Caricature Recognition in a Neural Network[1]

James W. Tanaka

Carnegie-Mellon University

Abstract

In a caricature drawing, the artist exaggerates the facial features of a person in proportion to their deviations from the average face. Empirically, it has been shown that caricature drawings are more quickly recognized than veridical drawings (Rhodes, Brennan, & Carey, 1987). Two competing hypotheses have been postulated to account for the *caricature advantage*. The *caricature hypothesis* claims that the caricature drawing finds a more similar match in memory than the veridical drawing because the underlying face representation is stored as an exaggeration. The *distinctive features hypothesis* claims that the caricature drawing produces speeded recognition by graphically emphasizing the distinctive properties that serve to individuate that face from other faces stored in memory. A computational test of the two hypotheses was performed by training a neural network model to recognize individual *face* vectors and then testing the model's ability to recognize both caricaturized and veridical versions of the face vectors. It was found that the model produced a higher level of activation to caricature face vectors than to the non-distorted face vectors. The obtained caricature advantage stems from the model's ability to abstract the distinctive features from a learned set of inputs. Simulation results were therefore interpreted as support for the distinctive features hypothesis.

Introduction

In a caricature drawing, the artist graphically exaggerates those features of a famous individual in proportion to their deviations from the normative or average face.

---

[1]This research was supported by a NIMH NRSA Training Grant #1T32 MH19102. I would like to thank Martha Farah, Steve Keele, Jay McClelland and Kris Taylor for their helpful suggestions and advice. Please address all correspondence to: Jim Tanaka, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA, 15213. E-mail address: tanaka@psy.cmu.edu.

For example, in political cartoons, George Bush is usually portrayed as having an especially long chin and Richard Nixon is recognizable by the distinctive slope of his nose. The goal of the caricaturist is to exploit the facial characteristics that serve to individuate that person from everyone else. Gibson (1973) noted that "the caricature may be a poor projection of his face but good information about it. The form of the face is distorted, but not the essential features of the face" (p.6). The relative ease with which caricatures are recognized presents a paradox of sorts. Why is it that these *distorted* renderings are as easily recognized as *veridical* drawings? One explanation has been to suggest that normal face recognition involves some type of caricature process. Like the caricature artist, face recognition processes exaggerate the distinctive features of an individual relative to a normative prototype before storing the representation in long term memory. According to the *caricature hypothesis*, faces are represented in memory not as veridical representations, but as caricatures. Hence, caricature drawings serve as better retrieval cues than veridical drawings because they are more similar to the underlying stored representation. Recent work by Rhodes, Brennan and Carey (1987) has provided some support for the caricature position. They found that caricature drawings of familiar individuals were recognized faster than veridical drawings. It has also been shown that subjects judge caricatures as depicting better likenesses of individuals than veridical drawings (Rhodes et al., 1987) and caricatures are just as likely to be falsely recognized as veridical drawings (Mauro & Kubovy, 1988). Thus, results showing a caricature *advantage* suggest that the caricature drawings may find a more similar match to the stored face representation than the veridical drawings.

The *distinctive features hypothesis* also claims that individual faces are encoded in terms of their distinctive deviations from the normative face. However, in contrast to the caricature hypothesis, the distinctive features position maintains that the deviations are not stored as exaggerated representations, but as veridical ones. In support of this view, past studies have shown that attending to distinctive facial characteristics improves face recognition (Winograd, 1981) and highly distinctive, atypical faces are better remembered than less distinctive, typical faces (Barlett, Hurry & Thorley, 1984; Light, Kayra-Stuart & Hollander, 1979). According to the distinctive features approach, the caricature advantage can be explained in terms of the drawing's stimulus properties, and not in terms of a distorted memory representation. Acting as a kind of "super stimulus," the caricature drawing directs the viewer's attention to distinctive properties of the face thereby facilitating quick access to the stored face representation.

Does the caricature advantage depend on the encoding and storage of a distorted caricature representation or can the advantage be shown based on distinctive feature information alone? In the following simulation, a test of the caricature and distinctive feature hypotheses was performed. A neural network model was trained to recognize the feature vectors describing three different *faces*. After learning, veridical

and caricature feature vectors were presented to the model and their output activations measured. Under these learning conditions, the caricature hypothesis would predict the absence of a caricature advantage because the exaggerated caricature representations were not explicitly encoded in memory. The distinctive features hypothesis predicts a caricature advantage provided that the network is able to abstract the distinctive feature information from the set of learned vectors.

Description of Caricature Model

As shown in Figure 1, the Caricature Model is a single layer connectionist network consisting of eight input units and three output units. The eight input units comprised the feature vector taking on the activation values of either 0 or +1. The three output units represented three hypothetical face units (i.e., Joe, Tom, Bob). Their activation was calculated by the linear activation function:

$$face_i = \sum_{j=1}^{8} i_j w_{ij}$$

where $face_i$ is the activation level of an individual face, i_j is activation of the input units, and w_{ij} is the strength of the connection weights between the feature units and the face unit.

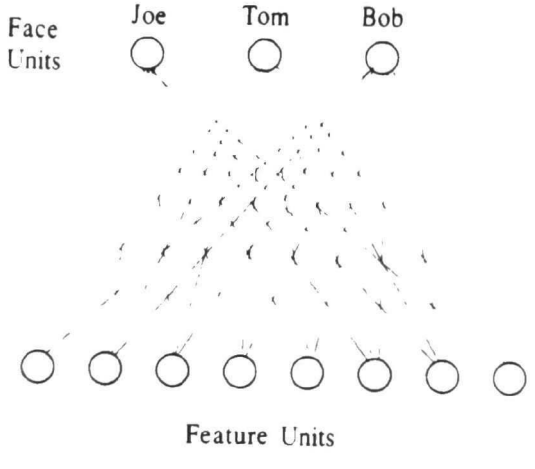

Figure 1. Caricature Model

The network was trained to recognize three prototypical feature vectors, [1,1,1,1,0,0,0,0], [1,1,0,0,1,1,0,0], [1,0,1,0,1,0,1,0], representing the face units of Joe, Tom, and Bob respectively. Instead of presenting the prototypical feature vectors of Joe, Bob and Tom for learning, eight permutations of the prototypical vectors were created by flipping the value of one of the eight input units either to one or zero. For example, a permutation of the first feature in Joe prototype vector would be the feature

vector of [0,1,1,1,0,0,0]. The permuted feature vectors represented the slight perturbations of an individual's face which may arise due to changes in facial expressions or viewing conditions. A total of 24 feature vectors, eight permutations of the three prototypical feature vectors, were presented randomly for learning to the network. Learning was accomplished by adjusting the connection weights between the feature units and the face units according to the standard delta rule (Rumelhart, Hinton & McClelland, 1986):

$$\Delta w_{ij} = \varepsilon (t_i - face_i) \; i_j$$

where Δw_{ij} is the change in connection weights, ε is the learning rate parameter, t_i is the expected output, $face_i$ is the activation of the face unit, i_j is the activation level of the feature unit. Learning of each vector continued until a combined absolute error value of .10 was achieved.

Simulation Results

An initial test was performed to explore whether the model was able to abstract the prototypical patterns from the set of permuted feature vectors. As a test of prototype abstraction, the network's output activations to the 24 learned feature exemplar vectors were compared to the output activations to three prototypical feature vectors. The three prototypical feature vectors produced higher output activations than 23 of the 24 learned feature vectors. Thus, consistent with similar distributed models of memory (McClelland & Rumelhart, 1987), the Caricature Model's was able to demonstrate prototype abstraction by computing a measure of central tendency from a set of given inputs.

An assumption common to the caricature and distinctive features hypotheses is the existence of a normative face representation. That is, both approaches assume that faces are encoded with respect to their deviations from an average representation. As a test for the presence of the normative face representation, the values of the eight feature units were averaged across the three prototypical feature vectors resulting in a "normative" feature vector of [1,.66,.66,.33,.66,.33,.33,0]. When presented to the network, the normative feature vector produced levels of activation that were roughly equivalent in the three face units. Specifically, Joe captured 44% of the available activation, Bob captured 25% of the available activation and Tom captured 30% of the available activation. Joe's higher level of activation was due to the "recency effects" of the model; that is, a Joe feature vector happened to be the last face learned and therefore, exerted a stronger effect on recognition. Thus, all three face units were partially activated by the normative feature vector with none of face units collecting the

majority of available activation. As a consequence of learning the pattern of activations particular to the faces of Joe, Bob and Tom, the Caricature Model indirectly encoded the pattern of activations that described the average face.

The critical test of the caricature and distinctive feature hypotheses was the model's response to the caricature feature vectors. The caricature hypothesis predicts no advantage for the caricature feature vectors because the distorted vectors were not encoded in memory. The distinctive features hypothesis predicts a possible caricature advantage depending upon the model's ability to identify distinctive features of the learned prototypes. As a test of these two hypotheses, caricature versions of the prototypical feature vectors were produced by the following equation:

$$caricature_j = i_j + \beta\ (i_j - norm_j)$$

where $caricature_j$ is the new caricature feature value, i_j is the activation value of the original feature unit, β is a caricature constant that controls the amount of exaggeration, and $norm_j$ is the value of the normative feature. The caricature equation is similar to the equation used in Brennan's Caricature Generator Program (Brennan, 1984). The caricature equation has the property of emphasizing features in proportion to the deviations from the norm. Features that show large deviations from the average face are weighted more heavily than those with small deviations. Applying the formula to the prototypical features with a β of .25 produced caricature feature vectors of [+1,+1.08, +1.08,+1.16,-.16,-.08,-.08,0], [+1,+1.08,-.16,-.08,+1.08,+1.16, -.08,0], [+1,-.16,+1.08,-.08,+1.08,-.08,+1.16,0] for the Joe, Bob, and Tom prototype vectors respectively.

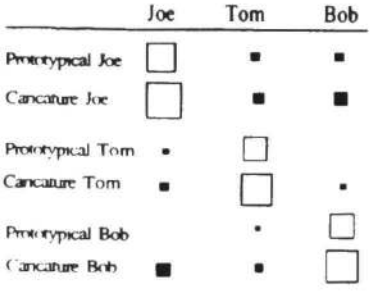

Figure 2. Output activations of the prototypical feature vectors and the caricature feature vectors. Unfilled squares indicate positive activation and filled squares indicate negative activation; the amount of activation is indicated by the area of the square.

Consistent with the prediction of the distinctive feature hypothesis, the caricature feature vectors generated a stronger activation level than the three original prototypical feature vectors (shown in Figure 2). Importantly, not only was the activation of the correct face unit enhanced, but suppression of the incorrect face units was also

increased. Thus, the model demonstrated a caricature advantage without explicitly encoding the caricature representation. What was the source of the caricature advantage? Inspection of the connection strengths between the feature units and the face units revealed that the model assigned a larger weight value to those connections that were the most distinctive of a given feature vector. In other words, those feature units that had the high discrimination value played a larger role in the face unit computation than those features that were less discriminating. For example, the fourth and fifth feature units distinguished the Joe prototype vector from the Bob and Tom prototype. Specifically, the fourth feature unit of the Joe prototype was 1 as compared to the fourth unit of Bob and Tom which was 0. Likewise, the fifth feature unit of Joe was 0 as compared to the fifth feature unit of Bob and Tom which was 1. Accordingly, the connections between the fourth and fifth feature units and the Joe face unit were weighted the most strongly. A strong positive weight connection between the fourth feature unit and the Joe face unit signaled strong evidence *for* the activation of the Joe face unit and a negative connection between the fifth feature unit and the Joe face unit signaled strong evidence *against* the activation of the Joe face unit. Thus, the distinctive feature information embodied in the connection weights of the model combined with the distinctive feature information found in the caricature vector produces the overall caricature advantage.

Implications and Limitations of the Model

The simulation results seem to be most consistent with the following interpretation of the caricature advantage: Faces are encoded in memory with respect to their distinctive properties. By emphasizing the same features in the stimulus that are distinctive in memory, caricature renderings can more strongly activate face representations than veridical drawings. Higher levels of activation lead to quicker access to face representations and faster recognition times.

The present simulation does not rule out the possibility that the recognition system encodes a distorted representation rather than a veridical representation. It only demonstrates that the extra computational step of distortion is not necessary to produce a caricature advantage. Therefore, the distinctive features hypothesis provides a more parsimonious account of the phenomenon. In fact, the Caricature Model would show a caricature advantage if the caricature vectors were learned instead of the veridical vectors. However, there may be additional reasons for treating caricature recognition as a special type of recognition rather than as a general model for face recognition. Theories accounting for the caricature advantage must also explain the ease with which the human recognition system can identify normal faces. If faces are stored as distorted caricatures, then normal, undistorted faces should be perceived as *anti-caricatures*; that is, a normal face would tend to de-emphasize its distinctive properties relative to the caricature representation. However, Rhodes et al. (1987) found that veridical drawings

of faces were better recognized than anti-caricatures indicating that normal faces are not treated as anti-caricatures.

Finally, a limitation of the current model is that the caricature advantage increases in proportion to the amount of exaggeration in the caricature vector (the β parameter in the caricature equation). This is not consistent with the empirical results of Rhodes et al. (1987) where they found that highly exaggerated drawings were not as quickly recognized as moderately exaggerated drawings. By adding an intermediate layer of hidden units to the model, it should be possible to simulate such non-linearities. Future simulation efforts will be directed toward investigating how modifications in the network's architecture might affect its behavior.

References

Barlett, J.C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory and Cognition*, **12**, 219-228.

Brennan, S.E. (1985). The caricature generator. *Leonardo*, **18**, 170-178.

Gibson, J.J. (1973). On the concept of formless invariants in visual perception. *Leonardo*. **6**, 3.

Light, L.L., Kayra-Stuart, F. &Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 212-228.

Mauro, R., & Kubovy, M. (1988). Caricature and face recognition. Unpublished manuscript.

McClelland, J.L. & Rumelhart, D.E (1986). A distributed model of human learning and memory. In D.E., J.L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* Cambridge, MA: Bradford Books.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, **19**, 473-497.

Rumelhart, D.E., Hinton, G.E, & McClelland, J.L. (1986). A general framework for parallel distributed processing. In D.E., J.L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 45-49.