Classification of Facial Photograph Sorting Performance Based on Verbal Descriptions

Daryl H. Hepting[1], Richard Spring[1], Timothy Maciag[1], Katherine Arbuthnott[2], and Dominik Ślęzak[3,4]

[1] Department of Computer Science, University of Regina
3737 Wascana Parkway, Regina, SK, S4S 0A2 Canada
{dhh,spring1r,maciagt}@cs.uregina.ca
[2] Campion College, University of Regina
3737 Wascana Parkway, Regina, SK, S4S 0A2 Canada
katherine.arbuthnott@uregina.ca
[3] Institute of Mathematics, University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
[4] Infobright Inc.
Krzywickiego 34 pok. 219, 02-078 Warsaw, Poland
slezak@infobright.com

Abstract. Eyewitness identification remains an important element in judicial proceedings. It is very convincing, yet it is not very accurate. To better understand eyewitness identification, we began by examining how people understand similarity. This paper reports on analysis of study that examined how people made similarity judgements amongst a variety of facial photographs: participants were presented with a randomly ordered set of photos, with equal numbers of Caucasian (C) and First Nations (F), which they sorted based on their individual assessment of similarity. The number of piles made by the participants was not restricted. After sorting was complete, each participant was then asked to label each pile with a description of the pile's contents. Following the results of an earlier study, we hypothesize that individuals may be using different strategies to assess similarity between photos. In this analysis, we attempt to use the descriptive pile labels (in particular, related to lips and ears) as a means to uncover differences in strategies for which a classifier can be built, using the rough set attribute reduction methodology. In particular, we aim to identify those pairs of photographs that may be the key for verifying an individual's abilities and strategies when recognizing faces. The paper describes the method for data processing that enabled the comparisons based on labels. Continued success with the same technique as previously reported to filter pairs before performing the rough sets analysis, lends credibility to its use as a general method. The rough set techniques enable the identification of the sets of photograph pairs that are key to the divisions based on various strategies. This may lead to a practical test for people's abilities, as well as to inferring what discriminations people use in face recognition.

1 Introduction

Eyewitness identification holds a prominent role in many judicial settings, yet it is generally not accurate. Verbal overshadowing [1] is an effect that can obscure a witness's recollection of face when he is asked to describe the face to create a composite sketch. Alternatively, if the witness is asked to examine a large collection of photos, her memory may become saturated and she may mistakenly judge the current face similar to another she has examined (i.e., inaccurate source monitoring) and not to the one she is trying to recall [2]. We hypothesize that if the presentation of images can be personalized, the eyewitness may have to deal with fewer images, minimizing both of the negative effects discussed. This research takes more steps along that path.

This paper discusses an analysis of data from a sorting study, which avoided verbalization completely while sorting. Each participant was asked to group a stack of 356 photos according to perceived similarity. One half of the photos ($n = 178$) depicted Caucasian males, taken in the southern United States of America. The other half of the photos depicted First Nations males, taken at different locales in the Canadian province of Saskatchewan. 'First Nations' is the term which has replaced 'Indian' in most cases. In Saskatchewan, there are 72 First Nations [5] governments or bands. As a participant encountered a photo, she could only place that photo and not disturb any existing piles. Indirectly, each participant made 63,190 pairwise similarity judgements. Once sorting was complete, each participant was asked to verbally label each pile according to the similarity used to create that pile. In this paper, we examine whether the occurrence of a label may be a good indicator of sorting performance.

We discuss the extension of previously published methods [3] by allowing the classification of facial photograph sorting performance based on verbal descriptions. The earlier work examined whether race (Caucasian/First Nations) had any impact on facial photograph sorting performance, which is also of interest because of the existence of a "cross-race" effect [4] which may make identifications of faces more difficult if those faces are not of the same race as the viewer.

Furthermore, we present several more successful examples of the filtering technique used to substantially reduce the processing time and effort needed to build a completely accurate classifier, if one exists.

Section 2 describes the method of making piles based on the presence or absence of a label. Section 3 describes the filtering technique developed to reduce the number of photo pairs needed as input to the attribute reduction methodology, and the results obtained for "ears/not-ears" and "lips/not-lips" label decision classes. Section 4 shows how to combine results from the previous study with those first reported here, to make a more complete test of participant performance. Section 5 presents conclusions and avenues for future work.

[5] source: http://fsin.com

Table 1. Labels for facial parts, listed from the top of head downwards, followed by general characteristic labels. Notice that many labels are used for every picture. This analysis only looked at the presence or absence of a label, so "ears" and "lips" were chosen (44%). We sought labels that were used for approximately 50% of the photos, because we required an equal number for which the label was not used. In this case, we randomly selected 157, of the 199, for which the labels were not used in order to perform our analysis. For the parts that were identified in all photos, such as "hair", "eyes", "head/face shape", or "skin/complexion", we might be able to use them to distinguish photos (as in "big head" compared to "small head"), but we did not record the data in this way.

Label	Photos	Percentage
hair	356	100
forehead	65	18
eyebrows	217	61
ears	157	44
nose	259	73
eyes	356	100
cheeks	25	7
lips	157	44
teeth	14	4
jaw/chin	318	89
neck	25	7
head/face shape	356	100
head/face size	125	35
skin/complexion	356	100
facial hair	243	68

2 Analysis of Verbal Descriptions

After sorting all 356 photos, all participants were asked to describe with a label the similarity embodied in each pile that they had created. The label or labels attached to each pile were then assigned to each photo with the pile. This process was repeated for all 25 participants. Table 1 shows the unique occurrences of various labels with photos.

The two labels occurring with approximately 50% of the photos ("ears" and "lips") were chosen for further analysis, following the procedure outlined in Hepting et al. [3]:

1. choose $N = 157$ of the photos to which the label was not attached (from 199 possible). Therefore, the label and not-label sets each have 157 photos
2. for each participant, exclude from the pile data all 42 of the photos not in the label/not-label sets
3. analyze the make-up of each pile, in terms of label (L) and not-label (NL) photos (only these photos remain in the pile). Our null hypothesis (H_0) is that each pile comprises the same proportion of L and NL photos. Using

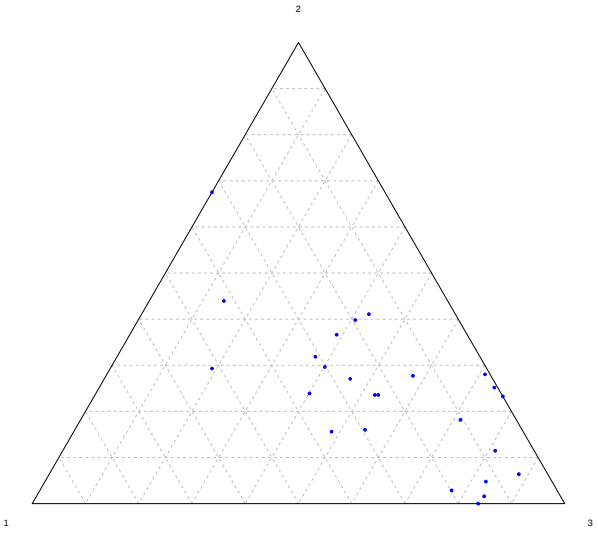

Fig. 1. Each point reflects the mix of photos classified by a participant. A point in the center of the triangle represents an equal mix of photos classified L (label), NL (not-label), and U (undecided). This figure shows the distribution of participants based on their classification of photos with respect to the "ears" label. Many points are located between Vertex 1 and Vertex 2, representing approximately equal numbers of photos classified as "ears" and "not-ears". If a point is near Vertex 3, that participant classified most of the photos as "Undecided". For the "ears" label, we constructed 2 decision classes, "uses-ears" and "uses-not-ears", based on the percentage of "undecided" photos. Participants with more than 60% "undecided" photos were put into "uses-not-ears" decision class and the others were put into "uses-ears".

the CHITEST function in Microsoft Excel, we test the independence of the observed ratio of L to NL (as a percentage) against an expected equal ratio (50%:50%). If p (returned from CHITEST) < 0.05, we rejected H_0 and either classified the pile as L, if L >NL or as NL if NL > L. The pile was classified as U (for undecided) if $p \geq 0.05$ (and we could not reject H_0). All pictures in that pile were then labelled as L, NL, or U. The total number of photos classified as L, NL, and U was expressed as a percentage (see Figure 1).

Figure 1 shows all participants plotted according to their percentages of photos classified as L, NL, and U for the "ears" label. Vertex 3 represents undecided (U) and points near this vertex represent participants who classified most photos as U. A threshold of 60% was set for the percentage of U and two groups were formed. We hypothesize that these groups correspond to different strategies for facial recognition, which we have labelled as "uses-ears" (U $< 60\%$, $n = 15$) and "uses-not-ears" (U $\geq 60\%$, $n = 10$). In other words, we hypothesize that

"ears" is being used by former group but not by the latter. In the same way for "lips", a threshold of 60% was set for the percentage of U and two groups were formed. We hypothesize that these groups correspond to different strategies for facial recognition, which we have labelled as "uses-lips" (U $< 60\%$, $n = 9$) and "uses-not-lips" (U $\geq 60\%$, $n = 16$). In other words, we hypothesize that "lips" is being used by former group but not by the latter.

We seek to find a simple way to classify participants according to these groups, which will allow for personalization of the eyewitness identification process. The strategy (uses-ears or uses-not-ears, uses-lips or uses-not-lips) then becomes the decision variable as we begin to apply the rough set attribute reduction methodology [5]. The objective is to reduce the number of pairs required as input to discriminate between the two strategies, as the original number of pairs is impractical.

3 Pair Filtering

For each participant, a decision is made (directly or indirectly) about whether a pair of photos is similar (same pile) or not (different piles). 63,190 pairs can be formed from the 356 photos used in this study, which is a very large input to the analysis stage. Thus, we have pursued a method to reduce the number of input pairs to the analysis stage, based on the following hypothesis (also discussed in Hepting et al. [3]): the pairs most useful in constructing reducts and rules will be those which are rated most differently between the decision classes, similar to the feature extraction/selection phase in knowledge discovery and data mining.

We used the following method to test the hypothesis: we compute the total distance for a pair within each decision class by normalizing the sum of all participant ratings. If all participants in the same decision class rate the pair as similar, the distance is 0. If all participants in the same decision class rate the pair as different, the distance is 1. In general, the distance is computed as the sum of similarity ratings (each one is either 0 or 1) divided by the number of participants in the decision class. We first look at the minimum of these two distances, $d = min(D_1, D_2)$, in order to find a pair that is rated as very similar by participants in one of the decision classes. If a pair is rated as very similar by participants in both decision classes (both D_1 and D_2 are small), that pair will not help to discriminate between the decision classes. Therefore, we also look at the gap between the two distances, $\Delta = |D_1 - D_2|$. The pairs which have a small d and a large Δ are those which meet the criterion of being rated most differently between the decision classes. Table 2 shows the collection of these values for ears and lips. The row values indicate the minimum distance (d) and the column values indicate the gap (Δ).

We used the Rough Set Exploration System (RSES) [6] to analyse the sets indicated by this filtering. We proceed through each table in Table 2 row by row from the top left to the bottom right, until a classifier with 100% accuracy and 100% coverage is found. Our procedure is outlined in the following:

Table 2. The values in the table indicate the number of pairs selected by each combination of minimum distance (row) and gap (column). The set of pairs used for further processing is indicated in bold. On the left, results of filtering for uses-ears/ uses-not-ears. On the right, results of filtering for uses-lips / uses-not-lips. One pair of photos from each set is illustrated Figure 2. The input to RSES (Rough Set Exploration System) [6] for uses-lips/uses-not-lips is shown in Table 3.

	Gap (Δ)				
d	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5
≤ 0.1	2	**9**	31	59	108
≤ 0.2	2	22	77	250	398
≤ 0.3	2	22	159	498	1246
≤ 0.4	2	22	159	675	1883
≤ 0.5	2	22	159	675	2314

	Gap (Δ)				
d	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5
≤ 0.1	0	0	0	2	2
≤ 0.2	0	3	14	**16**	46
≤ 0.3	0	3	24	78	179
≤ 0.4	0	3	24	166	762
≤ 0.5	0	3	24	166	1356

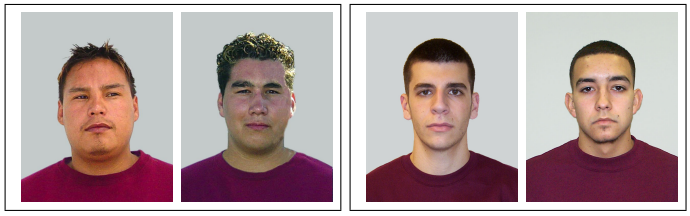

Fig. 2. On the left, one of the pairs of photos important in the classification of participants according to uses-ears/uses-not-ears. On the right, one of the pairs of photos important in the classification of participants according to uses-lips/uses-not-lips.

1. Split: Split input file (50/50): Each file in the analysis was split with 50% of participants in a training set (data from 12 participants) and 50% of participant's data (data from 13 participants) in a testing set. The files comprised objects each representing a pairwise comparison of facial photographs (0 if similar, 1 if dissimilar). The decision class was the strategy (either uses-ears/uses-not-ears (illustrated in Figure 1) or uses-lips/uses-not-lips).
2. Train: Calculate the reducts in training file using genetic algorithms in RSES. The genetic algorithms procedure calculates the top N reducts possible for a given analysis. For the purposes of our analysis, we chose N = 10 in order to pick the top 10 reducts possible (if indeed 10 top reducts could be found). Generate rules from these reducts.
3. Classify: Classify the 25 participants according to the generated rules, and observe the accuracy and coverage of the classifier.

We conducted k-fold cross-validation [7], with k = 10. If a classifier with 100% accuracy and 100% coverage is not found within 10 tries, it may still exist. Choosing 13 of 25 participants for training leads to a possible $\binom{25}{13} = 5,200,300$ combinations and classifiers.

Table 3. 16 pairs selected as input to RSES (Rough Set Exploration System) for classification based on uses-lips/uses-not-lips ($d = 0.2$, $\Delta = 0.6$).

	Photograph Pairs																
object	1969a-2094a	4211a-5893a	2660a-8127a	1296a-2811a	3722a-4158a	2094a-6682a	058-149	083-117	1716a-7001a	040-108	1032a-1867a	0011a-7453a	1296a-6682a	5241a-8164a	0079a-6524a	1032a-8831a	class
O:1	1	0	0	1	0	1	0	1	1	0	0	1	0	0	1	0	0
O:2	0	0	0	0	0	1	0	1	1	0	0	0	1	0	1	0	1
O:3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O:4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O:5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
O:6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
O:7	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
O:8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
O:9	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
O:10	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
O:11	0	1	1	1	1	1	0	0	0	0	0	0	1	1	1	0	0
O:12	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
O:13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
O:14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
O:15	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1
O:16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
O:17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
O:18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O:19	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
O:20	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
O:21	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
O:22	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0
O:23	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0
O:24	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
O:25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4. The distribution of the 25 participants between groups identified by combinations of strategies based on the apparent use of race, ears, and lips in their decision making.

Group	Members
uses-not-race, uses-not-ears, uses-not-lips	4
uses-not-race, uses-not-ears, uses-lips	1
uses-not-race, uses-ears, uses-not-lips	2
uses-not-race, uses-ears, uses-lips	4
uses-race, uses-not-ears, uses-not-lips	4
uses-race, uses-not-ears, uses-lips	1
uses-race, uses-ears, uses-not-lips	6
uses-race, uses-ears, uses-lips	3

4 Compare

We made groups based on strategies: uses-race/uses-not-race [3], uses-ears/uses-not-ears, and uses-lips/uses-not-lips. We found that the largest of these groups was uses-race, uses-ears, uses-not-lips. The distribution of the 25 participants between groups is shown in Table 4. Table 5 shows the relationship between the minimum distance and gap for this largest group in Table 4.

Table 6 presents a comparison of the classifiers discussed, based separately on different strategies identified (uses-race/uses-not-race, uses-ears/users-not-ears, and uses-lips/uses-not-lips) and on a combined strategy uses-race AND uses-ears AND uses-not-lips/NOT(uses-race AND uses-ears AND uses-not-lips) as identified in Table 4. The uses-race/uses-not-race classifier has been recomputed from the earlier paper [3], according to the algorithm described here. It is interesting to note that the average accuracy seems to be related to the first non-zero entry in the table of filtered pairs. Table 6 shows that in order of most to least accurate (with the position of the first non-zero entry, from the top-left in Tables 2 and 5, in parentheses), we have: ears (1), combined (2), lips (4), and race(4).

5 Conclusions and Future Work

Cross-race identification of faces is an important topic of ongoing research [8], and our sorting study seeks to contribute to this body of work. We have focused on the labelling of similarity judgements as a way to understand the way people perceive structure in the stimuli set.

Through this effort, we have found succinct tests to classify people into different strategy groups (ears/not-ears, lips/not-lips). Namely, we demonstrated that rough sets can help in accuracy and clarity of the results. It is interesting that the decision table for “ears” comprises almost exclusively First Nations pairs, and the decision table for “lips” comprises almost exclusively Caucasian pairs. Neither has any mixed pairs. Therefore, we hope that these results will help in our efforts to better understand the cross-race effect [4].

Table 5. The values in the table indicate the number of pairs selected by each combination of minimum distance (row) and gap (column). The set of pairs used for further processing is indicated in bold. On the left, results of filtering for uses-race,uses-ears, uses-not-lips / not(uses-race,uses-ears, uses-not-lips). On the right, results of filtering for race/not-race. The 7 pairs chosen for race here are different than those used in Hepting et al. [3], but they were selected according to the method outlined here.

	Gap (Δ)				
Dist.	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5
≤ 0.1	0	**10**	49	107	149
≤ 0.2	0	12	90	389	852
≤ 0.3	0	12	90	389	853
≤ 0.4	0	12	90	478	1486
≤ 0.5	0	12	90	478	1590

	Gap (Δ)				
Dist.	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5
≤ 0.1	0	0	0	1	2
≤ 0.2	0	0	0	**7**	11
≤ 0.3	0	0	17	82	197
≤ 0.4	0	0	17	130	401
≤ 0.5	0	0	17	130	798

Table 6. Accuracy (A) and Coverage (C) for each classifier over 10 trials (with mean and standard deviation following each). FA/FC indicates the number out of the 10 trials that had 100% accuracy and 100% coverage. This is followed by the pairs used to classify according to each strategy. Photos ending in 'a' are Caucasian, others are First Nations. None of the pairs is mixed. No pair repeats, though some individual photos are included with more than 1 pair. 4 First Nations and 3 Caucasian photos are used as input for the race strategy classification. For the ears strategy classification, almost all are First Nations photos, whereas for the lips and the combined strategy classifications, almost all the photos are Caucasian.

Race	Ears	Lips	Combined
A(92.38, SD : 4.38) C(99.60, 1.26) FA/FC: 1	A(99.20, SD : 1.69) C(100, SD : 0) FA/FC: 8	A(94.40, SD : 3.86) C(100, SD : 0) FA/FC: 3	A(97.20, SD : 2.70) C(97.60, 7.59) FA/FC: 3
004-050	033-121	0011a-7453a	0576a-8530a
039-125	037-176	0079a-6524a	062-178
050-176	038-068	040-108	1338a-6553a
0662a-4919a	058-157	058-149	1513a-1859a
087-142	095-106	083-117	1907a-9929a
2325a-8650a	111-121	1032a-1867a	4099a-4459a
6281a-9265a	146-172	1032a-8831a	4099a-6553a
	152-153	1296a-2811a	4488a-6553a
	4833a-9948a	1296a-6682a	6838a-8922a
		1716a-7001a	7297a-9860a
		1969a-2094a	
		2094a-6682a	
		2660a-8127a	
		3722a-4158a	
		4211a-5893a	
		5241a-8164a	

This work lends support to our filtering technique as a broadly applicable method. In general, all the classifiers have performed well, but the one based on the "ears" label is clearly the best among them. We still need to understand what strategy might be at work in these cases, but the accuracy of the classifier indicates a clear difference between the decision classes. Although we have not done any sort of exhaustive testing of all pairs to verify our selection criteria for filtering, that we have been able to generate consistently accurate classifiers from very small fractions of the total pairs is a very encouraging sign.

We have a test to classify participants. Further work will be devoted to validating it against the performance on eyewitness identification tasks, and to using it to help clarify the strategies being employed by participants.

Acknowledgements The first four authors were supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The fifth author was supported by the grants N N516 368334 and N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland.

References

1. Schooler, J.W., Ohlsson, S., Brooks, K.: Thoughts beyond words: when language overshadows insight. Journal of Experimental Psychology: General **122** (1993) 166–183
2. Dysart, J., Lindsay, R., Hammond, R., Dupuis, P.: Mug shot exposure prior to lineup identification: Interference, transference, and commitment effects. Journal of Applied Psychology **86**(6) (Sep 2002) 1280–1284
3. Hepting, D.H., Maciag, T., Spring, R., Arbuthnott, K., Ślęzak, D.: A rough sets approach for personalized support of face recognition. In: RSFDGrC. (2009) 201–208
4. Jackiw, L.B., Arbuthnott, K.D., Pfeifer, J.E., Marcon, J.L., Meissner, C.A.: Examining the cross-race effect in lineup identification using caucasian and first nations samples. Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement **40**(1) (January 2008) 52–57
5. Pawlak, Z.: Rough set approach to knowledge-based decision support. European Journal of Operational Research **99** (1997) 48–57
6. Bazan, J.G., Szczuka, M.: The Rough Set Exploration System. Number 3400 in LNCS. In: Transactions on Rough Sets III. Springer-Verlag (2005) 37–56
7. Maciag, T., Hepting, D.H., Hilderman, R.J., Ślęzak, D.: Evaluation of a dominance-based rough set approach to interface design. In: FBIT. (2007) 409–416
8. Platz, S., Hosch, H.: Cross-racial/ethnic eyewitness identification: A field study 1. Journal of Applied Social Psychology (Jan 1988)