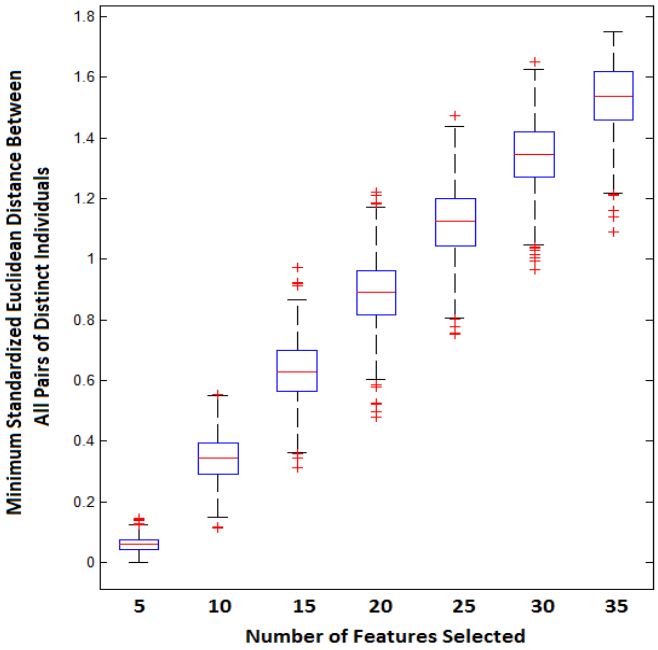

Figure 1. Box plots of the minimum standardized Euclidean distance among all distinct pairs of 2384 individuals in the gallery data for 1000 samples where the number of features shown in the x-axis was selected at random from the 43 available. The central mark in each box is the median and the edges are the 25 and 75 percentiles, the edges extend to the most extreme values not considered outliers and the crosses are outliers.

Figure 1 makes clear that, although some feature combinations provide better separation between all pairs of distinct individuals in the gallery, it is likely that any set with 10 or more features will be sufficient for perfect identification. Naturally, it is a much greater challenge to obtain estimates of gallery set measurements from data collected in a different, possibly uncontrolled way.

The re-identification question using two different sets of data becomes a feature selection problem. As Figure 1 shows, very few features are needed for perfect re-identification if the features in the gallery set are known or can be estimated with very high accuracy. Figure 1 also shows that, in general, using a larger number of features is better. If features in the gallery set can only be estimated from a secondary source of data with some degree of error, how many features are needed and what level of re-identification can we expect to achieve? Is there a subset of features that is better for re-identification purposes?

Because 2378 of the individuals are common to both the gallery and probe datasets, it is easy to study the relationship between pairs of features in the two sets. Simple linear correlation coefficients between every single feature in the gallery set and every single feature in the probe set range between -0.18 and 0.96, with most of the pairings (496/688) having simple linear correlation values under 0.6. This indicates that, with some exceptions, few features in the probe set are good linear predictors, on their own, of features

in the gallery set. Figure 1 shows that a relatively small number of features are sufficient to establish the identity of an individual in the gallery set so, a naïve approach would be to employ individual features in the probe set to predict gallery set values, and using those estimates for identification.

Gallery set feature values were estimated using, for each, the single most highly correlated feature in the probe set as a predictor. This approach is intuitively appealing because only the best possible individual predictor is used, and probe set features containing little or no useful information as predictors of gallery features are ignored to the extent possible. Unfortunately, results from this approach are disappointing, as only 120 individuals end up with a Rank of 5 or better.

The relationship between features in the probe and gallery sets may be more complex than predictions from one simple linear regression model may be able to convey. A more sophisticated approach involves the use of multiple linear regression models to obtain estimates of gallery set features. In this approach, a linear model relating a feature in the gallery set to one or more features in the probe set is built using a training set of randomly selected individuals. These multiple linear regression models are built using a forward stepwise procedure (p-value to enter of 0.05 and 0.1 to remove), one for each feature in the gallery set, and all using information from the same set of individuals. The models can then be used to predict a vector with gallery set feature estimates using probe set data as input.

As we have shown, accurate re-identification can be carried out with a relatively small number of gallery set features. This means that, as long as some predicted gallery measures are available, finding the best possible match with the gallery data is possible. Naturally, the accuracy of the match will depend on the quality and the quantity of the feature estimates. The problem then becomes one of finding an optimum set of features that will result in a maximum identification rate. Other parameters, such as the quality of predictions and the size of the training set used to build the multiple linear regression models, could also affect the correct identification rate.

Searching through this space for an optimal set of parameters in an exhaustive way is not practically feasible. The number of combinations of two or more features in the gallery set is of the order of 8.8×10^{12}. If we consider also the size of the training set used to create the prediction models and the quality of the fits for an estimate to be considered good as two more parameters to be optimized, the size of the problem space becomes even more intractable.

To find a solution, a genetic algorithm (GA) was implemented. Genetic Algorithms are a heuristic optimization technique, loosely based on the Darwinian theory of evolution, in which selective pressure is exerted on an evolving population of solutions (or chromosomes) through mechanisms of recombination, selection and mutation [8],[9]. Repeated application of the GA mechanisms forces improvement in the fitness (or objective function) value of the population until convergence is reached.

The form of a GA solution for the re-identification problem considered here consists of a vector (or chromosome) with 45 entries. The first 43 are binary (1/0) entries indicating whether the corresponding feature in the gallery set will be estimated and used in the matching process or not. The 44^{th} entry in the chromosome is an R^2 threshold, indicating that only multiple linear regression models that match or exceed this threshold with the training set will be used to create a vector of estimated gallery feature values. The last entry in the chromosome is the size of the training set (number of individuals) used to build the multiple linear regression models. To ensure that only multiple linear regression models with at least a moderately good fit were considered, it was decided to limit the search of R^2 threshold values to the [0.5, 1] range. The size of the randomly selected training set was also limited to remain between 100 and 500 individuals. The limits in the size of the training set were imposed to determine if it is possible to build models that produce reasonably good and reliable predictions without having to use all the data available.

The GA creates an initial population of solutions at random called the parent population. A population of offspring solutions is obtained by combining the contents of chromosomes in the parent population. Evaluation of every solution in the offspring population is carried out by choosing a training set of individuals, of the size indicated by the solution, selected at random. A multiple linear regression model for each of the gallery features that have a '1' in a solution is built, employing a stepwise procedure, using data in the training set while all the features in the probe set remain available to build the models. To avoid overfitting, the models are limited to purely linear terms. After the models have been built, data for every individual in the probe set is used to predict values for the appropriate features in the gallery set, provided that the multiple linear regression model for that feature has an R^2 value that is equal or greater than the threshold indicated by the GA solution. The resulting vector of gallery set feature estimates is compared to all the available gallery data by computing standardized Euclidean distances. The standardized Euclidean distances are ranked in ascending order and the position where the correct ID is found is stored. For example, if the top match corresponds to the correct identity, this individual has a Rank of 1. However, if the top four matches for an individual are wrong (the individual is not one of these four persons in the gallery set) and the correct identity is found as the fifth match, this individual has a Rank of 5. The fitness value of the chromosome is obtained by adding the number of individuals with Rank 5 or better.

After the offspring solutions have been evaluated, their fitness values are sorted and the best solutions are selected to become the new parent population. To ensure that all feasible chromosome entries remain available, and to help

avoid premature convergence, a mutation mechanism is applied to the new parent population. Mutation consists of making random changes to a small number of individuals, also selected at random, in the new parent population. The mechanisms of recombination, evaluation, selection and mutation are applied repeatedly until some measure of convergence is achieved. In general, the fitness value of the best solution (or solutions) is used to determine if the GA has converged. When the fitness value of the best solution remains unchanged generation after generation, we say that the algorithm has converged. The re-identification algorithm described here was implemented in MatLab (R2013)[11].

3. Results

Table 3 shows the gallery set features selected by the GA with multiple regression models that match or exceed the R^2 threshold selected by the algorithm. The values of R^2 threshold and training set size selected by the GA are 0.87 and 389 respectively.

Table 3. Identities of the Features in the North American 1D Standing Set Selected by the GA

Feature Name	Feature Name
Acromial Ht Stand Lt	Infraorbitale Ht Lt Stand
Acromial Ht Stand Rt	Infraorbitale Ht Rt Stand
Acromion-Radiale Length Lt	Knee Ht Stand Rt
Acromion-Radiale Length Rt	Sleeve Outseam Len Lt
Axilla Ht Lt	Trochanterion Ht Lt
Bitrochanteric Brth Stand	Trochanterion Ht Rt

Results from Table 3 show that only a subset of less than half of the features in the standing set are selected by the GA as optimal for building an anthropometric signature for the available individuals. Over repeated independently started GA runs, most of the features in Table 3 are selected again, indicating that there is a subset of features that are robust for re-identification purposes under the approach used in this work. Even though it appears as if some of the features in Table 3 are redundant -both the left and right measurements of four features are selected by the algorithm- removing even just one or two features from those shown in Table 3, results in an average decline of about 5% in the number of re-identifications with Rank 5 or better. Removing all four 'right' features in Table 3 and keeping only the 'left' when both are present, results in an average decrease of 16% in the number of re-identifications with Rank 5 or better.

Figure 2 shows the cumulative proportion of individuals plotted against their re-identification rank for feature vectors estimated in three different ways:

1) Ten GA solutions using the features shown in Table 3 and multiple linear regression models built using ten training sets of 389 randomly chosen individuals and an R^2 threshold of 0.87 (solid lines).

2) Estimates for all 43 features in the gallery set, each estimated using a multiple linear regression model where all individuals are used to build the models and all the probe set variables are potentially available (dashed line).

3) The approach described in the first part of this paper, using all 43 features and simple linear regression models built using all individuals. Each feature in the gallery set is estimated using the simple linear model with the best R^2 of those available (dot and dash line).

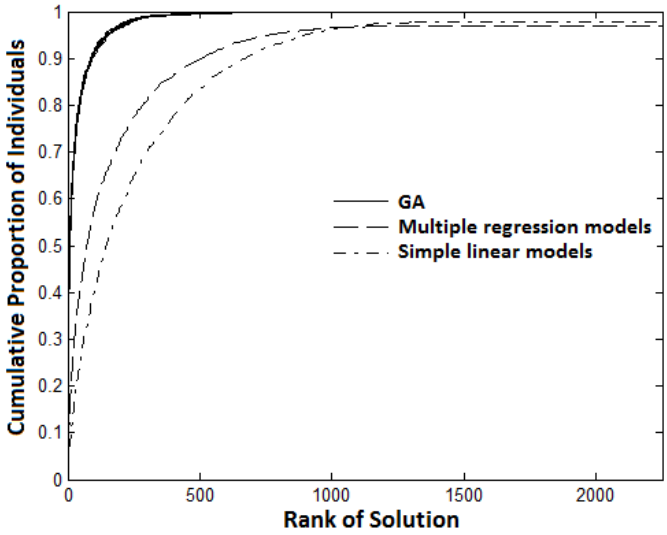

Figure 2. Plots of cumulative proportion of individuals (y-axis) for a given rank (x-axis) for ten solutions using the features found with the GA (solid lines), multiple linear regression models for all 43 gallery set features using information from all 2378 individuals (dashed line) and the best simple linear regression model for each feature in the gallery set (dot and dash line).

Figure 2 shows that the re-identification predictions found using the parameters reported by the GA exhibit better performance than multiple linear regression models obtained using all gallery set features and all individuals and much better than results obtained using the best simple linear regression model for each gallery set feature. This indicates that better quality re-identification results can be obtained by relying on a subset of accurately estimated features and that reliable predictive models for those features can be obtained using data from relatively few individuals.

Notice that, because the identity of the individuals selected to build the multiple linear regression models changes from generation to generation, even if the identity of the gallery set features chosen remains unchanged, the evolving solutions are robust against the particular subset of individuals used to build the models. Only solutions that perform well generation after generation, that is, solutions that maintain a high fitness value with relatively little

variability, will be maintained by the GA. This puts pressure on the algorithm to select gallery set features that can be reliably estimated without over-fitting.

Despite the encouraging results shown in Figure 2, the re-identification task remains challenging. Only an average of about 19% of the 2378 individuals receive a Rank of 1 using the GA solution over repeated runs using randomly selected subsets of 389 individuals to build the predictive models, and an average of 42% receive a Rank of 5 or better. Still, these results indicate that this re-identification approach may prove useful for greatly narrowing down the pool of individuals in a database that require closer inspection.

4. Conclusions and Future Work

We have presented a methodology to develop predictive models for biometric features linking two sets of distinct data involving 2378 individuals. Individuals in the gallery set can be unambiguously identified using only a few biometric measures if these measures are known, or can be estimated, with high accuracy. However, estimation of biometric features in a gallery set using as predictors data gathered under different circumstances presents a number of challenges. Investigating an adequate set of gallery features that can be predicted using features in a probe set is difficult because the combinatorial space is very large. In addition, the predictive models sought should be of enough quality (producing relatively accurate predictions) and should be robust to the particular subset of data used to build them.

A genetic algorithm (GA) was used to explore the problem space, searching for a group of gallery set features that could be linearly related to the features in the probe set. Results indicate that the GA selects less than half of the gallery set features to make a re-identification and that this approach produces better results than two other approaches that use information for all features and all individuals available.

The methodology presented in this paper could prove useful when incorporated into a re-identification system that is constantly updated. Biometric information from new individuals, or information from new biometric features, can be added and the algorithm trained again, helping in the development of a system that is robust and scalable.

In future work, we plan to investigate if re-identification performance can be improved by limiting the search to individuals that fit a profile consistent with an estimated vector of gallery set features. We are also exploring new modeling approaches, including feature transformations and different matching metrics. We are interested in studying the possibility of finding gallery set features that may be exchangeable to help in cases when one of the features selected by the GA is not available, and in determining the robustness of the multiple regression models when applied to a new set of data.

Acknowledgment

The research described in this paper is part of the Signatures Discovery Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. This research was also made possible by the Air Force Research Laboratory, who supplied the CAESAR data and provided valuable technical input.

References

[1] Han, J., Bhanu, B. (2006). Individual Recognition using Gait Energy Image, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 316-322.

[2] Satta, R. (2013). Appearance Descriptors for Person Re-Identification: a Comprehensive Review. arXiv e-print 1307.5748. URL http://arxiv.org/abs/1307.5748

[3] Ober, D.B., Neugebauer, S.P., and Sallee, P.A. (2010). Training and Feature-Reduction Techniques for Human Identification using Anthropometry. Biometrics: Theory Applications and Systems, Fourth IEEE International Conference on Biometrics. September 27-29, 2010, Washington, D.C.

[4] Godil, A., Grother, P., and Ressler, S. (2003). Human Identification from Body Shape. IEEE Fourth International Conference on 3-D Digital Imaging and Modeling. October 6-10, Banff, Canada.

[5] Bedagkar-Gala, A., and Shah, S.K. (2014). A Survey of Approaches and Trends in Person Re-Identification. Image and Vision Computing. Accepted Manuscript.

[6] Fouts, A., Rizki, M., Tamburino, L., Mendoza-Schrock, O. (2011). Evolving Robust Gender Classification Features for CAESAR Data. Proceedings of the IEEE 2010 National Aerospace and Electronics Conference. 20-22 July 2011, Dayton, OH.

[7] CAESAR: Civilian American and European Surface Anthropometry Resource Project. http://store.sae.org/caesar/

[8] Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading, MA.

[9] Holland, J.H. (1992). Adaptation in Natural and Artificial Systems. MIT Press, Cambridge, MA.

[10] Patrick, J., Clouse, H.S., Mendoza-Schrock, O., and Arnold, G. (2010). A Limited Comparative Study of Dimension Reduction Techniques on CAESAR. Proceedings of the IEEE 2010 National Aerospace and Electronics Conference. 14-16 July, 2010, Fairborn OH.

[11] MatLab. Version 8.1.0.604 (R2013a). The Mathworks Inc., Natick, MA, 2013.