Matched-Pair Classification Analysis on Siblings' Gait Biometric Data

Wan-Noorshahida Mohd-Isa*, Junaidi Abdullah, Chikkanan Eswaran, Amalina Ibrahim

Faculty of Computing and Informatics, Multimedia University,
63100 Cyberjaya, Selangor, Malaysia
*Corresponding author email: noorsha@mmu.edu.my

Abstract: This paper presents a supervised classification task on gait biometric of siblings' data sets. This task, which we refer to as matched-pair classification, evaluates the within pair differences in terms of the data set via jackknifing. A misclassification rate (MCR), which measures the percentage of misclassification of one sib compared to the other, gives an estimate on the potential uniqueness of gait for a person, particularly in twins. By this approach, the MCR values are mostly in the range of 90% for a data set of twins and in the range of 80% for a data set of non-twin siblings. When compared to the standard Leave-One-Out (LOO) classification, the MCR values of the proposed approach are higher than the LOO classification, which may suggest its potential use in machine learning with regard to biometric-based systems.

Keywords: matched-pair classification, jackknife-like, twins, siblings, machine learning, leave-one-out.

1. Background

The problem of classification in machine learning is to approximate an unknown class for an exemplar (or probe data) from a limited number of observations. In a supervised classification method, data sets are divided into a training (with known class label) and a test set. A classifier is then used to map the unlabeled exemplar from the test set to a labelled class, based on information learned from internal data structures of the training set. A common procedure, the Leave-One-Out (LOO) rule of matching one to a candidate of many is applied multiple times to measure classification performance.

Siblings are defined as a number of individuals who share at least one parent by blood. An individual, however, claiming to be unique, would have some shared features with his/her sib particularly in the case of twins. In such a case, the LOO rule may not be accurate when classifying a person, who may be a twin, into his/her unique class (or identity) in a biometric person identification system due to the shared features. The literature on face biometric, for example, has looked into unique features to uniquely identify a person with twin [1- 4], as have other researchers [5, 6].

Consequently, in statistics, siblings' biometric data are said to be dependent (or correlated). In statistics, a paired-test is a mechanism for comparing two dependent groups of data. Similarly, in psychology research, a matched-pair experimental design is most often used to analyze supposedly random people but who are believed to be related in some way (such as in studies of twins, husbands-wives). Thus, this paper adopts these concepts from these domains by proposing a matched-paired-test to measure the uniqueness of a sib when compared to his/her sib (in a pair) for person identification in siblings' biometric data sets.

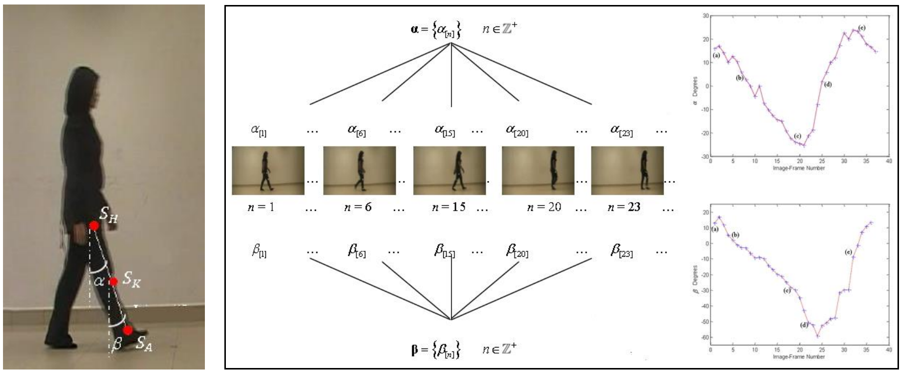
In this paper, the biometric that is investigated is the gait biometric. Gait is a behavioral biometric based on the way an individual walks [7]. Gait is believed to be unique to every person [8, 9]. It has the advantage that it can be observed from afar and hence suitable for use in video-based surveillance systems. Unlike the face, a gait biometric system does not require high resolution cameras to capture its detailed features since it captures only body movement patterns. Therefore, it is attractive to include gait in biometric-based systems to complement the shortcomings of face biometrics.

2. Methodology

2.1 Data Collection and Feature Extraction

To analyze the siblings' gait data in terms of supervised classification, the approach is to firstly collect the relevant data. Ten and 12 pairs of non-twins (NT) and twins (TW) siblings have volunteered to participate in this study. They are healthy young adults, aged between 16 and 28 years of age, with about the same body sizes. Videos of them walking indoors at their normal pace have been captured from a sideway's point of view, with a non-cluttered homogenous background. They have been asked to walk with flat shoes without carrying any objects from the left side to the right side and vice versa along a defined straight path at the filming location. These procedures have followed the data collection setup of other established gait research [10, 11].

By tracking each person's lower limbs from a sideway's point of view, the displacement values of the lower limbs with respect to his/her body's vertical axis throughout the video have been chosen as that person's gait feature, as illustrated in Figure 1(a). This follows successful similar research on unique individuals [12-15]. The forward movement of the lower limbs gives positive degree values and the reverse is a negative degree. The value of near to 0^o is set to be at the position when the lower limbs are about vertically straight. By gathering these displacement values from a video, these become a representative sinusoidal signal for a person's gait. These are illustrated in Figure 1(b).



(a) (b)

Figure 1. Sinusoidal feature extraction from a gait video.

2.2 Feature Selection

However, the length of each signal varies from person to person due to differences in walking speed. To establish invariance, the concept of a gait cycle as defined by Murray [8], can be used as a guideline for feature selection. After feature selection, signal resampling is performed in order to standardize the length of each signal. The resampling length value is chosen to be 30 based on the average length for all signals. Thus, the finalized feature (or **w** descriptor) is a concatenated sinusoid signal (Figure 2) of the two lower limbs' (thigh, $\boldsymbol{\alpha}$ – Figure 3(a) and lower leg, $\boldsymbol{\beta}$ – Figure 3(b)) displacement values that contain positive, zero and negative values identified within a gait cycle.

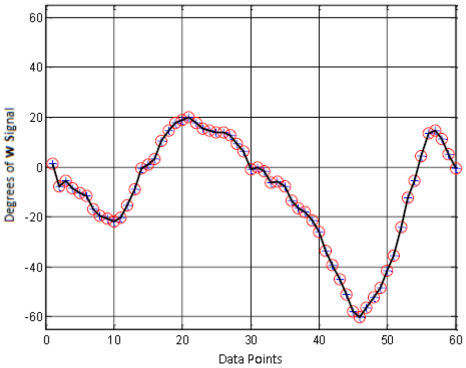
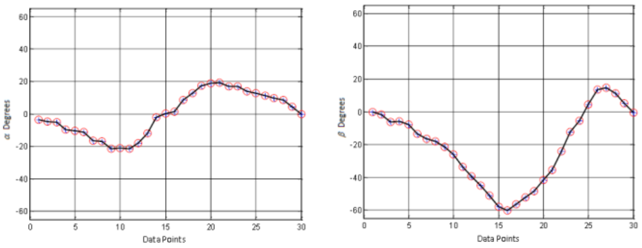


Figure 2. Thigh and lower leg signals for a person.



(a) (b)

Figure 3. Thigh and lower leg signals for a person.

The collection of all **w** signals is the gait biometric data for this paper. There are two data sets, one for non-twin siblings and another for twin siblings. Four **w** signals in total have been gathered from four videos of each person. They are defined as follows:

$$\mathbf{w}_i^{TW} = \{\boldsymbol{\alpha} \mid \boldsymbol{\beta}\} \quad \boldsymbol{\alpha},\boldsymbol{\beta} \in \left[-\pi/2, \pi/2\right]^{L_i} \tag{1}$$

$$\mathbf{w}_i^{NT} = \{\boldsymbol{\alpha} \mid \boldsymbol{\beta}\} \quad \boldsymbol{\alpha},\boldsymbol{\beta} \in \left[-\pi/2, \pi/2\right]^{L_i} \tag{2}$$

where,

$\mathbf{w}_i^{TW}$ is the **w** signal in the twin sibling data set

$\mathbf{w}_i^{NT}$ is the **w** signal in the non-twin siblings data set

$\boldsymbol{\alpha}$ is the thigh signal for a person

$\boldsymbol{\beta}$ is the lower leg signal for a person

$\mid$ is the concatenation operator

i is the video number $i = 1, 2, 3, 4$

L_i is the length of signal, in this paper $L_i = 30$.

2.3 Matched-Pair Classification with Jackknifing

To investigate the uniqueness of the descriptor for person identification, these data sets are tested through the use of supervised classification analysis within a matched-pair experimental setting. Due to the small sample size, only the linear discriminant (LDA) and *k*-nearest neighbor (*k*-NN) classifiers can be used. These classifiers are standard classifiers and have been performing well for small sample problem in previous gait recognition research [12, 14, 15].

The LDA finds discriminant functions that can give the maximum class separability via an eigen equation formed on a ratio of between class to within class scatter data. The class for an exemplar will be decided based on its nearest distance to the discriminant function in a data scatter space. The *k*-NN decides on the class label for an exemplar based on its nearest distance to some *k* data. The *k* distances create a pool of *k* class decisions that then form a class majority. The usual *k* values chosen relate to 1, 3, and 5 neighbors.

In our matched-pair setting, a paired-test is performed by looking into the within-pair differences. This can be done by omitting samples (signals) from data set of one sib and testing (classifying) an exemplar from this sib to the data set of its pair. The absence of samples from the *other* sib allows a paired-test for uniqueness. The concept of deleting samples from the data set is known as statistical jackknifing, and was invented by Quenouille [16]. This paper called it jackknife-like due to its difference from the original jackknife that calculates a pseudo-value, but ours do not estimate this value but instead we define a pseudo-class.

2.3.1 Jackknife-Like Paired-Test Formulation

Given a data set $X = \{P_1, P_2, ..., P_c\}$ with c number of classes and $P_i = \{B_1, B_2\}$ representing paired siblings, B_j for P_i of class i and sib j. Also each $B_j = \{s_1, s_2, ..., s_K\}$, denoted as s_k with a total of K number of samples. Let $C_n(X) = C_l(P_1, P_2, ..., P_c)$ be the class label for data set X at classification instance, n. The jackknife-like procedure defines a k^{th} pseudo-class for $C_n(X)$ to be,

$$ps^{k}_{i,j}(X) = C_n(P_1, P_2, ..., P_c) - C_n\left((P_1, P_2, ..., P_c)^{[k]}_{[i,j]}\right) + C_n\left((P_1, P_2, ..., P_c)^{[k]}_{[i,\sim j]}\right) \quad (3)$$

The $P^{[k]}_{[i,j]}$ is the sample $B_j = \{s_1, s_2, s_3, s_4\} \in P_i = \{B_1, B_2\} \in X = \{P_1, P_2, ..., P_c\}$ with the i^{th} class, j^{th} sibling, k^{th} sample, $s_k \in B_j \in P_i$ retained in the data set, but other K - 1 samples of i^{th} class and j^{th} sibling are omitted from the data set.

2.3.2 Jackknife-Like Paired-Test Algorithm

Using the jackknifing approach in a matched-pair classification with supervised data, the training set deletes any other sample data belonging to the class of the exemplar in the test set. This exemplar is assigned a supervised pseudo-class, which dictates that is belongs to the same class of his/her sibling. At each instance of a matched-pair classification, the hypothesis, which is if a sib's gait is unique, the pseudo-classes should give an inequality, $ps_{i,1} \neq ps_{i,2}$. If the equality fails to hold for the paired siblings, then this instance is counted as a misclassification. An average rate (in %) that calculates the misclassification instance is the average misclassification rate (MCR).

The algorithm to perform the jackknife-like paired-test within a matched-pair experiment is as follows:

```
Algorithm: Jackknife-Like_Matched-Paired-Test

Input  :  $X, C$
Output :  class, MCR

Initialization:  count_MCR = 0;

Loop   :

  for each value  $i$  of  $X$ , do
  for each value  $j$  of  $X$ , do
  for each value  $k$  of  $X$ , do
    set test =  $X_{[i,j,k]}$;

    set jackknife =  $\sim X_{[i,j,k]}$;
    set training = ~test - jackknife;
    set pseudoclass = $C\left(X_{[i,\sim j,k]}\right)$;
    set class_jackknife = $C\left(\sim X_{[i,j,k]}\right)$;
    set class_training = C(~test) - C(jackknife);
    solve classify;

    if class ~= pseudoclass
      count_MCR = count_MCR++;
    end if

  end for

MCR = count_MCR/$n$;
return MCR.
```

3. Results and Discussion

Table 1 lists the MCR for the matched-pair classification on both twin and non-twin siblings' data sets. Two standard similarity measures of Euclidean and City-Block distance are used with the k-NN classifier.

At the same time we have performed the standard LOO classification on the data sets for comparison. The difference between the LOO classification and our approach is that the LOO classification retains all the sample data and the class information as it is in the training set, and then performs classification on the test set. Our approach deletes all sample data belonging to the test set and assigns a pseudo-class to the exemplar in the test set. Thus, we are looking at a misclassification value because our hypothesis is that when the misclassification rate is high, the exemplar has been unable to be classified into his/her sib's class. This gives an estimate on a person's gait which may be unique when compared to that of his/her sib.

For the TW data set from the jackknife-like matched-pair classification, the k-NN classifier via the City-Block distance measure gives the MCR value of 95.8%. Similarly for the NT data set, the MCR value is 96.2% with k = 1 neighbor. This follows with the results of the City-Block k = 3 with 95%. However for the NT data set, the City-Block classifier of k = 5 attains only 83.7%, which is much lower than the result of LDA classifier with 86.2%. The second highest value for the TW data set from jackknife-like matched-pair classification is via the Euclidean classifier with 94.8% for all three k values. This is followed with 89.6% for the LDA.

As for the LOO classification, the highest MCR for the TW data set is 56.2% in terms of the LDA classifier. This is followed by 5-NN of both City-Block and Euclidean classifiers with 54.2% and 41.7%, respectively. Both k = 1 of City-Block and Euclidean classifiers give out the lowest MCR of 24% for the TW data set. Similarly both classifiers with k = 3 give MCR values of 34.4% for the TW data set.

The NT data set gives the highest MCR of 45% via the LOO classification using the Euclidean classifier with k = 5, followed by 35% with k = 3. The third highest is 30% via the City-Block classifier with k = 5. The LDA classifier gives a 25% MCR. The lowest MCR via the LOO classification is 5% by the 1-NN City-Block classifier.

Table 1. Comparison between the Misclassification Rate (MCR) of Jackknife-Like Matched-Pair Classification versus the Leave-One-Out (LOO) Classification

	MCR (%)			
	JACKKNIFE-LIKE MATCHED-PAIR CLASSIFICATION		**LEAVE-ONE-OUT (LOO) CLASSIFICATION**	
Classifiers	**TWINS (TW)**	**NON-TWINS (NT)**	**TWINS (TW)**	**NON-TWINS (NT)**
LDA	89.6	86.2	**56.2**	25.0
1-NN via Euclidean	94.8	83.7	24.0	12.5
3-NN via Euclidean	94.8	83.7	34.4	35.0
5-NN via Euclidean	94.8	83.7	54.2	**45.0**
1-NN via City-Block	**95.8**	**96.2**	24.0	5.0
3-NN via City-Block	**95.8**	95.0	34.4	23.7
5-NN via City-Block	**95.8**	83.7	41.7	30.0

When comparing the results of the jackknife-like matched-pair classification with the LOO classification, the MCR values of the LOO classification are mostly lower than 50%. Looking at the lowest MCR of the LOO classification on NT (5%), this means that only 95% of the LOO classification instances have correctly classified a unique person into his/her own class, as is expected for the NT data set. The LOO classification is expected to give an accurate estimate of a machine learning performance when using the NT data set, since the gait biometrics for a person has been shown to be unique in the literature. However, a misclassification error of higher than 5% for the TW data set may have not been acceptable in biometric-based systems design since we can say that 24% from the population of twins in the data set have been incorrectly identified, perhaps due to the similarity in their gait. With our approach, the MCR values are consistently high. This is where our approach can perhaps be used to investigate and estimate the uniqueness and hence viability of gait for use in biometric-based systems.

Nonetheless, there exists a small value of correctly classified instances, even though the MCR values are consistently high in the jackknife-like matched-pair classification. This indicates that the exemplar is nearest to one of his/her sib; thus it has been correctly classified to one of his/her sib's class in one of the classification instance. This may indicate a potential small similarity between each sib in the pairs of siblings of TW data sets.

4. Conclusion

This paper has presented a machine learning technique of supervised classification for data sets in pairs, which we refer to as a matched-pair classification via a jackknife-like approach. Jackknifing is a statistical technique that deletes samples from data set. In this paper, jackknifing has been applied in order to evaluate the within-pair difference with regard to siblings' gait biometric data in sets of twin and non-twin siblings. Misclassification values have been calculated to measure an estimate on the potential uniqueness of gait for an individual sib, and hence its viability for use in biometric-based systems. Comparing the MCR values of our approach to the standard LOO classification, it has been shown that our approach may have potential use in machine learning tasks related to the supervised classification of paired data.

5. Acknowledgements

This paper wishes to acknowledge the support from the Ministry of Higher Education (MOHE), Malaysia via its Exploratory Grant Scheme (ERGS).

References

[1] A.K. Jain, B. Klare, U. Park, "Face matching and retrieval in forensics application", *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 20-28, 2012.

[2] H. Li, D. Huang, L. Chen, Y. Wang, J.M. Morvan, "A group of facial normal descriptors for recognizing 3D identical twins", *Proc. 5th IEEE International Conf. Biometrics: Theory, Applications and Systems (BTAS)*, pp. 271-277, Sept. 2012.

[3] N. Srinivas, G. Aggarwal, P. Flynn, R.W. Vorder Bruegge, "Analysis of facial marks to distinguish between identical twins", *IEEE Trans. Information Forensics and Security*, vol. 7, no. 5, pp. 1536-1550, 2012.

[4] C.N. Dinakardas, S.P. Sankar, N. George, "Multimodal identification system in monozygotic twins", *International J. Image Processing (IJIP)*, vol. 7, no. 1, pp. 72-90, 2013.

[5] Z. Sun, A.A. Paulino, J. Feng, Z. Chai, T. Tan, A.K. Jain, “A study of multibiometric traits of identical twins”, *Proc. SPIE Biometric Tech. Human Identification*, vol. VII, no. 7667, pp. 76670T – 12, 2010.

[6] K. Hollingsworth, K. W. Bowyer, S. Lagree, S.P. Fenker, P.J. Flynn, “Genetically identical irises have texture similarity that is not detected by iris biometrics”, *J. Computer Vision and Image Understanding*, vol. 115, no. 11, pp. 1493-1502, 2011.

[7] M.S. Nixon, J.N. Carter, “Automatic recognition by gait”, in *Proc. The IEEE* vol. 94, no. 11, pp. 2013-2024, 2006.

[8] M.P. Murray, “Gait as a total pattern of movement”, *American J. Physical Medicine*, vol. 46, no. 1, pp. 290-333,1967.

[9] S.S. Stevenage, M.S. Nixon, K. Vince, “Visual analysis of gait as a cue to identity”, *Applied Cognitive Psychology*, vol. 13, no. 6, pp. 513-526, 1999.

[10] J.D. Shutler, M.G. Grant, M.S. Nixon, J.N. Carter, “On a large sequence-based human gait database”, *Applications and Science in Soft Computing, Adv. In Soft Computing*, vol. 24, pp. 339-346, 2004.

[11] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Gorther, K.W. Bowyer “The humanID gait challenge problem: data sets, performance, and analysis”, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 2, pp. 162-177, 2005.

[12] C. Yam, M.S. Nixon, J.N. Carter, “Automated person recognition by walking and running via model-based approaches”, *Pattern Recognition*, vol. 37, no. 5, pp. 1057-1072, 2004.

[13] I. Bouchrika, M.S. Nixon, “Gait recognition by dynamic cues”, in *Proc. 19th IEEE International Conf. Pattern Recognition (ICPR)*, pp. 1-4, Dec. 2008.

[14] N. Samudin, W.N. Mohd Isa, T. Maul, W. Lai, “Analysis of gait features between loaded and normal gait”, in *Proc. 5th IEEE International Conf. Signal-Image Technology and Internet-Based Systems (SITIS)*, vol. 27, no. 2, pp. 172-179, 2009.

[15] W.N.M. Isa, “Analysis on spatial and temporal features of gait kinematics”, in *Proc. 4th IEEE Workshop on Automatic Identification Advanced Technologies (AutoID)*, pp. 130-133, 2005.

[16] M.H. Quenouille, “A large-sample test for the goodness of fit of autoregressive schemes”, *J. Royal Statistical Society*, vol. 110, no. 2, pp. 123-129, 1947.