Neural Networks 21 (2008) 1238–1246

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Recognizing emotions expressed by body pose: A biologically inspired neural model

Konrad Schindler [a,*], Luc Van Gool [a,b], Beatrice de Gelder [c]

[a] BIWI, Eidgenössische Technische Hochschule, Zürich, Switzerland
[b] VISICS, Katholieke Universiteit Leuven, Heverlee, Belgium
[c] Cognitive & Affective Neurosciences Lab, Tilburg University, Netherlands

ARTICLE INFO

Article history:
Received 10 August 2007
Accepted 20 May 2008

Keywords:
Body language
Categorical recognition
Emotion recognition
Biologically inspired model

ABSTRACT

Research into the visual perception of human emotion has traditionally focused on the facial expression of emotions. Recently researchers have turned to the more challenging field of *emotional body language*, i.e. emotion expression through body pose and motion. In this work, we approach recognition of basic emotional categories from a computational perspective. In keeping with recent computational models of the visual cortex, we construct a biologically plausible hierarchy of neural detectors, which can discriminate seven basic emotional states from static views of associated body poses. The model is evaluated against human test subjects on a recent set of stimuli manufactured for research on emotional body language.

1. Introduction

The expression and perception of emotions have been studied extensively in psychology and neuroscience (Ekman, 1970, 1993; Frijda, 1986; Tomkins, 1962). A complementary body of work comes from the field of computational neuroscience, where researchers have proposed biologically plausible neural architectures for facial emotion recognition (Dailey, Cotrell, Padgett, & Adolphs, 2002; Fragopanagos & Taylor, 2005; Padgett & Cottrell, 1996). One important result, on which many (but not all, e.g. Ortony and Turner (1990) and Russell (1994)) researchers agree nowadays, is that the perception of emotion is at least to a certain degree *categorical* (Ekman, 1970; Izard, 1992; Kotsoni, de Haan, & Johnson, 2001; Tomkins, 1962), meaning that a perceived expression is assigned to one out of a small set of categories, which are usually termed the "basic" or "primary" emotions (although the precise number and type of basic emotions varies between theories). Categorical perception presupposes a sharp perceptive boundary between categories, rather than a gradual transition. At this boundary, the ability to discriminate between visually similar displays on different sides of the boundary is at its peak, so that stimuli can still be assigned to one of the categories. The most wide-spread definition of basic emotions since the seventies is due to Ekman, and comprises the six categories *anger, disgust, fear, happiness, sadness, surprise*. These seem to be universal across different cultures (Ekman, 1970) — in fact a theoretical motivation for emotion categories goes back to the notion that the same facial muscles are used to display emotions in widely different cultures.

The categorical nature of emotion recognition was established empirically, through carefully designed studies with human observers (Calder, Young, Perrett, Etcoff, & Rowland, 1996; de Gelder, Teunisse, & Benson, 1997; Ekman, 1992). However, there is also a computational argument for this capability: if a suitable set of categories can be found (suitable in the sense that they can be distinguished with the available data), then a categorical decision can be taken quicker and more reliably, because the problem is reduced to a forced choice between few possibilities, and because only those perceptual aspects need to be considered, which discriminate the different categories. In learning-theoretical terminology, categories can be represented by a *discriminative model*, which aims for large classification margins, rather than a *generative model*, which allows a complete description of all their aspects.

Over the last decades, most studies have concentrated on emotional signals in facial expressions. Recently, researchers have also turned to *emotional body language*, i.e. the expression of emotions through human body pose and/or body motion (de Gelder, 2006; Grezes, Pichon, & de Gelder, 2007; Meeren, van Heijnsbergen, & de Gelder, 2005; Peelen & Downing, 2007). An implicit assumption common to the work on emotional body language is that body language is only a different means of expressing the *same set of basic emotions* as facial expressions.

The recognition of whole-body expressions is substantially harder, because the configuration of the human body has more

* Corresponding author. Tel.: +41 44 6326670.
E-mail address: konrads@vision.ee.ethz.ch (K. Schindler).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.05.003

Fig. 1. Example of stimuli from the Tilburg University image set. Emotions are displayed by body language in front of a uniform background. The same emotion may be expressed by different poses.

degrees of freedom than the face alone, and its overall shape varies strongly during articulated motion. However, in computer vision and machine learning research, recent results about object recognition have shown that even for highly variable visual stimuli, quite reliable categorical decisions can be made from dense low-level visual cues (Dalal & Triggs, 2005; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2006).

In this work, we try to gain new insight into possible mechanisms of emotion recognition from body pose, by constructing a biologically plausible computational model for their categorical perception (plausible in terms of the high-level hierarchy, *not* in terms of low-level functionality such as information encoding). We stress that at present the neurophysiological data about the visual cortex is not complete enough for us to fully understand and replicate the underlying processes. Any computational model can therefore only strive not to contradict the available data, but remains in part speculative. Still, we believe that such an approach can be beneficial, both for machine vision, which is still far from reaching the capabilities of animal vision, as well as for neuroscience, where computational considerations can contribute new insights.[1]

We restrict ourselves to the analysis of body poses (form), as opposed to the dynamics of body language (optic flow). This corresponds to modeling only perception and recognition processes typically taking place in the ventral stream (Felleman & van Essen, 1991): we focus on the question, what categorization of single snapshots can contribute to the extraction of emotions from body pose, without including any motion information. Recent studies suggest that there are also recognition processes based on connections to areas outside the ventral stream (STS, pre-motor areas), which presumably explain sensitivity to implied motion (de Gelder, Snyder, Greve, Gerard, & Hadjikhani, 2004) (and also to action properties of objects (Mahon et al., 2007)). For the moment, we exclude these connections, as the corresponding computational mechanisms for extracting and encoding implied motion are not clear.

[1] An example is the MAX-pooling operation, which was postulated by computational models before it was actually observed, see Section 3.

Using a set of *emotional body language* stimuli, which was originally prepared for neuroscientific studies, we show that human observers, as expected, perform very well on this task, and construct a model of the underlying processing stream. The model is then tested on the same stimulus set. By focusing on form, we do not claim that motion processing is not important. The importance of motion and implied motion for the perception of human bodies is corroborated by several neurophysiological studies (Barraclough, Xiao, Oram, & Perrett, 2006; Bruce, Desimone, & Gross, 1981; Jellema & Perrett, 2006; Oram & Perrett, 1994), and we have taken care to keep our computational approach compatible with models, which include the dorsal stream. In particular, our model can be directly extended by adding a motion analysis channel as proposed by Giese and Poggio in their model of action perception (Giese & Poggio, 2003).

2. Stimulus set

The data we use for our study was originally created at Tilburg University for the purpose of studying human reactions to emotional body language with brain imaging methods.

The data consists of photographic still images of 50 actors (34 females, 16 males) enacting different emotions. All images are taken in a frontal position with the figure facing the camera, on a controlled white background. The stimulus set follows the list of six basic emotions originally inventorised by Ekman (1970): per subject 7 poses are recorded, corresponding to the emotional categories *angry, disgusted, fearful, happy, sad, surprised, neutral*, except for two subjects, where the image for *sad* is missing. Examples are shown in Fig. 1. The scale varies less than 25% between subjects, and subjects are centered, with variations of at most $\approx$10% of the image width. Since the images were created for perception studies, the background is homogeneous.

Actors' faces are visible in many images. We have opted to regard the coarse facial expression as part of the body pose, rather than mask out faces. The decision is motivated by the following observation: whether the face is visible correlates with the emotion category (e.g., in *sad* and *fearful* poses the face is often covered by hair or hands, respectively, in *surprised* poses, the

Table 1
Confusion matrix for recognition of 7 emotional categories by human observers

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Neutral
Angry	40	2	2	0	0	2	0
Disgusted	1	39	3	0	0	3	0
Fearful	4	8	44	0	1	2	0
Happy	1	0	0	49	0	0	0
Sad	1	0	1	0	42	1	2
Surprised	2	0	0	0	0	42	0
Neutral	1	1	0	1	5	0	48

Rows are the categories selected by test subjects, columns the "true" categories enacted in the stimuli (so for example, of 50 images depicting *angry*, 40 were classified correctly, 4 were miss-classified by test subjects as *fearful*, while 2 images showing *fearful* were miss-classified as *angry*).

mouth is often covered). Masking only some images, and with non-uniform masks (to avoid masking hand gestures), runs the risk of distorting the experiment, by introducing artificial categorization cues. For our setup, this danger seems bigger than the alternative of showing faces, which contribute only a small fraction of the image information, and, at the presented size (height of face less than 20 pixels) reveal only a coarse expression.

2.1. Emotion and attention

Neuroscientific research indicates that emotional content influences attention allocation by directing the observer to emotionally salient stimuli. This emotional bias on attention can already take place in the early stages of visual processing, even before the emotionally charged stimulus as such has been categorized (Eimer & Holmes, 2002; Pizzagalli, Regard, & Lehmann, 1999) (and even in patients with attentional disorders (Tamietto, Geminiani, Genero, & de Gelder, 2007)), although to which extent processing requires attention is still open to debate (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002). It is not yet clear, whether the so called "fast route" to amygdala, which is thought to be responsible for the first coarse emotion appraisal, runs through the low-level visual cortex, or a parallel sub-cortical route, nor has it been established, how the emotional nature of a stimulus is determined relatively independent of processing in the striate cortex. It does seem clear, however, that the first appraisal mainly prioritizes the emotional content. This is thought to happen either by directing high-level spatial attention (Pessoa, Kastner, & Ungerleider, 2002), or by increasing object saliency in the visual pathway through direct connections from the limbic system (Morris, de Gelder, Weiskrantz, & Dolan, 2001; Öhman, Flykt, & Esteves, 2001; Vuilleumier, Armony, Driver, & Dolan, 2001), or through a combination of both (Taylor & Fragopanagos, 2005). The detailed perception of the emotional stimulus then happens after the relevant cues have been extracted in the visual pathway (Adolphs, 2002).

The topic of the present work is the process, in which the visual cortex extracts cues from the visual input, which is then used to determine an emotion (and respond to it) in temporal and higher cognitive areas (in particular, the orbito-frontal and pre-frontal cortex). We note that in the given stimulus set, the figure dominates the image, and no distractors are present in the background. It is thus implicitly assumed that focal attention is directed to the person, and that figure-ground segmentation happens prior to, or in parallel with, categorization.

2.2. Categorization by human subjects

Enacting and recognizing emotion is not a trivial task even for human observers. To establish a baseline for computational approaches, we therefore conducted a validation study. 14 naive subjects (post-graduates and employees of our department with no involvement in the study of human emotions, 3 females and 11 males) were each shown 25 images from the data set and were asked to perform a forced-choice classification according to the 7 categories. The depicted emotions were selected randomly, but care was taken to present images from 25 different actors, in order to avoid learning biases. This precaution is required in a forced-choice categorization experiment, because previously seen emotions from the same actor rule out certain choices, and allow direct comparison of the two poses, while in real-world circumstances we need to recognize emotions of unknown subjects (the same conditions were created for the computational model, where the same subject could not appear in the training and test set).

The total rate of correct recognitions over all stimuli was 87% (the full confusion matrix is given in Table 1). Not surprisingly, certain emotional poses are quite unique, and allow almost perfect classification, while others are easily confused with each other (e.g. the pairs *disgusted–fearful* and *neutral–sad*). As will be seen later, this behavior is replicated by the computational model.

3. Neural model

Our model of the visual pathway for recognition has been inspired by the one of Riesenhuber and Poggio (1999) and Serre et al. (2006). It consists of a hierarchy of neural feature detectors, which have been engineered to fulfill the computational requirements of recognition, while being consistent with the available electro-physiological data. A schematic of the complete model is depicted in Fig. 2. As an important limitation, the model is purely feed-forward. No information is fed back from higher layers to lower ones. EEG experiments indicate that recognition tasks can indeed be accomplished with such low latencies that feedback from later processing stages and higher cortical areas is unlikely to play a key role (Thorpe, Fize, & Marlot, 1996). We do not claim that cortico-cortical feedback loops do not exist or are not important. Since abundant neural circuitry exists from higher back to lower layers (Salin & Bullier, 1995), it is quite likely that during longer observation information is fed back. Indeed, we cannot exclude at present that the performance gap between human observers and our model may be partly due to the unrestricted presentation times, which allow humans to use feedback for solving difficult cases. Note however that recent electro-physiological data supports rapid bottom up recognition: occipito-temporal vision processes are already sensitive to emotional expressions of face and body images (Meeren et al., 2005; Stekelenburg & de Gelder, 2004).

3.1. Low-level features

The first level of the hierarchy consists of a set of log-Gabor filters with physiologically plausible parameters, to extract local orientation at multiple scales from the input image. Gabor-like filtering is a standard way to approximate the responses of simple cells of Hubel and Wiesel (1962) in area V1 of the primary

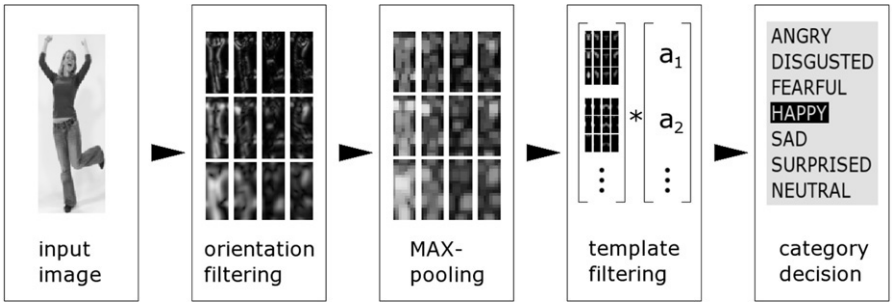


Fig. 2. Illustration of the neural model. From the raw image on the retina, local orientation is extracted (area V1), pooled over spatial neighborhoods (V2/V4), and filtered with learned complex features (V4/I).T The filter response serves as input into a discriminative classifier (IT). *Parameters were chosen for illustration purposes and are different from the actual implementation.*

visual cortex. Specifically, we use log-Gabor filters, which allow a better coverage of the spectrum than the standard (linear) version with fewer preferred frequencies, and are consistent with electro-physiological measurements (Field, 1987). The filter gain is proportional to the frequency, to compensate for the frequency spectrum of natural images, and give all scales equal importance. The magnitude of the log-Gabor filter response is computed at every pixel on the model retina, leading to strongly overlapping receptive fields. Details about the parameter settings used in our simulations are given in Section 4.

The next level consists of neurons with larger receptive fields, which pool the filter responses from spatially adjacent cells, yielding higher position-invariance, as observed for cells in area V2 (Hedgé & van Essen, 2000). Pooling is done with the MAX operator (sometimes referred to as "winner-takes-all"), meaning that the strongest response determines the output of the pooling neuron, separately for each orientation and scale. MAX pooling as a basic operation of the visual pathway has been proposed by Fukushima (1980) and has been strongly advocated and investigated in detail by Riesenhuber and Poggio (1999). It increases the position invariance and robustness to sensor noise, and has been observed electro-physiologically in areas V2/V4 (Gawne & Martin, 2002; Lampl, Ferster, Poggio, & Riesenhuber, 2004). The overlap between neurons in this layer is half the diameter of their receptive fields. The level therefore also yields a substantial data reduction, e.g. pooling with our standard setting of 5×5 pixels reduces the spatial resolution by a factor 2.5, and thus the amount of neurons on the next layer by 84%. In our experiments, the model proved robust against variations of the pooling resolution (see Section 4).

3.2. High-level features

The third level consists of more specific feature detectors sensitive to more complex structures (c.f. the "component-tuned units" of Riesenhuber and Poggio (1999), attributed to areas V4/IT). In our model, these structures are learned through principal component analysis (PCA) of a large set of responses from the previous level. Once learned, each basis vector ("eigen-image") of the reduced PCA-basis represents a detector. By projection onto the PCA-basis, the ensemble of detectors is applied to the incoming signal from the previous level, and the resulting coefficients form the output of the layer.

There are two possible interpretations of this process: the classical argument for PCA and related techniques in models of visual perception is to view it as an optimal compression algorithm, which reduces the amount of data by finding the linear basis, which retains most of the variance in the data (and thus most of the signal) with a fixed number of coefficients. There are multiple ways of learning principal components with a neural network, for example Bourlard and Kamp (1988) and Hinton and Salakhutdinov (2006).

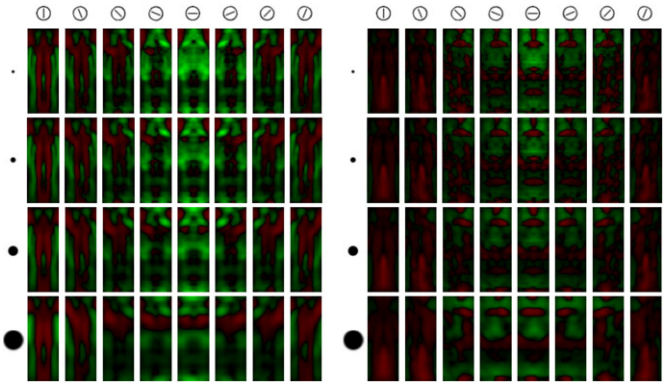


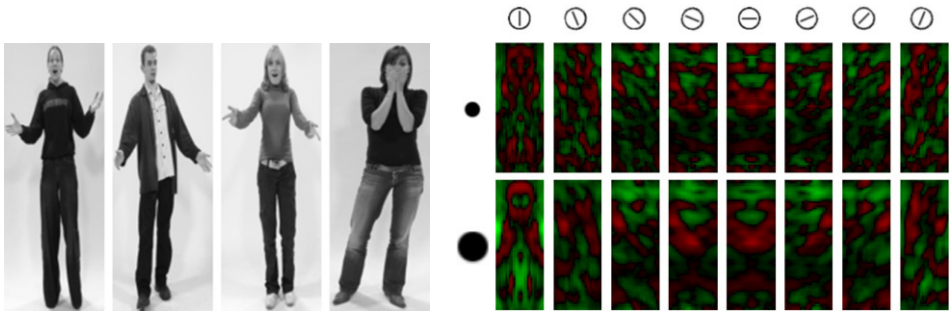
Fig. 3. PCA basis vectors are used as templates for complex features encoding components of the human pose. Shown are the first and second basis vector for the 32-channel filter-bank used in our simulations (4 scales $\times$ 8 orientations). Note how the basis captures different limb configurations. Positive values are printed red, negative values green, brighter means larger values. Symbols to the left and on top of the templates indicate the scale and orientation of the corresponding log-Gabor filter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We feel that a different view of PCA fits our model more naturally: the basis vectors $\{\mathbf{b}_i, i = 1 \ldots N\}$ are directly viewed as templates for relevant visual features — see Fig. 3 for an example of a basis vector. If the incoming signal $\mathbf{s}$ from the previous layer is scaled to have norm 1 (a simple form of normalizing the signal "energy" or "intensity"), then its projection $\langle \mathbf{s}, \mathbf{b}_i \rangle = \cos(\angle_{\mathbf{s}}^{\mathbf{b}_i})$ onto the basis vector can be directly interpreted as a measure of similarity, where 1 means that the two are perfectly equal, and 0 means that they are maximally dissimilar. In this way, the neurons on this layer compare the input to a set of learned "complex feature templates", in a similar way to the S2-detectors of Serre et al. (2006).

3.3. Decision level

At the top level, an emotional category has to be determined. Since the categories are represented by a range of body poses, the task is closely related to pose recognition, a functionality which neurophysiological experiments primarily attribute to area IT (Logothetis, Pauls, & Poggio, 1995). For simplicity, we directly classify into emotion categories — the alternative approach to first classify into a larger set of body poses, and then assign these to different emotions, is equivalent for our purposes.

The output of the previous level is converted to an emotion category with a *support vector machine* (SVM) classifier (Cortes & Vapnik, 1995). Again, once learning the classification has been accomplished, classification amounts to projection onto a template



(a) Example stimuli. (b) Learned weights.

Fig. 4. *Surprised* pose from different actors, and weights assigned to level 3 outputs for classifying *surprised* (only 2 scales are shown). Positive weight (red) is assigned to typical local configurations of the category, negative weight (green) to atypical ones. Note how the classifier emphasizes the arm and hand configurations, which are characteristic for *surprised*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(the normal vector of the separating hyperplane).[2] SVMs have their roots in statistical learning theory, and estimate the separating hyperplane between two categories as the solution of a closed-form optimization problem. This property has advantages for our simulation, because of the limited amount of available training data, but an equivalent classifier can also be learned with neural mechanisms, for example Anguita and Boni (2002). To extend the binary SVM concept to $N > 2$ categories, we use the "one-vs-all" method, i.e. for each category a binary classifier is learned, which discriminates it from all others. The input signal is then fed into all N classifiers, and the category with the strongest output (corresponding to the largest decision margin) is selected. Although there are theoretically more appealing multi-class extensions, the "one-vs-all" method gives comparable results in practice (Schölkopf & Smola, 2002), and has the most obvious biological interpretation: each of the N classifiers models one unit, which is selectively activated for a certain category, and the one with the strongest activation determines the category. These units are similar to the "snapshot neurons" of Giese and Poggio (2003) and to the "key frames" of Beintema and Lappe (2002), and could correspond to neurons tuned to body configurations as observed in Downing, Jiang, Shuman, and Kanwisher (2001), except that we allow assignment of several different poses to the same emotion. An illustrative example is shown in Fig. 4.

4. Experiments

The model has been tested on the stimulus set described in Section 2. All stimuli were used in their original orientation as well as mirrored along the vertical axis, to account for the symmetry of human body poses with respect to the sagittal plane. This gives a total of 696 images (for 2 out of the 50 actors the image for *sad* is missing). As explained earlier, we implicitly assume that attention has been directed to the person, because of the controlled imaging conditions (clean background, uniform scale). We therefore crop and rescale the stimuli to a uniform image size of 200 × 75 pixels for computational convenience. This normalization is common practice in work, which models the perception of faces, e.g. Dailey et al. (2002) and Giese and Leopold (2005).

For each simulation discussed in the following section, the reported result is the average obtained over 10 runs of 10-fold cross-validation: in each of the 10 runs, the data was randomly divided into 10 batches of 5 actors each. The model was trained on 9 of these batches, and then tested on the remaining batch. Splitting by actors ensures that the system learns all emotions equally well, but always has to generalize to unseen actors during training. The entire training was repeated for each run, i.e., the model had to learn the PCA basis as well as the SVM classifier from the training images, and apply them to the unseen test images. Overall, the correct recognition rate of the model is 82%.

4.1. Comparison with human subjects

The model on average miss-classified 5% more stimuli than the human test subjects (see Section 2), achieving 94% recognition rate relative to the human gold standard. A per-category comparison with the performance of human subjects is shown in Fig. 5, and the complete confusion matrix is shown in Table 2. We note the high correlation between the two results: the model predicts well which categories are "easy" ad which are "difficult" to recognize, and confuses the same pairs (e.g. *disgusted* and *fearful*).

4.2. Parameters

A notable property of the presented model is its stability to parameter changes. Results are comparable over a range of settings for the basic parameters (the number of scales and orientations of V1 neurons, the receptive field of pooling neurons, the number of component-tuned units), and vary gradually with changes to these parameters.

Responses of orientation-sensitive cells in the first level are calculated with a bank of log-Gabor filters. The response g at spatial frequency w is

$$g(w) = \frac{1}{\mu} \left\| e^{-\frac{\log(w/\mu)}{2\log\sigma}} \right\|,$$

with μ the preferred frequency of the filter, and σ a constant, which is set to achieve even coverage of the spectrum. Responses are computed at 8 orientations (spacing 22.5°), and 4 preferred frequencies. The high angular resolution is motivated by two sources: fMRI measurements of the human visual cortex have shown that a 22.5° orientation change produces significant changes in the fMRI signal (Tootell et al., 1998), and experiments in machine vision indicate that high angular resolutions around 20° are optimal for object recognition (Dalal & Triggs, 2005). To keep computations tractable, we use 4 preferred frequencies (8, 4, 2, 1 cycles per degree).

The receptive field size of second level neurons is chosen 5 times larger than those of the first level, in line with fMRI data from humans, and electro-physiological data from monkeys (Smith, Singh, Williams, & Greenlee, 2001). The response of each neuron is simply the maximum over the inputs from its 5 × 5 afferents:

[2] We use a linear SVM. Although a non-linear transformation is quite plausible in a biological neural architecture, standard kernels did not improve the results in our experiments.

Fig. 5. Classification performance of the model compared to human subjects.

Table 2
Confusion matrix for recognition of 7 emotional categories by the computational model (average over 10 iterations)

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Neutral
Angry	35	4	1	1	2	4	0
Disgusted	6	34	6	0	1	2	0
Fearful	2	5	41	0	0	0	0
Happy	2	3	1	47	0	0	0
Sad	3	1	0	0	40	0	3
Surprised	2	3	2	2	0	42	1
Neutral	1	0	0	0	5	1	46

Rows are the categories selected by test subjects, columns the "true" categories enacted in the stimuli. Compare to Table 1.

Fig. 6. Performance with varying receptive field of pooling neurons. Results are comparable over a range of receptive field sizes, and decline gracefully. Note the logarithmic scaling along the x-axis.

$$r_i = \max_{(x,y)\in \mathsf{G}_i} g(x, y),$$

where (x, y) is the spatial position of the afferent, and G_i denotes the receptive field of the ith neuron. Note that the exact choice of receptive field is not critical: the model gives comparable results over a range of values, see Fig. 6.[3]

[3] The good performance with no or very little increase in receptive field may be a bias due to the uniform background of the stimulus set. In the presence of background clutter, the critical responses along the actors' contours would be distorted.

For the third level, the local responses for each training stimulus over all orientations and frequencies are concatenated to form a response vector, and principal component analysis is performed on the set of vectors, to learn a bank $\{\mathbf{b}_i, i = 1 \ldots D\}$ of complex detectors (the first D principal components). When a new stimulus is presented, the response vector $\mathbf{r}$ from the previous layer is normalized, and then compared to each detector, to yield an output signal v_i:

$$v_i = \left\langle \frac{\mathbf{r}}{|\mathbf{r}|}, \mathbf{b}_i \right\rangle.$$