Pattern Recognition Letters 92 (2017) 25–32

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Joint facial expression recognition and intensity estimation based on weighted votes of image sequences

Siti Khairuni Amalina Kamarol[a,*], Mohamed Hisham Jaward[a], Heikki Kälviäinen[b], Jussi Parkkinen[c], Rajendran Parthiban[a]

[a] School of Engineering (Electrical and Computer Systems), Monash University Malaysia, 46150 Selangor, Malaysia
[b] School of Engineering Science, Lappeenranta University of Technology, 53850 Lappeenranta, Finland
[c] School of Computing, University of Eastern Finland, 80100 Joensuu, Finland

ARTICLE INFO

*Article history:*
Received 23 June 2016
Available online 4 April 2017

*Keywords:*
Video-based facial expression intensity analysis
Facial expression recognition
Voting scheme
Hidden markov models
Change-point detection
Computer vision

ABSTRACT

Facial behavior consists of dynamically changing properties of facial features as a result of muscle activation. Facial behavior analysis is a challenging problem due to complexity of emotions and variability of the facial expressions associated with the emotions. Most facial expression recognition systems attempt to recognize facial expressions without taking into account the intensity of the expressions. In this paper, a novel framework for facial expression recognition and intensity estimation with low computational complexity requirement is proposed. The algorithm constructs a representation of facial features based on a weighted voting scheme and employs Hidden Markov Models to classify an input video into one of the six basic expressions, namely anger, disgust, fear, happiness, sadness, and surprise. The temporal segments, neutral, onset, and apex, of an expression are then obtained by means of a change-point detector. Evaluations on subject-independent analysis was conducted using Cohn-Kanade dataset and Beihang University facial expression datasets. The proposed approach has demonstrated a superior performance in recognizing facial expressions and estimating expression intensities.



## 1. Introduction

Face and its features can be used to identify a person and to analyze the emotional state of a person. Facial expression analysis has been widely used and implemented across various fields such as computer technology, security, psychology, medicine, and pattern recognition [1,2]. One of the most important applications is in clinical investigations of neuropsychiatric disorders including affective disorders and schizophrenia [3]. With the rapid development of Human-Computer Interaction (HCI), it also plays an important role in affective computing technologies with many potential applications [4].

Many researchers have focused on facial expression recognition and recently researchers attempted to incorporate this task with estimation of facial expression intensities. Facial expressions are naturally dynamic and they can be segmented into four temporal segments: *neutral, onset, apex*, and *offset* [2,5]. *Neutral* means no expression is shown, *onset* is the instance when the muscular contraction occurs and increases in intensity, *apex* is the peak of the expression, and *offset* is the instance when the expression starts fading away. The dynamics of facial expressions are the crucial information required for interpreting facial behavior [6]. Examples of behavioral research related to facial expressions include the study of emotion, personality, social interaction, communication, anthropology, and child development [7]. Differences in terms of physical facial appearance such as wrinkles and skin texture of different individuals make facial expression intensity analysis a very challenging problem [3].

Most facial expression recognition systems attempt to recognize facial expressions of six basic emotions and only a few considered recognition of facial expressions and estimation of its temporal segments jointly [8]. Since each person may display a particular facial expression at different intensities, a system which can adapt to a particular individual by quantifying the expressions has an advantage over a system which only performs recognition of expressions. Automatic pain recognition from videos [9] is one of the significant applications of such a system.

In general, facial expression analysis comprises of recognition of facial expression of emotions and facial action units. Our work focuses on classifying a sequence of facial features in terms of emotions such as anger, disgust, fear, happiness, sadness, and surprise, as well as estimating the intensity of the expression. Re-

* Corresponding author.
*E-mail address:* siti.khairuni@monash.edu (S.K.A. Kamarol).

http://dx.doi.org/10.1016/j.patrec.2017.04.003
0167-8655/

search in estimating facial expression intensities is still in the early stage with few publications. Current approaches which perform facial expression recognition and intensity estimation jointly are reviewed in Section 2.

The main contribution of this paper is a novel framework for facial expression recognition and intensity estimation with low computational complexity. The algorithm is capable of classifying a sequence of features into an expression class and quantifying the intensity of the sequence as the expression changes in intensity from neutral to apex. A feature representation based on nearest neighbor search and a weighting scheme was also developed. An input video is represented by a weight vector which contains two significant information: the most likely expression class of the input sequence and the expression intensity (in terms of numerical rating) depicted by each frame in the sequence.

To take into account the dynamics of an expression from a video sequence, Hidden Markov Models (HMMs) [10] were used to classify the sequence into an expression class. A change-point detector [11] is then adopted to encode the expression intensities captured by the weight vector into one of three temporal segments: neutral, onset, and apex. In order to assess the performance of the proposed approach, comparisons were performed against three state-of-the-art algorithms: Hidden Conditional Random Field (HCRF) [12], Hidden Conditional Ordinal Random Field (HCORF) [13], and Variable-state Latent CRF (VSL-CRF) [14]. Evaluations were performed in a subject-independent manner in which videos from a subject can only appear in either training or test set. This paper also includes a study on the robustness of the proposed method against annotation errors of geometric feature.

This paper is organized as follows. Section 2 reviews existing approaches which perform facial expression recognition and intensity estimation jointly. Section 3 describes the details of the proposed approach. Section 4 describes the datasets, experimental methodology and parameters, along with a discussion on the experimental results. Finally, Section 5 concludes the paper.

## 2. Related work

Facial expression intensity estimation is currently an emerging field of research with only a few work addressing it so far [13–18]. The work on facial expression intensity estimation consist of approaches based on facial expression of emotions and facial action units. A review on intensity estimation of action units can be found in [19]. Girard et al. [20] proposed an approach for estimating the intensity of a specific expression, which is smile. In this section, we review existing work that combine recognition of facial expressions and estimation of the facial expression intensities into one system. Note that only the work based on estimating intensity of facial expression of emotions are reviewed in this paper.

In general, current approaches can be categorized into distance-based [3,21], cluster-based [16], regression-based [15,17], and methods based on probabilistic graphical models [12–14,18]. One of the earliest approaches on facial expression intensity estimation was a distance-based approach proposed by Verma et al. [3] in which high-dimensional shape transformations and regional volumetric difference maps were used to characterize and quantify facial expressions. Lee and Xu [21] adopted the Lucas–Kanade optical flow tracking algorithm to obtain the trajectory of facial feature points. The 1D manifold of expression intensity was then extracted from the trajectory of the feature points using isometric feature mapping. Quan et al. [16] proposed a cluster-based approach which employed a $k$-order emotional intensity model based on $k$-means clustering of facial features. Before the expression recognition, the task of intensity estimation of facial expression was carried out by the distance based and cluster-based methods.

On the other hand, the regression-based approach by Wu and Xiao [15] implemented facial expression recognition prior to intensity estimation. HMMs were adopted for classification and expression states generation. Then, energy value of each state was obtained based on the placement of landmark points. Incorporating the states variation and energy value gives the intensity curves for each expression using a linear regression algorithm. Chang et al. [17] proposed a framework that estimates facial expression intensity based on a single image. They implemented a scattering transform and a reduction framework with support vector machine for intensity estimation, omitting the task of recognizing facial expressions.

The existing algorithms based on probabilistic graphical models such as HCRF [12], HCORF [13], Laplacian Shared-parameter Multi-output Conditional Ordinal Random Field (LSM-CORF) [18] and VSL-CRF [14] incorporate facial expression recognition and intensity estimation in a single probabilistic framework. It has been shown in [13,14] that these models outperform other algorithms based on regular CRFs and static ordinal regression. HCRF was used to model temporal dynamics of facial expressions as a sequence of hidden states, which is a sequence of nominal values. HCORF, on the other hand, models the temporal dynamics as the dynamics of ordinal values. LSM-CRF is an extension of HCORF where the intrinsic topology of facial affect data is modeled and then incorporated into the HCORF framework. However, approaches based on HCRF and HCORF assume that the latent states are either nominal or ordinal for every class. Recently, Walecki et al. [14] proposed VSL-CRF which addresses these constraints by allowing both nominal and ordinal latent states to be used in modeling sequences across and within the classes. The algorithm outperforms HCRF and HCORF on four facial expression datasets for facial expression recognition and AU detection. Despite its superior performance in recognizing facial expressions, no results on intensity estimation were reported in [14]. HCRF and HCORF performed poorly as reported in [13,18].

## 3. Proposed facial expression recognition and intensity estimation framework

This section describes the details of the overall methodology for the proposed facial expression recognition and intensity estimation approach.[1]

First, we perform feature extraction using Active Appearance Model (AAM) [22] which marks a number of landmark points on facial images to represent the shape and locations of facial features such as eyebrows, eyes, nose, and mouth. These features are also known as geometric features. Note that in this work we do not consider the grey-level appearance of the facial images. The $x$ and $y$ coordinates of the landmark points are used to generate a feature vector $\mathbf{x}_f$ to represent each frame.

Features extracted from the video sequences are then partitioned into training and test sets. Both training and test sets contain feature vectors of different facial expressions. From the training set, a pool of apex features of different facial expressions is formed by selecting only features from the apex frames in the training set.

A feature representation based on $k$ Nearest Neighbor (kNN) and a weighting scheme is developed. Given a video with $t$ number of frames from the test set, we define $\mathbf{x}_f$ as the feature vector corresponding to frame $f$ where $f \in \{1, \ldots, t\}$. Using the pool of apex features created from the training set, we search for a set of $k$ nearest neighbors for $\mathbf{x}_f$. The kNN search is performed frame-by-frame

[1] MATLAB implementation of the proposed algorithm is available in the supplementary downloadable material.