Expectation maximization of prefrontal-superior temporal network by indicator component-based approach

T. Koshizen, Bernd Heisele, Hiroshi Tsujino

Honda Research Institute Co. Ltd.,
koshiz@jp.honda-ri.com

Abstract

In this paper, we attempt to propose top-down attention control system based on the outcome that we presented at CNS*02. That is, PFC is presumably interacted with superior temporal (ST) neuron in order to extract indicator component representing the whole view of face or object, by maximizing the expectant value where attention modulation can be taken into account of distinguishing different faces. The PFC-ST network indicates to compute the abstraction of face/object information based on the Expectation Maximization (EM) algorithm since the voluntary movements of facial viewpoints must play an important role of integrating spatial and temporal property.

Key words: Prefrontal and Superior-temporal Cortex, Distributed Cortical Neuronal Network, Spatiotemporal Attention, Cross Supra-modality, EM Algorithm

1 Introduction

PF cortex is involved in a broad array of cognitive functions, including learning, memory, attention, executive function, planning, and judgment [1]. It is known that PFC has also the executive committee by consolidating hippocampus and other cortical areas. In this paper, we propose the top-down attention control scheme that is hypothesized the computation with respect to the connection of PF and superior temporal (ST) cortex. ST cortex is presumably known a conjunctive point between the 'dorsal' stream specialized by motion (temporal attention) property and the 'ventral' stream (what) specialized by form (spatial attention) property. Additionally, the ST is known for the multimodal response as well as the cross-modal response when interacting with posterior parietal (PP) cortex. Importantly, recent physiological

results also demonstrated the neurons in ST involve in the computation of face perception [2]. This is because the two core areas of face processing engage in the categorization of a stimulus as a face, and the identification of a specific individual, by the ST neuron incorporated with inferior temporal (IT) neuron, in order to implement the facial perception consisting of the 'rough' and the 'finer' computations through their neurons belonging in PF and Orbito-Frontal (OF) cortex. The PF-based executive neuronal network involves ST, AC, Basal Ganglia (BG) and Hippocampus (HP). The key idea of our speculation is, PF cortex, which may expertise the top-down attention control where attention modulation allows the maximum expectant to yield the semantic abstraction representing specific face or object information in accordance with the supramodal computation. In this sense, the computational role of top-down attention may be implicated to calculate the expectation value where the visual motion may be crucial for maximizing it by adding the biasing signals to the visual form. It evokes generalization properties of biological motion perception using a new class of stimuli that were generated by the spatial and temporal characteristic of morphing among different viewpoint patterns. It has been demonstrated using several monkey's experiments shows that the PF neuron involves in the reward-based learning where spatial information becomes more accurate when reward outcome is expected; more accurate representations of spatial information would as a consequence lead to more accurate behavior, e.g., [3]. In this paper, we suggest the top-down attention control scheme based on the PFC-STS network related to a biological Expectation Maximization (EM) learning algorithm to extract the indicator component based on the top-down attention control.

2 Computational Model

In this section, we describe the computation model of our proposal, called indicator facial component-based learning approach. The overview of proposed system is shown in Fig.2 (left) where the single component classifier initially developed by [4] is hierarchically reorganized by the multiple component classifiers using Support Vector Machine (SVM) training algorithm which performs pattern recognition for a two-class problem by determining the separating hyperplane that has maximum distance to the closest points of the training set. The closest points, the maximum distance are called Support Vector (SV) and Margin respectively [5]. In our framework, the proposal system aims to extract the indicator facial component that is used to abstract the informative source to distinguish different faces. Face component-based learning algorithm aims to detect faces of different sizes and arbitrary positions in a gray value input image [4]. More precisely, Fig.2. (Right) shows the schematic drawing where component classifiers independently detect components of the face on the first

level. This classifier allows the components to extract the features around eyes, nose and mouth. On the second level, the geometrical configuration classifier performs the final face detection by linearly combining the results of the component classifiers. We eventually obtain the output of SVM component classifier indicates if there is a face inside the window or not. Generally, one

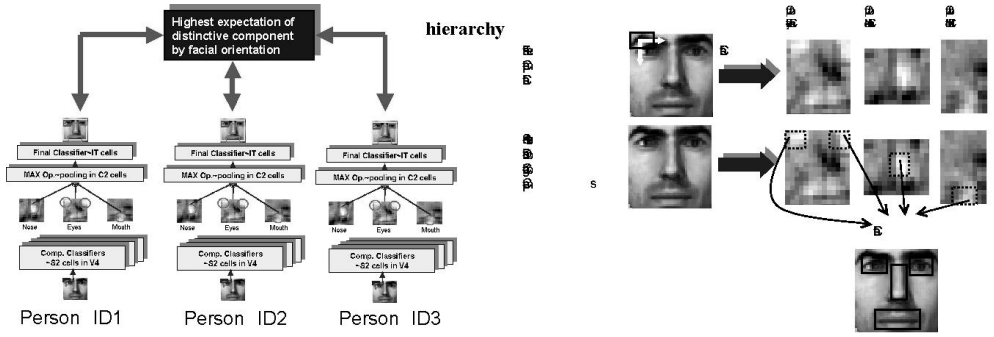

Fig. 1. Proposed multiple classifier system to attain indicator component(Left) and conventional two-level single classifier system (Right)

of the main problems in the component-based object recognition is the selection of the components; how to find discriminated components that allow to distinguish a particular object from rest. To address the question, we employ the proposed system shown in Fig.1 (Left) where top-down attention control must be taken into account of maximizing the expectation value where the viewpoint of faces are fluctuated by attention modulation to differentiate the semantic abstraction of indicator component.

Let the expectation value $\mathcal{E}$ be

$$\mathcal{E}(\Omega_\tau) \approx \int logP(\mathcal{O}, \Omega; \tau) d\mathcal{O} \tag{1}$$

where Ω represents the probabilistic attention variables in accordance with certain viewpoints τ. $\mathcal{O}$ is the outputs that are calculated from each SVM component classifier shown in Fig.1 (Left). Additionally, to calculate the probability density function $P(\mathcal{O}, \Omega; \tau)$ we assume the hypothetical mapping $h : \mathcal{O} \mapsto \Omega$. The mapping h will be mathematically defined as,

$$P(\Omega_\tau) \equiv P(y = 1 | \mathcal{O}_\tau) = \frac{1}{1 + exp(\mathcal{O}_\tau / \gamma_\tau)} \tag{2}$$

where, y is the classification label of positive examples and $\gamma \in \Omega$ is the predictive parameter of attention modulation to extract the indicator component that maximizes the expectation value $\mathcal{E}$. γ_τ represents the correspondance between the supramodalities such as visual form and motion. Importantly, in our framework the attention Ω can be regarded as the kind of 'hidden variables' calculating to determine the optimal subset of ranked features that is learned

from each SVM component classifier. The computation of Eq. (2) is initially suggested by [6]. Mathematically, γ_τ is somehow related to the differentiation of the expectation $\mathcal{E}$ over Ω_τ,

$$\gamma_\tau = \frac{\partial \mathcal{E}(\Omega_\tau)}{\partial \Omega_\tau} \tag{3}$$

Note that the right side of Eq.(3) computes the functional splope of the expectation value $\mathcal{E}$. In principle, we can maximize the logarithm of the joint distribution (which is proportional to the posterior):

$$\Omega_{\tau+1} = argmax_{\Omega_\tau} \;\; \mathcal{E}(\Omega, \Omega_\tau) \tag{4}$$

where τ denotes the parameter representing the viewpoint of a facial component. It is important to remind that the expectation value $\mathcal{E}$ is calculated in the **E**-step by evaluating the current guess Ω_τ where in the **M**-step we are optimizing $\mathcal{E}(\Omega, \Omega_\tau)$ with respect to the *free variable* τ (facial viewpoint) to accordingly obtain the new estimate $\Omega_{\tau+1}$.

To implement our indicator component-based approach, training images are captured over the circumstances of various illuminations and the unique (black) background. After the images are collected, pixel values are used as inputs to each layer of a SVM component classifier as shown in Fig.1. The cropped image is then converted into gray values and is re-scaled to 40×40 pixels. Histogram equalization is also applied to remove variations existing in image brightness and contrast. The 1,600 gray values of each face image are then normalized to the range between 0 and 1. Each image is represented by a single feature vector of length 1,600 - the total number of pixels in the image. These feature vectors serve as the inputs to the SVM face classifier during the training process. With respect to the training dataset it includes 974 images of all six subjects in our database. The rotation in depth is again up to about $\pm 42°$. Fig.2 (left) implicates the facial viewpoints has the rotation by right to left, or left to right within $-10°$ to $+10°$. By contrast, the rotation of facial viewpoints is only left by $+12°$ to $+42°$ in Fig.2 (right). Fig.3 represents the histgrams, which show the learning results computed by SVM in the proposed classifier system. Furthermore, Fig.4 shows the expectation values in both cases, and which are calculated based on Eq.(1). Conclusively, their results demonstrate that the expectation value is used for selecting component features by qualifying the input data structure in accordance with different facial viewpoints. We suggest the computation of top-down attention control by PF neuron may be originated to maximize the expectation value where attention class Ω is modulated over different component features, in order to restrict the quantity and quality of training data mapping into the feature space in order to find the optimal subset of selected features, as the indicator component.

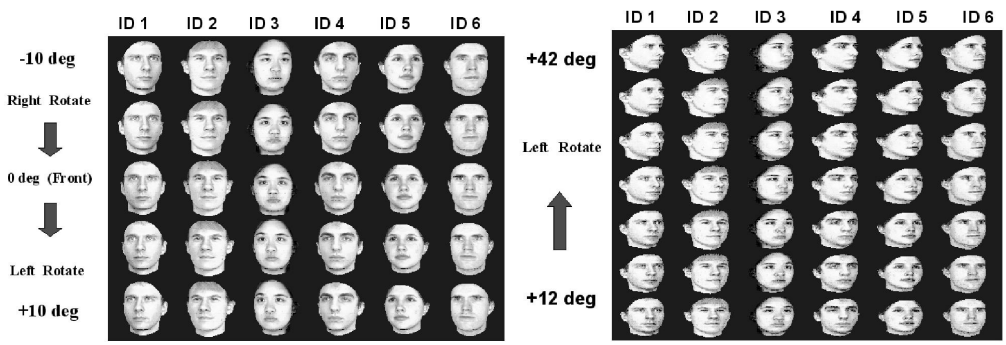

Fig. 2. Facial movement pattern by $-10°$ to $+10°$ (Left) and $+12°$ to $+42°$ (Right)

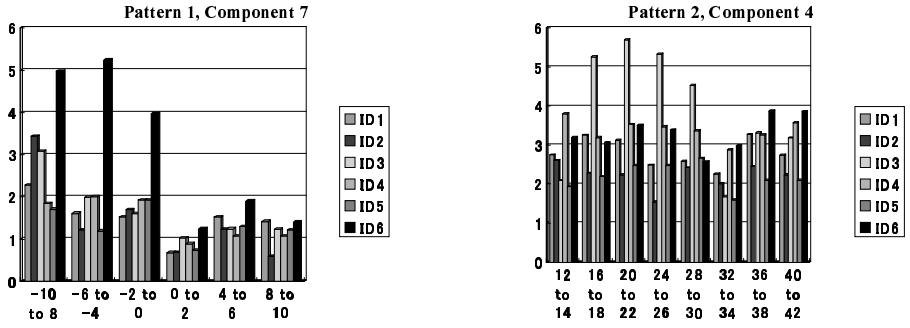

Fig. 3. Histogram shows the learning result by SVM. Vertical axis denotes the margin, while horizontal axis represents the viewpoint discretized by every 2°. ID6 shows the most distinction for the left-side nose component (Left) with $-10°$ to $+10°$. By contrast, ID3 shows the most distinction for the right-side eye component (Right) with $+12°$ to $+42°$

3 Discussion and Future work

In this paper, we showed the PFC-STS model suggests the EM algorithm that is basically originated from top-down attention control process biased by visual motion. In our framework, the top-down attention regulates each SVM component classifier by calculating the expectant value to extract indicator component. Future work may allow us the top-down attention to be applying in emotional perception and even in more general-domain for object perception

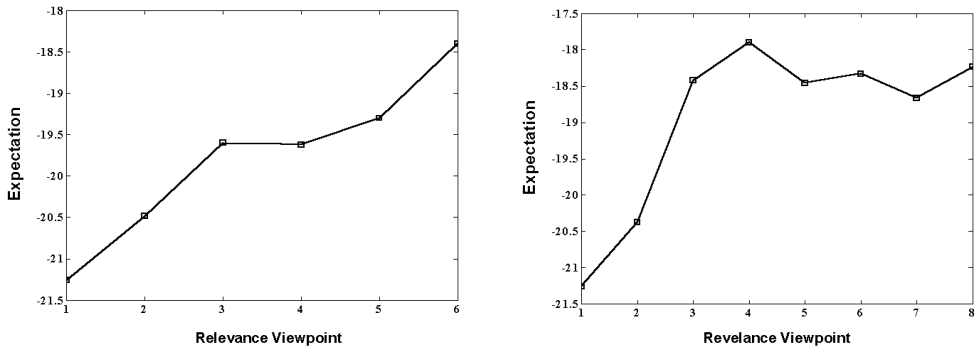

Fig. 4. Expectation value calculated by Eq.1 under the rotation of $-10°$ to $+10°$ (Right) and $+12°$ to $+42°$ (Left)

by extending the class of indicator component which is obtained from animated faces.

References

[1] Miller, E.K. and Cohen, J.D., (2001), *An integrative theory of prefrontal cortex function*, Annual Review Neuroscience, Vol.24, pp.167-202.

[2] Haxby, J.V. et al., (2000), *The distributed human neural system for face perception*, Trends Cognitive Science, Vol.4, pp.223-233.

[3] Kobayashi, S., Lauwereyns, J. Koizumi, M., Sakagami, M. and Hikosaka, O, (2002), Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex, J. of Neurophysiology, Vol.87, pp. 1488-1498.

[4] Heisele, B., Poggio, T. and Pontil, M., (2000), "Face detection in still gray images", A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA.

[5] Vapnik, V., (1998), "Statistical learning theory", John Wiley and Sons, New York.

[6] Wahba, G., (1999), *Support vector machines, reproducing kernel hilbert spaces and the randomized GACV*, In B. Scholkopf, C.J.C. Burges, and A.J.Smola, editors, Advances in Kernel Methods - Support Vector Learning, pages 69-88., Cambridge, MA, 1999. MIT Press.