Multimodal Intelligent Affect Detection with Kinect (Doctoral Consortium)

Yang Zhang
Northumbria University
United Kingdom
Yang4.zhang@northumbria.ac.uk

Dr. Li Zhang
Northumbria University
United Kingdom
li.zhang@northumbria.ac.uk

Prof. Alamgir Hossain
Northumbria University
United Kingdom
alamgir.hossain@northumbria.ac.uk

Abstract

Communication between human beings involves complex and rich means. In the past decades, computers have successfully supported human in a variety of tasks such as calculating and memorizing. However, when confronted with the demand of multimodal interaction with users, can these indispensable partners make us satisfied? This research might answer this question.

Categories and Subject Descriptors

I.2 [ARTIFICIAL INTELLIGENCE]: Miscellaneous;

Keywords

Affective computing, multimodal affect sensing and analysis, emotion theory.

1. Multimodal Affective Intercommunion

Before we try to teach a machine to detect and recognize human emotional signals, let us consider the following scenario: A little baby girl starts crying. Then her mother comes and sits beside. When the baby notices her mother is focusing on her, she points at a toy lying on the floor. The mother smiles sympathetically, bends over, picks up the toy and returns it to her baby. The baby stops crying, excited by the return of the treasure and grabs it greedily.

This scene appears to be the most common events in daily life. But please take down a few particulars: consider the modes of communication between the mother and the baby. When we try to implement a system where the mother or the baby or both of them are replaced by AI to simulate the communication, we can realize quickly the complexity of the communication methods utilized by the two participants. For example [1], in order to understand why the baby cries and what is her intention, we should employ facial expression recognition to infer the baby's emotion, gesture detection to comprehend the gesture of pointing to the toy and even audio processing to extract information contained in the cry. These actions are an indispensable aspect of human beings, and we have taken for granted the level of sophistication until we begin to train computers to deal with the same level of situations.

Let us take a closer look of another more complex scenario: You walk into an office where two people are in the middle of a conversation. One of them says a single word: 'excellent'. They invite you to join the conversation. How can you contribute? 'Excellent' means very well or agreeableness commonly, but it can be also an irony that implies totally different meanings. Without the benefit of the emotional status expressed by facial expressions, body gestures, and other nonverbal cues of both of the speaker and the audience and the context of the conversation, it is difficult to understand the meaning of a word or a sentence. In order to comprehend the intended meaning in a communication, there has to be some level of emotional and context understanding. This concept is the core goal and main motivation of this research.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. Kinect – A Powerful Research Tool

Microsoft Kinect and the Kinect for Windows SDK [2] provide researchers with incredible capabilities in a relatively inexpensive package. The Kinect device includes a color camera, a depth-sensing camera, and an array of four microphones. The following is a list of Kinect technical specifications:

- Color VGA motion camera: 1280x960 pixel resolution
- Depth camera: 640x480 pixel resolution at 30 fps
- Array of four microphones
- Field of view:
  - Horizontal field of view: 57 degrees
  - Vertical field of view: 43 degrees
  - Depth sensor range: 1.2m - 3.5m
- Skeletal Tracking System
- Face Tracking, track human faces in real time

3. Related Work

Recently, many researches and business applications aiming at making computers sensitive to users' emotional states have emerged. Some publicly available developments and applications have successfully collected and annotated a corpus of emotional states (usually segments of videos with labels). However, many of them focus on analyzing one or some particular gestures or expressions rather than addressing the issues of multimodal emotion recognition. Such as Cam3D [5], authors mainly analyze hand-over-face gestures and possible meanings.

Table 1 Overview of related work and databases

Works	CK+[3]	SAL[4]	Cam3D[5]	Ours
3D?	N	N	Y	Y
Video Data Type	C	C	C	C/D
Modalities	F	F/A	F/G/A	F/G/A
Spontaneity	P/S	S	S	P/S
Number of emotional states	6	N/A	12	7*
Number of subjects	210	24	7	5*
Number of videos	700	10 hours	108	350*
Emotional description	B	D	B/C	C

Table 1 lists some newest related work together with the databases used for easy comparison with our project. (*Because our research just started, the number of emotional states, subjects and videos could increase in future)

Symbols and letters shown in Table 1 are explained here: Video Data Type: C: color, D: Depth; **Modalities**: F: face, G: body gesture, A: audio; **Spontaneity**: S: spontaneous, P: posed; **Emotional description**: B: basic, C: complex, D: dimensional

4. Multi-modal Data Acquisition

We use Microsoft Kinect together with Kinect Studio v1.6.0 for multi-modal data collection (see figure 1). This novel tool can record all the data coming into an application from a Kinect unit, including raw data of color stream, depth stream and audio stream. We can then view, review, segment and store the data. Kinect Studio also has the capability of injecting the captured data streams back into a Kinect-enabled application, which enables easy data files sharing, testing and database establishment.

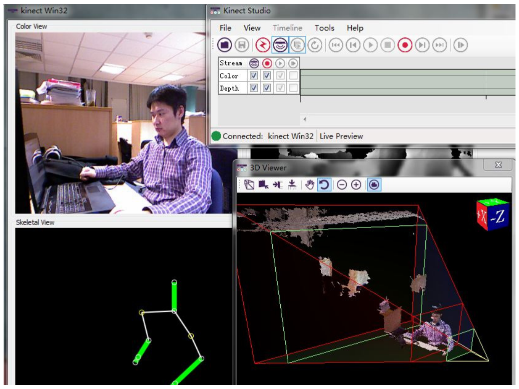

Figure 1 Kinect Studio v1.6.0

In the initial research, we choose the six universally recognizable emotions and painful as the results set (R_I = {disgust, sadness, happiness, fear, anger, surprise, painful}). For each emotion category, we record ten equal length segments of videos from five actors respectively (10 seconds per segment). From these posed videos, a total of 7×5×10 = 350 segments were collected.

In future search, this database can be extended by: adding more emotion categories in the results set, obtaining more video segments from different participants, and most importantly, exploring effective methods to segment spontaneous videos and validate them (e.g. each segment is labeled by a single event such as a change in facial expression or gesture [5]).

5. The First Stage - Facial Expression

We first look at facial expression in depth, because 1) some research shows that facial expression could contribute a large proportion of the whole affective communication [6]; 2) the meanings conveyed by languages or gestures can be totally different between different cultures, but facial expression has relatively universal interpretations [7].

Many existing facial expression recognition systems are based on either geometric or appearance facial features extraction. Both approaches take advantages of the most advanced Machine Vision Techniques, however, the lack of clinical and psychological foundation makes these systems failed in persuasiveness and proof when faced with so many critical issues raised by cognitive and psychological scientists [8].

We introduce an intermediary, known as Facial Action Coding System (FACS) [9] to bridge the gap. FACS is an objective and comprehensive system based on over 30 years' research of experimental psychologists, which initially aims to provide expert human observers rather than a computer with an objective measures of facial activities. Within the FACS, 44 individual action units (AUs) are defined to describe each facial action. Even though each individual AU does not have a strong link with emotions, the combinations of facial Action Units (over 7000 in total) can represent thousands of facial expressions.

We try to develop a FACS based model to map tracked facial features (mainly including 87 feature points in X and Y coordinates and 3D head pose) to AUs. We focus on only the actions that have been implicated as possibly related to the listed emotions. Each AU is represented as a numeric weight varying between -1 and +1. An initial input dataset with 8 elements (AUs = {AU2, AU4, AU10, AU13, AU16, AU20, AU26, AU27}) is tested by a Neural Network based facial emotion recognizer which appears to have a reasonable recognition result.

6. Future Work and Challenges

We present the collection and establishment of a 3D multi-modal database for emotion recognition. The database is demonstrated to be effective, thus adding more data to our corpus will be a part of future work. We develop a FACS based model with strong clinical background and have done some preliminary tests. In addition, the complement and optimization of this modal will be a key for effective facial expression interpretation. Therefore, we are exploring affect detection from other modalities: gesture and audio signals.

The multimodal fusion could also be a great challenge; the issues include the estimation of the reliability of each modality, the optimal level of merging different streams, and the construction of functions for the integration.

7. REFERENCES

[1] Holmquest, L. 2012. *Context-Aware Dialogue with Kinect. MSDN Magazine*, April 2012 Issue, Microsoft Corporation.

[2] Webb, J. and Ashley, J. 2012, *Beginning Kinect programming with the Microsoft Kinect SDK*, USA, Published by Apress.

[3] Lucey, P., 2010. *The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression*. In: CVPRW. pp. 94 - 101. IEEE.

[4] McKeown, G., Valstar, M.F., Cowie, R., and Pantic, M., 2010. *The SEMAINE corpus of emotionally colored character interactions*. In: ICME. pp. 1079 - 1084. IEEE.

[5] Mahmoud, M., Baltrusaitis, T., Robinson, P., and Riek, L. D. 2011. *3D Corpus of Spontaneous Complex Mental States. Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. pp. 205 - 214.

[6] Mehrabian, A. 1968. *Communication without Words, Psychology Today*. 2, 4. pp. 53 – 56.

[7] Ekman, P., and Friesen, W.V. 1971. *Constants across cultures in the face and emotion*. J. Personality Social Psychology. 17, 2. pp. 124 – 129.

[8] Kappas, A. 2010. *Smile When You Read This, Whether You Like It or Not*: Conceptual Challenges to Affect Detection. Affective Computing. 1, 1. pp. 38 – 42.

[9] P. Ekman, W.V. and Friesen, J.C. Hager. 2002. *Facial action coding system: Research Nexus. Network Research Information*. Salt Lake City.