Gait-based Gender Classification Considering Resampling and Feature Selection

Raúl Martín-Félez, Vicente García, and J. Salvador Sánchez
Institute of New Imaging Technologies/DLSI-Universitat Jaume I, Castelló de la Plana, Spain
Email: {martinr, jimenezv, sanchez}@uji.es

***Abstract*—Two intrinsic data characteristics that arise in many domains are the class imbalance and the high dimensionality, which pose new challenges that should be addressed. When using gait for gender classification, benchmarking public databases and renowned gait representations lead to these two problems, but they have not been jointly studied in depth. This paper is a preliminary study that pursues to investigate the benefits of using several techniques to tackle the aforementioned problems either singly or in combination, and also to evaluate the order of application that leads to the best classification performance. Experimental results show the importance of jointly managing both problems for gait-based gender classification. In particular, it seems that the best strategy consists of applying resampling followed by feature selection.**

***Index Terms*—gender classification, class imbalance, high dimensionality, resampling, feature selection**

I. Introduction

Gait is a term to describe a particular manner of moving on foot, mainly walking. The main interest in gait analysis comes from conclusions drawn by Cutting et al. [1], where the ability of humans to recognize their friends from their unique gait pattern was proved. Therefore, gait can be deemed as a behavioral biometric feature, which allows the estimation of other properties inherent to humans such as gender [2], [3], [4] and age [5]. In addition, gait has several important strengths in comparison to other biometric features (face, voice, fingerprint, iris, hand geometry, etc), being the possibility of a reliable perception at a distance without requiring contact with any capturing device the most relevant one. Nevertheless, there are also important factors that hinder the implementation of gait-based systems. For instance, gait analysis is very sensitive to deficient segmentation of silhouettes, but also to the so-called covariate conditions including variations in clothing, footwear, mood, walking speed, carrying conditions, etc.

One of the main computer vision approaches to gait analysis are the so-called model-free techniques [3], [4], [6]. These attempts to represent the subject appearance changes from a sequence of silhouettes, which implicitly contain dynamic information. Although they were devised for gait-based human identification, these methods have also been successfully applied to gender classification. The most widely used model-free approach is the *Gait Energy Image* (GEI) method [6], consists of obtaining an average silhouette image to represent both body shape and movements over a gait cycle. Another important method known as *Active Energy Image* (AEI) focuses only on representing movements, since subject appearance is important but very sensitive to covariate factors. AEI is computed as the average of the absolute differences between each two consecutive frames, representing thus the average movements or variations between those frames. These gait representations lead to a *high dimensional space* since the gray-value of each pixel is used as a feature. Besides, as can be seen in Fig. 1 most of these features do not represent a person appearance or their movements but the background. Therefore, a feature selection/extraction process could lead to a quite smaller and more representative feature space.

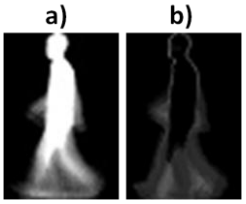

Figure 1. Two gait representations: a) Gait energy image (GEI) and b) Active energy image (AEI).

Apart from high dimensionality, another problem that must be tackled when using gait for gender classification is *class imbalance*. The best known benchmarking gait databases were created to study the problem of human identification, so they did not pay enough attention on the class distribution regarding gender. As a result, they are unequally distributed with respect to the number of men and women with a significant skew in favor of the men class (see Table I).

Learning from such a scenario without any pre-processing may lead to misleading results, generally because the class imbalance problem is fully ignored [3], [4], [7]. Despite several works [4], [7] have coped with this problem by using some small balanced subsets with an equal number of subjects per gender, their results strongly depend on the particular subsets selected and these might not consider all samples. A first attempt to properly deal with imbalance was proposed by Martín-

Manuscript received February 3, 2013; revised April 23, 2013.

doi: 10.12720/joig.1.2.85-89

Fdez *et al*. [8], where an ensemble of classifiers for learning from balanced subsets was implemented. Given an imbalanced data set, a number of balanced subsets are generated, each one containing all samples of the minority class (women) and as many randomly selected samples (with replacement) of the majority class (men) as needed to obtain a balanced subset. An odd number of individual classifiers are trained from the same number of balanced subsets, whose decisions are finally combined by simple majority voting.

So far as we know, no solution has been proposed to jointly address both high dimensionality and class imbalance in a gait-based gender classification domain. However, there exist several methods that have successfully been applied to a number of applications, including text categorization [9], medical diagnosis [10], remote sensing [11], intrusion detection [12] and software quality classification [13]. In comparison with the state-of-the-art, the main contribution of this paper is an insight about the suitability of some strategies to face the high dimensionality and imbalance problems either singly or in combination. Significant results are obtained from a thorough experimental study on two large gait databases.

II. Methodology

Given a gait video sequence of a person walking, a set of silhouettes are segmented from the corresponding frames. From these silhouettes, two gait representations are computed by using the GEI and AEI methods. As previously stated, any of these representations entail high dimensionality, which can be handled by some feature selection method (for example, those included in Sect. II-A). In addition, the benchmarking gait databases are biased with a higher number of men in comparison to that for women, so this imbalance can be tackled by some resampling techniques (see Sect. II-B).

Feature selection and resampling are considered as pre-processing steps that usually lead to a better feature space making easier the learning of the classifier. The effects of applying them singly or jointly (and, this case, the order of application) is here studied. It results in five strategies that are reflected in Fig. 2. The sixth one is a new proposal inspired in the work by Wang et al. [14] where some resampling method is applied to balance the data set before selecting features, and then the feature selection result is applied on the original imbalanced data. Note that, with this strategy, the resampling is only a mean to improve the feature selection process, thus avoiding the extra storage burden produced by some resampling techniques.

A. Feature Selection

Feature selection consists of reducing the dimensionality of data with the aim of allowing classifiers to operate faster and in general, more effectively. These methods can be categorized according to many criteria, for example: i) selection versus extraction of discriminant features and, ii) supervised versus unsupervised. Although feature extraction has been the common tool used in gait analysis, feature selection is here studied because its result can be visually understood and may produce improvements in performance. Therefore, the well-known unsupervised RELIEF method [15] and the supervised Threshold-based Feature Selection (TBFS) [14] have been chosen for the present analysis.

B. Resampling

One of the most common techniques to deal with imbalance consists of resampling the original data set either by over-sampling the minority class or by under-sampling the majority class until the class sizes are similar. In this work, two renowned resampling methods have been used to handle the class imbalance: Synthetic Minority Over-sampling TEchnique (SMOTE) [16] and Random Under-Sampling (RUS). The SMOTE algorithm generates new synthetic samples for the minority class. For each minority sample, this procedure computes the k intra-class nearest neighbors, and several new instances are created by interpolating the focused sample and some of its randomly selected neighbors. Its major drawback is an increase of the computational cost and the higher size of the resulting resampled data set. On the other hand, the RUS technique is a non-heuristic method that aims at balancing class distributions by randomly removing samples of the majority class. However, its primary shortcoming is that potentially useful data may be thrown out.

C. Evaluation Criteria

Unlike previous related works that only provide the overall accuracy to show their results [4], [7], this paper employs measures that are sensitive to the class imbalance effects. The plain accuracy, formulated as $Acc=(TP+TN)/(TP+FN+TN+FP)$, has been proved to be a measure that can be strongly biased towards the majority class [17]. Conversely, other performance measures are more suitable for imbalanced domains. For instance, the *True Positive* and *True Negative rates* (TPr and TNr) give the individual class performances, whereas the *Balanced Accuracy* (BA) provides a global unbiased performance assessment. In our scenario, TPr is defined as the percentage of women samples correctly classified, $TPr=TP/(TP+FN)$. Similarly, TNr is the percentage of men samples correctly classified, $TNr=TN/(TN+FP)$. Finally, the Balanced Accuracy uses both class classification rates to give a global clue about the classifier performance, $BA=(TPr+TNr)/2$. Note that this measure can be seen as the area under the ROC curve defined by a single point.

III. Experiments and Results

The empirical study is directed towards evaluating the six strategies previously introduced for tackling the high dimensionality and imbalance problems in gait-based gender classification. We are interested in investigating how each particular data complexity hinders classifier induction and discovering whether coping with them either singly or in combination is worthy. Also, when

both preprocessing tools are jointly used, we wonder which the best order of application is.

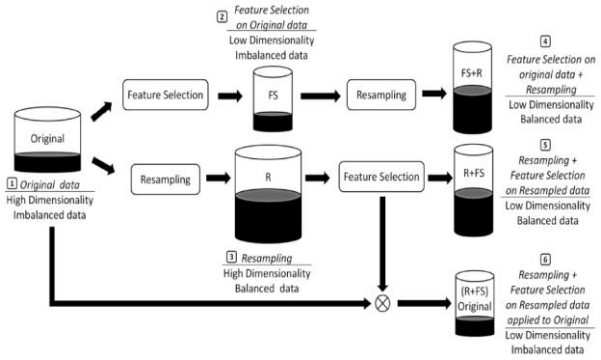

Figure 2. Two gait representations: a) Gait energy image (GEI) and b) Active energy image (AEI).

Experiments have been conducted on two public large gait databases: CASIA [18] - Data set B and the Southampton HID Database (SOTON) [19] - Set A. Table I summarizes their number of subjects (men and women) and total number of sequences. Note that the imbalance ratio regarding gender distribution of men with respect to women is 3:1 for CASIA and 4:1 for SOTON. Individuals recorded in both databases were filmed under controlled indoor conditions and walking in their side view. Only six sequences per person were considered without any covariate factor affecting the individual's gait. The well-segmented silhouettes provided by those databases are used as inputs to compute the corresponding GEI and AEI representations.

TABLE I. SOME CHARACTERISTICS OF THE LARGEST BENCHMARKING GAIT DATABASES

Databases	#Subjects	#Men	#Women	#Sequences Used
CASIA	124	93	31	744
SOTON	115	91	24	690

For each database, six different experiments are defined using a single sample per person. A stratified 5-fold cross validation scheme is performed on each experiment to estimate robust classification rates. It results in pairs of test and training sets, which are imbalanced according to the original class distribution. In addition, as only a sample per person is available, individuals in the training set are different to those in the test set, what demands a stronger generalization capability of the classifier. The widely-used *k*-Nearest Neighbor (*k*-NN) rule with *k=1* is employed in this analysis. In order to reduce the impact of subset singularities, the results correspond to the average of the classification rates across the six experiments for each strategy.

A. Results and Discussions

Fig. 3 shows bars whose height is the average balanced accuracy obtained with the 1-NN classifier for each strategy and each gait representation on the two databases. Table II reports the relative rank of each method, which has been computed according to the BA results. As there are six competing strategies the ranks are from 1 (best) to 6 (worst).

By analyzing Fig. 3 and Table II, some preliminary conclusions can be pointed out:

- As expected, the 1-NN classification on high-dimensional imbalanced data sets (without pre-processing) produces the poorest performance. This is not a surprising behavior because it is well-known that the *k*-NN rule is very sensitive to intrinsic data characteristics.
- The classification results using low-dimensional and/or balanced data sets are better than those with the original data set, what supports the previous assertion.
- When data complexities are faced in an individual way by using the FS or R approaches (see Fig. 2), the performance results are clearly inferior to those obtained when combining feature selection and resampling strategies (FS+R or R+FS).
- The results suggest that R+FS (resampling followed by feature selection) constitutes the best option for the high-dimensional imbalanced data because it shows a more stable behavior.
- It appears that the resampling methods help to provide more robust feature selection results. Such a behavior can be observed when comparing the FS and (R+FS)+Original approaches, where this shows a better ranking value. Note that in both cases the final subset of candidate features is generated from the original imbalanced data set. However, in the second case the feature selection process is done on a temporal balanced data set.

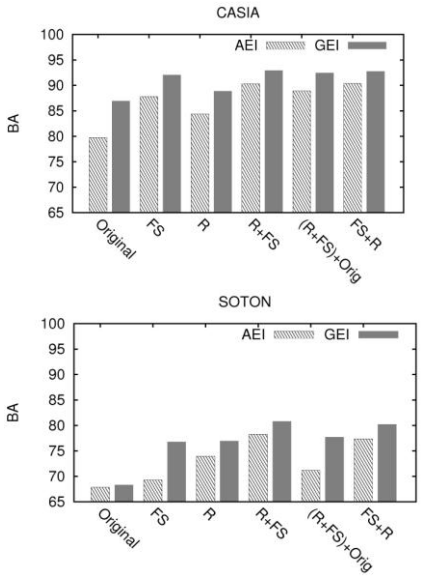

Figure 3. Average balanced accuracy obtained with 1-NN.

It is important to remark that all these findings can be observed for both gait representations and the two databases here experimented. Therefore the combined strategy of resampling followed by feature selection can be deemed as the most suitable solution.

Finally, it is worth pointing out that the SOTON database presents a higher variety of clothing and a wider

range of ages in comparison to CASIA. As the gait representations here used reflect not only the movement but also the appearance, the induced models are heavily affected by those factors. Consequently, SOTON generally achieves poorer results. Besides, it can be seen that resampling outperforms feature selection on SOTON but not on the CASIA database, probably because it presents a slightly higher imbalance ratio and also the aforementioned representation problem.

TABLE II. STRATEGIES SORTED BY THEIR RANK VALUES

	CASIA		SOTON	
Rank	AEI	GEI	AEI	GEI
1	FS+R	R+FS	R+FS	R+FS
2	R+FS	FS+R	FS+R	FS+R
3	(R+FS) + Original	(R+FS) + Original	R	(R+FS) + Original
4	FS	FS	(R+FS) + Original	R
5	R	R	FS	FS
6	Original	Original	Original	Original

IV. CONCLUSIONS

This paper has focused on gait-based gender classification when data sets present high dimensionality and class imbalance. Six different approaches have been analyzed when addressing both intrinsic data complexities. The experiments carried out on two public large gait databases with two different gait representations have demonstrated that resampling methods help to provide more robust feature selection results. This suggests that in high-dimensional imbalanced data sets is more important to balance the class distribution rather than to reduce the dimensionality. However, these conclusions should be taken just as a reference because it has been observed that results strongly depend on the particular characteristics of each database. Finally, future research will be mainly addressed to study other data complexities such as class overlapping, small disjuncts and data set shift.

ACKNOWLEDGMENT

This work has partially been supported by projects TIN2009-14205 from the Spanish Ministry of Innovation and Science and P1-1B2012-22 from Universitat Jaume I, and by PREDOC/2008/04 grant from Universitat Jaume I. Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences.

REFERENCES

[1] J. Cutting and L. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353-356, 1977.

[2] L. Kozlowski and J. Cutting, "Recognizing the sex of a walker from a dynamic point-light display," *Perception & Psychophysics*, vol. 21, pp. 575-580, 1977.

[3] X. Lin, S. Maybank, S. Yan, D. Tao, and D. Xu, "Gait components and their application to gender recognition," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 2, pp. 145-155, 2008.

[4] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Trans. on Image Processing*, vol. 18, no. 8, pp. 1905-1910, 2009.

[5] Y. Makihara, H. Mannami, and Y. Yagi, "Gait analysis of gender and age using a large-scale multi-view gait database," in *Proc. 10th Asian Conf. on Computer Vision*, vol. 2, Queenstown, New Zealand, 2011, pp. 440-451.

[6] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316-322, 2006.

[7] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. 5th IEEE Int'l. Conf. on Automatic Face and Gesture Recogn.*, Washingon, DC, 2002, pp. 148-155.

[8] R. Martín-Félez, R. A. Mollineda, and J. S. Sánchez, "A gender recognition experiment on the CASIA gait database dealing with its imbalanced nature," in *Proc. 5th Int'l Conf. on Computer Vision Theory and Applications*, vol. 2, Angers, France, 2010, pp. 439-444.

[9] M. Wasikowski and X.-W. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1388-1400, 2010.

[10] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 11, no. 1, pp. 523:1-523:7, 2010.

[11] X. Chen, T. Fang, H. Huo, and D. Li, "Semisupervised feature selection for unbalanced sample sets of VHR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 781-785, 2010.

[12] M. Alshawabkeh, M. Moffie, F. Azmandian, J. Aslam, J. Dy, and D. Kaeli, "Effective virtual machine monitor intrusion detection using feature selection on highly imbalanced data," in *Proc. 9th Int'l Conf. on Machine Learning and Applications*, Washington, DC, 2010, pp. 823-827.

[13] T. Khoshgoftaar, K. Gao, and A. Napolitano, "Exploring an iterative feature selection technique for highly imbalanced data sets," in *Proc. 13th IEEE Int'l. Conf. on Information Reuse and Integration*, Las Vegas, NV, 2012, pp. 101-104.

[14] H. Wang, T. Khoshgoftaar, and J. Van Hulse, "A comparative study of threshold-based feature selection techniques," in *Proc. IEEE Int'l. Conf. on Granular Computing*, San Jose, CA, 2010, pp. 499-504.

[15] K. Kira and L. A. Rendell, "A practical approach to feature selection", in *Proc. 9th Int'l. Workshop on Machine Learning*, Aberdeen, UK, 1992, pp. 249-256.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[17] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," in *Proc. 3rd Int'l. Conf. on Knowledge Discovery and Data Mining*, Newport Beach, CA, 1997, pp. 43-48.

[18] CASIA. (2005). CASIA Gait Database. [Online]. Available: http://www.sinobiometrics.com

[19] J. Shutler, M. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human," in *Proc. 4th Int'l. Conf. on Recent Advances in Soft Computing Gait Database*, Nottingham, UK, 2002, pp. 66-71.

Raúl Martín-Félez received the Engineering degree in Computer Science and the M.Sc. degree in Intelligent Systems both from the Universitat Jaume I, Castellón de la Plana, Spain, in 2007 and 2009, respectively.

He is currently a Ph.D. student at the Department of Computer Languages and Systems, Universitat Jaume I. His research mainly focuses on the area of biometric systems, specifically on using gait for human recognition and gender estimation. However, he has also addressed other related problems such as video surveillance and perimeter control. Eventually, he has also some experience with dealing with some complexity data, especially with dataset shift, high dimensionality and imbalance.

Mr. Martín-Félez is co-author of more than 10 scientific publications including conferences as ECCV, ICPR, IBPRIA, VISAPP, etc. and is a member of IAPR and AERFAI (Spanish Association of Pattern Recognition and Image Analysis).

Vicente García received the Engineering degree in Computer Systems from the Technological Institute of Villahermosa, Mexico, in 2000, the M.Sc. in Computer Science from the Technological Institute of Toluca, Metepec, Mexico, in 2002, and the Ph.D. degree in Computer Science from the Universitat Jaume I, Castellón de la Plana, Spain, in 2010.

He is now a Research Fellow at the Institute of New Imaging Technologies, Universitat Jaume I. Previously, he was with the Government of the Estado de Mexico as head of computer projects. He is co-author of more than 30 scientific publications and he has collaborated in several R+D projects, most of them related to pattern recognition methodologies and applications. His current research interests include classification, ensembles of classifiers, and analysis of data complexity. He is a member of IAPR and AERFAI.

J. Salvador Sánchez received the B.Sc. degree in Computer Science from the Technical University of Valencia, Spain, in 1990 and the Ph.D. degree in Computer Science Engineering from Universitat Jaume I, Castellón de la Plana, Spain, in 1998.

He is a Full Professor at the Department of Computer Languages and Systems, Universitat Jaume I, and is currently the head of the Pattern Analysis and Learning Lab. He is author or co-author of more than 140 scientific publications and co-editor of three books. His research lies in the fields of pattern recognition, machine learning and data mining, including classification, feature and prototype selection, ensembles of classifiers, data analysis, reinforcement learning and biometry.

Prof. Sánchez is a Senior Member of IEEE, IAPR, AERFAI, ECCAI, and AEPIA (Spanish Association for Artificial Intelligence).