J. Vis. Commun. Image R. 56 (2018) 182–187

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A discriminative dynamic framework for facial expression recognition in video sequences ☆

Xijian Fan [a,*], Xubing Yang [a], Qiaolin Ye [a], Yin Yang [a,b]

[a] Department of Computer Science, Nanjing Forestry University, Nanjing, China
[b] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, USA

ARTICLE INFO

Article history:
Received 21 August 2018
Revised 10 September 2018
Accepted 14 September 2018
Available online 15 September 2018

Keywords:
Histogram of gradients
Facial expression
Feature extraction

ABSTRACT

Facial expression involves a dynamic process, leading to the variation of different facial components over time. Thus, dynamic descriptors are essential 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. To enhance the spatial information, we divide the whole face region into a group of smaller local regions to extract local 3D features, and a weighting strategy based on fisher separation criterion is proposed to enhance the discrimination ability of local features. A multi-class classifier based on support vector machine is applied for recognising facial expressions. Experiments on the CK+ and MMI datasets using leave-one-out cross validation scheme show that the proposed framework perform better than using the descriptor of simple concatenation. Compared with state-of-the-art methods, the proposed framework demonstrates a superior performance.

1. Introduction

The automated recognition of facial expressions has drawn more and more attention recently because of its extensive applications such as surveillance, human–computer interaction and data-driven animation [1]. Other motivations include advancements in related research in face detection [2], tracking and recognition [3], as well as new developments in feature extraction algorithms [4] and machine learning [5]. Six primary facial expressions were first proposed in the work of Ekman et al. [6], which includes anger, disgust, fear, happiness, sadness and surprise. In recent years, broad research has been carried on the recogition of facial expression, and much progress has been made. However, accurate recognition of facial expressions is still a challenging problem due to the subtlety, complexity and variability of facial expressions [7,8].

Exsiting methods have mainly concentrated on attempting to capture expressions through either action units [7,9] or using the extraction techniques based on discrete frame [10]. 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 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 [11]. Also, not all facial parts contribute to recognising facial expression, where certain facial parts contain more important information than others. For example, the regions of cheek and mouth show more variations than those of forehead during the disgust expression [12]. Shan et al. [12] propose a weight strategy to give various weights to different facial regions to highlight those contains discriminative information. However, they compute weights empirically relying on their observation, which is impractical in real application. In this context, we propose to combine dynamic and discriminative information to improve the recognition: a novel spatio-temporal descriptor using the Pyramid Histogram of Gradients (PHOG) [13] to capture the changes in facial appearance, and an adaptively weighting strategy to represent the importance of different facial regions. In this context, an image sequence is regarded as a spatio-temporal volume, and temporal information describing the dynamic changes of appearance related to a facial expression is extracted. We extend PHOG descriptor which describes spatial variation of local shape to spatio-temporal domain to represent the variation of local shape

☆ This article is part of the Special Issue on TIUSM.

* Corresponding author.

E-mail addresses: xijian.fan@njfu.edu.cn (X. Fan), xbyang@njfu.edu.cn (X. Yang), yqlcom@njfu.edu.cn (Q. Ye), yangy@unm.edu (Y. Yang).

https://doi.org/10.1016/j.jvcir.2018.09.011
1047-3203/© 2018 Elsevier Inc. All rights reserved.

in the temporal dimension to form 3-dimensional (3D) descriptor. We refer this descriptor as PHOG_Three Orthogonal Planes (PHOG_TOP) [11]. In addition, by combining PHOG_TOP using weight function based on the discriminative information of facial regions, we develop a discriminative spatio-temporal features referred as weighted PHOG_TOP (WPHOG_TOP) to classify (recognise) facial expression. We summarised our main contributions as follows: (a) PHOG_TOP 3D descriptor, and (b) a discriminative framework that combines the dynamic information representing variation in facial appearance and discriminative information of various facial regions in spatial domain.

The remainder of paper is organised as follows. Section 2 presents a brief survey of previous related work. PHOG_TOP, discriminative weighting function, and WHOG_TOP descriptors are presented in Section 3. Section 4 presents the framework of proposed facial expression recognition and the experimental results, respectively. Conclusion are finally provided in Section 5.

2. Related work

A classical facial expression recognition framework consists of three steps: facial image pre-processing, feature extraction, and classification [14]. The step of feature extraction is crucial, which significantly affects the performance of recognition. There are numerous studies on features extraction, which can be classified into appearance-based and geometric-based methods. The features extracted using either approach aims to minimise intra-class variation of facial expressions, while maximising inter-class variations. Facial expression framework can be used in many intelligent system. For example, with the development of web, online financial transaction is becoming more and more popular. In addition, there will be probabilistic fraud. With the analysis of facial expression, an anti-fraud system can be designed according to the biology experiment that human expression can reflect human psychological activity. Thus, a state-of-the-art facial expression framework can reduce the probability of fraud effectively.

In the geometric-based method, shape and position information of facial landmarks or region are extracted to represent the face geometry [15-19]. Zhang et al. [16] utilise the geometric position of 34 fiducial points as facial features to represent the face geometry. Optical flow based methods have been widely used to detect the movements of facial landmarks by measuring the displacement of detected facial landmarks between two consecutive frames [20,21]. Geometric based methods are sensitive to noise, and strongly rely on tracking performance of facial landmarks.

In the appearance-based approach, facial texutre and appearance information can be represented using low level transformation to form feature vectors. Gabor wavelets [16] and local binary patterns (LBPs) [22] are two most popular representations in the appearace methods which can well describe the local appearance information of facial expressions. Gabor feature can be obtained by convolving the facial image with a group of filters, and are robust to alignment mistakes. However, the computation of Gabor feature is relativelly complex, and the dimensionality of the output might be large, which is in demand for extra dimensionality reduction [16]. 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 [23]. Zhao and Pietikainen [24] 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. Histogram of gradients (HOG) [25] was originally developed for person detection and object recognition, and then used for face recognition [26]. In the work of Lazebnik et al. [27], HOG descriptors are extracted from face image using a dense grid, and are used for face recognition. The PHOG proposed in [13] 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.

There are several approaches that subdivide a face image into a number of local regions, and adopt some feature extraction methods (e.g., LBP [12], scale invariant feature transformation (SIFT) [28], non-negative matrix factorisation [29], etc.) from local regions. All these methods simply concatenate the local features to form the final feature. However, the features extracted from different facial regions might have various contribution to the recognition of facial expression, where some regions such as mouth contain more discriminative information than others (e.g., nose). Thus, simply concatenating local features could ignore such discriminative information, and accordingly affect the recognition performance.

3. Methods

The proposed framework uses discriminative dynamic features referred as WPHOG _TOP, which gives a robust and accurate recognition of facial expressions.

3.1. PHOG_TOP descriptor

PHOG is a descriptor using edge information, which is first proposed for object classification [13]. Inspired by HOG [25], the PHOG descriptor not only takes the edge information, but also exploits the spatial layout information of the local shape using the image pyramid [27]. More specifically, edge contours of an image are extracted at different pyramid resolution level, and occurrences of gradient orientation of edges are counted to construct a gradient histogram [13]. The histograms from selected pyramid levels are then concatenated to form the final PHOG descriptor. Fig. 1 shows an example of a typical PHOG.

Facial expression is usually performed dynamically, thus its dynamic information is essential for its recognition. We propose a spatio-temporal descriptor PHOG_TOP to capure such dynamic information, which is part of our previous work [11]. This descriptor concatenates the three orthogonal planes XY, XT and YT to give PHOG_TOP, taking into account the co-occurrence statistics in these three planes [11]. For a easy understanding, we give the details on constructing this descriptor in this sub-section. The XY plane is used to extract the local spatial information, and the XT and YT planes are used to extract temporal information. A video sequence can be regarded as a stack of XY slices in the temporal dimension, and similarly for XT and YT slices but in the Y and X dimensions, respectively. The spatio-temporal PHOG over each slice in three orthogonal axes (i.e., XY_PHOG, XT_PHOG, and YT_PHOG) are separately obtained and then combined. Take XY_PHOG for instance, we compute PHOG descriptor in every single image from a video sequence, i.e., along the temporal axis. First, the Canny edge detector is employed to capture the edge information. The image region is divided into a set of spatial grid by repeatedly doubling the number of divisions along each axis. Thus, the grid at resolution level l has 2^l cells along each dimension [13].

The orientation of gradient for each grid at each resolution level are computed using a Sobel mask [25]. The histogram of edge orientations within an image sub-region is quantized into K bins, and magnitude information is added as a weight. The PHOG descriptor