Pattern Recognition 48 (2015) 3407–3416

Contents lists available at ScienceDirect

Pattern Recognition

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

A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences

Xijian Fan, Tardi Tjahjadi*

School of Engineering, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form
31 March 2015
Accepted 27 April 2015
Available online 8 May 2015

Keywords:
Histogram of gradients
Facial expression
Optical flow
Feature extraction

ABSTRACT

Facial expression causes different parts of the facial region to change over time and thus dynamic descriptors are inherently more suitable than static descriptors for recognising facial expressions. In this paper, we extend the spatial pyramid histogram of gradients to spatio-temporal domain to give 3-dimensional facial features and integrate them with dense optical flow to give a spatio-temporal descriptor which extracts both the spatial and dynamic motion information of facial expressions. A multi-class support vector machine based classifier with one-to-one strategy is used to recognise facial expressions. Experiments on the CK+ and MMI datasets using leave-one-out cross validation scheme demonstrate that the integrated framework achieves a better performance than using individual descriptor separately. Compared with six state of the art methods, the proposed framework demonstrates a superior performance.

1. Introduction

The automated recognition of facial expressions has been a widely researched topic in recent years due to its wide range of applications such as surveillance, human–computer interaction and data-driven animation. Other motivations include advancements in related research in face detection [1], tracking and recognition [2], as well as new developments in feature extraction algorithms and machine learning [3]. Six prototypical facial expressions were first formalised in [4], namely anger, disgust, fear, happiness, sadness and surprise. Although much progress has been made since then, accurate recognition of facial expressions is still a challenging problem due to the subtlety, complexity and variability of facial expressions [5,6].

Most existing works on facial expression recognition focus on analysing and extracting facial features in a single image or one frame in an image sequence, i.e., recognition of static expression. Previous methods have mainly concentrated on attempting to capture expressions through either action units [5,7] or via discrete frame extraction techniques [8]. All of these methods require either manual selection of facial features in order to determine where the particular changes in the facial region occur, or the subjective thresholding for feature selection. This means that any classification is highly dependent on subjective information in the form of a threshold or other a priori knowledge.

A facial expression involves a dynamic process, and the dynamic information such as the movement of facial landmarks and the change in facial shape contains useful information that can represent a facial expression more effectively. Thus, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Previous recognition methods on video sequences tend to only focus on the movement of facial landmarks, not analysing the variation of facial shape. In this paper, we utilise two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on the pyramid histogram of gradients (PHOG) [9] to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. We view an image sequence as a spatio-temporal volume, and use temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. In this context, we extend PHOG descriptor representing spatial local shape to spatio-temporal domain to capture the changes in local shape of facial sub-regions in the temporal dimension to give 3-dimensional (3D) facial component sub-regions of forehead, mouth, eyebrow and nose. We refer this descriptor as PHOG_three orthogonal planes (PHOG_TOP). By combining PHOG_TOP and dense optical flow of the facial region, we exploit the fusion of discriminant features for classifying and thus recognising facial expressions.

The main contributions of this paper are: (a) a framework that integrates the dynamic information extracted from variation in facial

* Corresponding author. Tel.: +44 24 76523126; fax: +44 24 76418922.
E-mail address: t.tjahjadi@warwick.ac.uk (T. Tjahjadi).

http://dx.doi.org/10.1016/j.patcog.2015.04.025

shape and movement of facial landmarks, (b) PHOG_TOP 3D facial features, (c) a means of fusing weighted PHOG_TOP with dense optical flow, and (d) an analysis on the contribution of different facial subregions using the proposed framework.

This paper is organised as follows. Previous related work is presented in Section 2. Section 3 presents PHOG_TOP, the dense optical flow descriptor, and the fusion of these descriptors. The proposed facial expression recognition framework and the experimental results are, respectively, presented in Sections 4 and 5. Finally, Section 6 concludes the paper.

2. Related work

There are two main approaches to recognising facial expressions: (1) recognition based on facial action coding system (FACS) [10] action units (AUs) and (2) direct content-based recognition (non-AU). The weakness of the AU based approach is that errors in the AU classification affect the recognition rate. Thus, the framework proposed in this paper adopts the non-AU approach.

A typical facial expression recognition system comprises three modules: image pre-processing, facial feature extraction, and facial expression classification. The feature extraction module is important and thus numerous methods for facial features extraction have been proposed. These methods can either be appearance-based or geometric-based methods. The features extracted using either approach should minimise intra-class variation of facial expressions, while maximising inter-class variations.

In the appearance-based approach, transformations and statistical methods are used to determine the feature vectors that represent textures and are thus simple to implement. Gabor wavelets [11] and local binary patterns (LBPs) [12] are two representative feature vectors of such approach that describe the local appearance models of facial expressions. Gabor magnitudes are commonly adopted as features as they are robust to misalignment of corresponding image features. However, computing Gabor filters has a high computational cost, and the dimensionality of the output can be large, especially if they are applied to a wide range of frequencies, scales and orientations of the image features. The LBP descriptor is a histogram where each bin corresponds to one of the different possible binary patterns representing a facial feature, resulting in a 256-dimensional descriptor. However, it has been shown that some of the patterns are more prone to encoding noise. The most popular LBP is the uniform LBP [13]. Zhao and Pietikainen [14] proposed a method which extends LBP to spatio-temporal domain so as to utilise the dynamic information, which results in a significant improvement in the recognition rate. One drawback of appearance-based approach is that it is difficult to generalise appearance features across different persons.

In the geometric-based approach, shape and position information of facial landmarks or region are extracted to represent the face geometry [11,15–17]. For example, the method in [11] uses the geometric position of 34 fiducial points as facial features to represent the face geometry. In image sequences, optical flow analysis has been applied to detect the movements of facial components which are qualified by measuring the geometric displacement of facial feature points between two consecutive frames [18,19]. Although geometric-based methods are sensitive to noise, and require accurate tracking of facial features, geometric features alone can provide sufficient information for efficient recognition of facial expressions.

Histogram of gradients (HOG) [20] was originally developed for person detection and object recognition. In [21], HOG descriptors are extracted from face image using a dense grid, and are used for face recognition. The PHOG proposed in [9] is an extension of HOG and is used to represent the local shape of facial region. However, all these methods only analyse individual frames of a video sequence, i.e., not taking the dynamics of a facial expression into account.

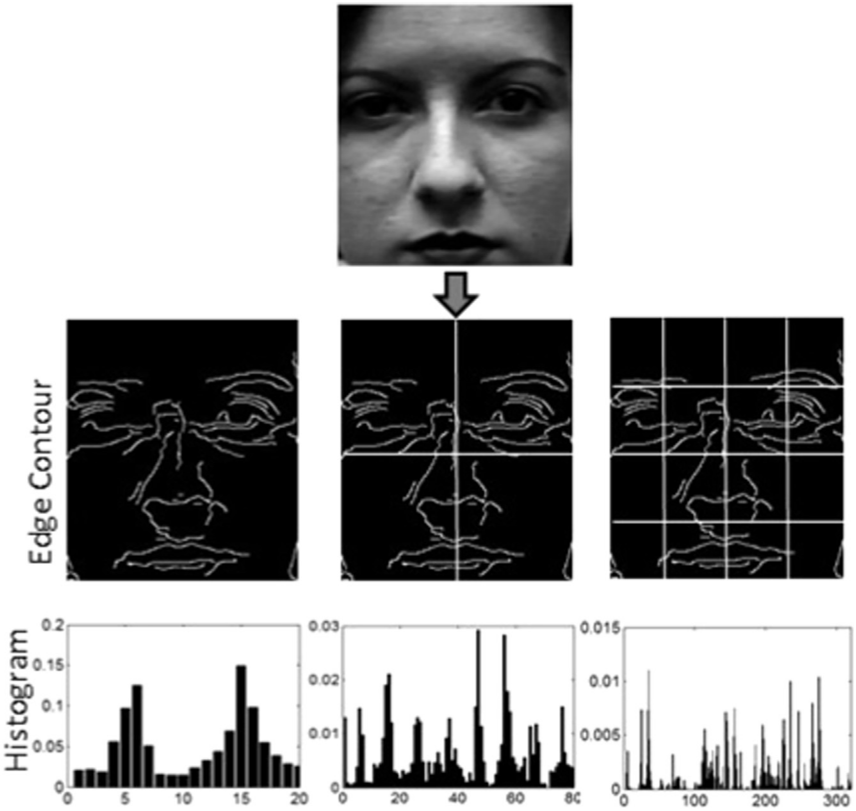

Fig. 1. PHOG descriptor of a face.