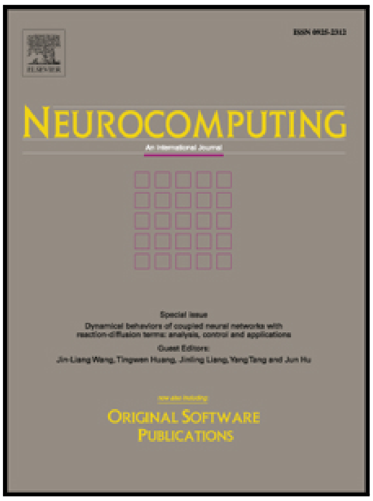
ISSN 0925-2312
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
ORIGINAL SOFTWARE PUBLICATIONS

Towards an Audiovisual Attention Model for Multimodal Video Content

Naty Sidaty, Mohamed-Chaker Larabi, Abdelhakim Saadane

PII: S0925-2312(17)30260-6
DOI: 10.1016/j.neucom.2016.08.130
Reference: NEUCOM 18056

To appear in: *Neurocomputing*

Received date: 24 February 2016
Revised date: 17 August 2016
Accepted date: 18 August 2016

Please cite this article as: Naty Sidaty, Mohamed-Chaker Larabi, Abdelhakim Saadane, Towards an Audiovisual Attention Model for Multimodal Video Content, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.08.130

Towards an Audiovisual Attention Model for Multimodal Video Content

Naty Sidaty, Mohamed-Chaker Larabi*, Abdelhakim Saadane

XLIM, University of Poitiers, France

Abstract

Visual attention modeling is a very active research field and several image and video attention models have been proposed during the last decade. However, despite the conclusions drawn from various studies about the influence of human gazes by the presence of sound, most of the classical video attention models do not account for the multimodal nature of video (visual and auditory cues). In this paper, we propose an audiovisual saliency model with the aim to predict human gaze maps when exploring video content. The model, intended for videoconferencing, is based on the fusion of spatial, temporal and auditory attentional maps. Based on a real-time audiovisual speaker localization approach, the proposed auditory map is modulated depending of the nature of faces in the video, *i.e.* speaker or auditor. State-of-the-art performance measures have been used to compare the predicted saliency maps with the eye-tracking ground truth. The obtained results show the very good performance of the proposed model and a significant improvement compared to non-audio models.

Keywords: Audiovisual saliency, Talking faces, Visual attention, Eye-tracking, Audio-visual synchrony, Fusion strategies.

1. Introduction

Visual attention is a selective process and a clever mechanism of the human visual system (HVS). It allows selecting the most attractive areas of a visual scene, called salient regions. Visual attention studies permitted to use this HVS property in various applications such as: computer vision (recognition and object detection, tracking, compression, . . .), computer graphics (image rendering, dynamic lighting, . . .) and robotics. On the research side, important efforts have been devoted to studying and modeling visual attention, leading to the introduction of numerous image and video saliency models. The reader can refer to the very comprehensive review made by *Borji et al.* [1]. Although visual and auditory systems exhibit a significant anatomical difference, neuronal researches have shown that their sensory mechanism is very similar [2]. Therefore, different auditory saliency models have been proposed with encouraging

*Corresponding author
Email address: chaker.larabi@univ-poitiers.fr (Mohamed-Chaker Larabi)