A NEURAL-NETWORK-BASED APPROACH TO ADAPTIVE HUMAN COMPUTER INTERACTION

George Votsis, Nikolaos Doulamis, Anastasios Doulamis, Nicolas Tsapatsoulis and Stefanos Kollias

Image, Video and Multimedia Systems Laboratory
Department of Electrical and Computer Engineering
National Technical University of Athens
Zografou, 15773, Greece
stefanos@cs.ntua.gr

Abstract. A neural-network-based approach is proposed in this paper providing multimedia systems with the ability to adapt their performance to the specific needs and characteristics of their users. Two HCI applications are used to illustrate the performance of this approach. In the former, adaptive non-linear relevance feedback is proposed for content-based retrieval of multimedia information. In the latter, facial analysis is used to derive cues about the emotional state of a PC user, creating a more friendly and effective interaction.

1. INTRODUCTION

The widespread use of the Internet, mobile telephony and inexpensive computer equipment has made human computer interaction (HCI), and in general, man machine communication, a usual activity in everyday life. Making this interaction as efficient and friendly as possible, by including intelligence to the machines that interact with people, is a crucial aspect which is currently attracting large R&D efforts worldwide.

Neural networks have not played a significant role in the development of audiovisual coding standards, such as MPEG-1 and MPEG-2. Nevertheless the MPEG-4 and MPEG-7 standards [1], which refer to content-based audio-visual coding, analysis and retrieval, are based on physical object extraction from scenes, focusing on handling multimedia information with an increased level of intelligence. Neural networks, with their superior non-linear classification abilities, can play a major role in these standards, especially due to their ability to learn and adapt their behavior based on real actual data [4]. Probably the most important issue when designing and training artificial neural networks in real life applications is network generalization. Despite the achievements obtained during the last years, most real life applications do not obey some specific probability distribution and may significantly differ from one case to another, mainly due to changes of their environment. In such cases, the training set is not able to represent all possible variations or states of the operational environment to which the network is to be applied. Instead, it would be desirable to have a mechanism, which would provide the network with the capability to be on-line retrained, when its performance is not acceptable. The retraining algorithm should update the network weights taking into account both the former network knowledge and the knowledge extracted from the current input data.

In this paper we show that neural networks can form a crucial component of HCI systems responsible for content-based retrieval of multimedia information based on user relevance feedback [2] and for recognition of users' emotional state based on their facial expressions. The major issue that is explored is the ability of the system to adapt its performance to its specific user preferences and aims, expressions and emotional states [3]. In the former case, the system evaluates its user's selections, following each query he/she addresses to the system, and tackles his/her current intentions by refining its former knowledge (i.e., network architecture/weights); in this way, it makes information search and retrieval more user-centric and consequently more efficient. In the latter case, a well trained expression/emotion recognition system, which is, for example, sold by service providers to their customers, is on-line retrained, adapting its former knowledge to the specific characteristics or expressions of its owner.

2. RETRAINING STRATEGY OF THE NEURAL SYSTEM

Let us first consider that a neural-network-based system has been created by a service provider and is being supplied to customers, so as to be used, e.g., for facial expression recognition; it is able to classify each image, or video shot of a human to one, or more specific categories, such as happy, angry or neutral. The neural classifier has obtained the necessary knowledge to perform the task, having been trained with a carefully selected training set, say, S_b. Let us then consider that a specific customer includes the system in his/her own PC and starts to use it. The system will now be facing its true owner, and thus should adapt to its owner's specific characteristics and behavior, while keeping up with its former knowledge.

Let us consider network adaptation through retraining. Let vector $\underline{w}_b$ include all weights of the network before retraining, and $\underline{w}_a$ the new weight vector which is obtained through retraining. A retraining set S_c is assumed to be extracted from the current operational situation composed of, say, m_c feature vectors. The retraining algorithm should compute the new network weights $\underline{w}_a$, by minimizing the following error criterion with respect to the weights,

$$E_a = E_{c,a} + \eta E_{f,a} \tag{1}$$

where $E_{c,a}$ is the error performed over training set S_c ("current" knowledge), $E_{f,a}$ the corresponding error over training set S_b ("former" knowledge). Parameter η is a weighting factor accounting for the significance of the current training set compared to the former one.

In most real life applications, training set S_c is initially unknown; consequently selection of S_c, as well as detection of the need for training should be provided to the system, either through user interaction, or automatically, when this is possible.

The goal of the training procedure is to minimize (1) and estimate the new network weights $\underline{w}_a$. Let us first assume that a small perturbation of the network weights (before retraining) $\underline{w}_b$ is enough to achieve good classification performance. This assumption leads to an analytical and tractable solution for estimating $\underline{w}_a$, since it permits linearization of the non-linear activation function characterizing each neuron, using a first order Taylor series expansion.

It can be shown [5] that, through linearization, solution of this problem with respect to the weight increments is equivalent solving a set of linear equations

$$\underline{c} = \mathbf{A} \cdot \Delta \underline{w} \tag{2}$$

where vector $\underline{c}$ and matrix **A** are appropriately expressed in terms of the previous network weights $\underline{w}_b$. In particular, $\underline{c}$, indicates the difference between network outputs after and before retraining for all input vectors in S_c. Among all possible solutions that satisfy (2), the one which causes a minimal degradation of the previous network knowledge is selected as the most appropriate. The network weights before retraining, i.e., $\underline{w}_b$, have, however, been estimated as an optimal solution over data of set S_b. Furthermore, the weights after retraining provide a minimal error over all data of the current set S_c. Thus, minimization of the second term of (1), which expresses the effect of the new network weights over data set S_b, is equivalent to minimization of the absolute difference of the error over data in S_b with respect to the previous and the current network weights. This means that the weight increments are minimally modified, resulting in minimization of the following error criterion

$$E_S = \left\| E_{f,a} - E_{f,b} \right\|_2 = \frac{1}{2} (\Delta \underline{w})^T \cdot \mathbf{K}^T \cdot \mathbf{K} \cdot \Delta \underline{w} \tag{3}$$

where $E_{f,b}$ is defined similarly to $E_{f,a}$ and the elements of matrix **K** are expressed in terms of the previous network weights $\underline{w}_b$ and the training data in S_b. Thus, the problem results in minimization of (3) subject to constraints of (2).

The error function defined by (3) is convex since it is of squared type, while the constraints are linear equalities. Thus, the solution should lie on the hyper-surface defined by (2) and simultaneously minimize the error function given in (3). The gradient projection method has been used to solve this problem. The gradient projection method starts from a feasible point and moves in a direction, which decreases E_S and simultaneously satisfies the constraints; a point is called feasible, if it satisfies all constraints. The computational complexity of the retraining algorithm is very small. The total cost of retraining is of order of a few seconds, allowing the efficient use of the proposed scheme to real-life interactive multimedia systems.

3. ADAPTIVE CONTENT INFORMATION RETRIEVAL

In order to implement, a content-based image retrieval (CBIR) system, initially several descriptors are extracted from the images providing a more efficient representation of the visual content. Let us focus on user's queries that are submitted to the system in the form of images, icons or sketches. The goal in all cases is to retrieve the best *M* images from the database, the visual content of which is closer to the user's query. However, even an experienced user will not be in a position to know all types of images or other information included in the database. Consequently, the results following his/her query are expected to only partially cover what he/she really is looking for. For this reason, the system will follow an iterative procedure, taking advantage of its continuous interaction with the user and using the retraining methodology described in section 2 to discover the real ‘semantics’ of the user query.

To accomplish this, the provided image is first analyzed similarly to the database images, and then a distance or similarity measure is used to find the set of images in the database that best match the user's query.

Following this first retrieval, the user selects to see one or more of the returned icons in more detail, assigning a degree of appropriateness to each image, which actually indicates the degree of similarity of the respective image to the semantics of the image query. The system will take advantage of this information, so as to generate a new (re)training set for the neural network classifier. Following this, we can apply the retraining strategy described in section 2, so as to let the neural network (having initially been trained, e.g., with the available user profile) learn to classify the above feature vectors in the correct category. After retraining, the neural classifier will be applied to the whole part of the multimedia database, so as to select and present a new set of results/images to the user. Interaction with the user will be then repeated, so that the system iteratively identifies what the user is searching for.

Let us denote by $\underline{f}_q$ the feature vector extracted from the query image, and by $\underline{f}_i$ the respective vector of the *i*th image in the database. A feedforward neural network is used, in the following, to model the adopted non-linear similarity distance $d_{NL}(\underline{f}_q, \underline{f}_i)$. The network input is the difference $\underline{e}_{q,i} = \underline{f}_q - \underline{f}_i$, while the network output vector, say $\underline{z}(\underline{e}_{q,i})$, including two or three elements, correspondingly indicates whether the user is interested, indifferent, or rejects, the *i*th database image.

Let us denote by S_b the initial set used to train the network and by I the set containing all indices of the images selected by the user at a specific iteration. Then, a new training set, say S_c, is created, $S_c = \{(\underline{e}_{qk}, d_k)\}\ k \in I$. Let us also denote by $\underline{w}_b$ and $\underline{w}_a$ the total network weights, which for the specific architecture include the weights v_k and $w_{i,k}$, before and after the updating process respectively. Weights $\underline{w}_a$ can be estimated by minimizing the error criterion in (1), where E_c indicates the error over all samples of training set S_c formed by the users' selection, while E_f represents the error over all elements of the initial training set.

4. RECOGNITION OF USER'S EMOTIONAL STATE

Analysis of the emotional expression of a human face requires a number of pre-processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face. Most of the above techniques are based on a well known system for describing "all visually distinguishable facial movements", called the Facial Action Coding System (FACS). The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard. In particular, the Facial Definition Parameter (FDP) set and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, as well as the animation of faces reproducing expressions, emotions and speech pronunciation [1]. By monitoring facial gestures corresponding to FDP and/or FAP movements over time, it is possible to derive cues about user's expressions/emotions. Facial anatomy as well as social issues make the emotional analysis of faces to be more or less user dependent.

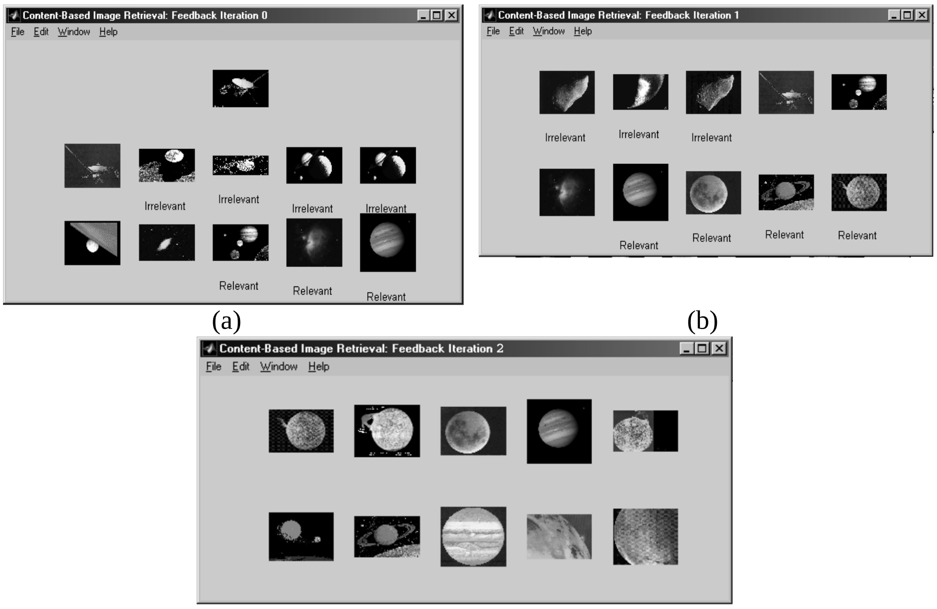
Let us assume that prominent MPEG-4 FDP points have been extracted and used to define corresponding FAPs. The latter are normalised according to some fixed face distances and provide a user independent description of the movement of particular facial areas. In the following we use a subset of the FDP feature points that we have identified as relevant to emotion [3]. Based on that, we have created a relation between FAPs that can be used for the description of the archetypal facial expressions and the selected FDP subset [6].

In the next section, we use the above-mentioned techniques and a database generated in [3] with presence of realistic, non-extreme emotional states to extensively train a neural network classifier so as to be able to discern among different human expressions. We will, however, show that, when providing a new user with the trained classifier (also providing the training feature data set, which can serve as the initial set for retraining), the obtained results will greatly improve, if we permit retraining of the system, as described in section 2 of this paper. Retraining will be performed, at the first time that the customer uses the system and, possibly, when a significant change of the environment takes place, that results in deterioration of the system's performance; an interactive framework has been created, where the system captures its user's expressions and interacts with him to get his/her own categorisation of them, so as to create the 'current' (re)training data set.

5. EXPERIMENTAL RESULTS

5.1 A Study on Adaptive Non-Linear Relevance Feedback

In this section we present experimental results related to the application of the proposed neural network architecture to the relevance feedback case. We apply the proposed adaptive neural network architecture for optimal non-linear relevance feedback to a real-world image database, consisting of a variety of images (about 10,000) organized in 39 different categories, such as nature, space, animals and cartoons. The system performance is subjectively analyzed, by depicting the actually retrieved images as a result of the user's selection of relevant or irrelevant images.



(a) (b)

(c)

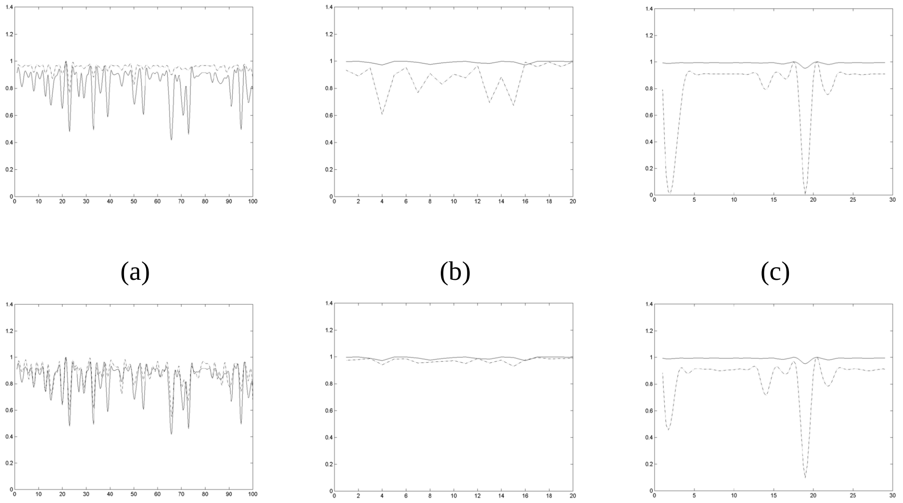
Fig. 1. The performance of the proposed non-linear relevance feedback scheme using the adaptive neural network architecture on a real-life database.

Figure 1(a) presents the first user's query submitted to the system, showing a white-red small space equipment, as well as the system response to the user's query. Let us assume that the actual user's information needs concern middle-sized red planets. Since, however, an image of such a type is not available in the system response, he/she selects the white-red space equipment, which resembles his/her needs. For this reason, the last three images are selected as relevant while the second to the fifth, are marked as irrelevant. The system response after the first relevance feedback iteration is presented in Figure 1(b). As is observed, in this case, more red, middle-sized planets have been retrieved. The system response in the second iteration is depicted in Figure 1(c). As is observed, most of the retrieved images are in accordance with the user's needs.

5.2 A Study on Adaptive Facial Expression Recognition

An experimental study has been conducted in the emotion recognition case using well known databases [3] such as the MediaLab database, the database of the UCSF

Human Interaction Lab, and the database with natural video sequences developed in [3]. The initial training set (S_b) includes faces of various people expressing six archetypal – type emotions, i.e. anger, sadness, happiness, disgust, fear and surprise, as well as their neutral state. More than one thousand images have been incorporated in this set. The emotional content of each image is described through the feature vector described in section 4. We then assumed that these data have been used to train a commercial neural network product, that needs to be adapted to its final end user. A second data set has been created with 20 images, portraying a specific end user, in each of the above seven emotional states. This data pool serves as retraining set (S_c) for the weight adaptation procedure. Following retraining, the product will be ready for use by its end user; a third data set composed of 170 other images with expressions of the same end user constitutes the test material (S_t), on which the performance of the retrained neural network is to be examined.



(a) (b) (c)

(d) (e) (f)

Fig. 2. "Happy" expression. (a) Performance of networks Z_b (dashed) and Z_c (solid) over set S_b . (b) Performance of Z_b (dashed) and Z_c (solid) over set S_c . (c) Performance of networks Z_b (dashed) and Z_c (solid) over set S_t . (d) Performance of networks Z_o (dashed) and Z_c (solid) over set S_b . (e) Performance of networks Z_o (dashed) and Z_c (solid) over set S_c . (f) Performance of networks Z_o (dashed) and Z_c (solid) over set S_t . Horizontal axis in (a)-(f) shows the sample numbering, while vertical axis shows the relevance to the specific class.

Using the above data sets, two kinds of experiments were carried out. In the first, the scenario of the retraining strategy was tested. In the second, for comparison purposes, the retraining data set was used together with the initial training set to train the neural network de novo. Let us denote by Z_b the initially trained network (on S_b), by Z_c the retrained network (on S_c according to the presented procedure) and by Z_o the overall trained network (on the union of S_b and S_c).

The results for the “happy” expression are presented in Figure 2, verifying the good performance of the proposed approach.

6. CONCLUSIONS

Content-based information retrieval through non-linear relevance feedback, and emotion recognition through facial expression analysis have been examined in this paper, using a novel adaptive framework based on artificial neural networks. In particular, an efficient scheme for on-line retraining of neural network classifiers has been proposed, which provides the above systems with the ability to adapt their behavior to the specific needs or characteristics of their users, while retaining their former knowledge.

REFERENCES

1. L. Chiariglione, “MPEG and Multimedia Communications,” *IEEE Trans. on Circuits and Systems for Video Techn.*, vol. 7, pp. 5-18, 1997.
2. A. Doulamis, Y. Avrithis, N. Doulamis and S. Kollias, “Interactive Content-Based Retrieval in Video Databases Using Fuzzy Classification and Relevance Feedback,” *Proc. of IEEE Inter. Conf. on Multimedia, Comp. & Syst. (ICMS'99)*, Florence, Italy, June 1999.
3. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, Y. Votsis, S. Kollias, W. Fellenz and J.Taylor, “Emotion Recognition and Human Computer Interaction,” *IEEE Signal Processing Magazine*, No.1, January 2001.
4. S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: Prentice Hall, 2nd edition, 1999.
5. N. Doulamis, A. Doulamis and S. Kollias, “On-Line Retrainable Neural Nets: Improving Performance of Neural Networks in Image Analysis Problems,” *IEEE Trans. on Neural Networks*, vol. 11, no 1, pp. 137-155, 2000.
6. N. Tsapatsoulis, Y. Avrithis and S. Kollias, “Facial Image Indexing in Multimedia Databases,” *Pattern Analysis and Applications, Special Issue on Image Indexation,* Springer Verlag, (to appear).