Identifying Emotions Using Topographic Conditioning Maps

Athanasios Pavlou and Matthew Casey

Department of Computing, University of Surrey,
Guildford, Surrey, GU2 7XH, UK
{a.pavlou,m.casey}@surrey.ac.uk
http://www.cs.surrey.ac.uk

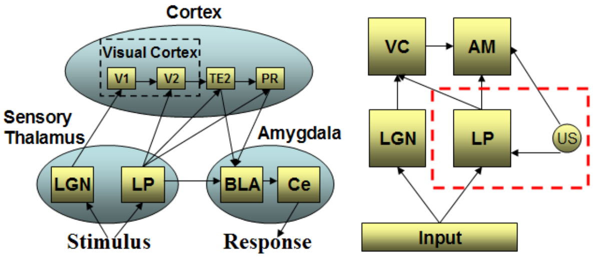
Abstract. The amygdala is the neural structure that acts as an evaluator of potentially threatening stimuli. We present a biologically plausible model of the visual fear conditioning pathways leading to the amygdala, using a topographic conditioning map (TCM). To evaluate the model, we first use abstract stimuli to understand its ability to form topographic representations, and subsequently to condition on arbitrary stimuli. We then present results on facial emotion recognition using the sub-cortical pathway of the model. Compared to other emotion classification approaches, our model performs well, but does not have the need to pre-specify features. This generic ability to organise visual stimuli is enhanced through conditioning, which also improves classification performance. Our approach demonstrates that a biologically motivated model can be applied to real-world tasks, while allowing us to explore biological hypotheses.

1 Introduction

Emotions are an essential survival tool. The amygdala is the critical neural structure that acts as an evaluator of potentially threatening stimuli, priming our bodies for action [1]. As such, the amygdala is an attractive area to study because simulating such emotions may be an effective way for an artificial system to interact with humans, adjusting its responses to different events. One way in which such emotions can be modelled is through studying the elicitation mechanisms of the brain at the anatomical and behavioural levels. Fear was the first emotion that allowed such studies to be conducted, because of its evolutionary significance for survival. In addition, fear is easy to elicit or even artificially create (classical conditioning) [2]. This occurs when repeatedly pairing a neutral stimulus, the conditioned stimulus (CS), with an unconditioned stimulus (US), typically a loud burst of noise or electric shock [3]. The neural pathways of this process have been extensively studied for the auditory modality of rats [3]. However, several aspects such as their role for processing a wider spectrum of emotions [4] as well as anatomical interconnectivity of participating structures [5] are still under investigation for the visual modality. In particular, a dual neural pathway has been identified through which visual stimuli flow through thalamic areas and feed directly (from the lateral posterior nucleus) and indirectly (though the lateral geniculate nucleus and visual cortices) to the amygdala [1].

M. Köppen et al. (Eds.): ICONIP 2008, Part I, LNCS 5506, pp. 40–47, 2009.

This study presents a model of visual fear conditioning that explores the dual cortical and sub-cortical visual pathways leading to the amygdala (Fig. 1). This model captures the basic properties of the participating structures, and is the first that can condition on complex visual stimuli by extending the work of Armony et al [6] to the visual domain. We first evaluate the capability of our extended algorithm on processing abstract stimuli, and then we test the efficiency of the model's sub-cortical pathway to categorize the emotions expressed in face images. By these enhancements we aim to examine the degree to which a biologically constrained architecture can tackle real-world problems and gain insight into how such emotional detection takes place at early stages of sensory processing in the brain. The architecture of our model enables us not only to use it as an emotion detector but also, by applying conditioning, we can embed significance evaluation (learning higher responses when a particular stimuli is presented). In contrast with other techniques (see [7] for a review on facial emotion detection) our architecture does not rely on any manual feature extraction or particularly imposed geometric relations, suggesting that it can be used on a generic class of visual inputs. In addition, we aim to provide a more detailed framework that can assist and verify results of current neurobiological experiments of this field.



a) b)

Fig. 1. Visual fear conditioning pathways [1] a) lateral geniculate nucleus (LGN), lateral posterior nucleus (LP), basolateral amygdala complex (BLA), central amygdaloid nucleus (Ce), primary (V1) and secondary (V2) visual cortices, temporal cortical areas (TE2), and perirhinal cortex (PR), b) model schematic with LGN, LP, early visual cortices (VC) and the amygdala (AM). The area within the dashed line denotes the part of the model used for the evaluation using real emotional expressions.

2 Method

We represent the anatomical structures participating in visual fear conditioning using a series of feedforward neural modules that are trained competitively. Armony et al [6] modelled auditory fear conditioning using one-dimensional modules. To successfully model visual pathways, the topographic relationship between regions within a single stimulus is crucial. Our work extends Armony et al's to have two-dimensional topographic maps. We introduce the notion of lateral inhibition using a neighbourhood as formulated by Kohonen [8]. Conditioning on these maps can then occur via an additional input (equivalent to the US), which is active for the CS, affecting the plasticity of all neurons in the map. This combination of Hebbian learning in a topographic map is similar to the Laterally Interconnected Synergetically Self-Organizing Map (LISSOM) technique developed from studies on V1, which has successfully been used to model selectivity in the visual cortex [9]. Whereas such techniques offer more plausible models of receptive fields, our model builds on the key strength of Armony et al's approach in that it allows us to condition on arbitrary inputs, while still maintaining the ability to successfully model different layered brain structures at a sufficient level of detail (cf. [6]).

Fig. 1b) shows a schematic of our model of visual fear conditioning. The cortical pathway is represented as connections from the LGN and LP to the VC, which then feed the AM. The sub-cortical pathway feeds the output of the LP directly to the AM. The visual stimulus is input to the LGN and LP and the US is input to the LP and AM, taking a value of 1 when conditioning, and 0 at all other times. This US always has a fixed weight value, which for us is 0.7. Each module consists of a lattice of neurons that are fully connected to the input, such that a neuron (i, j) has an output y corresponding to an m-dimensional input x:

$$u_{ij} = \sum_{k=1}^{m} x_k w_{kij}(t), \tag{1}$$

$$y_{ij} = \begin{cases} f(u_{ij}) & \text{if } \|c_{ij} - c_{win}\| < h(t) \\ f(u_{ij} - y_{win}) & \text{otherwise} \end{cases}, \tag{2}$$

$$f(u) = \begin{cases} 1 & u \geq 1 \\ u & 0 < u < 1 \\ 0 & u \leq 0 \end{cases}, \tag{3}$$

where $w_{kij}(t)$ is the weight from input k for neuron (i, j) in the lattice at time step $t \geq 0$, initialized with uniformly distributed small random values. Note in equation 2 that a neuron is considered to be in the winner area if the distance from the neuron (i, j) to the winner in the lattice is less than the current radius value $h(t)$. Here we use c_{ij} and c_{win} to denote the lattice co-ordinates of the two neurons. All neurons outside this area are inhibited by the activation value of the winning neuron $y_{win} = max_{ij} f(u_{ij})$.

Competitive learning is achieved by updating each weight, except those fixed for the US, and then normalizing all weights to prevent exponential growth:

$$w'_{kij}(t+1) = w_{kij}(t) + \epsilon(t) x_k y_{ij}, \tag{4}$$

$$w_{kij}(t+1) = \frac{w'_{kij}(t+1)}{\sum_{l=1}^{m} w'_{lij}(t+1)}, \tag{5}$$

where $\epsilon(t)$ is the learning rate at time step t, corresponding to the presentation of a single input. This differs slightly from Armony et al's [6] formulation in that all weights are updated, not just those that have an input that is above average.

3 Experiments and Evaluation

The experiments presented in this section aim to evaluate the model's capabilities in terms of biological relevance and then application to more complex tasks. By the use of abstract stimuli we can obtain a clear insight into the model's ability to form topographic representations, as well as its parameter requirements. By then focusing on recognising emotions in faces, we can examine the model's potential for more challenging tasks, especially those already associated with the amygdala [4].

3.1 Experiments on Abstract Stimuli

The first phase of the evaluation aims to determine if the model can correctly form topographic maps. Here we use overlapping patterns of Gaussian activation in varying locations to represent visual stimuli. We use 266 patterns as the training examples corresponding to a series of spatial locations within an input representing azimuth [-90, 90] and elevation [-65, 65] (similar to a human's visual field). Within this, we allow the positioning of an object at a discrete interval of 10, so that we can encode 19 different positions for azimuth and 14 for elevation. For a stimulus at azimuth p and elevation q, we have an input x as:

$$x_{pq} = \lambda e^{-(\frac{p^2-q^2}{\sigma^2})}, \tag{6}$$

where λ is the maximum amplitude and σ the radius, chosen as $\lambda = 1$ and $\sigma = 10$ for these experiments.

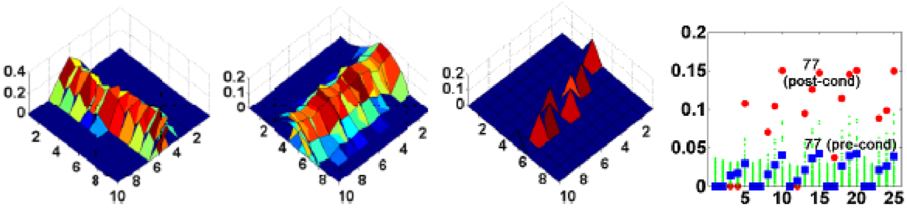
Training takes place by layers so that the input feeds into the LGN and LP first. After training is finished these modules' outputs are used to train the VC. In turn, when the VC has finished training, then its combined outputs with the LP train the AM in a similar fashion (Fig. 1b). The LGN, LP and VC are represented by a 10 by 10, and the AM by a 5 by 5 neuron lattice. The smaller size of the AM reflects biology [1]. The chosen sizes of the maps provide detailed representation of the inputs while remaining computationally efficient. All 266 inputs were presented during each epoch in uniformly random order. A decreasing per epoch Gaussian neighbourhood radius function and an exponential learning rate function were used:

$$h(t) = r_{min} + (r_{max} - r_{min})e^{-(\frac{(t/t_e)^2}{2r_s^2})}, \tag{7}$$

$$\epsilon(t) = l_{min} + (l_{max} - l_{min})e^{-(\frac{(t/t_e)}{2l_s^2})}, \tag{8}$$

where r_{min} and r_{max} are the minimum and maximum radius for the neighbourhood and r_s is the bandwidth; similarly, l_{min}, l_{max} and l_s for the learning rate. The values only vary per epoch, thus t_e defines the number of time steps per epoch (266). Through trial and error the chosen parameter values produced stabilized map organization over the training period. All modules used the same set of parameters differing on the minimum neighbourhood radius. Here, r_{max}

was equal to the lattice width (10 or 5), r_{min} was 2 for the LGN, 3 for the LP, 1 for the VC and 4 for the, AM corresponding to the differing visual coarseness in each module, while $r_s = 300$, $l_{max} = 0.1$, $l_{min} = 0.001$ and $l_s = 13$. A stable topographic organization was achieved after 700 epochs of training (Fig. 2).



a) LGN b) LP c) VC d) AM

Fig. 2. Topographic organization using stimuli with azimuth 0 and elevation in the range [-65 65] for the a) LGN ($r_{min} = 2$), b) LP ($r_{min} = 3$), and c) VC ($r_{min} = 1$). Observe that the LGN and LP have different topographic orientations for the same locations, but these have successfully been aligned in the VC. Neuron activations in the AM pre- and post-conditioning (squares, circles) are shown in d) in response to CS 77, together with responses from the rest of the inputs (dots).

For the second phase we evaluate the effect of conditioning on the maps, selecting a single stimulus as the CS. Training differs from the pre-conditioning phase in that now all the layers are trained concurrently. We note that the map radii have already reached their minimum values, however the learning rate continues to drop after each epoch (continuing from the value it had on the last epoch of pre-conditioning). This time training occurred for 530 epochs until the map activation patterns were again stable.

The results from the model's pre-conditioned phase indicate successful map organization with different specificity of input representation on each map depending on the minimum neighbourhood radii used (more specific representation of inputs on the VC compared to the rest). A large radius for the AM was selected to ensure that no topographic organization will occur since the amygdala is not known to have such capability. As we see from Fig. 2, the post-conditioning activations of stimulus 77 on the AM have significantly increased compared to their pre-conditioning values, as have the activations corresponding to the locations surrounding 77. Overall, these results show both that the maps correctly organise the stimuli by similarity in the input (location), while allowing us to condition a particular output to gain a higher activation, as per Armony et al's results on one-dimensional modules [6].

3.2 Experiments on Emotional Expressions in Face Images

Having established the model's capability of handling abstract two dimensional input we proceed to examine its efficiency on the real-world task of emotion recognition using real face images taken from the MMI Facial Expression

Database collected by Pantic & Valstar [10]. The experiments are restricted to the sub-cortical pathway (LP module) since this is adequate for observing map organisation as well as conditioning effects, and allows us to explore the efficacy of this 'coarse' visual processing pathway computationally.

The same method of training (pre- and post-conditioning phases) was followed as before. This trial differs from the previous experiment in that the map size is now scaled up to a 32 by 32 lattice to correspond with an increased input size, while all other parameters remain the same. For training we use 598 frames of a single person (training subject) taken from the MMI database. The frames are taken from video files containing transitions from a neutral state to a gradually increasing emotional expression (anger, disgust, fear, happiness, sadness and surprise) that reaches a peak (peak frame) and from that returning to the original (neutral) expression. Fig. 3 shows the map activations for the training data for the peak frame of each emotion. The only pre-processing of the images was cropping to have approximate upper and lower boundaries the eyebrows and chin, resizing this to 32 by 32 pixels and finally grey-scaling them. For testing we used frames from videos of two other subjects from the MMI database. From the first subject (S002) we acquired frames depicting gradations of anger, happiness and sadness whilst from the second subject (S031) frames depicting gradations of disgust, fear and surprise. In total we used 121 testing frames taking 21 consecutive frames for each emotion that include the peak emotion frame.

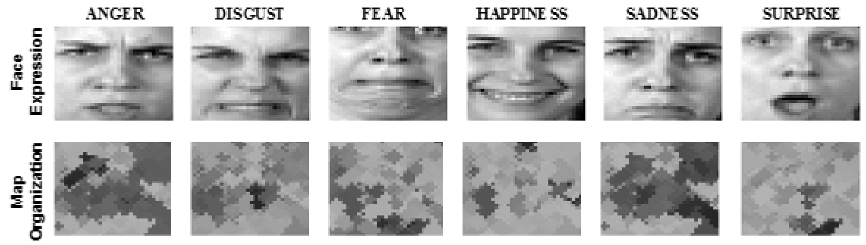


Fig. 3. Example trianing inputs for each emotion (peak video frame) and corresponding map activations (pre-conditioning). Higher activations are denoted by darker shading. Note that the activated areas are distinct and differ for each emotion.

To label the map on the training data, for each peak frame for the six emotions we recorded each neuron's activation. Since each emotion produced a cluster of highly active neurons (Fig. 3), we labelled progressively larger clusters to determine the ability of the map to have the same pattern of activity during testing: the top 0% (just the most active neuron), 10%, 20% and 30%. To obtain a class for a test image, we determined the number of neurons that overlapped in these areas to those that had been lablled for each emotion, picking the label with the majority of overlap. The top 20% of active neurons gave the best results, as shown in Table 1. Classification accuracy is 76% or more for fear, happiness, sadness and surprise, whereas anger and disgust are misclassified (as sadness, and as anger, disgust and sadness, respectively), with an overall accuracy of 62%.

Table 1. Confusion matrix of the number of successfully classified frames for the top 20% of active neurons, shown for pre- and post-conditioning (on the preak frame of the angry face). The training peak frames were compared to 21 testing frames per emotion. Unclassified patterns are shown as NK.

	Pre-conditioning								Post-conditioning							
	An	Di	Fe	Ha	Sa	Su	NK	Accuracy	An	Di	Fe	Ha	Sa	Su	NK	Accuracy
Anger (S002)	**0**	0	0	0	21	0	0	0%	**21**	0	0	0	0	0	0	100%
Disgust (S031)	3	**3**	0	0	12	0	3	14%	3	**4**	0	0	8	0	6	19%
Fear (S031)	0	0	**16**	0	0	0	5	76%	0	0	**21**	0	0	0	0	100%
Happiness (S002)	0	0	0	**19**	2	0	0	90%	0	0	0	**21**	0	0	0	100%
Sadness (S002)	0	0	0	0	**21**	0	0	100%	0	0	0	0	**21**	0	0	100%
Surprise (S031)	0	0	2	0	0	**19**	0	90%	0	0	12	0	0	**9**	0	43%

This performance can be attributed to the similarity of the training peak frames for anger, disgust and sadness (Fig. 3). While these results do not approach the performance of feature-based techniques (overall 89% [11]), they at least show that the model is capable of distinguishing between facial expressions. Our biologically motivated question is whether conditioning on an arbitrary stimulus can enhance this recognition rate?

To understand this, we chose to condition the model on the peak frame of the anger examples (CS 55). After conditioning, the map activations both of training and testing subjects for anger were increased, while activation levels for the rest of the emotions remained similar. This matches the results seen on the abstract stimuli and shows that the model is capable of behaving in the same way on more complex input. Further, conditioning affected classification performance (Table 1), with anger and disgust improved (100% and 19%), while the performance for surprise decreased (43%). This follows from the localised increase in activations in the map, where anger, disgust and surprise are similarly located. Conditioning provokes an increased level of activation from the neurons most active for the CS, and because of the neighbourhood, a similar increase in surrounding neurons. As a result, the CS activations are separated from those that they were previosuly close. Overall the model achieved an accuracy of 77%. This is comparable with feature-based techniques such as [11] which for the emotions of happiness (joy), surprise, disgust, anger, sadness and fear achieved 100%, 63%, 100%, 89%, 95% and 89% respectively, overall 89%, albeit using a larger test sample with 62 subjects (1500 frames). Here, we show that a non-feature based technique can gain comparable results just through a process of conditioning on topographic maps, motivated from biology and without any hard-coded feature extraction.

4 Conclusion

In this paper, we presented a topographic conditioning map (TCM) model of the cortical and sub-cortical visual pathways leading to the amygdala. We evaluated the model on abstract stimuli to determine whether it could learn topographic relationships, and how conditioning on an arbitrary input affected the responses. Our motivation is whether taking inspiration from biology can help both neuroscience through developing an understanding of the different pathways computationally, while providing new techniques to tackle complex problems, particularly those associated with vision which offer significant challenges. While in this paper we have not directly addressed the former except to provide a model with which we can explore neuroscientific hypotheses in the future, for the latter we have demonstrated that the TCM is capable of successfully recognising emotional facial expressions. Our model differs from other work in this area [7,11] in that none of the features of each emotional class were chosen manually. Rather, the model developed distinct activation clusters for each of the six emotions it was trained on during pre-conditioning. Conditioning the model then enhanced the map activation levels to improve classification performance for a selected emotion. Future work involves testing on a larger number of stimuli, and to extend the model to tackle the important challenge of object invariance (cf. [12]).

References

1. Morris, J.S., Öhman, A., Dolan, R.J.: A subcortical pathway to the right amygdala mediating “unseen” fear. Proceedings of the National Academy of Sciences 96, 1680–1685 (1999)
2. Pavlov, I.P.: Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxford University Press, London (1927)
3. LeDoux, J.E.: Emotion, memory and the brain. Scientific American Special Edition 12(1), 62–71 (2002)
4. Murrey, E.A.: The amygdala, reward and emotion. Trends in Cognitive Sciences 11(11), 489–497 (2007)
5. Pessoa, L.: To what extent are emotional visual stimuli processed without attention and awareness. Current Opinion in Neurobiology 15, 188–196 (2005)
6. Armony, J.L., Servan-Schreiber, D., Romanski, L.M., Cohen, J.D., LeDoux, J.E.: Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats. Cerebral Cortex 7(2), 157–165 (1997)
7. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
8. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69 (1982)
9. Miikkulainen, R., Bednar, J.A., Choe, Y., Sirosh, J.: Computational Maps in the Visual Cortex. Springer Science+Business Media, New York (2005)
10. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: IEEE International Conference on Multimedia and Expo. (ICME 2005) (July 2005)
11. Hupont, I., Cerezo, E., Baldassarri, S.: Facial emotional classifier for natural interaction. Electronic Letters on Computer Vision and Image Analysis 7(4), 1–12 (2008)
12. Stringer, S.M., Perry, G., Rolls, E.T., Proske, J.H.: Learning invariant object recognition in the visual system with continuous transformations. Biological Cybernetics 94, 128–142 (2006)