Neural Networks 105 (2018) 36–51

Contents lists available at ScienceDirect

Neural Networks

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

Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy

Bo Sun, Siming Cao, Jun He *, Lejun Yu

College of Information Science and Technology, Beijing Normal University, Beijing, 100875, China

ARTICLE INFO

Article history:
Received 7 May 2017
Received in revised form 21 November 2017
Accepted 28 November 2017
Available online 7 December 2017

Keywords:
Affect recognition
Deep learning
Convolutional neural network
Bilateral long short-term memory recurrent neural network
Deep spatio-temporal hierarchical feature
Multi-modal feature fusion strategy

ABSTRACT

Affect presentation is periodic and multi-modal, such as through facial movements, body gestures, and so on. Studies have shown that temporal selection and multi-modal combinations may benefit affect recognition. In this article, we therefore propose a spatio-temporal fusion model that extracts spatio-temporal hierarchical features based on select expressive components. In addition, a multi-modal hierarchical fusion strategy is presented. Our model learns the spatio-temporal hierarchical features from videos by a proposed deep network, which combines a convolutional neural networks (CNN), bilateral long short-term memory recurrent neural networks (BLSTM-RNN) with principal component analysis (PCA). Our approach handles each video as a "video sentence." It first obtains a skeleton with the temporal selection process and then segments key words with a certain sliding window. Finally, it obtains the features with a deep network comprised of a video-skeleton and video-words. Our model combines the feature level and decision level fusion for fusing the multi-modal information. Experimental results showed that our model improved the multi-modal affect recognition accuracy rate from 95.13% in existing literature to 99.57% on a face and body (FABO) database, our results have been increased by 4.44%, and it obtained a macro average accuracy (MAA) up to 99.71%.

1. Introduction

The affect recognition ability is an important aspect of computer intelligence. It primarily influences the computer's response to an operator or interlocutor. And it has a wide range of applications in entertainment, industry, transportation, medicine, military and many other fields. Over the past few decades, several affect recognition methodologies have been proposed. The research has led to two key trends toward greater practicality, i.e., use of multi-modal information instead of mono-modal, and dynamic video instead of static images.

In this research, initially, American psychologists Ekman et al. defined six basic categories of emotions, i.e. angry, disgust, fear, happy, sad and surprise (Ekman & Friesen, 1978). Several years later, they developed the Facial Action Encoding System (FACS) (Ekman & Friesen, 1971). In this system, a facial expression is deemed the result of facial muscles and the combination of many action units (AUs), which display the corresponding relationship between facial movement and expression. The two works marked key milestones in the field and have continued to serve as the basis of emotion recognition, especially in facial emotion recognition research. Subsequently, many related algorithms and systems have been proposed (Cheon & Kim, 2009; Glowinski, Dael, Camurri, et al., 2011; Liu, Han, Meng, et al., 2014; Nicolaou, Gunes, & Pantic, 2012; Taheri, Patel, & Chellappa, 2013; Zhong, Liu, Yang, et al., 2014).

Recently, deep learning methods have become widely used in the field of computer vision. The convolutional neural network (CNN) and bilateral long short-term memory recurrent neural network (BLSTM-RNN) are state-of-the-art machine-learning techniques in this area. Fan, Lu, Li, et al. (2016) present a video-based emotion recognition system using CNN–RNN and C3D hybrid networks. Chen, Zhang, and Allebach (2015) explored two simple, yet effective deep-learning-based methods for image emotion analysis. Noroozi et al. (1949) applied a CNN to obtain key frames for summarizing videos.

However, human emotion expression manifests in multi-modal, not mono-modal, information, such as facial movements, body gestures, voice utterances, etc. Each mono-modal is often ambiguous, uncertain, and incomplete. Metallinou et al. (2012) thus examined context-sensitive schemes for emotion recognition in a multi-modal, hierarchical approach referred to as a bidirectional long short-term memory (BLSTM) neural network. In addition, Kret, Roelofs, Stekelenburg, and de Gelder (2013) performed a psychological analysis of body movements for body expression recognition. They showed that using only facial expressions can be

* Corresponding author.
E-mail addresses: tosunbo@bnu.edu.cn (B. Sun), caosiming@mail.bnu.edu.cn (S. Cao), hejun@bnu.edu.cn (J. He), yulejun@bnu.edu.cn (L. Yu).

https://doi.org/10.1016/j.neunet.2017.11.021
0893-6080/

misleading, whereas combining them in some way can improve the emotional state recognition accuracy. Moreover, Neverova et al. (2016) proposed a method for adaptive multi-modal gesture recognition. They showed that fusing multiple modalities leads to a significant increase in recognition rates and that information items from the individual channels have complementary characteristics.

In recent years, to advance multi-modal affect recognition, many multi-modal emotion recognition competitions have been organized, such as the Emotion Recognition in the Wild Challenge (EmotiW) (Dhall, Ramana Murthy, Goecke, Joshi, & Gedeon, 2015), Audio/Visual Emotion Challenge (AVEC) (Valstar, Gratch, Schuller, et al., 2016), and Multimodal Emotion Recognition Challenge (Li, Tao, Schuller, et al., 2016). Furthermore, the idea of combining multi-modalities for affect recognition has generated a new research topic, specifically determining which modalities should be used, and how to effectively integrate them. Some researchers have initially focused on fusing visual and audio modalities (Bengio, 2004; Castellano, Kessous, & Caridakis, 2008; Pan, Levinson, Huang, & Liang, 2004; Sebe, Cohen, Gevers, & Huang, 2006). Later, others have explored utilizing audio, visual, and physiological signals synchronously for recognizing affects (Brady, Gwon, Khorrami, et al., 2016; Chen & Jin, 2016; He, Jiang, Yang, et al., 2015).

In this regard, Ambady and Rosenthal suggested that visual channels, i.e., facial movement and body gestures, are the most important cues for classification of human behavior (Ambady & Rosenthal, 1992). Consequently, some researchers have suggested that fusing cues can produce better affect recognition results. Accordingly, corresponding studies have been conducted (Chen, Tian, Liu, et al., 2011; Gunes & Piccardi, 2007, 2009; Kapoor & Picard, 2005; Karpouzis et al., 2007; Shan, Gong, & Mcowan, 2007). To date, two major fusion strategies exist, namely feature-level fusion and decision-level fusion. Feature-level fusion directly combines the discriminative ability of multiple features, which is assumed to be more suitable for modalities that are almost synchronous in the timescale (e.g., speech and lip movements) (Wu, Oviatt, & Cohen, 1999). Decision-level fusion combines the discriminative results of multiple features. This approach is assumed to be more suitable for modalities that do not simultaneously occur in the timescale. (e.g., speech and body gestures) (Wu et al., 1999).

In studies based on the face and body (FABO) database, Gunes & Piccardi (2009) separately applied feature-level fusion and decision-level fusion. Meanwhile, Barros, Jirak, Weber, et al. (2015) used a fully connected layer to fuse each multi-modal stream, while Chen et al. (2011) used only feature-level fusion. A means of integrating facial movements and body gestures needs further development.

The actions of facial movements or body gestures comprise a dynamic process, which can be described by four temporal phases: neutral, onset, apex, and offset (Ekman, 1979). Affect recognition based on videos contains spatio-temporal information compared with affect recognition based on static images. In studies based on the FABO database (Gunes & Piccardi, 2006b), Barros et al. (2015) selected several apex frames for spatio-temporal feature extracting. Moreover, Chen et al. (2011) proposed a framework to extract the temporal dynamic features of face and body gestures from the whole video. However, a method of exploring effective spatio-temporal features requires further exploration.

To address the above issues, we propose a spatio-temporal fusion model, which not only extracts the high-level spatio-temporal hierarchical features, but it also includes a multi-modal hierarchical fusion strategy. This paper provides the following key contributions:

(1) To extract effective spatio-temporal hierarchical features, we propose a temporal selection approach to obtain expressive materials. First, we employ the onset–apex–offset sequence as a video-skeleton. Then, using the sliding window strategy, we obtain several video-words from the video-skeleton. We describe conducted experiments that confirmed that our proposed temporal selection approach is notably more effective than previous methods.
(2) Based on the expressive materials, we extract deep spatio-temporal features from the video-skeleton and video-words by respectively using a proposed network, which combines CNN, BLSTM-RNN, and principal component analysis (PCA).
(3) We propose a hierarchical fusion method by combining feature-level fusion and decision-level fusion for visual multi-modal affect recognition. The method is based on facial movements and body gestures. It showed excellent performance in conducted experiments. We evaluated the proposed method on the FABO database. The proposed method performed better than existing state-of-the-art methods for visual multi-modal affect recognition.

The remainder of this paper is organized as follows. Section 2 introduces related works. Section 3 describes the details of the overall methodology we proposed. The performed experiments and extensive experimental results are detailed in Section 4. Finally, our conclusions are given in Section 5.

2. Related work

In this section, we firstly review some existing methods of mono-modal affect recognition from facial movements or body gestures. Secondly, we review some existing works on visual multi-modal recognition.

2.1. Mono-modal affect recognition

As above mentioned, the original studies on the affect recognition algorithm are based on single-mode static images, especially face images. To date, many studies have been conducted on facial expressions. Zhong et al. (2014) proposed a method to divide a face image into blocks of different scales. Then, similar and special ones were selected from among different expressions through learning to identify the most representative areas. Cheon and Kim (2009) proposed an algorithm for facial expression recognition based on a differential active appearance model and manifold learning. Liu et al. (2014) proposed an improved depth learning method, which can extract a series of representative facial features through repeated learning and training. A strong boosted classifier with statistical properties is then formed. In addition, Sun et al. (2016a) extracted several CNN features for continuous affect recognition. They also extracted acoustic features, LBP from three orthogonal planes (LBPTOP), Dense SIFT and CNN-LSTM features to recognize the emotions of film characters (Sun et al., 2016b). Guo et al. (2017) proposed a multi-modality convolutional neural networks (CNNs) based on visual and geometrical information for micro emotion recognition. Schwan et al. (2017) described an advanced pre-processing algorithm for facial images and a transfer learning mechanism for face emotion recognition.

Compared to research on facial expressions, few body expression studies have been undertaken. This may be because it is difficult to accurately and reliably define the corresponding relationships of various body gestures to emotional categories. Some researchers in psychology, cognitive science, and computer science have studied emotion recognition based on body gestures, and effective systems have been presented for body emotion recognition. Glowinski et al. (2011) studied the association between gestures and affective changes. They first coded the upper extremity changes of the human body, and then expressed the emotion through specific gestures. Nicolaou et al. (2012) studied head movements. They mapped the angle and direction of the head motion to the emotions in the emotional space to produce emotional