Deciphering the Face

Aleix M. Martinez
The Ohio State University
aleix@ece.osu.edu

Abstract

We argue that to make robust computer vision algorithms for face analysis and recognition, these should be based on configural and shape features. In this model, the most important task to be solved by computer vision researchers is that of accurate detection of facial features, rather than recognition. We base our arguments on recent results in cognitive science and neuroscience. In particular, we show that different facial expressions of emotion have diverse uses in human behavior/cognition and that a facial expression may be associated to multiple emotional categories. These two results are in contradiction with the continuous models in cognitive science, the limbic assumption in neuroscience and the multidimensional approaches typically employed in computer vision. Thus, we propose an alternative hybrid continuous-categorical approach to the perception of facial expressions and show that configural and shape features are most important for the recognition of emotional constructs by humans. We illustrate how these image cues can be successfully exploited by computer vision algorithms. Throughout the paper, we discuss the implications of these results in applications in face recognition and human-computer interaction.

1. Introduction

Faces are one of the most important objects we see and interact with everyday. Faces tell us the identity of the person we are looking at and provide information on gender, attractiveness and age, among many others. Of primary interest is the production and recognition of facial expressions of emotion. Emotions play a fundamental role in human cognition [5] and are thus essential in studies of cognitive science. Facial expressions of emotion could also play a pivotal role in human communication [26]. And, sign languages use facial expressions to encode part of the grammar [30]. It has also been speculated that expressions of emotion were relevant in human evolution [6]. Models of the perception of facial expressions of emotion are thus important in many scientific disciplines.

A first reason computer vision research is interested in creating computational models of the perception of facial expressions of emotion is to aid studies in the above sciences [17]. Furthermore, computational models of facial expressions of emotion are important for the development of artificial intelligence [19] and are essential in human-computer interaction (HCI) systems [22].

Yet, as much as we understand how facial expressions of emotion are produced, very little is known on how they are interpreted by the human visual system. Without proper models, the scientific studies summarized above as well as the design of intelligent agents and efficient HCI platforms will continue to allude us. A HCI system that can easily recognize expressions of no interest to the human user is of limited interest. A system that fails to recognize emotions readily identified by us is worse.

This paper defines what a computer vision system for the recognition of facial expressions of emotion in the above applications should look like and presents a biologically-inspired computational model. In particular, we provide the following results.

- We show that there are (at least) three types of expressions, those used *a*) for communication between a sender and multiple receivers, *b*) in proximal, personal interactions, and *c*) (primarily) for other than communication purposes.
- We define a shape-based model consistent with the available data on human subjects. Configural features are powerful but specific to the two groups used in communication (*a* and *b* above). Configural features are defined as a non-rotation invariant modeling of the distance between facial components; e.g., the vertical distance between eyebrows and mouth.
- We argue that to overcome the current problems in face recognition software (including identity and expressions), the area should make a shift toward a more shape-based modeling. Under this model, the major difficulty for the design of computer vision and machine learning systems is that of precise detection of the features, rather than classification.

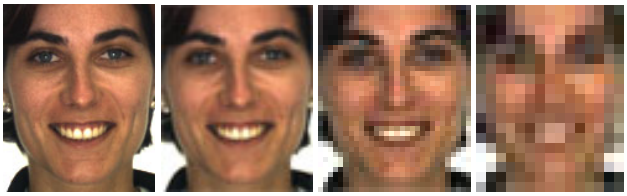

Figure 1. Four images of the same face expression joy as seen at different resolutions. Each image was reduced to one half the size of the image on its left; with the first one at about 180×150 pixels. The images are displayed at the same size to facilitate comparison and analysis.

The rest of the paper is organized as follows. Section 2 reviews relevant research on the perception of facial expressions of emotion by humans. Section 3 illustrates the importance of configural and shape features for the recognition of emotions in face images. Section 4 defines a computational model consistent with the results reported in the previous two sections. Section 5 argues that the real problem in computer vision is a detection one and emphasizes the importance of research in this domain before we can move forward with improved algorithms of face recognition. Conclusions are in Section 6

2. Facial Expressions: From production to perception

The human face is an engineering marvel. A large number of muscles allow us to produce enormous configurations. The face muscles can be summarized as Action Unit (AU) [9] defining positions characteristic of facial expressions of emotion. With proper training, one can learn to move most of the face muscles independently. Otherwise, the face seems to take predetermined configurations.

There is debate on whether these predetermined configurations are innate or learned (nature vs. nurture) and whether the expressions of some emotions is universal [12]. By universal, we mean that people from different cultures produce similar muscle movements when expressing some emotions. Facial expressions typically classified as universal are joy, surprise, anger, sadness, disgust and fear [6, 9]. Universality of emotions is controversial, since it assumes facial expressions of emotion are innate (rather than culturally bound). It also assumes emotions are categorical. That is, there is a finite set of predefined classes such as the six listed above. This is known as the categorical model.

An alternative to the categorical model is the continuous model [25, 23]. Here, each emotion is represented as a feature vector in a multidimensional space given by some basis characteristics common to all emotions. One such model is Russell's 2-dimensional circumplex model [24], where the first basis measures pleasure-displeasure and the second valance. This model can easily justify the perception of many expressions, whereas the categorical model would need to have a class (or classifier) for every possible expressions. Yet, morphs between expressions of emotions are generally classified to the closest class rather than to intermediate labels [1]. These results make the categorical versus continuous debate even more intense than that of the universality of some emotion labels. Note that, in fact, the two controversies are related to one another.

In neuroscience, the multidimensional view of emotions constitutes the limbic hypothesis [3]. Under this model, there should be a neural mechanism responsible for the recognition of all facial expressions of emotion, which was assumed to take place in the limbic system. Recent results have however uncovered dissociated networks for the recognition of most emotions. This is not necessarily proof of a categorical model, but it strongly suggests that there are at least distinct groups of emotions, each following distinct interpretations.

Computer vision researchers may believe humans are extraordinarily good at recognizing facial expressions of emotion and that we are yet to master a way to imitate this capacity. However, humans are only very good at recognizing a number of expressions. The most readily recognized emotions are happiness (joy) and surprise [8]. It has been shown that joy and surprise can be robustly identified extremely accurately at almost any resolution [8]. Fig. 1 shows a happy expression at four different resolutions. The reader should not have any problem recognizing the emotion in display even at the lowest of resolutions. However, humans are not as good at recognizing anger and sadness. Recognition drops from above 90% in joy and surprise to below 70% in anger and sadness even at the largest resolution. Fig. 2 summarizes the classification rates for images at the specified resolutions. Interestingly, we also note that while joy and surprise are quite invariant to the image size, recognition of sadness and anger degrades with the resolution. More surprisingly, we see that recognition of fear and disgust is below 50%. These emotions are nonetheless robustly recognized under image manipulations [17].

A major question of interest is the following. Why are some facial configurations more easily recognizable than others? One possibility is that expressions in the first group (e.g., joy and sadness) involve larger face transformations than the others [17]. This has recently proven to not be the case [8]. While surprise does have the largest deformation, this is followed by disgust and fear (which are poorly recognized). Learning why some expressions are so readily classified by our visual system should facilitate the definition of algorithms that can more easily recognize other expressions (such as fear and disgust) to outperform human perception where needed.

The search is on to determine which are the features ex-

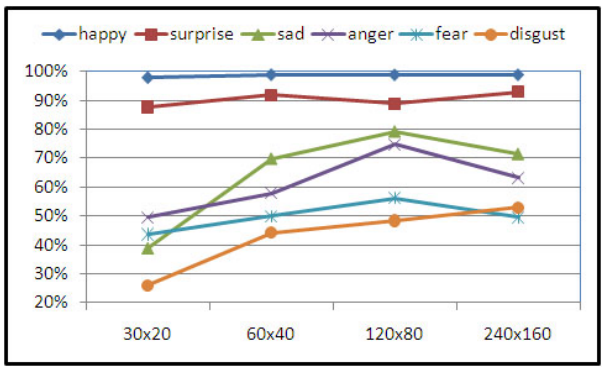

Figure 2. Correct classification of images displaying six typical emotions (happy, surprise, sad, angry, disgust, fear). Results averaged over thirty-four human subjects. Adapted from [8].

Figure 3. The three face images and schematics shown above all correspond to neutral expressions (i.e., the sender does not intend to convey any emotion to the receiver). Yet, most human subjects interpret these faces as conveying anger, sadness and surprise [20, 21].

tracted by the visual system which make the recognition of expressions such as happy and surprise easy but others more challenging. We turn to this topic in the next section.

3. Deciphering the Algorithm

In the early years of computer vision, researchers derived several feature- and shape-based algorithms for the recognition of objects and faces [13, 16, 15]. In these methods, geometric, shape features and edges were extracted from an image and used to build a model of the face. This model was then fitted to the image. Good fits determined the class and position of the face. Then, the so-called appearance-based approach, where faces are represented by their pixel-intensity maps, was studied [27]. A metric is defined to detect and recognize faces in test images [28]. Advances in pattern recognition and machine learning have made this the preferred approach in the last two decades [2].

Inspired by this success, most algorithms developed in computer vision for the recognition of expressions of emotion have also used the appearance-based model. The appearance-based approach has also gained momentum in the analysis of AUs from images of faces. The main advantage of the appearance-based model is that one does not need to predefine a feature or shape model as in the earlier approaches. Rather, the face model is inherently given by the training images.

The appearance-based approach does provide good results from near-frontal images of a reasonable quality, but it suffers from several major inherent problems. The main drawback is its sensitivity to image manipulation. Image size (scale), illumination changes and pose are all examples of this. Most of these problems are intrinsic to the definition of the approach since this cannot generalize well to conditions not included in the training set. One solution would be to enlarge the number of training images [18]. However, learning from very large datasets (in the order of millions of samples) is, for the most part, unsolved [14]. Progress has been made in learning complex, non-linear decision boundaries, but most algorithms are unable to accommodate large amounts of data – either in space (memory) or time (computation).

This bears the question as to how the human visual system solves the problem. One could argue that, throughout evolution, the homo genus (and potentially before it) has been exposed to trillions of faces. This has facilitated the development of simple, yet robust algorithms. In computer vision and machine learning, we wish to define algorithms that take a shorter time to learn a similarly useful image representation. One option is to decipher the algorithm used by our visual system. Research in face recognition of identity suggests that the algorithm used by the human brain is not appearance-based [29]. Rather, over time, the algorithm has identified a set of robust features that facilitate rapid categorization.

This is also the case in the recognition of facial expressions of emotion [21]. Fig. 3 shows three examples. These images all bear a neutral expression, that is, an expression associated to no emotion. Yet, human subjects perceive them as expressing sadness, anger and surprise [20]. The most striking part of this illusion is that these faces do not and cannot express any emotion, since all relevant AUs are inactive. This effect is called over-generalization [32], since human perception is generalizing the learned features defining these classes over to images with a different label.

The images in Fig. 3 do have something in common – they all include a configural transformation. What the human visual system has learned is that faces do not usually

look like those in the image. Rather the relationship (distances) between brows, nose, mouth and the contour of the face is quite standard. They follow a Gaussian distribution with small variance [21]. The images shown in this figure however bear uncanny distributions of the face components. In the sad-looking example, the distance between the brows and mouth is larger than normal [20] and the face is thinner than usual [21]. This places this sample face, most likely, outside the 99% confidence interval of all Caucasian faces on these two measures. The angry-looking face has a much-shorter-than-average brow to mouth distance and a wide face. While the surprise-looking face has a large distance between eyes and brows and a thinner face. These effects are also clear in the schematic faces shown in the figure.

Yet, configural cues alone are not sufficient to create an impressive, lasting effect. Some shape changes are also needed. For example, the curvature of the mouth in joy or the opening of the eyes – showing additional sclera – in surprise. Note how the surprise-looking face in Fig. 3 appears to also express disinterest or sleepiness. Wide-open eyes would remove these perceptions. But this can only be achieved with a shape change. Hence, our computational model should include both – configural and shape features. It is important to note that configural features can be obtained from an appropriate representation of shape. Expressions such as fear and disgust seem to be mostly (if not solely) based on shape features, making recognition less accurate and more susceptible to image manipulation.

4. Computational Model

Many computer vision algorithms define a single face space to represent all possible emotions. This would be consistent with the continuous view in cognitive science defined earlier. This approach has a major drawback – it can only detect one emotion from a single image. Yet, everyday experience demonstrates that we can perceive more than one emotional label in a single image. For instance, one can express a joyful surprise or a sad surprise. If we were to use a continuous model, we would need to have a very large number of labels (or multiple combinations of them) represented all over the space. This would require a very large training set, since each possible combination of labels would have to be learned. On the other had, if we define a computational (face) space for each emotion label we wish to consider, we will only need sample faces of those few emotions. This is thus the approach we will follow in this section.

Note that although the approach just described may be thought to fall in the categorical class of models, each categorical space will be defined as a continuous feature space. This allows for the perception of each emotion at different intensities (e.g., less happy to exhilarant). Furthermore, linear combinations of two or more spaces are possible. By doing so, we avoid a major problems of the categorical model – that many categories (i.e., face spaces) need to be learned and to add non-innate expressions or those used in specific cultures. Thus, *the proposed hybrid model bridges the gap between the categorical and continuous ones and resolves most of the debate facing each of the models individually*. A main issue we will leave for future work is to determine which are the basic, essential categories that, once combined, can produce the large variety of expressions of emotion we observe. For now, we will assume these correspond to those seen in most cultures – happy, surprise, sadness, anger, disgust and fear.

Each of the six emotions listed above is represented in a shape space as follows. First the face and the shape of the major facial components are detected. This includes delineating the brows, eyes, nose, mouth and jaw line. The shape is then sample with d fiducial points; equally spaced. The mean (center of mass) of all the points is computed. The $2d$-dimensional shape feature vector is given by the x and y coordinates of the d shape landmarks subtracted by the mean and divided by its norm. This provides invariance to translation and scale. 3D rotation invariance can be achieved with the kernel of [11]. The dimensions of each emotion category can now be obtained with the use of an appropriate discriminant analysis method. We use the algorithm defined in [10] because it minimizes the Bayes classification error regardless of the number of classes.

As an example, the approach detailed in this section yields the following 2-dimensional space of anger and sadness: distance between the brows and mouth and width of the face. It is important to note that, if we reduce the computational spaces of anger and sadness to 2-dimensions, they are almost indistinguishable. Thus, it is possible that these two categories are in fact connected by a more general one underlying the two. The results of Section 2 also suggest this, since anger and sadness formed the second group of emotions.

The space of anger and sadness is illustrated in Fig. 4 where we have also plotted the feature vectors of the face set of [9].

In summary, readily identified expressions are classified using a small number of configural features. Other image cues are not as good for recognition. Thus, computer vision systems that attempt to achieve high recognition rates in all emotions should be mostly based on configural features.

5. Precise Detection of Faces and Facial Features

As seen thus far, human perception is extremely tuned to small configural and shape changes. If we are to develop computer vision systems that can emulate this capacity, the

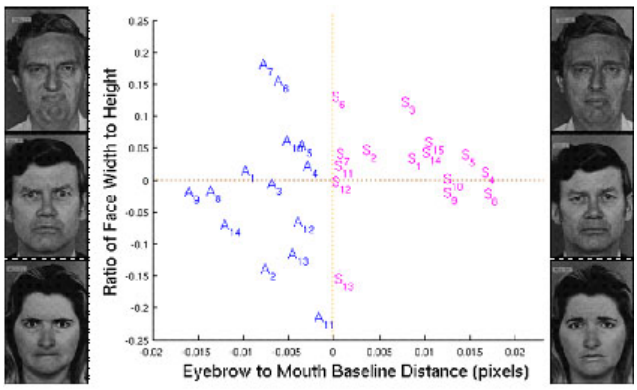

Figure 4. Shown here are the two most discriminant dimensions of the face shape vectors. We also plot the images of anger and sadness of [9]. In dashed are simple linear boundaries separating angry and sad faces according to the model. The first dimension (distance between brows and mouth) successfully classifies 100% of the sample images. The second dimension correctly labels 70% of the images. Adapted from [21].

real problem to be addressed in that of *precise detection of faces and facial features* [7]. Classification is less important, since this is embedded in the detection process; i.e., we want to precisely detect changes that are important to recognize emotions.

Most computer vision algorithms defined to date provide, however, inaccurate detections. One classical approach to detection is template matching. In this approach, we first define a template (e.g., the face or the right eye). This template is learned from a set of sample images; for example, estimating the distribution or manifold defining the appearance (pixel map) of the object [31]. Detection of the object is based on a window search – the learned template is compared to all possible windows in the image; if the template and the window are similar according to some selected metric, bounding box defining the window marks the location and size of the face. The major drawback of this approach is that it yields imprecise detections of the learned object, because a window of an non-centered face is more similar to the learned template than a window with background (say, a tree).

A solution to the above problem is to learn to discriminate between non-centered windows of the objects and well centered ones [7]. In this alternative, a non-linear classifier (or some density estimator) is employed to determine the region of the feature space defining well-centered windows of the objects from the non-centered ones. The non-centered windows are referred to as the context of the object, in the sense that these windows provide the information typically found around the object and necessary to achieve a good detection. The same approach can be applied to other detection and modeling algorithms, such as Active Appearance Models (AAM) [4].

Fig. 5 shows some sample results of accurate detection of faces and facial features with this approach.

By now we know that humans are very sensitive to small changes. But we do not know how sensitive (accurate). Of course, it is impossible to be pixel accurate when marking the boundaries of each facial feature, because edges blur over several pixels. This can be readily observed by zooming in the corner of an eye. To estimate the accuracy of human subjects, we performed the following experiment. First, we designed a system that allows users to zoom in at any specified location to facilitate delineation of each of the facial features manually. Second, we asked three people (herein referred to as judges) to manually delineate each of the facial components of close to $4,000$ images of faces. Third, we compared the markings of each of the three judges. The within-judge variability was (on average) 3.8 pixels, corresponding to a percentage of error of 1.2% in terms of the size of the face. This gives us an estimate of the accuracy of the manual detections. The average error of the algorithm of [7] is 7.3 pixels (or 2.3%), very accurate but still far short of what humans can achieve. Thus, further research is needed to develop computer vision algorithms that can extract even more accurate detection of faces and its components. This problem becomes exacerbated when the resolution diminishes, Fig. 1.

6. Conclusions

We propose a new hybrid continuous-categorical model of the perception of facial expressions of emotion – a linear combination of computational spaces defining a set of basic emotions. The model is consistent with our current understanding of human perception and can be successfully exploited to achieve great recognition results for computer vision and HCI applications. The dimensions of the computational spaces encode configural and shape features.

We conclude that to move the state of the art forward, face recognition research has to focus on a topic that has received little attention in recent years – precise detection of faces and facial features. We base our argument on recent advances in the understanding of human perception of faces. Although we have focused our study on the recognition of facial expressions of emotion, we believe that the results apply to most face recognition tasks.

Acknowledgement

This research was supported in part by the National Science Foundation, grant 0713055, and the National Institutes of Health, grants R01 EY 020834 and R21 DC 011081. The author is indebted to his students who have made the research reviewed above possible. Particular thanks to Don Neth, Liya Ding, Onur Hamsici and Shichuan Du.

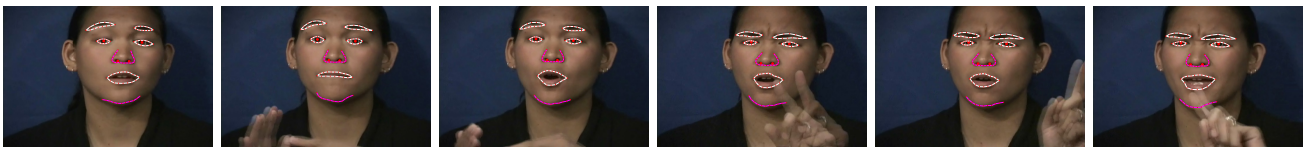

Figure 5. Sample precise detections of faces and facial features using the algorithm of [7].

References

[1] J. M. Beale and F. C. Keil. Categorical effects in the perception of faces. *Cognition*, 57:217–239, 1995. 8

[2] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993. 9

[3] A. J. Calder, A. D. Lawrence, and A. W. Young. Neuropsychology of fear and loathing. *Nature Review Neuroscience*, 2:352–363, 2001. 8

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 11

[5] A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons, New York, 1995. 7

[6] C. Darwin. *The Expression of the emotions in man and animal*. J. Murray., London, 1872. 7, 8

[7] L. Ding and A. M. Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32, 2010. 11, 12

[8] S. Du and A. M. Martinez. The size of facial expressions of emotion. 8, 9

[9] P. Ekman and W. Friesen. *Pictures of Facial Affect*. Consulting Psychologists Press, Palo Alto, CA, 1976. 8, 10, 11

[10] O. C. Hamsici and A. M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30:647–657, 2008. 10

[11] O. C. Hamsici and A. M. Martinez. Rotation invariant kernels and their application to shape analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31, 2009. 10

[12] C. E. Izard. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual Review of Psychology*, 60:1–25, 2009. 8

[13] T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, Japan, 1973. 9

[14] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, (6):1783–1816, 2005. 9

[15] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1983. 9

[16] D. Marr. Early processing of visual information. *Philosophical Transactions of the Royal Society of London*, 275(942):483–519, 1976. 9

[17] A. M. Martinez. Matching expression variant faces. *Vision Research*, 43:1047–1060, 2003. 7, 8

[18] A. M. Martinez and A. C. Kak. Pca versus lda. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001. 9

[19] M. Minsky. *The Society of Mind*. Simon & Schuster, New York, N.Y., 1988. 7

[20] D. Neth and A. M. Martinez. Emotion perception in emotionless face images suggests a norm-based representation. *Journal of Vision*, 9(1):1–11, 2009. 9, 10

[21] D. Neth and A. M. Martinez. A computational shape-based model of anger and sadness justifies a configural representation of faces. *Vision Research*, 50:1693–1711, 2010. 9, 10, 11

[22] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, 2000. 7

[23] E. T. Rolls. A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, 4:161–190, 1990. 8

[24] J. A. Russell. A circumplex model of affect. *J. Personality Social. Psych.*, 39:1161–1178, 1980. 8

[25] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, 2003. 8

[26] K. Schmidt and J. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression. *Yearbook of Physical Anthropology*, 44:3–24, 2001. 7

[27] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Optical Soc. Am. A*, 4:519–524, 1987. 9

[28] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3:71–86, 1991. 9

[29] D. A. Wilbraham, J. C. Christensen, A. M. Martinez, and J. T. Todd. Can low level image differences account for the ability of human observers to discriminate facial identity? *Journal of Vision*, 8(5):1–12, 2008. 9

[30] R. B. Wilbur. Nonmanuals, semantic operators, domain marking, and the solution to two outstanding puzzles in asl. In *Nonmanuals in Sign Languages*. John Benjamins, in press. 7

[31] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002. 11

[32] L. Zebrowitz, M. Kikuchi, and J. Fellous. Are effects of emotion expression on trait impressions mediated by babyfaceness? evidence from connectionist modeling. *Personality and Social Psychology Bulletin*, 33:648–662, 2007. 9