C17 癒し工学システムの開発
Development of伊ashi Engineering System

ディアゴ・ルイス・東工大
Luis DIAGO, Tokyo Institute of Technology

北岡哲子・東工大
Tetsuko KITAOKA, Tokyo Institute of Technology

豊原一郎・東工大
Ichiro HAGIWARA, Tokyo Institute of Technology

Key Words: facial expression analysis, holographic neural networks, type-II fuzzy quantification, 伊ashi expressions

論文要旨

1. Introduction

The human senses are sight, touch, hearing, smell, and taste. The input modalities of many computer input devices can be considered to correspond to human senses: cameras (sight), haptic sensors (touch), microphones (hearing), olfactory (smell), and taste. Despite important advances in multimodal human-computer interaction (MMHCI) see e.g. [1], further research is still required to investigate fusion models able to efficiently use the complementary cues provided by information channels from multiple modalities. This is precisely the way human read other people's feelings. As a result of our investigations, firstly, to build human-like multimodal systems we must provide to the machines a framework able to deal with the information distributed between different information channels.

Recent researches have included a physiological monitoring as part of the human–computer interface (see [2]). Using a thermal camera as a computer peripheral almost the full range of vital signs can be extracted, including localize blood flow, cardiac pulse, and breath rate. On the other hand, since the original demonstration that electrical activity generated by ensembles of cortical neurons can be employed directly to control a robotic manipulator, research on brain–machine interfaces (BMIs) has experienced an impressive growth. According to the review of BMIs made by Lebedev and Nicolelis [3], today BMIs designed for both experimental and medical studies can translate raw neuronal signals into motor commands that reproduce arm reaching and hand grasping movements in artificial actuators. Based on these recent developments, instead of controlling the machine using BMIs, we extract cognitive state information from neural signals to produce appropriate feedback to the machine and increase classification accuracy. From the fusion of facial expressions, body gestures and vocal expressions of the persons with their brain signals, the machine predicts forward states to induce positive emotions. According to the survey about MMHCI by James and Sebe [1], most researchers process each modality (e.g. visual, audio) independently, and multimodal fusion is still in its infancy.

Our novel hypothesis states that “what lights up in the brain can be visualized and complemented by analyzing the information in the face, voice and gestures of observed person”. In order to go beyond a human-like multimodal analysis of multiple input signals acquired by different sensors, compared with similar studies (see details in [4]), the input data is processed in a joint feature space using a centralized architecture according to a new proposed fusion model. In this paper, we present a new fusion model based on holographic neural networks. The rest of the paper is organized as follows. Section 2 shows the details of proposed fusion model. Section 3 explains the main features corresponding to each information channel and Section 4 presents preliminary results from the experiments developed models. Finally, section 5 presents the conclusions and future works.

2. Holographic fusion model

The main concept towards achievement of the goal is in the development of new models based on a new type of Holographic Neural Networks (HNN) [5] which includes only one neuron by each information channel. Holographic neurons are represented by circles in Figure 1. The first two channels ($S_1$ and $S_2$) use the features of facial expressions presented as stimulus and facial expressions of observed person to predict the inner state of the person (represented by $R$). Proposed model adds a new channel ($S_3$) related with brain signals of observed person. The outputs of all channels are synchronized using one more HNN neuron. The new feature space initially includes information from three channels. Machine learning is performed with HNN by using the information of three channels. Most important features from each channel (i.e. $S_1^*$, $S_2^*$ and $S_3^*$) are selected by summing the response output of three holographic neurons ($\Sigma R_k$) and adjusting the parameters of holographic neurons based on the difference ($R_{dif}$) with the inner state provided by the user.

Figure 1 - The new proposed context-dependent model for fusing facial information with brain signals.

Problems faced by current technologies include large dimensionality of the required joint feature space, differing feature formats, and time-alignment. Because HNN learning is developed simply by finding the inverse of a complex matrix, the learning time is considerably smaller than learning time in other fusion technologies (e.g.[6]) enabling proposed fusion model achieves a real time synchronization of all information channels.

2.1 Synchronization of three channels of information

When we deal with images acquired from a high-speed video camera, one of the main bottlenecks is the introduction of the huge amount of data from the device to the computer with considerable speed in order to control the evaluation time of observed expressions. As video acquisition devices produce higher detailed...
information that contains huge number of frames, image processing algorithms need large memory and running time. Clearly, the sampling by these automatic video acquisition devices is not dependent on the image features, and hence the data sets exhibit a large proportion of redundant information. Compression of such data sets is one of the important pre-processing techniques for making storage, transmission, computation, and display more efficient. Other than these applications, the time consuming processes, such as re-sampling and de-interlacing, involved in the reconstruction of the image from its samples using implicit representations benefit a lot by the reduction of the number of features used to represent each image and the input image set itself. We have proposed an efficient preconditioning technique [7] for image reconstruction with Compactly-Supported Radial Basis Functions (CSRBF) that is under current investigation for lossy and lossless image coding in our proposed system.

Figure 2 shows the diagram of time used for synchronization of three channels of information.

![Diagram of time used for synchronization of three channels of information.](image)

Figure 2 - Diagram of time used for synchronization of three channels of information.

Ten seconds used for the evaluation of one image have been shown for convenience in Figure 2. Each image was shown for 5 seconds (tEQ) and removed for 5 seconds (tREP) to leave a margin to the observed person to think about the state induced by the system. Brain signals (denoted as B) and facial expressions of observed person (denoted as F) are acquired during evaluation of each image. Facial information from the camera is captured several times during the acquisition of brain signals as it is shown in Figure 2 because facial acquisition time is smaller than that of brain signals. In addition, several operations are necessary between two acquisitions. In this way each input modality is synchronized from the moment that the stimulus picture is presented by taking facial expressions and brain signals of the users to increase classification accuracy.

2.2 Extraction of cognitive information from neural signals

Compared with former HNN and SVM classifiers, the proposed combination of HNN with I-Q-II achieved the best classification accuracy using less training iterations than SVM classifiers (see [5] for details). However, there were ambiguous cases where the subject himself could not decide the state that has been induced by the system. Sparked in part by the growing interest in the evaluation of multiple communication channels to determine the inner states of a person, we combine the facial expressions of the subject with signals from the brain to clarify the state induced on the person in ambiguous situations. The extraction of cognitive state information from neural signals will produce appropriate feedback to the system and it will increase the classification accuracy.

In the past decades, pioneer researchers have tried to apply to EEG data analysis techniques developed in electrical engineering and information theory, including time/frequency analysis and Independent Component Analysis (ICA). These techniques have revealed EEG processes whose dynamic characteristics are also correlated with behavioral changes, though they cannot be seen in the averaged ERP. For example, short-term changes in spectral properties of the ongoing EEG in specific frequency bands may be correlated with cognitive processes, e.g. expectancy of a target stimulus and with visual awareness. To date, the majority of BCI systems rely on EEG recordings. In this paper we research some possible brain features for learning and predicting subjects evaluations about a small dataset of facial images in order to gain a clear idea of what it means for a face to be lyashi according to the computational models.

3. Tracking Facial Features and Brain Signals

This paper is a continuation of our previous works [5] in which we introduced fuzzy quantification to build a MF describing how subjects classify lyashi-stimulus into fuzzy groups.

3.1 Correlation between facial features

The main goal of the simulations is to gain a clear insight into the reasoning made by HNN. The same 20 images represented by the same number of parameters presented in [5] were evaluated by one hundred and fourteen subjects between 15 and 70 years old (102 Japanese and 12 non-Japanese, 47 females and 67 males) and were used to train different neuro-fuzzy classifiers. The participants rated each stimulus on the scale '0'-NO, '1'-DON'T KNOW, '2'-YES to express whether or not they feel an lyashi-stimulus. Figure 3 shows the set of parameters used to describe the 20 faces.

![Facial parameters and Facial images](image)

Figure 3 - Face Parameters and Facial images

Figure 4 shows the pictures obtained by the system during the evaluations made by one of the subjects (Japanese male in the lower right part of each panel). We have confirmed also that there is a high correlation between stimulus facial expressions, the evaluation of the subject and his facial expression shown during stimulus evaluation.

![Application of the system for the definition of lyashi expressions](image)

Figure 4 - Application of the system for the definition of lyashi expressions [5]. Left and right stimulus (shown in the upper part) were evaluated as '0'-NO and '2'-YES respectively.

3.2 Acquisition of brain signals

The increase in classification accuracy in current systems due to the inclusion of the new modality of brain signals is not evident. To measure brain activities, noninvasive measurement of the brain functions such as functional Magnetic Resonance Imaging (fMRI), Magneto-encephalography (MEG), electro-encephalogram (EEG) and Optical Tomography have been developed. Emotion Spectrum Analysis (ESA) [8] has been developed to evaluate human emotion such as Anger, Joy, Sadness and Relaxation using EEG signals. Recently, a Generalized Emotion Near Infrared-spectroscopy Analysis System (GENIASS) [9] has been introduced. GENIASS allows estimating quantitatively human

1 the subject cannot decide if yes or no
emotion by using Near Infrared-spectroscopy (NIRS) signal. NIRS is a non-invasive functional neuro-imaging method that measures regional brain activations in terms of e.g. oxygenated or deoxygenated hemoglobin with low cost and high time resolution of 0.1 second. However, the dimensionality of features coming from brain signals increases with the number of acquisition channels. For example, GENIAS uses more than 4000 features obtained from second and third order correlation coefficients between signals coming from 24 channels. In comparison with GENIAS, in which the subject was asked to recall each emotion and the affect recognizer is linear, proposed system uses paintings and photograph for emotion induction and it is based on a nonlinear HNN affect recognizer.

Currently, most EEG researchers still interpret their data by measuring peaks in event-locked ERP averages. Free availability of more general and easy-to-use signal processing software for EEG data may encourage the wider adoption of more inclusive approaches. In the following we report the preliminary results using EEGLAB software toolbox for Matlab (freely available from http://www.sccn.ucsd.edu/eeglab/) to extract brain features during the evaluation of Iyashi-stimulus.

4. Preliminary Results

In order to keep the emotional state of the subjects as uniform as possible, in the experiment the subjects performed a calculation task for 60 seconds, rested for 90 seconds with their eyes closed. And the stimulus pictures were presented continuously at 10-second intervals. Then, they rested again with their eyes closed during 90 seconds. The experiment was repeated twice for two sets of stimulus images previously classified as Iyashi as Non-Iyashi stimulus by a set of subjects that did not participate in the experiment. The EEG signals were recorded using the ESAM system. Figure 5 shows one of the subjects, which participated in the experiment and the position of the electrodes in the standard 10-20 System. Only 10 channels (Fp1, Fp2, F3, F4, P3, P4, O1, O2, T3 and T4) are used.

Figure 6 show one example image classified by the subject as Non-Iyashi stimulus. Stimulus and brain signals corresponding to each channel (y-axis) during 10 seconds (x-axis) used for evaluating the image are shown in the upper part. In the lower part, the channels spectra and topographic scalp maps are shown. Each colored trace represents the spectrum of the activity of one data channel. The leftmost scalp map shows the scalp distribution of power at 6 Hz, which in these data is concentrated on the back midline (i.e. O1 and O2). The other scalp maps indicate the distribution of power at 10 Hz and 22 Hz. The same information is represented in Figure 7 for an Iyashi stimulus classified by the subject.

5. Conclusion and Future Works

After observing the spectrum and topographic scalp maps derived from brain signals reflected by the presentation of each of the 20 images, we found that those images that are evaluated as 2 (stimulus Iyashi) show a peak greater than 10db in the spectrum below 5 Hz (see Figure 7 in comparison with Figure 6). Also note that the topographic maps in the bands 5 and 10 Hz show greater symmetry for Iyashi stimuli than for non-Iyashi stimuli. As the distribution of power for the 22 Hz band is similar in both cases more research is need to remove noises affecting the high frequency band of 20Hz.

6. References