Multi-modal Open-Set Person Identification in HRI

Irfan, B

http://hdl.handle.net/10026.1/16699

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.
Multi-modal Open-Set Person Identification in HRI

Bahar Irfan  
CRNS, Plymouth University, UK  
bahar.irfan@plymouth.ac.uk

Michael Garcia Ortiz  
SoftBank Robotics Europe, France  
mgarciaortiz@softbankrobotics.com

Natalia Lyubova  
SoftBank Robotics Europe, France  
nlyubova@softbankrobotics.com

Tony Belpaeme  
IDLab - imec, Ghent University, Belgium  
CRNS, Plymouth University, UK  
tony.belpaeme@plymouth.ac.uk

ABSTRACT
In this paper, we describe a multi-modal Bayesian network for person recognition in a HRI context, combining information about a person’s face, gender, age, and height estimates, with the time of interaction. We conduct an initial study with 14 participants over a four-week period to validate the system and learn the optimal weights for each of the metrics. Several normalisation methods are compared for different settings, such as learning from data, face recognition threshold and quality of the estimation. The results show that the proposed network improves the overall recognition rate by at least 1.4% comparing to person recognition based on face only in an open-set identification problem, and at least 4.4% in a closed-set.

KEYWORDS
Person recognition; Bayesian network; multi-modal data fusion; soft biometrics; personalisation

1 INTRODUCTION
Recognising a person is an essential step in establishing a personalised long-term human–robot interaction (HRI). In contrast to verification problems, where a user would state her identity and the system confirms or rejects it, in an HRI scenario, automatic recognition is desired for a natural interaction. In addition, the user might not be encountered before, in which case, the robot is expected to “meet” the user, i.e. enroll the user into the system. This problem is classified as an open-set identification problem, which is more difficult than closed-set identification or verification problems [5].

Biometric systems generally perform user recognition based on face recognition (FR). However, most FR challenges such as Face Recognition Vendor Tests1 evaluate algorithms that perform verification. To this date, the only available open-set identification challenge is the Unconstrained Face Detection and Open Set Recognition Challenge2, which shows that the algorithms achieve good identification accuracies at high false identification rates [6].

Similarly, during a real-time interaction, FR could be unreliable due to a number of reasons including changing facial features, expressions, and lighting conditions [16] (see Fig. 1). Another example is the recent release of a smart phone with the built-in FR system for unlocking the phone that struggle to distinguish family members due to similarity of their facial features [4]. This issue increased awareness of the security and privacy problems that might arise from using a uni-modal biometric system that might not be as reliable as using a pass code.

Moreover, a biometric system may not be able to obtain meaningful data in some cases, resulting in a failure-to-enroll (FTE) error [13]. For instance, a face may not be detected in a blurry image where a person is moving. In addition, the upper bound on identification accuracy would limit the matching performance of a uni-modal biometric system. However, multi-modal biometric systems can improve the matching accuracy of a recognition by fusing information from multiple sources that could reduce the effects of noisy data, decrease FTE error, and eliminate the upper bound issue for a better determination of the identity. Robots, due to the rich sensor suite they carry, lend themselves well to multi-modal recognition.

In this paper, we explore a multi-modal Bayesian network (BN) for integrating soft biometric information, such as a user’s gender, age, height, and time of interaction, together with the primary biometric information provided by face recognition. These biometric modalities are non-intrusive and can be obtained using the camera embedded on the robot. We designed a pilot study to validate our system in a real-time HRI scenario. We compare performances of several normalisation methods using optimised weights for each comparison. The proposed recognition system is intended to be used in a real-world application in Cardiac Rehabilitation therapy with a personalised robot [10].

1https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt
2http://vast.uccs.edu/OpenseFace

Figure 1: Examples of unreliable face recognition from our study: (a) a blurry image; (b) an oblique viewing angle; (c) occlusions, e.g. glasses; (d) lighting condition, e.g. direct light.
2 RELATED WORK
Several post-classification fusion methods have been proposed for integrating multi-modal information in biometrics. They can be classified into three main categories: decision, rank, and confidence level fusion [8]. Decision level fusion (e.g. majority voting, AND/OR rule) is based on combining individual best matches from each biometric matcher. Rank level methods (e.g. highest rank, logistic regression) are used when the output of each biometric matcher consists of ranked matches. Confidence level fusion is the most common approach, as it allows a weighted decision from multiple biometric classifiers. There are two main approaches for combining the scores for confidence level fusion: classification methods and combination approaches. Classification methods treat the classifications of individual classifiers as input for a new classifier, such as ANNs and Support Vector Machines, which allows combining non-homogeneous data without preprocessing. The combination approach on the other hand consist of three steps: (1) normalisation of scores from different modalities into a common domain, (2) combination of scores through a method such as sum or product rule, and (3) thresholding to obtain the identification results. Performances of the combination approaches depend on the method and threshold chosen at each of the steps.

Although most biometric systems utilise primary biometrics for person recognition, such as fingerprint or face, other attributes of an individual such as age, gender, and clothing – referred to as soft biometrics-- can provide additional information to improve the recognition performance [2]. In [9], the authors proposed combining a primary biometric trait (fingerprint) with soft biometric traits, such as a person’s gender, ethnicity, and height, using a BN. In the weighting scheme, the traits with smaller variability and larger distinguishing capability were given more weight in the computation of the final matching probabilities. Furthermore, smaller weights were assigned to the soft biometric traits, so that if a soft biometric trait is measured incorrectly (e.g. a male user being identified as a female) the rejection probability is decreased. They achieved a 4% improvement in the genuine acceptance rate, however, the fusion weights were not optimised.

To the best of our knowledge, our approach is the first in combining soft biometrics with a primary biometric to identify a user in real-time, in the field of HRI. Moreover, it is the first time that the presented modalities (face, gender, age, height and time of interaction) are fused together, although they have shown improvement when fused separately or with other biometrics [2, 11, 15].

3 METHODOLOGY
We developed a BN based on [9] integrating multi-modal biometrics for reliable recognition in real-time human-robot interaction. We fuse face recognition (F) information (primary biometric) with gender (G), age (A), and height (H) estimations and the time of the interaction (T) (soft biometrics). Conditional independence is assumed between nodes, given the identity (I). The pyAgrum [3] library is used for implementing the structure.

3.1 Structure
The states of each node are determined by: the number of known users (for F and I), the available range of the modality (for A and H), and the pre-defined values (for G, "female" and "male", and for T, the day of the week combined with the time of the interaction). The data available about each user are converted to probabilities for each state within a node. These values are used as evidence in the network, and the maximum posterior for the I node determines the estimated identity of the user.

FR values are assumed to be similarity scores, such that, each score gives the percentage of similarity of the current user to the faces in the database. These scores are normalised to find the probabilities of the states. Age, height, and time are considered as discrete random variables (e.g. age is taken as 26, between 26 and 27). We estimate the probabilities of the remaining states by assuming a discretised and normalised normal distribution, \(N(\mu, \sigma^2)\), defined by Eq. 1, where \(X\) is the estimated value, \(Z\) is the z-score, and \(C\) is the confidence of the biometric indicator for the estimated value.

\[
\mu = X, \quad P(-0.5 < Z < 0.5) = C \tag{1}
\]

Generally, in FR systems, if the highest similarity score or probability is below a given threshold, the user is declared as "unknown". However, in a BN, the posterior probabilities can be quite low due to the multiplication of probabilities during inference and the value can decrease with increasing number of states in a node. Thus, instead of using a fixed threshold in our system, we use the quality of the estimation (\(Q\)), in which we compare the highest probability \(P_w\) to the second highest probability \(P_s\), as shown in Eq. 2. The difference is multiplied by the number of people in the database \(n_p\), because the difference between the probabilities decreases as the number of people increases (the sum of probabilities is 1.0). Initially \(Q = 0\), which eliminates the cases where the first and the second highest probabilities are the same.

\[
Q = \left[ P_w(I[F,G,A,H,T]) - P_s(I[F,G,A,H,T]) \right] \times n_p \tag{2}
\]

The FR threshold \(\theta_{FR}\) is maintained in the system through the introduction of the "unknown" (U) state in face and identity nodes. The similarity score of U for FR is set to the \(\theta_{FR}\), hence, when normalised, the similarity scores below the threshold have lower probabilities than U. Similarly, those that have higher similarity scores than the threshold will have higher probabilities than U.

3.2 Learning
Our hypothesis is that the recognition could be improved by learning the likelihoods of the system through evidence. Hence, as our contribution, we propose a BN where the likelihoods of the system are learned from data.

A possible solution could be to create a model that depends on time-series data, like a dynamic BN. However, in a dynamic BN, only the immediate prior value at the previous time step is used, which differs from an open-set identification problem, where the previous state can contain values that applied to another user. For example, user "1" might be encountered right before user "2"; in which case, the evidence for user "1" might not have an effect on the identity estimation for user "2".

Therefore, we designed our own approach in learning from data. We initially use the prior knowledge in setting the likelihoods for each variable (e.g. \(P(F|I), P(G|I)\)). When the robot identifies a user, and the user confirms her identity, the recognition information and the identity of the user are fed as evidence to the network, and the
We smooth the recognition results of each modality by using the weights as an exponent to the results for the evidence, due to using a good normalisation method should be insensitive to the outcome, (c) posterior using the evidence, (d) updated posteriors. Initial P(F|I="1")

Likewise, if the recognised user was not previously enrolled into the system, then posterior of P(F|I="0") is updated. However, as probabilities within the BN: min-max, tanh [7], softmax [1], and norm-sum (dividing by the sum of values).

BNs use the product rule for combining the results of each node, hence, if a probability of a classifier is zero, it results in an overall zero probability for a class irrespective of the results from other classifiers. In order to overcome this problem we used a small cut-off probability threshold as $p_t = 10^{-6}$.

3.5 Extendability

Our approach relies on FR primarily, but the described system can be extended with other primary biometric traits such as voice and fingerprint, and soft biometric traits, such as the location of the interaction, and ethnicity. It is intended to increase the recognition rate from a single image, and tracking is not applied between images. In order to increase the reliability of the system, multiple images (3 images here) are taken in succession during the pilot study and the results are normalised to estimate the identity of the user, which allows discarding the images without a face detected.

The system does not require heavy-computing, hence, it is suitable for use on commercially available robots. We use a Pepper robot in our study with Naoqi software modules (providing a user’s face ID, gender, age, and height that we used as input modalities), however, the network is applicable to any recognition software.

4 STUDY IN USER IDENTIFICATION

The objective of this study is to gather data for finding the optimal weights for the proposed network for user identification.

4.1 Protocol

The user initially enrolls to the system by entering his/her name, gender, age, and height, and then the robot takes photos of the user. During next encounters, the robot predicts the identity of the user and asks for confirmation.

We ran the study with 14 participants (4 female, 10 male, of age range 24-40) and collected a total of 66 images per user over the four weeks period. The recognition process took approximately 5 seconds: ~2-3s for user detection, ~1.5s (0.5s each) for image capture, ~1s to load the network parameters, ~0.6s (0.2s each) for recognition from modalities, ~0.9s (0.3s each) for estimation of the identity using the network. The robot stayed in a fixed position before the interaction, and only when a user was identified, it would become animate to ensure a natural interaction. We aimed to achieve better quality of images by keeping the robot fixed, however, since the robot did not notify participants when taking images, some of the captured images include people looking sideways, smiling, partially covering their faces or moving (see Fig. 1).

We use single-user recognition within the images, that is, only one user is assumed to be present in front of the camera. Hence, the image database was cleaned of images with multiple people or any other user rather than the claimed identity for cross-validation. However, in the future, the position of each user can be considered for multiple people recognition and interaction.
4.2 Results

In order to validate our system, 5-fold cross-validation is applied with 13 images per user in each bin with a different randomised initial ordering of the users, and the results are averaged. Detection and identification rates (DIR) and false alarm rates (FAR) are reported for the pilot study along with receiver operating characteristics (ROC) curves (see Appendix and Fig. 3), which are the performance measures for the open-set identification problem [12]. The average failure to enroll error (FTE) is 0.214 (0.008), which corresponds to the fraction of images where a face cannot be detected. The identity was not estimated by the network in those cases because the only primary biometric in our system is FR and soft biometrics do not have the deterministic characteristic to estimate the identity on their own.

The optimised weights (see Appendix) show that in our study the age is the least effective soft biometric in determining the identity, whereas height is the most effective one. However, this might be due to the characteristics of the population in our pilot study, as the participants’ ages are close to each other. Another important factor is the reliability of the age recognition software. The standard deviation of the estimated age of a user on average was 9.3. Hence, we cannot conclude that age should not be used to supplement the FR in general, but if used, the accuracy of the software used should be high, especially in a population with a narrow age range. On the other hand, the effectiveness of the height can also be explained by the nature of the population (3 relatively tall (> 180 cm) and 2 relatively short (< 160 cm) users), even though the average standard deviation of estimated height was 6.3 cm. A more balanced dataset would allow observing the true effects of these parameters.

The cross-validation results for the optimised weights (see Appendix) show that combining soft biometrics using our proposed BN can increase the DIR, depending on the inner settings. It can be observed that although norm-sum and min-max methods provide good results without learning, the recognition rate drops below FR with learning, whereas softmax and tanh methods are not affected.

However, the FAR of the network for any normalisation method is greater than the FAR of FR. This is caused by the combination of multimodal data. For example, if the highest face similarity score is greater than the FAR of FR. This is caused by the combination of other sensor input, where errors might increase FAR. In order to compare the effects of learning, we chose the min-max method without learning and with a cut-off threshold (N_{\text{minmax}}) and the softmax method with learning and no cut-off threshold (NL_{\text{softmax}}), because the former provides the second highest DFR but with lower FAR than that of the best methods (highlighted in blue), and the latter provides the best DFR in learning in both training and test sets. The results are presented in Fig. 3.

The trade-off between DFR and FAR can be observed in Fig. 3a. The ideal FR threshold (\theta_{\text{FR}}) should maintain a low FAR with a good DFR. For example, at \theta_{\text{FR}} = 0.7, the FAR is very low for both FR and NL_{\text{softmax}}, however, the DFR has also decreased substantially. If we compare the results in the range where \text{FAR}_{\text{FR}} \leq 0.5 (\theta_{\text{FR}} = 0.3) and \text{DFR}_{\theta_{\text{FR}} = 0.3} \geq 0.8 (\theta_{\text{FR}} = 0.6), then \text{N}_{\text{minmax}} is better in identification (0.93 \leq \text{DFR}_{\text{minmax}} \leq 0.949) than \text{NL}_{\text{softmax}} (0.873 \leq \text{DFR}_{\text{softmax}} \leq 0.946) and FR (0.801 \leq \text{DFR}_{\text{FR}} \leq 0.933). However, \text{NL}_{\text{softmax}} misidentifies the “unknown” users much less (0.286 \leq \text{FAR}_{\text{softmax}} \leq 0.543) than \text{N}_{\text{minmax}} (0.457 \leq \text{FAR}_{\text{minmax}} \leq 0.571).

\theta_{\text{FR}} = 0.4 gives the highest detection rate for both \text{N}_{\text{minmax}} and \text{NL}_{\text{softmax}} with lower FAR, hence, we compared the effects of quality of estimation (Q) at this rate (see Fig. 3b). The area of improvement for the open-set identification problem is where \text{FAR} \leq \text{FAR}_{\text{FR}} and \text{DFR} \geq \text{DFR}_{\text{FR}}. \text{NL}_{\text{softmax}} does not provide a value in this range, hence, we can conclude that the proposed learning method performs worse than the method without learning. \text{DFR}_{\text{minmax}} is 1.4% higher than \text{DFR}_{\text{FR}} where the FAR is equal (Q = 0.31), and \text{FAR}_{\text{minmax}} is 1.4% lower than \text{FAR}_{\text{FR}} where DIR is equal (Q = 0.41). On the other hand, if the problem was treated as a closed-set problem (where all the users are enrolled into the system), Q = 0 would be sufficient and the increase in DIR would be 4.4%.

5 CONCLUSION

Our results suggest that the use of soft biometrics increases the recognition rate, however, it can also increase the misidentification rate of unknown users. Increasing \theta_{\text{FR}} and Q can indeed decrease the FAR, but it can decrease the DFR as well. On the other hand, our proposed learning method mostly performs worse than the traditional BN on this dataset.

Furthermore, the results indicated that our dataset might be biased due to the small population size and the characteristics of the population. To our knowledge, the only publicly available database that contains the soft biometrics used in our system (except the time of interaction) with a face database is the recently released BioSoft [14]. However, the number of subjects is limited to 75, and the height is defined in labels instead of numeric values. Therefore, as a future extension, we will generate an artificial database with a higher amount of subjects with differing soft biometrics, which would also allow setting the noise level in the modalities. We aim to compare the performance of our system with other classification methods such as Support Vector Machines on the artificial dataset.

In the near future, we plan to use the proposed user recognition system in Cardiac Rehabilitation (CR) therapy, during which the robot will recognise the patients and personalise the interaction based on the information about the patients’ previous sessions and their progress during the therapy [10]. The study will allow us to evaluate our system in a real-world application, and to observe the effects of personalisation in a long-term HRI.
Appendix: 5-fold cross validation mean (with standard deviation) of false alarm rates (FAR) on the training set, and identification rates (DIR) for rank 1 for training and test sets. Highlights in red show the chosen methods for the comparison of learning from data.

<table>
<thead>
<tr>
<th>Learning</th>
<th>$p_t$</th>
<th>Normalisation</th>
<th>FAR</th>
<th>DIR$_1$ (Training)</th>
<th>DIR$_1$ (Test)</th>
<th>$w_G$</th>
<th>$w_A$</th>
<th>$w_H$</th>
<th>$w_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>none</td>
<td>FR</td>
<td>0.443 (0.078)</td>
<td>0.933 (0.004)</td>
<td>0.945 (0.015)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>none</td>
<td>none</td>
<td>norm-sum</td>
<td>0.629 (0.032)</td>
<td>0.951 (0.004)</td>
<td>0.967 (0.013)</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>none</td>
<td>none</td>
<td>min-max</td>
<td>0.629 (0.032)</td>
<td>0.951 (0.005)</td>
<td>0.965 (0.015)</td>
<td>0.2</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>none</td>
<td>none</td>
<td>softmax</td>
<td>0.571 (0.072)</td>
<td>0.947 (0.004)</td>
<td>0.965 (0.014)</td>
<td>0.1</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>none</td>
<td>none</td>
<td>tanh</td>
<td>0.571 (0.051)</td>
<td>0.942 (0.005)</td>
<td>0.955 (0.012)</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>norm-sum</td>
<td>0.529 (0.081)</td>
<td>0.943 (0.003)</td>
<td>0.956 (0.015)</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>min-max</td>
<td>0.586 (0.060)</td>
<td>0.949 (0.005)</td>
<td>0.965 (0.014)</td>
<td>0.2</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>softmax</td>
<td>0.571 (0.072)</td>
<td>0.946 (0.003)</td>
<td>0.959 (0.015)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>tanh</td>
<td>0.543 (0.039)</td>
<td>0.942 (0.003)</td>
<td>0.957 (0.013)</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>evidence none</td>
<td>norm-sum</td>
<td>0.629 (0.032)</td>
<td>0.782 (0.063)</td>
<td>0.694 (0.093)</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence none</td>
<td>min-max</td>
<td>0.629 (0.032)</td>
<td>0.776 (0.064)</td>
<td>0.692 (0.090)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence none</td>
<td>softmax</td>
<td>0.586 (0.060)</td>
<td>0.946 (0.005)</td>
<td>0.961 (0.017)</td>
<td>0.1</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>evidence none</td>
<td>tanh</td>
<td>0.571 (0.051)</td>
<td>0.943 (0.007)</td>
<td>0.955 (0.012)</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>norm-sum</td>
<td>0.571 (0.072)</td>
<td>0.75 (0.082)</td>
<td>0.632 (0.127)</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>min-max</td>
<td>0.643 (0.0)</td>
<td>0.776 (0.061)</td>
<td>0.697 (0.089)</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>softmax</td>
<td>0.586 (0.060)</td>
<td>0.946 (0.005)</td>
<td>0.954 (0.025)</td>
<td>0.1</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>evidence</td>
<td>1e-6</td>
<td>tanh</td>
<td>0.543 (0.039)</td>
<td>0.943 (0.006)</td>
<td>0.961 (0.019)</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENTS

This work has been supported by the EU H2020 Marie Sklodowska-Curie Actions ITN project APRIL (grant 674868). The authors would like to thank Valerio Biscione for his constructive suggestions in the design of the Bayesian network, and the participants of the user study for their time and efforts.

REFERENCES