Image and Vision Computing 32 (2014) 107–119

Contents lists available at ScienceDirect

Image and Vision Computing

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

Facial expression recognition experiments with data from television broadcasts and the World Wide Web☆

Ligang Zhang [a,b,*], Dian Tjondronegoro [b], Vinod Chandran [b]

[a] Faculty of Computer Science and Engineering, Xi'an University of Technology, 5 South Jinhua Road, Xi'an 710048, China
[b] Science and Engineering Faculty, Queensland University of Technology, 2 George St, Brisbane 4000, Australia

ARTICLE INFO

Article history:
Received 9 March 2013
Received in revised form 14 November 2013
Accepted 13 December 2013

Keywords:
Facial expression recognition
Realistic
Texture
Geometry
Experiment

ABSTRACT

Facial expression recognition (FER) systems must ultimately work on real data in uncontrolled environments although most research studies have been conducted on lab-based data with posed or evoked facial expressions obtained in pre-set laboratory environments. It is very difficult to obtain data in real-world situations because privacy laws prevent unauthorized capture and use of video from events such as funerals, birthday parties, marriages etc. It is a challenge to acquire such data on a scale large enough for benchmarking algorithms. Although video obtained from TV or movies or postings on the World Wide Web may also contain 'acted' emotions and facial expressions, they may be more 'realistic' than lab-based data currently used by most researchers. Or is it? One way of testing this is to compare feature distributions and FER performance. This paper describes a database that has been collected from television broadcasts and the World Wide Web containing a range of environmental and facial variations expected in real conditions and uses it to answer this question. A fully automatic system that uses a fusion based approach for FER on such data is introduced for performance evaluation. Performance improvements arising from the fusion of point-based texture and geometry features, and the robustness to image scale variations are experimentally evaluated on this image and video dataset. Differences in FER performance between lab-based and realistic data, between different feature sets, and between different train-test data splits are investigated.

1. Introduction

The natural expression of an emotion on the face is a highly reliable component of human perception in social interactions. An angry look conveys more than words can. Facial expression recognition (FER) with computers has potential applications in human computer natural interfaces, video surveillance etc. For these applications to be realized, FER algorithms must work on real emotion data, which reflects the real reaction of humans and real environmental conditions more closely than lab-based data, which is normally collected under a pre-set simplified laboratory environment [1] and is a product of acting.

The vast majority of FER studies have focused on lab-based data. It is difficult to obtain facial expression data from real world events on a large scale for benchmarking because of privacy laws. If data are sourced from TV broadcasts and web sources, privacy is less of an issue but copyrights must be obtained. There are also challenges because such data are not constrained and will have large variations in lighting, pose, etc. Current FER methodologies developed on lab collected data are not designed to generalize to such scenarios. However, even before these processing challenges are addressed there are some questions to be answered:

(1) Are facial expressions in video obtained from TV and World Wide Web sources different from those in lab collected data? In what aspects?
(2) Is the recognition performance on such data significantly different using different types of features, and different FER algorithms?
(3) Does the performance differ between different types of realistic data?

In this work, a database from television (TV) broadcasts and the World Wide Web (referred to as the QUT FER dataset) has been collected. Two public lab-based databases—FEEDTUM and NVIE, and one realistic SFEW database are used for the purpose of comparisons of feature and performance differences between lab-based and realistic data and between different realistic data. The evaluation is based on an automatic approach that is specifically designed using fusion of texture (scale invariant feature transform—SIFT) and geometry (facial animation parameters—FAP-based distances). The main objective is to explore the differences in FER performance to provide new insights as

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.

* Corresponding author at: Faculty of Computer Science and Engineering, Xi'an University of Technology, 5 South Jinhua Road, Xi'an 710048, China. Tel.: +61 7 3138 5074.
E-mail addresses: ligzhang@gmail.com (L. Zhang), dian@qut.edu.au (D. Tjondronegoro), v.chandran@qut.edu.au (V. Chandran).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.12.008

Fig. 1. Samples for lab-based and realistic facial expressions (from the CK, NVIE and QUT databases).

the data moves from lab-based to more realistic. The contributions are original primarily in the range covered by the data and the comparisons.

The rest of the paper is organized as follows. Section 2 gives a brief review of existing realistic FER databases. Section 3 describes the collection of the QUT FER database and explores the difference in feature distributions between lab-based and realistic data. Section 4 introduces a benchmark approach and reports its performance across different datasets, feature sets and train-test splits. Performance is also compared with previous approaches. Conclusions are drawn in Section 5.

2. Overview of realistic FER databases

This section reviews available realistic databases. A recent comprehensive survey of FER can be found in [2].

Existing FER databases approximately fall into two categories: *lab-based* data where the emotions are intentionally expressed by selected actors or deliberately induced by outside perceived stimuli in a pre-set laboratory environment, and *realistic* data where the emotions naturally occur in real-world conditions. The majority of current FER databases belong to the first category, and popular examples include the Cohn–Kanade (CK), CK+, JAFFE, FEEDTUM, BU-3DFE, Semaine, MMI, UT-Dallas, SAL, NVIE, RU-FACS, AAI, PETS2003 and GEMEP-FERA. Compared with lab-based emotions, realistic expressions (samples shown in Fig. 1) are also often accompanied by big variations in face pose, size, illumination, facial occlusions etc., and thus they are more challenging to classify and hold more significance in real-world applications.

Table 1 reviews typical examples of existing realistic databases. Early attempts at collecting realistic data, such as luggage lost and Belfast, tried to capture facial reactions in real scenarios, such as interviews and conversations between subjects. Recent studies have collected emotional video segments from TV broadcasts or the World Wide Web to more closely mimic real situations, yielding databases including HUMAINE [1], VAM, TV data [3], acted facial expressions in the wild (AFEW) [4], static facial expressions in the wild (SFEW) [5], happy people images (HAPPEI) [6], GENKI-4K [7] and Gv [8]. Video frames in the HUMAINE and VAM databases are not directly annotated with basic emotion categories. TV broadcast data are presented by Yeasin et al. in [3] but details have not been provided. The AFEW is a relatively new dataset comprising of both a video set and an image set collected from 37 movies. The SFEW database contains 700 images collected from frames of AFEW videos. The HAPPEI database is composed of 4600 images collected from the Flickr website and was built for happiness intensity estimation for groups of people. The GENKI-4K images are primarily for smile, while details of Gv images have not been given. All the above databases are collected *from only one source* (i.e. either TV programs or movies, or the web), or are annotated with a *limited number of emotion types* (such as only happiness in the HAPPEI database). In this work, a realistic QUT FER database has been compiled with some unique characteristics:

(1) Data collected from three sources (i.e. TV drama, TV news and the web), which contain variations in illumination, pose etc. closer to those expected in real applications.
(2) More types of emotions (i.e. positive, negative and neutral), which are particularly significant to specific applications such as sentiment analysis, as not all realistic emotions can be categorized into the six basic emotions.
(3) Intensity for nine emotions, which can be used for evaluating and estimating the level of emotional reactions of subjects to outside stimulus.

3. Compiling the realistic QUT FER database

This section describes the QUT FER realistic database collected from web-based and broadcast video recordings. Feature distributions for three emotions in the database are compared with those in two public lab-based FER databases—FEEDTUM and NVIE.

3.1. Video selection and segmentation

To collect video segments from real multi-media materials, we chose to use three sources: TV News, TV drama, and YouTube. These sources are publicly available, represent real day-to-day situations, and contain lots of interpersonal communications and performances, which lead to a wide range of emotional states. As a result, these materials can provide a lot of basic and non-basic emotions in various real conditions, and cover a wide range of people with different ages. Table 2 shows the number of videos and contents for each source.

All video segments are converted to the AVI format with the original resolution ranging from 480 × 360 to 1024 × 576 pixels and a frame rate from 14 fps to 25 fps. The Video Splitter software is used to

Table 1
Overview of realistic FER databases.

Database	Source	Emotion	Subject	Data size
Luggage lost [9]	Airport	Humor, sadness, anger, stress and indifferences	109	209 videos
Belfast [10]	TV	Activation and evaluation	125	298 videos
Yeasin et al. [3]	TV	Six basic emotions	N/A	108 videos
VAM [11]	TV	Valence, activation and dominance	104	1421 videos/1872 images
HUMAINE [1]	TV	Emotion words, intensity, activation and valence etc.	48	48 videos
SFEW [5]	TV	Six basic emotions and neutral	N/A	700 images
AFEW [4]	TV	Six basic emotions and neutral	330	1426 videos
HAPPEI [6]	Flickr	Happiness (six stages)	N/A	4600 images
GENKI [7]	Web	Smile	N/A	4000 images
Gv [8]	Web	Six basic emotions and neutral	328	350 images
QUT (this work)	TV/web	Six basic emotions, neutral, positive and negative	219	458 videos/2927 images