Science in China Series F: Information Sciences

© 2009 SCIENCE IN CHINA PRESS

Springer

www.scichina.com
info.scichina.com
www.springerlink.com

Comparison of human face matching behavior and computational image similarity measure

CHEN WenFeng[1], LIU ChangHong[2], LANDER Karen[3] & FU XiaoLan[1†]

[1] State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China;

[2] Department of Psychology, University of Hull, Hull HU6 7RX, UK;

[3] School of Psychological Sciences, University of Manchester, Manchester M13 9PL, UK

Computational similarity measures have been evaluated in a variety of ways, but few of the validated computational measures are based on a high-level, cognitive criterion of objective similarity. In this paper, we evaluate two popular objective similarity measures by comparing them with face matching performance in human observers. The results suggest that these measures are still limited in predicting human behavior, especially in rejection behavior, but objective measure taking advantage of global and local face characteristics may improve the prediction. It is also suggested that human may set different criterions for "hit" and "rejection" and this may provide implications for biologically-inspired computational systems.

image similarity, human behaviour, face matching, objective measure, subjective measure

1 Introduction

Image similarity measurement is a fundamental issue in both computer and human vision including many real world applications. For example, many matching algorithms of face and object recognition systems are built on image similarity measurements such as Gabor jet similarity[1].

Perhaps the simplest way to quantify the similarity between two images is the mean squared error. This method is appealing because it is easy to compute and is mathematically convenient in the context of optimization. However, the method is not robust enough to deal with image variations caused by changes in illumination, pose, and facial expression. Gabor features are known for being more robust to small variations in scaling, rotation, distortion, illumination, poses, and expressions[1]. For this reason, they are frequently employed in face and object recognition[1−4]. Gabor-based approach mimics the spatial filtering operations of the neurons in the primate striate cortex[2]. Because of this biologically inspired nature, Gabor based approaches have often been chosen to model face recognition in humans[1−5].

While Gabor jet similarity is based on the local facial features[1,5], human observers are known to

Received May 23, 2008; accepted November 25, 2008
doi: 10.1007/s11432-009-0044-6

†Corresponding author (email: fuxl@psych.ac.cn)

Supported by the National Basic Research Program of China (Grant No. 2006CB303101), the National Natural Science Foundation of China (Grant Nos. 60433030, 30600182 and 30500157), and the Royal Society

employ both local and global information for face processing[6,7]. Accordingly, the Gabor jet similarity approach cannot adequately account for human performance for some face recognition tasks. For example, although image analysis shows that illumination accounts for more image variation than pose, a change in illumination seems to matter less than a change in pose for human observers[8,9].

Due to the limitations of approaches based on local features, a great deal of work in recent years has gone into the development of similarity measurement methods that take advantage of known characteristics of the human visual system. For example, Wang et al.[10] suggest that the pixels in natural images exhibit strong dependencies, especially when they are spatially proximate, and these dependencies carry important information about the structure of the objects in the visual scene. The luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Based on this observation, they proposed a top-down approach to compute image similarity called structural similarity index (SSIM).

The approach has so far been evaluated in a variety of ways. Its performance in computer vision tasks has been tested. Its biological plausibility has mainly been assessed at a relatively low level. However, there have been a limited number of attempts to validate its computational measures based on a high-level, cognitive criterion of objective similarity. For example, Wang et al.[10] provided a behavioral validation with higher prediction power of subjective rating score based on objective SSIM score than that of other models. However, high-level, cognitive behavior is beyond simple similarity rating. Furthermore, both local and global information is important for face recognition[6,7]. Most computational similarity measures only resemble some aspects of these known characteristics of the human visual system.

In this paper, we validate computational similarity measures by examining their predictive power for face matching performance in human observers. Our study is situated in the context of a larger framework elaborated to identify synergies between the development of computer vision systems and our understanding of human visual systems. Specifically, our paper is motivated by a number of issues. Firstly, SSIM is derived from image quality method. Although it can be thought of as a similarity measure for comparing any two signals[10], more evidence is needed to examine whether it is a robust image similarity measure. Secondly, although machine and human face recognition have been compared in the literature, the human performance data in these comparisons were rarely based on face matching (see ref. [3] for a review). While the prior comparisons have provided some very useful information about the possible sources of information used in remembering faces, they are not necessarily informative about the process of matching faces[3]. Thirdly, unlike the principal components analysis (PCA), new similarity measures such as SSIM have not been compared with human face recognition. Finally, because humans employ both local and global information for face processing, it is important to compare computational measures based on the same information[6,7].

To assess the structures of the objects in SSIM and the use of global information, we asked observers to undertake two tasks: (1) a different-expression face matching task, whereby observers judged two face images with different expressions of same person, thus the structures of the two images were similar; (2) a same-expression face matching task, whereby observers judged two face images with same expression of different person, thus the structures of the two images are less similar.

2 Methods

2.1 Stimuli

The face database was obtained from Binghamton University. It contained 100 3D faces and photographical texture maps captured from real people without facial hair or spectacles. We used the original texture maps only without the 3D data. The texture maps are 2D photographs taken in a strictly controlled environment. More details about this database can be found in ref. [11]. We used 75 models from the database. There were 51 Caucasian models (21 for male, 30 for female), 24 East Asian models (11 for male, 13 for female). Nine additional models were used in the practice session. Each face model was rendered against a black background in the full frontal pose. Each face had three facial expressions (neutral, happy, sad). The faces were saved as grey-level bitmap images. To minimize the low-level image cues for the task, the luminance and root-mean-square contrast of the images were scaled to the grand means.

2.2 Sequential face matching task as behavioral similarity measure

The performance of sequential face matching task was used as human behavior validation of computational similarity measure. Each participant completed 6 practice trials and 50 experimental trials. Practice trials were used to help participants familiarize the task procedure. The performance in practice trials was not used in data analysis.

Each matching trial consisted of a pair of faces presented one after the other in the centre of the screen (see Figure 1). A trial began with a 500 ms central fixation cross and a 500 ms blank screen. The first face was then presented for 3 s. The second face appeared after a 500 ms blank screen. The expression on the first face was randomly chosen from one of the three categories: happy, sad, or neutral. The three expressions occurred equally often. In the different-expression matching task, the second face was always shown in neutral expression. In the same-expression matching task, the second face was shown in the same expression as the first face. In half of the trials, the second face was the same identity as the first face. In the remaining trials, the second face was different from the first face. The pair of faces were also presented in different sizes. Either one of them was 512 by 512 pixels, whereas the other 384 by 384 pixels. Difference sizes were used to prevent participants from using the position information alone for the matching task. Participants were instructed to judge whether the pair of face images were of the same person. They were told to give their answer as quickly and accurately as possible by pressing one of the two keys labeled "Yes" or "No". The second face remained on screen until the participant responded.

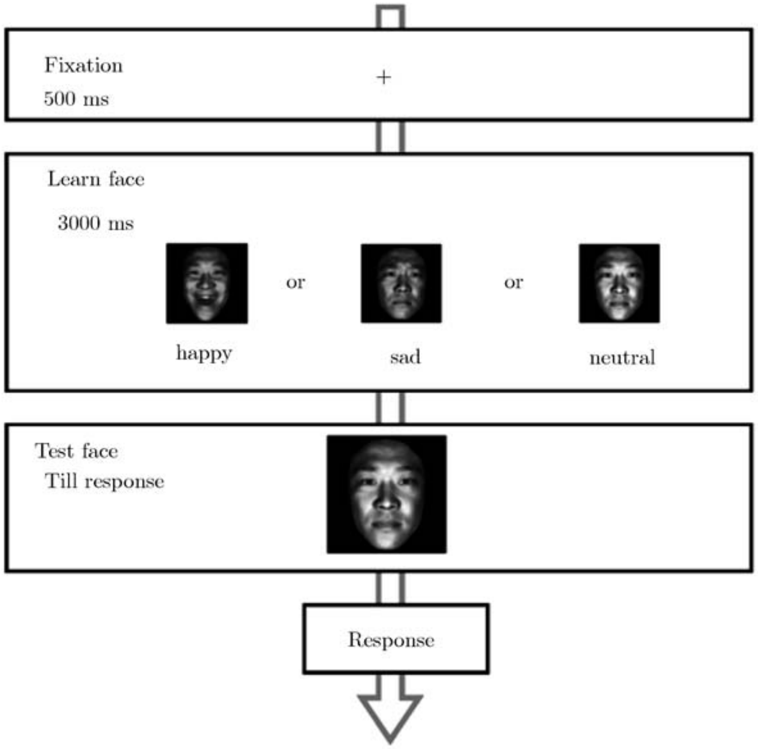

Figure 1 Illustration of the procedure used in the tasks.

Twenty-three Caucasian undergraduate students participated in the different-expression matching task, and forty Caucasian undergraduate students participated in the same-expression matching task. Their ages ranged from 19 to 35 years. All participants had normal or corrected-to-normal vision.

2.3 Computational similarity measure

The SSIM index is a similarity measure that employs global information[10]. In contrast, the Gabor similarity measure (GaborSim) is based on local information[1]. We computed GaborSim score based on Gabor wavelets feature vectors[12]. We also computed a similarity score called Gabor-SSIM as a measure of both local and global information. The latter was an SSIM score computed from Gabor features of the face[12]. For the different-expression matching task, the similarity score was computed from face pairs of where a face with the neutral expression is matched with a face with sad/happy expression. For the same-expression matching task, the similarity score was computed from face pairs where the two faces had the same expression. The face pairs in both tasks were of the same race/gender.

3 Results

3.1 Similarity data

Mean computational similarity scores for the two types of face pair are shown in Tables 1 and 2. An analysis of variance (ANOVA) revealed that for all similarity measures, there was a significant main effect of image similarity ($Fs > 16,\ ps < 0.001$), where sad-neutral face pairs were more similar than happy-neutral face pairs when the pairs were of the same person. When the pairs were of different persons, the happy face pair was more dissimilar than the sad face pair for both SSIM and GaborSim scores ($Fs > 4,\ ps < 0.05$). However, no significant difference was found when the Gabor-SSIM measure was used ($F = 0.65, p = 0.429$).

A pair of face images is more likely to be judged as the same person if they are similar based on a set of criteria. Otherwise, they would be judged as different persons. Therefore, the criteria based on SSIM, GaborSim and Gabor-SSIM scores would predict a higher hit rate for the sad-neutral face pairs. For the happy-neutral face pairs, SSIM and GaborSim results would predict a higher correct rejection rate, whereas the Gabor-SSIM result would predict comparable matching performance or the correct rejection rate for these face pairs. We used our behavioral data to test these predictions.

3.2 Behavioral data

Mean matching performance (hit rate and correct rejection rate) for the different types of face image pair are shown in Tables 3 and 4 respectively. An ANOVA on hit rate data revealed no significant main effect of race or interaction ($Fs < 1, ps > 0.45$). However, the main effect of expression pair was significant ($F = 4.32, p < 0.05$).

The main effect of race or interaction for the rejection rates was also not significant ($Fs < 2.2$, $ps > 0.16$). However, there was a significant main effect of expression ($F = 7.89$, $p < 0.01$), where the correct-rejection rate for the happy pair was lower than the sad pair ($p < 0.05$). This result was contrary to the predictions based on the computational similarity scores.

Table 1 Mean similarity scores for image pairs of the same person where happy/sad expression was compared to the neutral expression

Face-pair	SSIM	GaborSim	Gabor-SSIM
Happy-neutral	0.38(0.33)	0.72(0.05)	0.84(0.03)
Sad-neutral	0.79(0.15)	0.77(0.05)	0.87(0.04)

Table 2 Mean similarity scores of image pair for two different faces with the same expression

Face-pair	SSIM	GaborSim	Gabor-SSIM
Happy	0.43(0.17)	0.79(0.04)	0.83(0.02)
Sad	0.77(0.08)	0.81(0.04)	0.82(0.02)

Table 3 Matching performance (hit rate) for different expressions of the same person

Race	Happy-neutral	Sad-neutral
Asian	0.79(0.15)	0.85(0.16)
Caucasian	0.82(0.17)	0.87(0.14)
Total	0.81(0.16)	0.86(0.15)

3.3 Correlation analysis

Correlations between matching performance (hit rate and correct rejection rate) and similarity scores for the different combinations of face image pair are shown in Table 5. The results show that Gabor-SSIM performed best among all the three indexes, whereas GaborSim performed second best. Overall, the similarity measures predicted the hit rates more accurately than for the rejection rates. Remarkably, only Gabor-SSIM resulted in correct predictions with negative correlations, i.e., the more similar the two images of different faces, the poorer the rejection rates in matching performance.

Table 4 Matching performance (correct-rejection rates) for image pair of faces with the same expression

Race	Happy	Sad
Asian	0.79(0.07)	0.82(0.07)
Caucasian	0.81(0.11)	0.89(0.09)
Total	0.80(0.09)	0.87(0.09)

Table 5 Correlations between matching performance and similarity scores

Performance	Expression	SSIM	GaborSim	Gabor-SSIM
Hit	happy-neutral	0.070	0.549	0.549
	sad-neutral	−0.205	0.355	0.505
Rejection	happy-neutral	0.219	0.370	−0.250
	sad-neutral	−0.199	0.213	−0.228

4 Discussion

In this paper, we used face-matching performance in humans as a criterion to study how well computational image similarity measures predict human face recognition based on image similarity. According to our comparison of behavioral and computational similarity data, a local-based image similarity measure can efficiently predict human performance in matching faces with similar structure information. However, it can lose its predictive power when faces contain dissimilar structural information. Human observers are known to employ both local and global information for face processing, and the contour information and configural relationships of faces are known as important global information[6,7]. For this, an image similarity measure based on both local feature and structure information is more likely to be successful.

SSIM is derived from image quality method. Although SSIM can be thought of as a similarity measure for comparing any two signals, our data suggest that it still is quite limited and may be improved by adding local information.

In a purely behavioral and top-down approach, a face is represented as a point in an abstract psychological space where the features are interpreted so that they are related to the physical appearance of the face[5]. In a purely computational approach, a face may be represented as a collection of explicitly derived physical features[5]. More work by behavioral and engineering scientists is needed to integrate these approaches. In our attempt, we integrated the global characteristics of human face recognition into the computation of similarity measure and found the relatively higher correlation of Gabor-SSIM score and human performance. It indicates human face recognition performance can be better predicted when the computational measures incorporate both global and local information.

Our results also indicate that responses for the same-face and different-face pairs can be predicted quite differently by the computational measures. This may result from independent "same"/"different" decisions of human observers in "same"-"different" judgments, in which the "different" decision is based on a serial, self-terminating process and the "same" decision is based on a fast identity reporter[13,14]. Stewart and Brown[15] showed that humans classify novel stimuli into a category based on similarity, but reject novel stimuli from a category based on dissimilarity. This is consistent with Tversky[16], who observed that judgments of similarity and dissimilarity in humans are asymmetric. Similarity between objects is determined both by common and distinctive features, whereas dissimilarity is determined by their distinctive features. This is in clear contrast to the simple linear transform of similarity and dissimilarity commonly used in computer vision. Our results suggest that the human face recognition system may set different criteria for "hit" and "rejection" decisions. This difference may be useful for the future generation of biologically-inspired computer vision system.

Clearly, future studies must address generalization of these results by varying factors such as lighting conditions, viewpoint, and part configuration including the assessment of other computational measure. Further studies may implement a wider range of computational and behavioral measures.

We thank Dr. Lijun Yin of Binghamton University for providing the BU-3DFE face database and Dr. Jianke Zhu of Chinese University of Hong Kong for sharing the code of Gabor wavelet transform. We also thank the anonymous reviewers for their helpful comments on an earlier version of this paper.

1 Lades M, Vorbrüggen J C, Buhmann J, et al. Distortion invariant object recognition in the dynamic link architecture. IEEE T Comput, 1993, 42(3): 300–311

2 Biederman I, Kalocsai P. Neurocomputational bases of object and face recognition. Phil Trans R Soc B, 1997, 352: 1203–1219

3 Burton A M, Miller P, Bruce V, et al. Human and automatic face recognition: A comparison across image formats. Vision Res, 2001, 41: 3185–3195

4 Russell R, Biederman I, Nederhouser M, et al. The utility of surface reflectance for the recognition of upright and inverted faces. Vision Res, 2007, 47: 157–165

5 Steyvers M, Busey T. Predicting similarity ratings to faces using physical descriptions. In: Wenger M, Townsend J, eds. Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges. Lawrence Erlbaum Associates, 2000

6 Farah M J, Wilson K D, Drain H M, et al. What is "special" about face perception? Psychol Rev, 1998, 105: 482–498

7 Rakover S S. Featural vs. configurational information in faces: a conceptual and empirical analysis. Brit J Psychol, 2002, 93(1): 1–30

8 Adini Y, Moses Y, Ullman S. Face recognition: the problem of compensating for changes in illumination direction. IEEE Trans Patt Anal Mac Intel, 1997, 19(7): 721–732

9 Liu C H, Bhuiyan A -A, Ward J. Transfer between pose and illumination training in face recognition. Percept, 2007, 36(Suppl.): 148

10 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13(4): 600–612

11 Yin L, Wei X, Sun Y, et al. A 3D facial expression database for facial behavior research. In: IEEE 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 2006. IEEE Computer Society TC PAMI, 2006. 211–216

12 Zhu J, Vai M I, Mak P U. A new enhanced nearest feature space (ENFS) classifier for Gabor wavelets features-based face recognition. ICBA 2004. Lect Notes Comput Sci, 2004, 3072: 124–131

13 Farell B. "Same"-"different" judgments: A review of current controversies in perceptual comparisons. Psychol Bull, 1985, 98(3): 419–456

14 Bamber D. Reaction times and error rates for "same"-"different" judgments of multi-dimension stimuli. Percept Psychophys, 1969, 6: 169–174

15 Stewart N, Brown G D A. Similarity and dissimilarity as evidence in perceptual categorization. J Math Psychol, 2005, 49(5): 403–409

16 Tversky A. Features of similarity. Psychol Rev, 1977, 84: 327–352