# A Linear Model for Three-Way Analysis of Facial Similarity

Daryl H. Hepting  $^{[0000-0002-3138-3521]}$ , Hadeel Hatim Bin Amer  $^{[0000-0003-2515-2329]}$ , and Yiyu Yao  $^{[0000-0001-6502-6226]}$ 

Department of Computer Science University of Regina, Regina, SK, S4S 0A2, Canada hepting@cs.uregina.ca, binamerh@uregina.ca, yyao@cs.uregina.ca

Abstract. Card sorting was used to gather information about facial similarity judgments. A group of raters put a set of facial photos into an unrestricted number of different piles according to each rater's judgment of similarity. This paper proposes a linear model for 3-way analysis of similarity. An overall rating function is a weighted linear combination of ratings from individual raters. A pair of photos is considered to be similar, dissimilar, or divided, respectively, if the overall rating function is greater than or equal to a certain threshold, is less than or equal to another threshold, or is between the two thresholds. The proposed framework for 3-way analysis of similarity is complementary to studies of similarity based on features of photos.

Keywords: similarity, three-way decision, card sorting, linear model

#### 1 Introduction

A basic idea of three-way decisions (3WD) is thinking and problem solving in threes [15]. According to a trisecting-and-acting model of 3WD, we divide a whole into three parts and devise strategies to process the three parts [14,15]. While each part captures a particular aspect of the whole or consists of elements of particular interest, their integration reflects the whole. By thinking in threes, 3WD may provide a simplification of processing the whole through processing three parts. The theory of 3WD has been applied in many fields [6–9, 13, 16].

In a previous paper [3], we presented some preliminary results on applying the 3WD theory to a card sorting problem. Card sorting has been successfully applied to gain insight about the structure of information in different contexts [1, 2, 10, 11]. For our card sorting problem, we have a set of facial photos and a group of raters. Each rater was instructed to sort similar photos into the same pile. A pair of photos is similar if both photos are sorted into the same pile. A pair of photos is dissimilar if the photos are sorted into different piles. An analysis of rating results shows that there is a large variance amongst raters in terms of the number of piles and the sizes of those piles (see Figure 1).

To arrive at an overall rating of similarity of photos by combining judgments from individual raters, it seems unrealistic to consider only two values for each rater, i.e., similar or dissimilar. Following the philosophy of 3WD, we take three values: a pair of photos is considered to be similar if the group agrees on their similarity at or above a certain degree and to be dissimilar if the group agrees on their similarity at or below another degree; otherwise, the pair is considered to be divided or undecided.

In our earlier paper [3], we only considered a simple function to synthesize ratings from a group of raters. The main objective of the current study is to propose a general 3-way framework for analyzing facial similarity. We suggest and study a more general function for pooling together individual ratings, in order to arrive at an overall 3-way rating of similar, dissimilar, and divided. The results of such a 3-way analysis would be useful for an in-depth understanding of and further applications of group ratings.

The purpose of this work is to understand what, if any, differences can be reliably extracted from the card sorting data so that raters who may define similarity in different ways can be identified, classified, and perhaps quantified. Three-way classification of similarity may be viewed as a first step towards a more comprehensive model of similarity analysis.

Once we have a 3-way classification, we can attempt to extract features that contribute to the similarity and dissimilarity of photos. A more in-depth analysis of undecided photos may also reveal possible reasons that raters are divided. Similarity analysis based on card sorting by people is complementary to similarity analysis based on photo features. It will be interesting to compare and integrate the two types of approaches. For example, based on photo features, we may be able to ask raters to sort a sample set of photos, rather than the entire set. Alternatively, card sorting by people may provide insights into the design of a feature-based similarity measure. Results from the present study serves as a basis for these future investigations on similarity.

# 2 A Simple Linear Model of Three-Way Analysis

A group of raters (N=25) were asked to sort a set of facial photographs (M=356) into an unrestricted number of piles based on how they judged similarity of the photos. Through this activity, each rater contributed to the assessment of the similarity amongst the photos. It is not possible for any rater to directly consider the similarity of all  $\binom{356}{2}=63,190$  pairs of photos. Data about which comparisons were made directly (between the photo being sorted and the top photo on each pile) and which indirectly was not recorded. Therefore, a means of analysing the similarity judgements is sought in order to reduce the number of photos under consideration in further studies. If there is more than one strategy being used by different raters to judge similarity, we seek to focus our efforts to understand these different strategies on the photos about which there is possibly disagreement, those photo pairs whose similarity score is between the two thresholds,  $\alpha$  and  $\beta$ .

An algorithmic approach such as Eigenfaces, popularized by Turk and Pentland [12], provides a feature-based calculation of facial similarity. In contrast, our card sorting approach to similarity attempts to understand the human perception of facial similarity. Ideally, the piles made by each rater represent equivalence classes. More pragmatically, the boundaries between the piles are likely not so clear. Intuitively, the larger the pile the more difficult it is to maintain the same high threshold for inclusion of photos in that pile.

Raters were instructed to not create any pile with only a single photo, because such a pile conveys no similarity information, only that such a photo is unrelated to all others. A small number of piles with single photos were removed from further consideration. Also, during data entry, a small number of photos were not recorded. Therefore, not all raters made judgements based on all 356 photographs. Table 1 presents the total number of photos considered by each rater.

The stimuli used in the card sorting activity combined two sets of facial photos: one set of 178 Caucasian male subjects and the other set of 178 First Nations male subjects. All photos are identified by a 4 digit code, which is a departure from earlier publications describing the card sorting study (see Hepting et al. [4,5]). The first digit indicates the stimulus set (1: Caucasian, 2: First Nations) followed by 3 digits to indicate the sequence number in that stimulus set (0-177).

Let us begin with an expression for the similarity, S, of two photos A and B, according to a rater, r, who has made  $n_r$  piles  $P_1, \ldots, P_{n_r}$ . For each rater, we can obtain a binary interpretation of similarity in terms of piles, namely, photos in the same pile are similar and photos in different piles are dissimilar.

Let P denotes the set of photos. Formally, we define a function  $s_r: P \times P \longrightarrow \{0,1\}$  for rater r as follows:

$$s_r(A, B) = \begin{cases} 1, & A \text{ and } B \text{ are in the same pile,} \\ 0, & A \text{ and } B \text{ are in two different piles.} \end{cases}$$
 (1)

In order to obtain an overall evaluation of similarity, we can synthesize ratings from all raters. A simple fusion function is a summation of ratings of individual raters, that is,

$$S(A,B) = \frac{1}{N} \sum_{r=1}^{N} s_r(A,B).$$
 (2)

It is simply the average of the similarity values given by individual raters. We have  $0 \le S(A, B) \le 1$ , S(A, B) = 0 if all raters put A and B in different piles, and S(A, B) = 1 if all raters put A and B in the same pile.

Raters may provide different classifications of photos, in terms of piles. It seems reasonable to expect that a pair of photos is similar if a majority of raters view the pair as similar and dissimilar if a majority of raters view the pair as dissimilar. If the raters are divided in the middle, we introduce the case of divided ratings. In this way, we have a 3-way interpretation of similarity. Given a pair

Rater	$M_r$ : number of	Identifiers of photos
	photos considered	not considered
1	356	
2	356	
3	356	
4	356	
5	356	
6	352	1063, 2043, 2095, 2170
7	353	1085, 2007, 2131
8	354	1164, 2001
9	356	
10	356	
11	352	1067, 2002, 2005, 2036
12	356	
13	356	
14	355	2133
15	350	1018, 1050, 1067, 2055, 2133, 2157
16	356	
17	356	
18	356	
19	356	
20	356	
21	355	2133
22	356	
23	356	
24	356	
25	355	2175

**Table 1.** The number of photos,  $M_r$ , that each rater considered. As described in Section 2, every rater may not have considered all 356 photos. The photo identifiers listed in the third column are as described in Section 2. Of the photos not considered, only 2133 and 1067 appear more than once (respectively 3 times and 2 times).

of thresholds  $(\alpha, \beta)$  with  $0 \le \beta < \alpha \le 1$ , a 3-way rating of similarity is given by:

$$\mathbb{S}(A,B) = \begin{cases} \text{Dissimilar,} & S(A,B) \leq \beta, \\ \text{Similar,} & S(A,B) \geq \alpha, \\ \text{Divided,} & \beta < S(A,B) < \alpha. \end{cases}$$
 (3)

The pair of thresholds reflect our confidence in deciding similarity and dissimilarity. When  $\alpha$  approaches 1, we become more confident about similarity, and when  $\beta$  approaches 0, we become more confident about dissimilarity. In previous studies, we used  $\beta=0.4$  and  $\alpha=0.6$ .

### 3 A Generalized Linear Model of 3-Way Analysis

This section looks at the large variance amongst individual raters and suggests a linear function for combining ratings. The linear function takes into consideration the variance amongst the raters.

#### 3.1 Variance of Raters

From the card sorting data, we have two important observations: that the numbers of piles made by different raters are very different (see Figure 1(a)), and that the sizes of piles made by each rater differ to a large extent (See Figure 1(b)). The expected inverse relationship between Figure 1(a) and Figure 1(b) holds in general. In Figure 1(c), the histogram of the sizes of piles for all raters illustrates that small pile sizes clearly predominate. It suggests that similarity judgments from different raters are of different strengths.

#### 3.2 A Linear Function of Fusion

The analysis of last subsection show two types of variance amongst raters, one is the number of piles and the other is the sizes of piles. These two types of difference affect the strength of similarity. The simple linear function (2) does not reflect these differences. Accordingly, we introduce a generalized linear model to account for both. More specifically, we propose the following linear function:

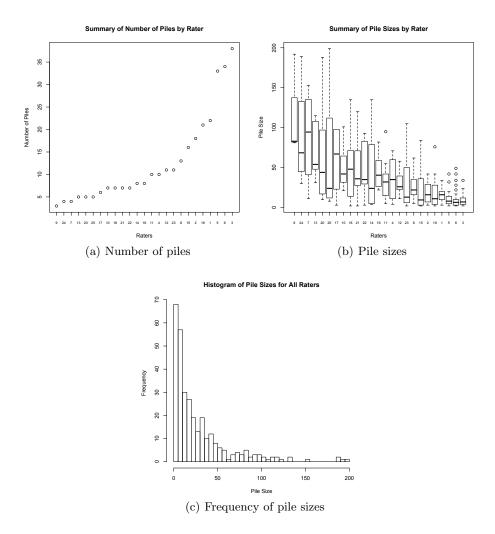
$$S(A,B) = \frac{1}{N} \sum_{r=1}^{N} w_r \cdot s_r(A,B).$$
 (4)

The weights,  $w_r$ , reflect the differences of individual raters. The similarity function  $s_r$  is generalised as  $s_r: P \times P \longrightarrow [0,1]$ , from the set  $\{0,1\}$  to the closed unit interval [0,1]. The values 0 and 1 indicate full dissimilarity and full similarity, respectively. A value between 0 and 1 indicates partial similarity and partial dissimilarity. The generalised function  $s_r$  reflects the differences in pile size.

For the weights, we assume that a rater who made more of piles, and consequently, with smaller pile sizes is more informative and confident in assessing similarity. This rater's judgements should be assigned a larger weight. For the similarity function, we assume that pairs in a pile of smaller size are more similar than pairs in piles of larger size. These two assumptions are in fact two different forms of an underlying assumption that, when deciding different piles, a rater implicitly uses a threshold on a perceived degree of similarity. A pair of photos is put into the same pile if the similarity is above the threshold.

#### 3.3 Determining the Weights

Without considering the number of photos in each of the piles that a rater made, a first attempt to quantify the differences between raters can be made by looking



**Fig. 1.** Raters and the piles they made, organized by increasing number of piles. In (a), the number of piles made by each rater is plotted on the vertical axis. In (b), the sizes of the piles made by each rater is summarized with a boxplot, where the median pile size for each rater is indicated by the bold line. The expected inverse relationship between number of piles and sizes of piles per rater holds in general, but with plenty of variability. In (c), the histogram summarizes the sizes of piles for all raters and small pile sizes clearly predominate.

at the number of piles,  $n_r$ , made by rater r. We assume that the more piles made by a rater, the more reliable the rater's ratings. This suggests that any positive monotonic increasing function of  $n_r$  will serve the purpose.

A very simple weighting function, based on the number of piles that each rater made, is given by:

$$w_{1r} = 1 - \frac{1}{n_r}. (5)$$

The function has the minimum value of 0 when  $n_r = 1$ , that is, the rater does not do any rating by putting all photo into one pile. As the number of piles increases, the weight increases and approaches 1. The maximum value of 1-1/M is achieved when  $n_r = M$ , where M is the number of photos. In this case, each pile only contain a single photo. The rater considers a photo to be only self-similar and all non-identical photos are pair-wise dissimilar. Although the two extreme cases are theoretically interesting, they normally will not happen in real ratings.

We can also directly use the number of piles as weights for the raters. By normalisation, we have the following weighting function:

$$w_{2r} = \frac{n_r}{\max\limits_{1 \le i \le N} (n_i)},\tag{6}$$

where N is the number of raters. The maximum weight of 1 is assigned to a rater who produces the largest number of piles. A rater who produces the least number of piles is assigned the minimum weight of  $\min_{1 \le i \le N}(n_i)/\max_{1 \le i \le N}(n_i)$ . Like Equation (5), the theoretical minimum value is  $1/\max_{1 \le i \le N}(n_i)$ , which may not actually happen in real rating. In contrast, the maximum weight of 1 is actually realizable.

The two weighting formulas (5) and (6) do not consider the sizes of different piles. In general, we can design a weighting formula by considering detailed information about similarity ratings. One can look at the ratio of the number of photo pairs rated similar by a rater over the number of all possible pairs. Raters who made fewer, larger piles will have a larger number of similar pairs than raters who made more, smaller piles. That is, any positive decreasing transformation of the ratio may serve as a weighting formula. A possible formula is given by:

$$w_{3r} = 1 - \frac{\sum_{i=1}^{n_r} \binom{|P_i|}{2}}{\binom{M_r}{2}},\tag{7}$$

where  $M_r$  is number of photos that rater r actually considered when rating similarity. This formula has the same value as the formula given by Equation (5) for the two extreme cases. More specifically, the minimum weight 0 is obtained if a rater only has one pile. The maximum weight of 1 is obtained when a rater puts each photo as a separate pile. In general, the fewer the similar pairs, the closer to 1 will be the weight.

For our data set of 356 photos and 25 raters, Figure 2 summarizes the weights obtained from the three formulas. It can be observed that the three formulae behave similarly: the weights generally increase as the number of piles increases.

#### **Weighting Functions for Raters**

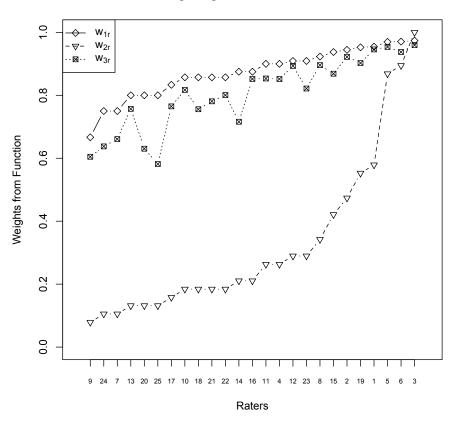


Fig. 2. Plots of the three weighting functions, from Equations 5, 6, and 7.

#### 3.4 Determining the Individual Similarity Values

For a specific rater, we can generalize the binary similarity function (1) by taking the sizes of the different piles into consideration.

Let  $k_1, \ldots, k_{n_r}$  denote the sizes of the  $n_r$  piles,  $P_1, \ldots, P_{n_r}$ , made by rater r. We assume that the strength of similarity for a pair photos depends on the size of pile into which the pair is sorted. More specifically, the larger the pile, the the

weaker the degree of similarity. If we assume that the degree of similarity is 1 for pairs in a smallest pile, the similarity will decrease as the size of a pile increases. This immediately leads to the following generalised similarity measure: for two photos A and B,

$$s_{1r}(A,B) = \begin{cases} \frac{\min\limits_{1 \le i \le n_r} (k_i)}{k_j}, & A \text{ and } B \text{ are in the same pile } P_j, \\ 0, & A \text{ and } B \text{ are in two different piles.} \end{cases}$$
(8)

It can be seen that  $s_{1r}$  has the maximum value of 1 for pairs in a smallest pile and increases as the pile increases. The binary similarity defined by Equation (1) is a special case. When  $k_1 = \ldots = k_{n_r}$ , that is, all piles given by rater r are of the same size, we have  $s_r(A, B) = 1$  if A and B are in the same pile and  $s_r(A, B) = 0$  if A and B are in different piles.

Formula (8) uses the size of a pile. When studying similarity, we consider the number of pairs produced by a pile. Alternatively, we can use the following pair-based similarity measure: for two photos A and B,

$$s_{2r}(A,B) = \begin{cases} \min_{1 \le i \le n_r} \binom{k_i}{2} \\ \frac{\binom{k_j}{2}}{\binom{k_j}{2}} \end{cases}, \quad A \text{ and } B \text{ are in the same pile } P_j, \qquad (9)$$

$$0, \qquad A \text{ and } B \text{ are in two different piles.}$$

Again, the maximum similarity value is assigned to a pair from a smallest pile. However, similarity value of Equation (9) will decrease faster than the value of Equation (8) when the size of a pile increases. Their minimum values are normally different.

Once we introduce weighting formulas for individual raters and generalize similarity value into unit interval [0, 1], we can form different linear models as defined by Equation (4). Furthermore, we can immediately apply Equation (3) to obtain 3-way division of similarity, dissimilarity, and divided. An important question is how to interpret and determine a pair of thresholds in a general linear model, which will be a topic of a future paper.

#### 4 Conclusion

Three-way analysis of facial similarity offers a new viewpoint for interpretation. While the similarity or the dissimilarity of a pair of photos is the result of a large degree of agreement of a group of raters, the divided judgment is the result of disagreement. This 3-way classification provides an effective method for dealing with uncertainty involved in facial similarity judgments. This study only considers the step of 3-way division of similarity. As future research, we plan to investigate each of these three divisions with pairwise comparisons in order to obtain a better understanding of facial similarity. Furthermore, it will be interesting to combine the similarity analysis based on human perception with that based on feature calculation to create a more comprehensive model.

## Acknowledgements

The authors would like to thank Dominik Ślęzak for his encouragement and the anonymous reviewers for their constructive comments. This work has been supported, in part, by two NSERC Discovery Grants.

#### References

- Deibel, K., Anderson, R., Anderson, R.: Using edit distance to analyze card sorts. Expert Systems 22(3), 129–138 (July 2005)
- 2. Faiks, A., Hyland, N.: Gaining user insight: a case study illustrating the card sort technique. College & Research Libraries 61(4), 349–357 (July 2000)
- Hepting, D.H., Bin Amer, H.H., Yao, Y.: Three-way analysis of facial similarity judgements. In: Proceedings of ISFUROS 2017 (October 2017)
- 4. Hepting, D.H., Maciag, T., Spring, R., Arbuthnott, K., Ślęzak, D.: A rough sets approach for personalized support of face recognition. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. pp. 201–208. Springer Berlin Heidelberg, Berlin, Heidelberg (December 2009)
- Hepting, D.H., Spring, R., Ślęzak, D.: A rough set exploration of facial similarity judgements. In: Peters, J.F., Skowron, A., Hiroshi, S., Chakraborty, M.K., Ślęzak, D., Hassanien, A.E., Zhu, W. (eds.) Transactions on Rough Sets XIV, Lecture Notes in Computer Science, vol. 6600, pp. 81–99. Springer (November 2011)
- B.Q., Wong, H., Yiu, K.f.C.: On two novel types of threeway decisions inthree-way decision spaces. International Jour-285 nal of Approximate Reasoning 82, 306 (March 2017), http://www.sciencedirect.com/science/article/pii/S0888613X1630319X
- 7. Li, H., Zhang, L., Zhou, X., Huang, B.: Cost-sensitive sequential three-way decision modeling using a deep neural network. International Journal of Approximate Reasoning 85, 68 78 (June 2017), http://www.sciencedirect.com/science/article/pii/S0888613X17302086
- 8. Li, X., Sun, B., She, Y.: Generalized matroids based on three-way decision models. International Journal of Approximate Reasoning 90, 192 207 (November 2017), http://www.sciencedirect.com/science/article/pii/S0888613X17304784
- 9. Liang, D., Xu, Z., Liu, D.: Three-way decisions decision-theoretic rough sets with dual hesitant fuzzy information. Information Sciences 396, 127 143 (August 2017), http://www.sciencedirect.com/science/article/pii/S002002551730539X
- 10. Martine, G., Rugg, G.: That site looks 88.46% familiar: quantifying similarity of web page design. Expert Systems 22(3), 115–120 (July 2005)
- 11. Soranzo, A., Cooksey, D.: Testing taxonomies: Beyond card sorting. Bulletin of the American Society for Information Science and Technology 41(5), 34–39 (June 2015)
- 12. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience  $3(1),\,71–86$  (January 1991)
- Yang, X., Li, T., Liu, D., Chen, H., Luo, C.: A unified framework of dynamic three-way probabilistic rough sets. Information Sciences 420, 126 – 147 (December 2017), http://www.sciencedirect.com/science/article/pii/S0020025517308939

- 14. Yao, Y.Y.: An outline of a theory of three-way decisions. In: Yao, J., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) Rough Sets and Current Trends in Computing: 8th International Conference, RSCTC 2012, Chengdu, China, August 17-20, 2012.Proceedings. pp. 1–17. Springer Berlin Heidelberg (August 2012), https://doi.org/10.1007/978-3-642-32115-3\_1
- 15. Yao, Y.Y.: Three-way decisions and cognitive computing. Cognitive Computation 8(4), 543-554 (Aug 2016), https://doi.org/10.1007/s12559-016-9397-5
- Y., Wang, 16. Yu, Yao, Н., Jiao, P., G.: Detecting and refinregions in complex networks overlapping with Sciences 21 - 41decisions. Information 373,(December 2016), http://www.sciencedirect.com/science/article/pii/S0020025516306703