SENSORIAL SUBSTITUTION SYSTEM WITH ENCODING OF VISUAL OBJECTS INTO SOUNDS

Damien Lescal[1], Jean Rouat[1], and Stéphane Molotchnikoff[2]
[1]NECOTIS, Dept. GEGI, Université de Sherbrooke, Quebec QC, Canada, J1K 2R1
[2] Dept. de sciences biologiques, Université de Montréal, Quebec QC, Canada, H3C 3J7

1. INTRODUCTION

Visual and auditory prostheses involve surgeries that are complex, expensive and invasive. They are limited to a small number of electrodes and can only be used when the impairment is peripheral. Non invasive prostheses (sensorial substitution systems) have existed for more than 40 years but have not been well accepted in the disability sector. Severals systems have been developped since the emergence of this concept. Paul Bach-Y-Rita proposed a substitution system from vision to touch (1969) in which pictures captured by a camera were converted into electrical stimulation of the tongue. Other studies have shown that some simple tasks such as localization (Jansson, 1983), shape recognition (Sampaio et al., 2001; Kaczmarek and Haase, 2003) and reading (Bliss et al., 1970) can be achieved using vision-to-touch substitution devices.

More recently, substitution systems from vision to audition have been proposed: (Akinbiyi, 2007; Merabet, 2009; Hanneton 2010). The following systems are the most important to have been developed so far: the vOICe (Meijer, 1992), PSVA (Prosthesis for Substitution of Vision by Audition) (Capelle et al., 1998), the device developed by Cronly-Dillon (Cronly-Dillon, 1999) and the Vibe (Hanneton et al., 2010). These systems encode a full image with no prior analysis of the visual scene. Thus, they overload the ears of the patient with wasteful sounds that carry useless characteristics of the image. Usually these systems encode the luminosity of all pixels from the image in the amplitude of modulated sounds. The vOICe and the Cronly-Dillon device use left-to-right time scanning to encode horizontal position. The Vibe and PSVA encode the entire image in one complex sound. The PSVA uses frequencies that are associated with each pixel and increase from left to right and from bottom to top of the image. The Vibe splits the image into several regions that are equivalent to receptive fields. Each receptive field is associated with a single sound and the sum of all sounds forms a complex sound transmitted to the two ears. The receptive fields design is inspired by work on the retina.

The challenge in this project resides in the design of a suitable encoding of the visual scene into auditory stimuli such that the content of the sound carries the most important characteristics of the visual scene. These sounds should be shaped in a way that the subject can build mental representations of visual scenes even if the information carrier is the auditory pathway. A sensorial substitution system using an object-based approach is described. An image segmentation algorithm using a spiking neural network combined with a sound generation is proposed. That neural system has a good potential for object based image and visual scene analysis. Also, human auditory features such as interaural time difference (ITD) and interaural level difference (ILD) are used to synthesize sounds that better characterize the visual scene for real-life environments.

2. DESCRIPTION

The image analysis system has to be able to find and track objects in image sequences. For this purpose, a spiking neural network that offers a good potential is used for object-based image processing.

2.1 Mapping between image analysis and sound generation

An image is presented to a one layer self-organized spiking neural network (Figure 1).The positions of neurons from a given group are encoded into the same sound.

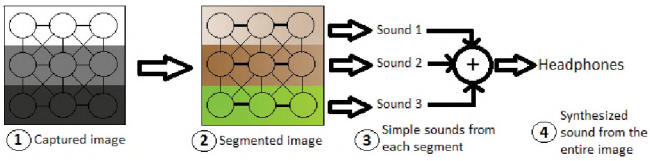

Figure 1. Pixels are input to the neural network (1). After synchronization, segments appear (2) and are encoded into sounds (3). In (2) color encodes moments of spike times and thickness of lines between neurons encodes weight values.

In the two following sections, the segmentation and the sound generation are described.

2.2 The spiking neural network

The neural network described by Molotchnikoff and Rouat (2011) is used. The neuron model is the conventional simplistic integrate and fire neuron. The sub-threshold potential of the neuron with a constant input is:

—— - - (1)

is the transmembrane potential, is the input current. When crosses, at time , a predetermined threshold θ, the neuron fires. Then is reset to (resting potential). C is the membrane capacitance; τ has the dimension of a time constant expressed in seconds. In this case – must be superior to θ for the neuron to be able to fire.

Each neuron characterizes a pixel of the image and is connected to eight neighbours. A synaptic weight between two neurons encodes the similarity between the respective features (here gray levels of pixels). The weight between neuron i and neuron j is computed according to:

$$w_{i,j} = 1 - \frac{1}{1+e^{-\alpha(|f_i - f_j| - \Delta)}} \qquad (2)$$

α=0.1, 0.05 or 0.015, Δ=100, 128 or 200 and $|f_i - f_j|$ is the absolute value of the gray level difference between neuron i and neuron j. A segment represents a group of neurons that spike at the same time (thus being synchronized). So, neurons associated to pixels with a small difference in gray level will spike at the same time and are identified as belonging to the same group (segment). Results of segmentation with this neural network are shown in Figure 1 and 2.

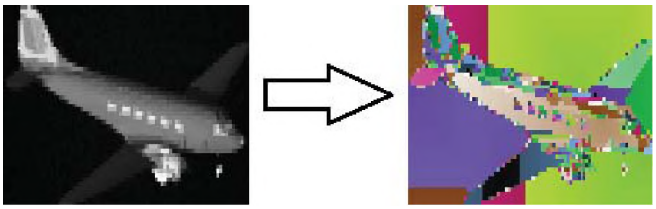

Figure 2. Input (left) and time moments of spiking (right).

2.3 Sound generation

The segmented image is then encoded into sounds. A single sound is generated for each segment using the averaged gray level, the size and the position of the segment in the image. $S_j^R(t)$ is the sound from the segment to be played in the right ear and $S_j^L(t)$ is the one in the left ear.

$$S_j^R(t) = A_j^R \sin(w_j t + \Phi_j^R) \qquad (3)$$
$$S_j^L(t) = A_j^L \sin(w_j t + \Phi_j^L) \qquad (4)$$

- $\boldsymbol{w_j = \overline{g_j}}$ with $\overline{g_j}$: average level of gray of the segment j
- $\boldsymbol{A_j^L = \alpha . S_j}$ with S_j : size of the segment j and α : average distance of the segment j from the center of the image (ILD)
- $\boldsymbol{\Phi_j^L = \frac{\alpha}{\alpha_{max}} * \frac{\pi}{2}}$ with α_{max} the half of the width of the image (ITD)
- The expression of $\boldsymbol{\Phi_j^R}$ and $\boldsymbol{A_j^R}$ are the same than $\boldsymbol{\Phi_j^L}$ and $\boldsymbol{A_j^L}$ except that the reference locations in the image are different.

The complex sound is the sum of all single sounds from each segment. One complex sound is generated for the right ear and another one for the left. In short, the differences between the right and the left sound reside in the size of the objects and their positions in the image.

3. DISCUSSION AND CONCLUSIONS

The approach described in this paper is very promising. The next step of this project is to adapt and modify the neural network to identify highly textured regions of an image. In others words, highly textured portions of an image would be isolated and the most homogeneous segment would be identified. Using this neural network, it would be possible to identify textured objects (natural objects) and non-textured objects (objects man-made objects). The strength of this approach is the combination of an object-based image analysis with the sound generator so that mental visual representations can be carried via an auditory stimulation. Indeed, this system does not convert the entire image into sound but only parts corresponding to important features of the image. This is not the case in the literature. Furthermore, the sound generation is based on human auditory features like ITD, ILD and the size of the object for the depth in the image. Using this approach, it might be possible to help people with visual disabilities with tasks like localisation or shape recognition.

REFERENCES

Akinbiyi, T. et al., (2006). "Dynamic augmented reality for sensory substitution in robot-assisted surgical system."*Proc.: Annual I. C. of the IEEE Eng. in Med. Bio. Society.*.

Bach-y Rita, P. et al., (1969). "Vision substitution by tactile image projection." *Nature,* vol. 221, p. 963-964.

Bliss, J. C. et al., (1970). "Optical-to-tactile image conversion for the blind" *IEEE Transactions on Man-Machine Systems*

Capelle, C. et al., (1998). "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution." *Biomedical Engineering, IEEE Trans.*

Cronly-Dillon, J. et al., (1999). "The perception of visual images encoded in musical form: a study cross-modality information transfer." *Proc. Biological sciences / The Royal Society*

Hanneton, S. et al., (2010). "The vibe : a versatile vision-to-audition sensory substitution device." *Applied Bionics and Biomechanics,* vol. 7, num. 4, p. 269-276

Jansson, G. (1983). "Tactile guidance of movement" *International Journal of Neuroscience,* vol. 19, p. 37-46

Kaczmarek, K. A. and Haase, S. J. (2003). "Pattern identification and perceived stimulus quality as a function of stimulation current on a fingertip-scanned electrotactile display" *Transaction on Neural System Rehabilitation Engineering,* vol. 11, p. 9-16

Meijer, P. (1992). "An experimental system for auditory image representations" *Biomedical Engineering, IEEE Trans.*

Merabet, L. B. et al., (2009). "Functional recruitment of visual cortex for sound encoded object identification in the blind" *Neuroreport,* vol. 20, num. 20, p. 132-138

Molotchnikoff, S. and Rouat, J. (2011). "Brain at work: Time sparness and multiplexing/superposition principles", *Frontiers in Bioscience*, in press.

Sampaio, E. et al., (2001). "Brain plasticity: Visual acuity of blind persons via tongue" *Brain Research,* vol. 908, p. 204-207

ACKNOWLEDGEMENTS

Sean Wood for English corrections, FQRNT of Québec and ACELP/CEGI of Univ. de Sherbrooke for funding this project.