Pattern Recognition Letters 49 (2014) 55–61

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# A novel phase congruency based descriptor for dynamic facial expression analysis ☆

Seyedehsamaneh Shojaeilangari [a,*], Wei-Yun Yau [b], Eam-Khwang Teoh [a]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] Institute for Infocomm Research, A*STAR, Singapore

ARTICLE INFO

*Article history:*
Received 25 March 2013
Available online 1 July 2014

*Keywords:*
Phase congruency
Spatio-temporal descriptor
Emotion recognition
Facial expression

ABSTRACT

Representation and classification of dynamic visual events in videos have been an active field of research. This work proposed a novel spatio-temporal descriptor based on phase congruency concept and applied it to recognize facial expression from video sequences. The proposed descriptor comprises histograms of dominant phase congruency over multiple 3D orientations to describe both spatial and temporal information of a dynamic event. The advantages of our proposed approach are local and dynamic processing, high accuracy, robustness to image scale variation, and illumination changes. We validated the performance of our proposed approach using the Cohn-Kanade ($CK^+$) database where we achieved 95.44% accuracy in detecting six basic emotions. The approach was also shown to increase classification rates over the baseline results for the AVEC 2011 video subchallenge in detecting four emotion dimensions. We also validated its robustness to illumination and scale variation using our own collected dataset.

## 1. Introduction

Detection of human facial expression from video sequences is a challenging problem due to real-world constrains such as background clutter, partial occlusion, viewpoint variations, scale changes, and lighting conditions. Finding a suitable feature representation is a vital step to model the facial expression and subsequently recognize it in a video sequence.

The traditional method for video representation is the extension of successful techniques used in static image analysis to support the dynamic requirement for video processing. 3D Scale Invariant Feature Transform (SIFT), spatio-temporal Local Binary Pattern (LBP), and spatio-temporal descriptor based on 3D gradient are typical examples of video representation successfully applied to facial affect analysis or human action recognition [1,2]. In this paper, we followed the same idea for feature representation by extending the phase congruency (PC) concept. We applied our proposed approach to dynamic facial emotion recognition from video sequences.

Local energy-based and phase-based models have emerged as successful tools to detect various image patterns such as step edges, corners, valleys, and lines. The phase-based feature extraction model proposes that the features of a signal are observed at locations where its Fourier components are in harmony. Such concept is also seen in the human visual system that the image features are perceived at points where the phase values of its Fourier components are maximally in congruence [3].

There are some advantages for PC-based feature extraction approaches over gradient-based techniques [4]. The gradient operators such as Prewitt, Sobel, Laplace, and Canny edge detector may fail to precisely identify and localize all image features, especially in region affected by illumination changes. Unlike the gradient-based approaches which look for sharp changes of image intensity, PC is a dimensionless quantity which is robust to image contrast and illumination changes.

In Fig. 1, we show the advantages of PC-based line detection over Canny and Sobel methods. This figure illustrates that PC is able to localize the sharp line similar to the gradient operators. However, for features that are not sharp (gradual intensity variation), PC is able to detect such feature better than the traditional gradient operators as shown in Fig. 1. Indeed, PC captures the discontinuities even at small intensity differences which might be missed by the typical image gradient-based edge descriptors. It can thus be useful for facial features detection including skin folds due to aging and expression.

This paper explores the effectiveness of PC-based feature representation for facial expression recognition from video sequences. The proposed descriptor, named Histogram of Dominant Phase

☆ This paper has been recommended for acceptance by G. Borgefors.

* Corresponding author. Address: 50 Nanyang Ave, S1-B4C-14, Electrical and Electronic Engineering Department NTU, Singapore 639798, Singapore. Tel.: +65 84534486; fax: +65 67933318.

*E-mail address:* seyedehs1@e.ntu.edu.sg (S. Shojaeilangari).

http://dx.doi.org/10.1016/j.patrec.2014.06.009
0167-8655/

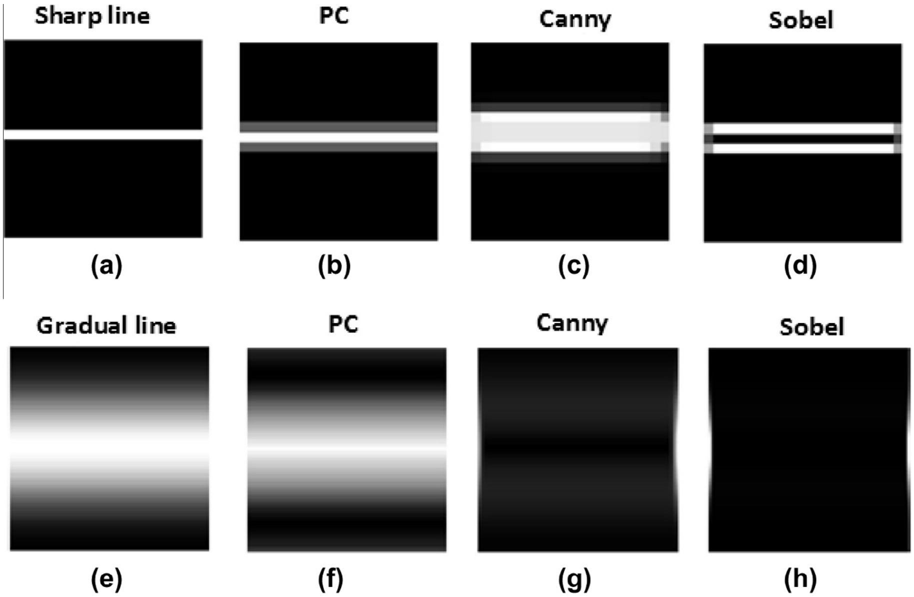

**Fig. 1.** Comparison of methods for line detection; (a) sharp line; (b) line detection based on phase congruency; (c) line detection based on Canny; (d) line detection based on Sobel; (e) gradual line with intensity range of [0 3]; (f) line detection based on phase congruency; (g) line detection based on Canny; (h) line detection based on Sobel.

Congruency (HDPC), comprises histograms of dominant PC over multiple 3D orientations to describe both the spatial and temporal information of a dynamic event.

To construct HDPC descriptor, the spatio-temporal PC values are calculated for multiple orientations. Therefore, each pixel of a video is characterized by multiple oriented PC values. Thus the PC values are able to encode various features at different scales and orientations for both spatial and time domains. After calculating the oriented PC values, the next step is to find the maximum PC for each pixel while preserving the dominant orientation. In other words, each pixel is represented by a vector where its length is equal to maximum PC, and its direction is determined by the dominant orientation. Keeping the dominant PC and its orientation information will preserve the key feature contributing to a dynamic event. The final step of our novel descriptor is building a local histogram of PC directions over all pixels over a spatio-temporal patch.

The novelties of our proposed approach are:

(1) Extending the PC concept to spatio-temporal domain to extract both static and dynamic information from a video sequence by applying the 3D log-Gabor filter.
(2) Designing a bank of oriented 3D log-Gabor filters to detect the image features at various orientations.
(3) Selecting dominant PC to capture the most significant motion information while preserving its direction.
(4) Proposing histogram of dominant spatio-temporal PC to summarize the acquired information of each local 3D region.

This paper is organized as follows: Section 2 summarizes the literature review. The proposed method for feature extraction is explained in Section 3, including computing the 3D PC for sequenced images, the proposed HDPC algorithm, and summary of the algorithm's properties. Sections 4 and 5 describe the experimental results and conclusion respectively.

## 2. Related works

Automated analysis of facial expression has been the subject of many researches due to its potential applications such as human–computer interaction, automated tutoring systems, image and video retrieval, smart environments, and driver warning systems. Although considerable progress has been reported in the literature, there are still challenges for a robust and automated analysis. Most previous works focused on facial emotion recognition via static images. The static analysis systems ignore the dynamics of facial expression due to expensive computational time involved or the complicated temporal mode [5–9]. However, it is confirmed by human visual system that the judgement about an expression is more reliable when its temporal information is also taken into account [10].

To exploit the temporal information of facial expression, different techniques have been developed. There were several reported attempts to track the facial expression over time for emotion recognition via Hidden Markov Models (HMM). A multilevel HMM is introduced by Cohen et al. to automatically segment the video and perform emotion recognition [11]. Their experimental results indicated that the multilevel HMM have better performance than the one layered HMM. Cohen et al. introduced a new architecture of HMMs for automatic segmentation and recognition of human facial expression from live videos [12].

Dynamic Bayesian Networks (DBN) is another successful method for sequence-based expression analysis. Ko and Sim developed a facial expression recognition system based on combining the Active Appearance Model (AAM) for feature extraction and DBN for modeling and analyzing the temporal phase of an expression [13,14]. They claimed that their proposed approach is able to achieve robust categorization of missing and uncertain data and temporal evolution of the image sequences.

Optical flow (OF) is also a widely used approach for facial features tracking and dynamic expression recognition. Cohn et al. developed an OF based approach to automatically discriminate the subtle changes in facial expression [15]. They considered sensitivity to subtle motion when designing the OF which is crucial for spontaneous emotion detection.

Methods based on local features or interest points such as SIFT have shown to perform well for object recognition. These methods were further extended to video analysis. Camara-Chavez and Araujo proposed a method for event detection in a video stream by combining Harris-SIFT with motion information in the context of human action recognition [16]. They used the Harris corner detection for key-point extraction and the phase correlation method was used to measure the motion information.