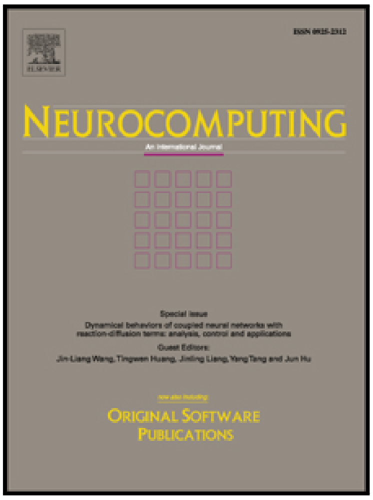
Communicated by Dr. Qingshan Liu

Accepted Manuscript


ISSN 0925-2312
ELSEVIER
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
now also including:
ORIGINAL SOFTWARE PUBLICATIONS

Facial Expression Intensity Estimation using Siamese and triplet networks

Motaz Sabri, Takio Kurita

PII: S0925-2312(18)30792-6
DOI: 10.1016/j.neucom.2018.06.054
Reference: NEUCOM 19732

To appear in: *Neurocomputing*

Received date: 22 December 2017
Revised date: 7 May 2018
Accepted date: 17 June 2018

Please cite this article as: Motaz Sabri, Takio Kurita, Facial Expression Intensity Estimation using Siamese and triplet networks, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.06.054

Facial Expression Intensity Estimation using Siamese and triplet networks

Motaz Sabri * **Takio Kurita** **

* *Department of Information Engineering, Hiroshima University,*
Hiroshima, Japan
(e-mail: d151350@hiroshima-u.ac.jp)
** *(e-mail: tkurita@hiroshima-u.ac.jp)*

Abstract: This paper investigates the Siamese and triplet networks abilities of emotional intensity estimation in facial image sequence. In our method, we extract the sequential relationship in the temporal domain that appears due to the natural onset apex offset variation in pattern of facial expression. Siamese and triplet networks are shown to perform better than the earlier convolutional neural networks in such task. The branches of the Siamese and triplet networks help in leading to an output that is more definite. Compared with Siamese network, the triplet network internal representation of learned features emerges clearer and more accurate localizations of those features appear with training. This property improves the network generalization when dealing with similar sequential images. We confirmed this by experiments on Cohn-Kanade, MUG and MMI datasets for intensity estimations and CASME, CASME II and CAS(ME)^2 datasets on micro-expressions detection.

Keywords: Intensity Estimation, Affection computing, Siamese Network, Triplet Network, Feature localization, Machine learning.

1. INTRODUCTION

Every day we perform hundreds of nonverbal behaviors to exchange information through body language, facial expressions and paralanguages. Each behavior is delivered in parallel with an emotion paired to it (Kaltwang (2012)). Understanding those emotions deepens our communication and enriches our discussions. Such emotional awareness is important for computers as well to allow more natural human-machines interaction (Ambadar (2005)).

Many researches have tackled the recognition of emotions in videos and pictures (Nicholson (2000); Kuilenburg (2005); Battiti (1994)). Subject facial expressions are used to predict an emotional state. Another trending emotional analysis area is facial expression intensity estimation (Yang (2009); Zhao (2016)). It helps defining how much a person is influenced by experiences such as discomfort in a medical treatment or joy while watching an advertisement.

The diversity of patterns in facial expressions beside the absence of standard way to label intensities make intensity estimation a challenging task (Bartlett (1999)). Many researches propose methods based on Facial Action Coding System proposed by Ekman (1978) (FACS) which describes the facial expression using facial actions such as contracting glabella or squinting outer canthus. Such researches aim to classify an expression in an image or video as one of the six basic expressions (Batty (2003); Kobayashi (1992)). Regardless of those methods good performance, they neglect the dynamics of an expression which is critical in interpreting it's meaning (Mavadati (2013); Rudovic (2015); Savran (2012)). In this paper, we propose a Convolutional Neural Networks (CNN) based method to interpret the dynamics of a facial expression through it's intensity in the temporal domain.

CNN's have been successfully used to achieve state-of the-art performance in a variety of pattern recognition tasks such as object classifications, feature localization, image regeneration, and speech recognition. However, with weak supervision on the training, predictions given by CNN tend to break down. Koch (2015) et al. proposed generalization method to recognize these unfamiliar categories through Siamese networks. This network structure has improved the network learning capabilities by exploiting additional information about how the training samples are related. With their abilities to minimize the distance, those networks were used in many fields such as signature matching (Bromley (1993)), Image originality verification and media search retrieval (Lorenzo (2015)).

However, the representations learned by these models provide below average results when used as features to learn explicit representations. Triplet network model learns representations that are comparable with a network that was trained explicitly to classify samples (Hoffer (2015)). This model doesn't require knowing the class of the processed input, it needs to know that only two out of three inputs are sampled from the same class, rather than knowing what that class is. This model learns using only comparative measures instead of labels (Wang (2015)).

Inspired by Siamese and triplet network abilities, we analyze facial expression intensity estimation against the basic emotion states of anger, happiness, sadness, surprise, contempt and disgust. We also analyze the evolution of the emotion in terms of the emotional features locality