Towards a Neural-Based Theory of Emotional Dispositions

J. G. TAYLOR, W. A. FELLENZ, R. COWIE and E. DOUGLAS-COWIE

Department of Mathematics, University of London, King’s College, Strand, WCR2, UK
School of Psychology, Queen’s University, Belfast, UK

Abstract: We present underpinning neural structures to represent various components of emotional dispositions. After a description of simple models of the dynamics, an even simpler multi-layer perceptron model is presented with three outputs. This leads to a high level of success in recognition of emotional dispositions from a database of faces, as well as an interpretation in terms of the underlying neural system. By extending the discrete classification approach to continuous variables in a three-dimensional state-space the recognition performance can be improved and the causes for classification errors can be studied.

Key-Words: - facial analysis, emotion recognition IMACS/IEEE CSCC'99 Proceedings, Pages:5341-5346

1 Introduction

Recent developments in the analysis of emotions have side-stepped the old and unresolved problem of delineating primary from secondary or higher emotions by turning to the social uses to which the emotions are put. These are in terms of the dispositions for action that the emotions arouse. For example the emotional epithet ‘exasperated’ would lead, with high probability, to withdrawal from ongoing actions and with a lower probability to destructive actions. This is the approach of ‘social constructivism’ [1]. One of its effects is to shift the priorities of brain-based theories of emotions. Modelling the conscious experience associated with emotions is a thorny problem, and it remains so. However, social constructivism suggests that modelling dispositions is an equally important part of the task, and it is much more tractable.

In this paper we explore the possiblity of developing a neurally-based theory of emotional dispositions. We consider the main neural substructures in the brain which lead to dispositions to act from an emotional origin, and from that construct a simple neural model for emotional dispositions. The model can be trained using a data-base of realistic emotion-based dispositions arising from facial or speech images.

The crucial circuits involved with emotionally-based dispositions to act are those at the basis of motivation. This is known to be the limbic circuitry, involving the hippocampus and amygdala as well as the dopamine sources of motivation (ventral tegmental area or VTA) and the output ‘gate’ from the limbic system composed of the nucleus accumbens. There are also crucial contributions from prefrontal sites as well as posterior cortical regions. It is this system of nuclei that will be modelled by simple neural modules to attempt to explore the neural basis of emotional dispositions in the brain.

In more detail, the basic nuclei in the brain involved in the production of emotional dispositions are:

- the hypothalamus (HYP): to control autonomic and endocrine responses, and to gate inputs in terms of the internal state of the system;
- the amygdala (AMYG): to learn the salience of inputs, both positive and negative;
- the ventral tegmental area (VTA): to produce dopamine as a signal of a rewarding input;
- the prefrontal cortex (PFC): to encode novel inputs, and excite the VTA to broadcast relevant reward;
- the hippocampus (HC) to gate inputs from the amygdala as to the salience of an input, in terms of memory of past encounter;
- the nucleus accumbens (NACC): the outlflow of motivation, from the limbic circuitry (amygdala, hippocampus) to action orchestrated by the basal ganglia and brainstem motor centres.
- the anterior cingulate (ACG): as an overall executive controller of actions being taken.

There are numerous other sites also concerned with production of emotional dispositions, but the above are the basic circuits [2,3]. The overall connectivity of these sites (excluding the anterior cingulate) is indicated in figure 1.

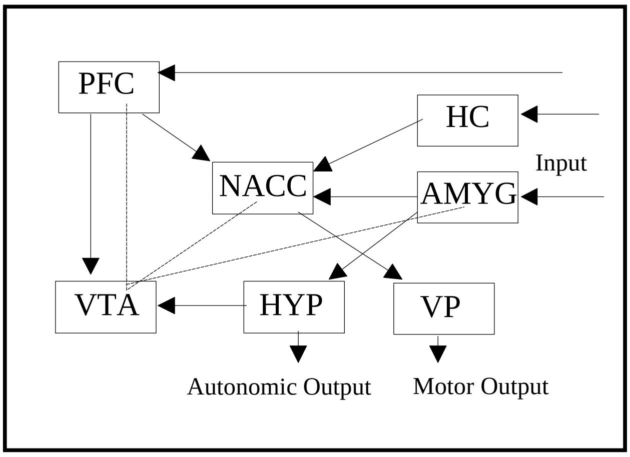

Figure 1. The connectivity between the main sites involved with the creation of dispositional states in the brain. PFC = prefrontal cortex; HC = hippocampus; NACC = nucleus accumbens; AMYG = amygdala; VTA = ventral tegmental area; VP = ventral pallidum; HYP = hypothalamus. The dashed lines in the figure denote dopamine modulation from the VTA used in control and learning of synapses at the projection sites.

The accompanying paper [4] indicates that a set of primitive emotion words (amused, angry, afraid, affectionate, bored, confident, content, disappointed, excited, happy, interested, loving, pleased, relaxed, sad, worried) is obtained from studying a group of subjects, who can rate these words on a two-dimensional activation-evaluation plane for a continuous range of levels. We add to these two dimensions the third one of action (action is described as positive if it is directed towards the object of emotion). This third dimension allows separation of emotion descriptors, such as fear and anger, poorly separated in the activation-evaluation plane (see figures 3 and 4 of Cowie et al, 1999 [4]). Thus we wish to understand how the neural system of figure 1 can produce activations along three different dimensions which can be recognised as activation, action and evaluation.

1.1 Outlook

Having described the known neural circuitry in the brain, and the problem we need it to solve of giving outputs recognisable in three dimensions, we propose in the next section an identification process for the outputs of the system of figure 1. Following that in section 3 we give a neural implementation of the architecture of figure 1. This is then trained to give the three-dimensional dispositional state outputs for the faces of a simple data-base. The paper concludes with a discussion.

2 The Identification of Emotional Dispositions

The problem we face first in this program is to identify which outputs of the overall system of figure 1 correspond to which of the three basic emotional dimensions. We introduced these in the previous section as:

- activation
- evaluation
- action

These were the three dimensions which were recognised from the study of Cowie et al [4] to be important determinants or descriptors of emotional words used by subjects studying them. The first two of these dimensions were used in the first phase of the study [4]; the third was used in the study of emotional schema employed by the same subjects when assessing the same emotional words. There are other dimension of relevance that were employed in the study, but action direction was found to be a useful third dimension to separate otherwise indistinguishable words, such as anger and fear. We now turn to possible neural identification of each of the above emotion dimensions in turn.

Let us first consider evaluation. That can be assessed immediately by means of overall activities of neurons in all of the sites of figure 1. Such an assessment would, however, not take account of known results from brain imaging that indicate the amygdala is the main site of activation for giving the emotional salience of a particular input (be it external or imagined). This is supported by the well-known Kluver-Bucy syndrome in subjects with loss of amygdala. For them affect is lost and all inputs have similar value. Monkeys without amygdala are not frightened by the appear-ance of their keeper or by a snake, an otherwise fearful object. Thus we initially propose to take the net output of the amygdala as giving the level of evaluation for the input, in other words its salience.

A result now available from brain imaging is that the amygdala can be activated for inputs which possess either positive or negative salience to a given subject. Thus while the amygdala level of activation gives an overall absolute value of activation, as determined by its feedback to cortical sites leading to conscious experience, there must also be a signature factor determining whether or not this salience is positive or negative. This signature is taken to be the value of the difference between the amygdala ongoing activity and its mean level.

We employ the autonomic system response to give the level of activation, the second of the three emotional dimensions in our analysis. This total autonomic system response is highly complex. To simplify we will solely take the output of the hypothalamus as this activation level of the corresponding emotion.

Finally we turn to consider the action response in the emotional situation. This is given by the overall output of the system for action. We take this to be the overall output of the VP, which is the controlling system for motor responses. Again there is a signature to be attached to this action response, be it either towards the object under emotional evaluation or away from it. We take this action signature to have the same sign as that arising in the evaluation of the amygdala signal. Such an identification has a strong prediction: that all emotion words will reside in the upper right-hand or lower left hand quadrants of the action-evaluation plane (ignoring the other dimension of activation level). This is satisfied by the results from the Cowie et al study [4] with the eight emotional words worried, afraid, angry, sad, happy, loving, exited and interested being only in these two quadrants.

We have now given a complete description of the manner in which the outputs from the neural system of figure 1 are to be related to the three dimensions of emotion words, and the related underpinning by a three-dimensional space of emotions. It is now the appropriate time to turn to the possible neural implementation of this identification given above of the neural system of figure 1.

3 Neural Implementation

3.1 General Neural Systems

It is clearly not feasable to attempt at this stage to implement all of the modules in figure 1. Before we turn to stripping the model down to its barest essentials, let us consider what is presently available in terms of known neural network architectures. We already have several models of the hippocampus in terms of quite different principles. One class of models is based on the attractor net storage of information [5,6]; this regards the CA3 cell field in the hippocampus as having enough lateral connectivity to qualify as such a network. Another class uses randomly generated sparse subsets in a suitable hidden layer, such as CA3, to encode inputs by changing efferent and afferent weights, and gives a considerably higher retrieval capacity, especially for degraded inputs [7]. The amygdala has connectivity allowing it to be treated as a feed-forward network [8]. The prefrontal cortex, the nucleus accumbens and ventral pallidum, form a recurrent system of modules which constitute the so-called ACTION network which can learn response sequences [9,10]. There are various approaches to modelling the dopamine system using TD-learning. Thus altogether there would appear to be neural models of all of the modules present in figure 1 except for the hypothalamus. That will be modelled in terms of distance from a set point, and a signal given proportional to the distance of the system from that set point value, such as for temperature or water.

3.2 The '2-Weights Model' of DA Learning in VTA

Not all of the above models are appropriate. Thus we wish to model the dopamine output from VTA in figure 1 to act as a reward system. However it has to have novelty value from the PFC, and needs also the gating action of the hippocampus on the output of the amygadala to the NACC, since otherwise response would occur to novel inputs with no, as yet known, salience value. To prevent that we use the fact that the AMYG output to NACC is gated by HC. Such gating prevents a novel input (giving no output from HC) from being able to activate NACC incorrectly. Only after learning of the reward value, arising from HYP to VTA, will the reward value of the input be learnt by VTA neurons and thence allowing the learning by AMYG of the salience of the input. At the same time we note that the AMYG has direct output to the hypothalamus, corresponding to the primitive response patterns learnt without cortical support.

Given the above structure, the crucial component is that involving dopamine. We model that by the simple model of figure 2, which incorporates several features of an earlier model [11]. The activation and training rules are:

$$y=\tanh|ws-r| \quad (1)$$

$$z=vs \quad (2)$$

$$dw/dt=ys-w \quad (3)$$

$$dv/dt=f(y)s-v \quad (4)$$

where f is a function of its variable which is negative for small values and then turns positive to become a positive constant for large values of its variable.

This behaviour arises from experiemental analysis of the modulatory effects of dopamine on learning: LTD arises for low dopamine levels but LTP occurs when dopamine levels become strong enough. The results of this set of activation/learning rules are consistent with observations made in various parts of the limbic system.

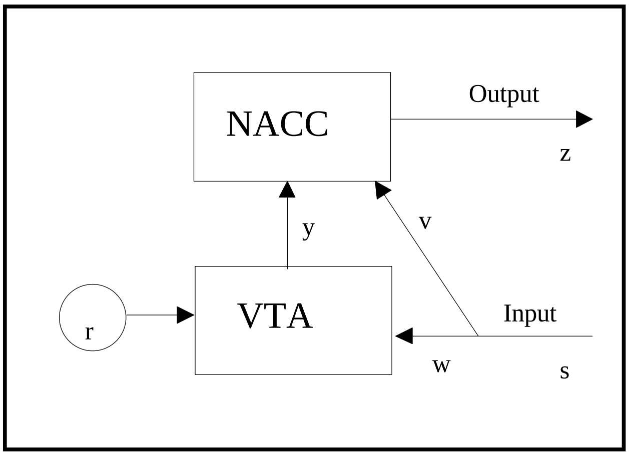

Figure 2. Model of dopamine (DA) learning of the reward value of an input by the VTA, using the ‘2 weights model’ of Krekelberg and Taylor [11]. The primary reward r is input from the hypothalamus, and is a measure of the intrinisic value of reward obtained by consummatory actions made in association with the input. The weights v and w are trained by learning rules to learn the reward value r of the relevant input.

In order to apply the identification of the three dimensions of emotions we developed in the previous section to the model of figure 2 we must augment it by three modules: the VP, HYP and Amygdala. The last of these we take to be an MLP, as noted earlier; the first and second we simplify to single nodes.

3.3 The Extended Model of Salience/ Evaluation Learning

We now have to extend the 2-weights model so as to allow the salience of the input to be learnt by the amygdala. This is achieved, in terms of figure 1, by learning the input to the amygdala, with learning rate the response of VTA. The amygdala output to NACC has two components: one passing through the VTA, and giving the reward value of any input. The other, not through VTA, is trained by using the DA output from VTA, as previously it was for figure 2. We do not include the VP, but take its output to be given by that of the NACC. The resulting network system is shown in figure 3.

The net system of equations is now:

$$y=\tanh|ws-r| \quad (5)$$

$$z=vs \quad (6)$$

$$dw/dt=ys-w \quad (7)$$

$$dv/dt=f(y)s-v \quad (8)$$

$$du/dt=f(y)s-u \quad (9)$$

$$A=us \quad (10)$$

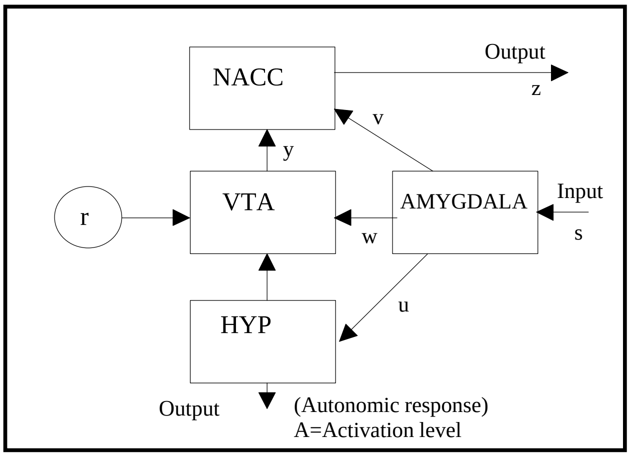

Figure 3. The augmented network of figure 2, with inclusion of the amygdala and hypothalamus. There is an overall weight u for inputs to the amygdala, also trained by us of DA from the VTA.

Again the results of this system of equations will provide a model of modulated learning in which rewards are transferred from the primary reward value r to the VTA dopamine output, in a similar manner to the system of figure 2; we do not repeat the details.

3.4 Dispositions

We finally come to the output side of the motivational circuit of figure 1. This arises from the VP as a control system acting on the motor decision systems, firstly involving the prefrontal cortex, and then moving dorsally to the motor cortex and motor components of the thalamus. At the same time there are dispositional states arising from the control by hypothalamus, and related sub-cortical sites of the autonomic and endocrine system mentioned earlier, and which are aroused by the amygdala [8]. The final determiner of the dispositional state is the salience of the input, as coded in the amygdala.

As noted above, we model the amygdala by a feed-forward neural network with several hidden layers. This allows work on learning the emotional coding of inputs, such as faces, by such a multi-layer perceptron, to be recognised in terms of modelling of the amygdala. However the resulting models use coding of emotions directly, and not of dispositions to act. It would be possible to train the MLP directly on dispositions as outputs, instead of emotions, for given face inputs. We use three separate outputs (action, activation, evaluation) to give a catoon version of the models of figure 3.

4 Training on Face Emotional Expression with Continuous Values.

In a previous study we explored the generalization capabilities of several architectures to classify novel emotional face images into the four classes neutral, happy, angry, and sad. The comparison showed, that only the multi-layer perceptron (MLP), trained by the backpro-pagation of error algorithm and using a cross-validation procedure, is able to generalize well to novel face images. However, the generalization performance depends on a good alignment of facial keypoints like the eyes and the facial shape. A limitation of the classification approach is the exclusive assignment of the facial expression to one class, which does not allow the coding of combined and secondary emotions. Furthermore, the strength ot the exposed expression is not coded by the classification scheme and some of the face images which show only subtle changes during a facial expression were incorrectly classified as neutral due to the small changes of the facial shape.

To remedy these limitations we modified a MLP-network to continuous output variables resembling the three dimensions evaluation, activation and action (see [13] for a minimal „mnemonic“ network for action responses). The training procedure using a cross-validation scheme was the same as in the previous study [14,15] except for the use of continuous target values for the exposed expressions. The target values for the expressions happy, angry and sad (Table 1) were determined in a psychological study of emotional words [4] relative to the baseline (0,0,0) for a neutral word.

	Neutral	happy	angry	sad
Evaluation	0.0	+ 0.5	- 0.7	- 0.8
Activation	0.0	+ 0.5	+0.65	- 0.15
Action	0.0	+ 0.75	- 0.1	- 0.8

Table 1 Target values for the primary emotions.

The data-set consists of 20 persons showing the four primary facial expressions. After normalizing and cropping the faces to 35x37 pixel images, four persons were excluded due to a large misalignment or missing data. The remaining 16 images for each expression were split into 12 training images, three validation images and one test image.

Figure 4 shows the final activations of the 48 training examples from a single run in the three-dimensional activation-evaluation-action space. Although the activations show some spread around the target values, a clear separation into four clusters can be observed.

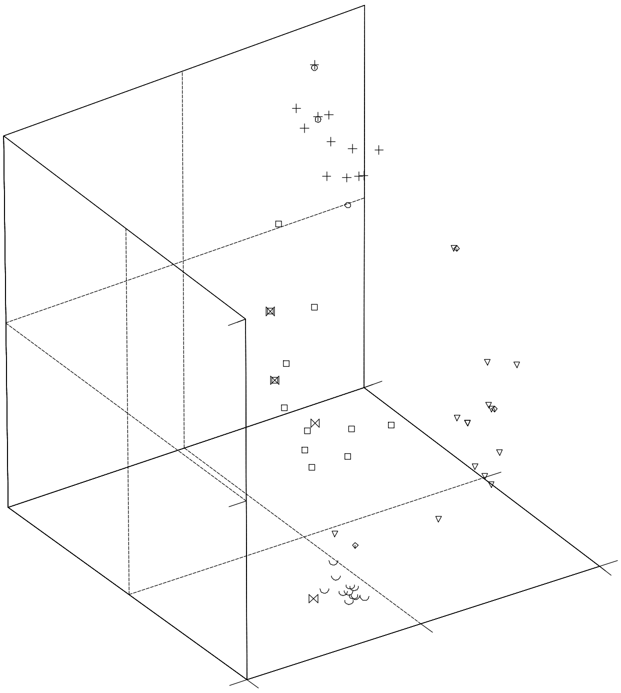

Fig. 4 3D-plot of the Activation-Evaluation-Action Space for the MLP-network trained on neutral (square), happy (plus), angry (triangele), and sad (half-circle) face images.

Figures 5 to 7 show the activations of four trials with 12 training examples for each facial expression along the evaluation-activation plane (Fig. 5), the evaluation-action plane (Fig. 6) and the activation-action plane (Fig. 7). In figure 5 the target values are overplotted as stars and the network responses to the novel images are plotted as crosses and plusses. The generalization performance for the depicted four runs was almost perfect but reached lower levels for the remaining runs of the cross-validation procedure.

5 Conclusion

We have shown that interpreting the output of a neural network for facial expression recognition as continuous variables of a three-dimensional state-space leads to an improved recognition performance compared to binary classification. Furthermore, the classification scheme allows for a better evaluation of the network performance since no thresholding or maximum search is necessary and all variables contribute to the classification of the expression.

The proposed scheme can be directly mapped onto limbic brain structures discussed earlier by identifying the amygdala with a part of the multi-layer perceptron, associated to the evaluation output, the NACC/VP as associated to the action output, and the HYP/PFC to th activation output.

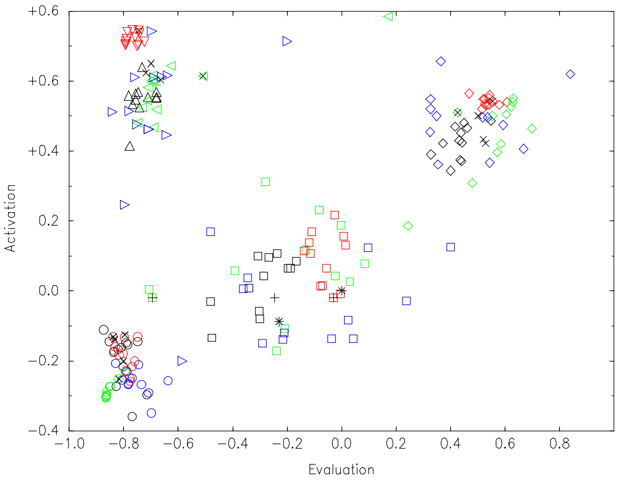

Fig. 5 Evaluation/Activation plane for the MLP-network trained on neutral (square), happy (diamond), angry (triangele), and sad faces (circle). The star marks the target values, plusses and crosses mark generalization output for three correctly classified test images. See text for further details.

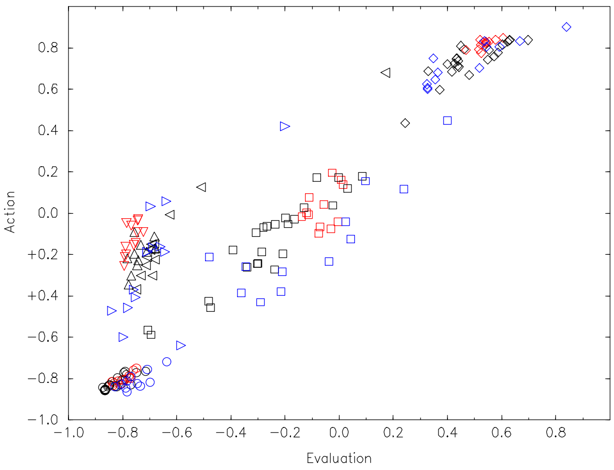

Fig. 6 Evaluation/Action plane; same as Fig. 4.

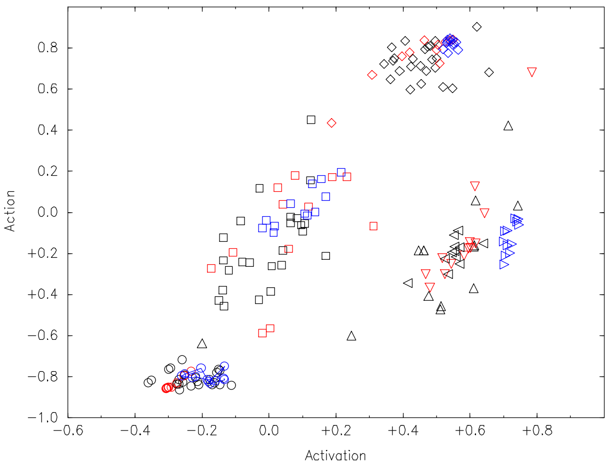

Fig. 7 Activation/Action plane; same as Fig. 4.

References:

[1] Review of existing techniques for human emotion understanding. *Report for TMR Physta project*, Research contract FMRX – CT97 – 0098 (DG12-BDCN)

[2] P. W. Kalivas and C. D. Barnes, *Limbic Motor Circuits and Neuropsychiatry*, Baton Rouge, CRC Press, 1993

[3] P. Willner and J. Scheel-Kruger (Eds.), *The Mesolimbic Dopamine System: From Motivation to Action*, Wiley, 1991

[4] R. Cowie, E. Douglas-Cowie, B. Appoloni, J. Taylor, A. Romano and W. Fellenz, What a neural net needs to know about emotion words, 1999

[5] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 79, 1982, pp 2445-2558

[6] E. T. Rolls, Functions of neuronal networks in the hippocampus and neocortex in memory. In J.H. Byrne and W.O. Berry (Eds.) *Neural models of plasticity: Experimental and theoretical approaches*, pp. 240-265, San Diego, Academic Press, 1989

[7] W. A. Fellenz and J. G. Taylor, Enhancing the storage capacity of linear associative memories, submitted, 1999

[8] J. P. Aggleton (Ed.) *The Amygdala*, Chichester UK: Wiley, 1989

[9] O. Monchi and J. G. Taylor, A hard wired model of coupled frontal working memories for various tasks, Information Sciences, Vol. 113, 1999, pp. 221-243

[10] N. Taylor and J. G. Taylor, Experimenting with models of the frontal lobes, In D. Heinke, G. W. Humphreys and A. Olson (Eds.) *Connectionist Models in Cognitive Neuroscience*, pp 92-103, London, Springer

[11] B. Krekelberg and J. G. Taylor, The two weights model of reward learning, EC TMR Project Report

[12] J.-M. Fellous and C. Linster, Computational models of neuromodulation, Neural Computation Vol. 10, 1998, pp. 771-805.

[13] J. Z. Young, The Organization of a Memory System, The Croonian Lecture, *Proceedings of the Royal Society B*, Vol. 163, 1965, pp. 285-320

[14] Development of Feature Representations from Emotionally coded Facial Signals and Speech, *Report for TMR Physta project*, Research contract FMRX – CT97 – 0098 (DG12-BDCN)

[15] W. A. Fellenz, J. G. Taylor, N. Tsapatsoulis and S. Kollias, Comparing Template-based, Feature-based and Supervised Classification of Facial Expressions from Static Images, 1999