A. Brömme, C. Busch, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 1

Predicting Missing Demographic Information in Biometric Records using Label Propagation Techniques

Thomas Swearingen[1] and Arun Ross[2]

Abstract: Biometric systems use biological attributes such as face, fingerprint, or iris to automatically recognize an individual. In many law enforcement applications, the biometric record of a person in the database is often supplemented with biographic and demographic data such as name, address, age, ethnicity, gender, etc. In such applications, some of the records may have missing or incorrect demographic data. In this work, we explore the potential of a label propagation method to impute demographic data to partially incomplete biometric records. The proposed method utilizes a graph-like structure to capture similarities between biometric records based on the available data. This structure is then used by the label propagation method to predict missing data. Experiments using the face image, name, gender and ethnicity of subjects in two datasets confirm the efficacy of the scheme in predicting missing data in biometric records.

Keywords: Biometrics, Face Recognition, Missing Data, Demographic Prediction, Label Propagation, Graphs, Gender, Ethnicity

1 Introduction

In classical biometric recognition, a biometric probe sample is compared against a set of biometric gallery samples in order to recognize an individual. For example, an unknown face image (the probe) may be compared against a set of known face images in a database (the gallery) in order to recognize it. The gallery samples pertaining to multiple individuals are often assumed to be independent of one another. Therefore, the probe sample is compared against every gallery sample (or a subset of gallery samples) *independently* in order to generate match scores. These match scores are then used to either *verify* the claimed identity of the probe sample, or to *determine* the identity associated with the probe sample [JRN11].

We explore the use of a graph structure to model the relationship *between* gallery samples. In this graph, each node corresponds to an identity and the edges between pairs of nodes describe the similarity between identities. Each identity (node) is a combination of biometric, biographic (*e.g.,* name, occupation) and demographic (*e.g.,* ethnicity, gender) data of a person. The relationship between two nodes (manifested as an edge or a set of edges) is defined by the similarity between their biometric, biographic and demographic data. Utilizing a graph has several advantages. The output of the identification process could be a subgraph consisting of not only "matching" candidates whose face images look similar to the probe image, but also other candidate images that are "related" to the probe. For example, when searching for the identity of a probe sample in the graph, the output may consist of gallery identities that are in social or professional proximity to the individual

[1] Graduate Student, Michigan State University, swearin3@msu.edu
[2] Professor, Michigan State University, rossarun@msu.edu

(such as a close friend or a co-worker). This would be useful in cases where the identity of the probe is not in the gallery, but related identities are present in the graph.

In this work, we focus on one advantage of such a graph — **the ability to deduce missing (or incorrect) data in a node**. Some nodes are likely to have incomplete information (*e.g.,* there may be missing demographic information). Deducing missing information can improve the integrity of the gallery database. Since a similar pair of nodes are likely to have a stronger edge between them, the confluence of information from neighboring nodes can be used to impute missing values to an incomplete node.

In the graph formulation, every node can be viewed as a *record* consisting of several *fields* such as name, age, gender, ethnicity, face image, *etc.* The edge weights in the graph is a function of the degree of similarity between two participating nodes and is based on the available fields in the nodes. Some nodes may be complete, in that all their field values are available, while other nodes may be incomplete.

Many operational biometric systems store the biographic and demographic data of a person in addition to the biometric data. Examples include the UIDAI Aadhaar program in India, and the OBIM and TWIC programs in the United States. In such programs, the gallery database could be viewed as consisting of "records" of individuals. The main contribution of this work is a method for predicting demographic attributes of a biometric record that does not rely on the face image or name alone, but exploits the existing relationship between records. The method is based on the label propagation technique. Section 2 provides a review of related literature. Section 3 presents methods to predict demographic information from faces and names, which will be used as a baseline for comparison to the label propagation method. Section 4 details the label propagation method. Section 5 reports the experiments and results. Section 6 discusses the results and Section 7 offers concluding remarks.

2 Literature Review

2.1 Predicting Demographic Information

The field of soft biometrics, amongst other things, has focused on deducing demographic information from biometric data. There is a rich literature on this topic and we refer the interested reader to [DER16]. In particular, the problems of age [FGH10], race [FHH14] and gender [MR08] prediction from face images have been studied in detail. In addition, there have been some preliminary attempts to predict a person's occupation or name from a face photo [CC14, CGG13], but the success of such methods has been extremely limited compared to gender/race prediction.

2.2 Label Propagation

Label propagation algorithms are examples of semi-supervised learning techniques. Label propagation operates under the assumption that points on the same manifold are likely to have the same label. Since both labeled and unlabeled data are available, the goal is to induce labels on the unlabeled data from the labeled data using the natural structure of the manifolds within the data. This is accomplished by constructing a graph of nodes and

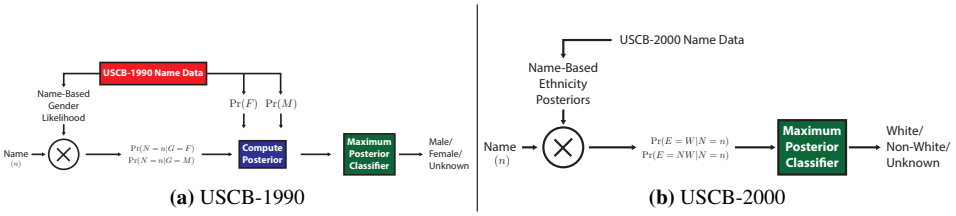
edges. The nodes are simply the data points while the edge weights represent the similarity between these data points. In the past, label propagation has been used to improve the quality of labeling in datasets where there are missing or incorrect labels [Li11, Ta15].

3 Predicting Demographic Information from a Single Attribute/Record

In this work, we consider two types of demographic attributes: gender and ethnicity. Each attribute has two labels associated with it. The gender labels are *Male* and *Female*, while the ethnicity labels are *White* and *Non-White*. However, this work is applicable even when other types of attributes and labels are used. Before we describe the label propagation method used in this work, we first establish *baseline methods* where gender or ethnicity is deduced from a *single* attribute of a *single* record.

3.1 Demographic Prediction from Name

We use two data sources from United States Census Bureau (USCB) for predicting gender and ethnicity from name. The first dataset, USCB-1990, is used to predict gender from a forename. The USCB published a list of forenames that were reported in the 1990 census and their corresponding frequencies for male and female categories [Fr95]. The second dataset, USCB-2000, is used to predict ethnicity from a surname. The USCB published a list of surnames that were reported in the 2000 census and their corresponding ethnicity posteriors [DLWK]. Figure 1 shows an overview of these methods.



(a) USCB-1990 **(b)** USCB-2000

Fig. 1: Overview of the USCB-1990 Gender-from-Name Demographic Classifier and the USCB-2000 Ethnicity-from-Name Demographic Classifier. In Figure 1(a), a forename, *n*, is input into the system and a gender label, *Male/Female/Unknown*, is output. In Figure 1(b), a surname, *n*, is input into the system and an ethnicity label, *White/Non-White/Unknown*, is output.

3.2 Demographic Prediction from Face Image

As stated earlier, there are a number of publications that discuss the possibility of deducing gender and ethnicity from face images. In this work, we used the *Intraface* SDK to deduce ethnicity and gender from face images. *Intraface* is a face attribute extractor that includes functionality for determining ethnicity and gender [To15]. Given an input face image, the software outputs gender and ethnicity labels. The authors in [To15] tested their software on the PubFig dataset [Ku09] where it obtained an F1 score of 96.1% for gender prediction and an average F1 score of 91.8% for ethnicity prediction. Therefore, it is an appropriate choice for automatic demographic attribute extraction from face images.

4 Predicting Demographic Information Using Multiple Records

In contrast to the baseline demographic prediction schemes discussed in the previous section, here we employ a graph-based method that uses evidence from *multiple* records (nodes) to estimate missing demographic values in a biometric record. In the proposed method, the gallery data is organized in a graph structure with nodes corresponding to biometric records and edge weights corresponding to degrees of pair-wise similarity between nodes. In such a graph, we identify two types of nodes:

1. Complete Node: A nodal record which has no missing fields.
2. Incomplete Node: A nodal record that has one or more missing demographic fields.

In order to propagate demographic labels from the complete nodes to the incomplete nodes, we use a label propagation method [Zh04]. Suppose that we are given a set of records/nodes $\mathscr{R} = \{R_1, \ldots, R_v, R_{v+1}, \ldots, R_n\}$ where the first v records are complete and the remaining $n-v$ records are incomplete. Each record has 4 fields: face, name, gender and ethnicity. Therefore, let $R_i = \{F_i, N_i, G_i, E_i\}$. Here, gender and ethnicity are viewed as binary attributes who values are in the label set $\mathscr{L} = \{0,1\}$. For ethnicity, 0 is non-white and 1 is white. For gender, 0 is female and 1 is male. Let $\{y_1, y_2, \ldots, y_v\}$, $y_i \in \mathscr{L}$, be the gender (or ethnicity) labels of the complete nodes.

Algorithm 1 Demographic Label Propagation

1: **procedure** PROPAGATELABELS($\mathscr{R}, Y, \sigma, \alpha$)
2: **for** $i, j \in [1,n]$ **do**
3: **if** i = j **then**
4: $W_{ij} = 0$
5: **else**
6: $W_{ij} = \exp\left(-\frac{f_{\text{diff}}(R_i,R_j)^2}{2\sigma^2}\right)$ ▷ Edge weights are based on record similarity.
7: **end if**
8: **end for**
9: $D_{ii} = \text{zeros}(n)$
10: **for** $i \in [1,n]$ **do**
11: $D_{ii} = \sum_{j=1}^{n} W_{ij}$ ▷ Diagonal entries are the sum of the corresponding row in W.
12: **end for**
13: $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$
14: $F^* = (I - \alpha S)^{-1} Y$
15: **for** $i \in (v,n]$ **do**
16: $l_i = \text{argmax}_{0 \le j < k} F^*_{ij}$
17: **end for**
18: **return** l_i's ▷ Labels for incomplete nodes.
19: **end procedure**

Algorithm 1 details the label propagation method. To propagate the demographic labels, we first calculate the affinity matrix for the graph. Note that the $f_{\text{diff}}(R_i, R_j)$ function on Line 6 produces a scalar difference score between records R_i and R_j using the match scores between available fields. We then normalize the affinity matrix with the sum of each row resulting in the similarity matrix S. Next, we construct a matrix Y of size $n \times k$, where k is the cardinality of the label set. Without loss of generality, let us assume that the *gender* labels are missing for the incomplete nodes, R_{v+1} to R_n. Since $k = 2$, Y will be of size $n \times 2$. The i^{th} row of Y pertains to record R_i, and has two columns: $Y_{i,1} = 1$, if $G_i = 0$ and $Y_{i,2} = 1$, if $G_i = 1$. However, for incomplete records that are missing gender labels, both

$Y_{i,1}$ and $Y_{i,2}$ are set to 0. We then use Y to let information "flow" from the complete nodes to the incomplete nodes by using the node relationships manifested as values in S.

As noted by [Zh04], rather than iteratively pushing label information between nodes, we can compute the final values directly as $F^* = (I - \alpha S)^{-1} Y$. The $(I - \alpha S)^{-1}$ term can be seen as a diffusion kernel that diffuses the complete node labeling from the upper part of Y onto the incomplete nodes in the lower part of Y. A label for an incomplete node is primarily dictated by the subset of nodes that sent the most information to it. The labels for the incomplete nodes can be directly derived from F^*. The i^{th} row of F^* has two columns; if $F^*_{i,0} > F^*_{i,1}$ then incomplete node i is predicted to have label value 0, else it is predicted to have label value 1.

5 Experiments and Results

5.1 Dataset

Most of the publicly available face datasets do not include information about the name, gender and ethnicity of the subjects. Therefore, we assembled two datasets from images downloaded from the Web: (1) Knox County Arrest Dataset and (2) Online Celebrity Dataset. Two subsets are next appropriated from each dataset, one for gender prediction and one for ethnicity prediction. In each subset, one of the two demographic attributes has equal-sized cohorts. Figure 2 shows examples from both datasets and Table 1 summarizes the demographic statistics of the datasets.

(a) Knox County Arrest Dataset **(b)** Online Celebrity Dataset

Fig. 2: Example of biometric records from the two datasets assembled in this work.

Knox County Arrest Dataset: The Knox County Sheriff's Office posts information about arrested individuals every 24 hours. We downloaded arrestee information using an automated script in order to compile the Knox County Arrest Dataset. The Knox County Arrest Dataset consists of 1,422 records each of which includes forename, surname, gender, ethnicity, and a face image.

Online Celebrity Dataset: We also assembled another dataset, that we refer to as the Online Celebrity Dataset, which contains biographic and demographic details of several celebrities. It consists of 521 records and each record contains forename, surname, gender, ethnicity, and 2 face images.

5.2 Demographic Prediction

Single Record Prediction: Table 2 and Table 3 show the baseline results of gender prediction and ethnicity prediction, respectively, based on only names or faces from the Online

Tab. 1: Demographic counts of datasets by male (*M*), female (*F*), white (*W*), and non-white (*NW*).

Dataset	Gender		Ethnicity	
	M	*F*	*W*	*NW*
Knox County Arrest Dataset	1001	421	1099	323
KC-G: Knox County Subset G (for Gender Prediction)	**421**	**421**	678	164
KC-E: Knox County Subset E (for Ethnicity Prediction)	498	148	**323**	**323**
Online Celebrity Dataset	246	275	399	122
OC-G: Online Celebrity Subset G (for Gender Prediction)	**246**	**246**	375	117
OC-E: Online Celebrity Subset E (for Ethnicity Prediction)	140	104	**122**	**122**

Celebrity Dataset. Table 2 shows results on the Online Celebrity G dataset. Table 3 show results on the Online Celebrity E dataset.

Graph-Based Multi-Record Prediction: For the label propagation method, the KC-G Dataset is used as the complete node set and the OC-G Dataset is used as the incomplete node set for gender prediction. For ethnicity prediction, the KC-E Dataset is used as the complete node set and the OC-E Dataset is used as the incomplete node set. The optimal parameter values for the label propagation method are found by performing a grid search over the σ and α parameters. The value for each parameter is varied in the interval $[0.01,\ 0.99]$ in steps of 0.01. The overall test classification accuracy as a function of σ and α is shown in Figure 3. For gender prediction, the optimal value of σ is 0.12 and the optimal value of α is 0.91. For ethnicity prediction, the optimal value of σ is 0.15 and the optimal value of α is 0.13. Table 4 shows the results of gender and ethnicity prediction using the label propagation method with these optimal values. The label propagation method uses all of the fields except the (missing) field that is being predicted. When finding similarity (edge weights) between records in the complete node set, the face, name, gender, and ethnicity fields are used. *When predicting gender* and comparing nodes where at least one node is from the incomplete node set, then only the face, name, and ethnicity fields are used. *When predicting ethnicity* and comparing nodes where at least one node is from the incomplete node set, then only the face, name, and gender fields are used.

Tab. 2: Results of gender prediction via name (USCB-1990) and face (Intraface) on the OC-G Dataset.

	USCB-1990 (%)	Intraface (%)
Overall	97.8	97.6
Female	98.8	97.6
Male	96.7	97.6

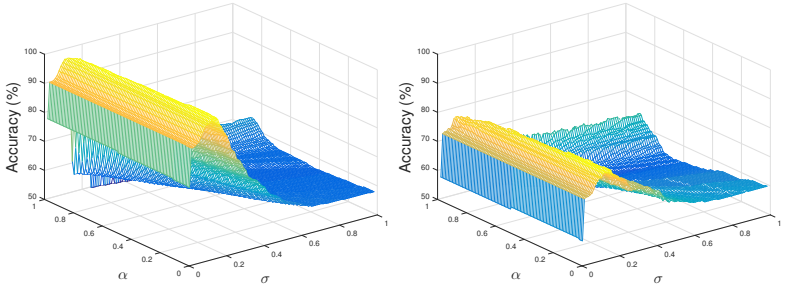
Tab. 3: Results of ethnicity prediction via name (USCB-2000) and face (Intraface) on the OC-E Dataset.

	USCB-2000 (%)	Intraface (%)
Overall	60.7	85.7
Non-White	40.2	73.0
White	81.1	98.4

The face similarity scores are obtained using a COTS face matcher. The name similarity is calculated using the levenshtein distance between the two names. For gender (ethnicity), the score is set to 1.0 if the gender (ethnicity) matches, and 0.0 if it does not.

6 Discussion

Based on the results shown in Section 5, we observe that gender prediction from face and name works well (>95% in all cases). However, ethnicity prediction exhibits a lower performance: only 85.7% from face and 60.7% from name.



(a) Gender prediction accuracy as a function of σ and α

(b) Ethnicity prediction accuracy as a function of σ and α

Fig. 3: Impact of label propagation parameters, σ and α, on demographic prediction.

Tab. 4: Results of gender and ethnicity prediction using the label propagation technique on the Knox County Dataset G and E Subsets, respectively, as the complete nodes and Online Celebrity Dataset G and E subsets, respectively, as the incomplete nodes.

Demographic	**Overall Accuracy (%)**	**Female Accuracy (%)**	**Male Accuracy (%)**
Gender	98.37	97.56	99.19
		Non-White Accuracy (%)	**White Accuracy (%)**
Ethnicity	80.33	61.48	99.18

We observe that the graph-based label propagation method results in similar performance as that of single attribute-based approaches. The label propagation prediction accuracy is 98.37% for gender and 80.33% for ethnicity. The difference between the two approaches is that the graph-method based is more easily extensible to other attributes. The name-based and face-based methods have a limited number of attributes for which they can be effective predictors (*e.g.,* gender, ethnicity, age, *etc.*).

The advantage of the record-based label propagation method is that it utilizes the evidence from multiple records and multiple fields in order to predict missing values. Thus, prediction is based on *relationships* that exist between records. This type of relationship is not captured in single attribute-based classifiers, but is more easily captured in graph-based approaches. When predicting a full range of attributes (e.g. occupation, education-level, *etc.*), we believe a graph-based method will yield better results than single attribute-based predictors.

7 Conclusion

In this work, we demonstrated the benefit of structuring a biometric gallery using a graph structure. Such a structure not only captures the relationship between gallery records, it can also be used to deduce missing (or overwrite incorrect) information in these records. A label propagation scheme was adopted to illustrate the possibility of imputing missing gender and ethnicity information. Experiments on two datasets demonstrated that the proposed method is (a) capable of imputing missing information and (b) generalizable across datasets. Future work would involve testing the method on demographic attributes that have more than two labels.

Acknowledgement

This project was supported by Award No. 2015-R2-CX-0005, from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

References

[CC14] Chu, W. T. ; Chiu, C. H. : Predicting Occupation from Single Facial Images. In: Proceedings of the IEEE International Symposium on Multimedia. ISM '14, IEEE Computer Society, Washington, DC, USA, pp. 9–12, 2014.

[CGG13] Chen, H. ; Gallagher, A. C. ; Girod, B. : What's in a Name? First Names as Facial Attributes. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3366–3373, 2013.

[DER16] Dantcheva, A. ; Elia, P. ; Ross, A. : What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Transactions on Information Forensics and Security, 11(3):441–467, March 2016.

[DLWK] Word, D. L. ; Coleman, C. D. ; Nunziata, R. ; Kominski, R. : Demographic Aspects of Surnames from Census 2000. Technical report, United States Census Bureau.

[FGH10] Fu, Y. ; Guo, G. ; Huang, T. S. : Age Synthesis and Estimation via Faces: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(11):1955–1976, Nov 2010.

[FHH14] Fu, S. ; He, H. ; Hou, Z. G. : Learning Race from Face: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(12):2483–2509, Dec 2014.

[Fr95] Frequently Occuring First Names and Surnames From the 1990 Census. Technical report, United States Census Bureau, 1995.

[JRN11] Jain, A. ; Ross, A. ; Nandakumar, K. : Introduction to Biometrics. Springer US, 2011.

[Ku09] Kumar, N. ; Berg, A. C. ; Belhumeur, P. N. ; Nayar, S. K. : Attribute and Simile Classifiers for Face Verification. In: IEEE International Conference on Computer Vision (ICCV). Oct 2009.

[Li11] Liu, D. ; Yan, S. ; Hua, X. S. ; Zhang, H. J. : Image Retagging Using Collaborative Tag Propagation. IEEE Transactions on Multimedia, 13(4):702–712, Aug 2011.

[MR08] Makinen, E. ; Raisamo, R. : Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3):541–547, March 2008.

[Ta15] Tang, J. ; Li, M. ; Li, Z. ; Zhao, C. : Tag Ranking Based on Salient Region Graph Propagation graph propagation. Multimedia Syst., 21(3):267–275, 2015.

[To15] De la Torre, F. ; Chu, W.-S. ; Xiong, X. ; Vicente, F. ; Ding, X. ; Cohn, J. : IntraFace. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. pp. 1–8, May 2015.

[Zh04] Zhou, D. ; Bousquet, O. ; Lal, T. N. ; Weston, J. ; Schölkopf, B. : Learning with Local and Global Consistency. In: Advances in Neural Information Processing Systems 16, pp. 321–328. MIT Press, 2004.