Benchmark Driven Framework for Development of Emotion Sensing Support Systems

Senya Polikovsky*, Maria Alejandra Quiros-Ramirez*, Yoshinori Kameda*, Yuichi Ohta* and Judee Burgoon†
*Department of Intelligent Interaction Technologies, University of Tsukuba
senya@image.iit.tsukuba.ac.jp, alejandra@fhuman.esys.tsukuba.ac.jp, {kameda,ohta}@iit.tsukuba.ac.jp
†Eller College of Management, University of Arizona, jburgoon@cmi.arizona.edu

***Abstract*—Emotion sensing support system to assist human decision making during interview scenario is a developing research field. This paper presents a new framework for the development of emotion sensing support systems that is a complete, easily extendible, flexible, and configurable environment with intensive benchmark capabilities. The design of the framework was inspired by behavior-driven development, agile software development technique. It provides: 1) effective collaboration platform between technological and psychological researches, and 2) intensive benchmarking capabilities to test the performance of the entire system as well as individual algorithms.**

I. Introduction

It is still an ongoing debate in the psychological community if physiological changes are reliable and universal indicators of human emotional state. However, in practice, police and border control officers use behavioral and physiological clues in their daily work.

In this paper, we describe a framework for the development of emotion sensing support systems (ESSS) to be used for assisting human in decision-making process during interview scenarios. Such systems aim to support police and border control officers during questioning and doctors during medical diagnosis, among others. In addition to assist the specialists to reach the correct conclusion, the use of ESSS will improve the skills of the specialists who use it.

ESSS apply psycho-physiological models based on data extracted from a wide range of sensors, for example hi-speed camera, infrared camera, array of high resolution cameras, pressure sensors, microphones, etc. Challenges in the development of such multi-modal sensory systems include synchronization of data followed by their integration and analysis using a variety of algorithms.

The development of a fully automatic virtual agent for questioning procedures during checkpoint crossing is also under investigation [1]. Nevertheless, we consider ESSS systems are an essential step in the development of fully automatic systems. Operation of ESSS will allow the collection of a large amount of real-world data required in order to develop more sophisticated psychophysiological models for fully automatic systems.

Basic characteristics of a framework for ESSS development are modularity, simple addition and replacement of new algorithms, support of several fusion strategies and configurability. A recent overview of the emotion sensing field is presented in [2], which introduces the importance of creating realistic multi-modal emotional databases. This review also presents the newest trends in feature extraction algorithms and multi-modal emotion classification strategies. A survey on development environments for real time multi-modal programs is presented in [3]. HephaisTK [4], HCI2 [5], and Open-Interface [6] are recent implementations of multi-modal programming environments, they are useful tools for rapid development and they include facilities for easier configuration as well as different levels of data fusion.

However, these environments do not focus on introducing standardized benchmarks for neither the individual algorithms nor the complete process of emotion detection. There are several examples of fields where the introduction of well-organized benchmarks led to strong improvement: supercomputing, CCTV monitoring [7] and recently face detection. In our opinion, the establishment of benchmarks is essential for the development of ESSS.

Our framework design was mostly inspired by Behavior-Driven Development (BDD) which is an agile software development technique devised by Dan North [8]. It describes a cycle of interactions with well-defined outputs, resulting in the delivery of working, tested software.

The originality of our framework consists in organizing its development into separate steps, each associated with a standardized benchmark. This organization allows measuring the significance of every step of emotion detection and evaluating the performance of all individual algorithms implemented in the system. It will support the coordination between technological and psychological researchers, and simplify the search for the right balance between the complexity of feature tracking algorithms and the quality of psychophysiological models.

II. Benchmark Driven Framework

This section introduces a benchmark driven framework for development of ESSS. An overview of the framework is presented in figure 1. The framework is divided into four steps: data capturing, feature extraction, feature analysis and classification. Each step is associated with a standardized benchmark that includes variety of ground truth notations and a programmable interface for related data access.

The basic concept behind the benchmarks is the supply of input data to algorithms treated as interchangeable black box programs followed by comparative analysis between outputs and ground-truth. To increase the flexibility of the tests

978-0-7695-4782-4/12 $26.00 © 2012 IEEE
DOI 10.1109/EISIC.2012.27

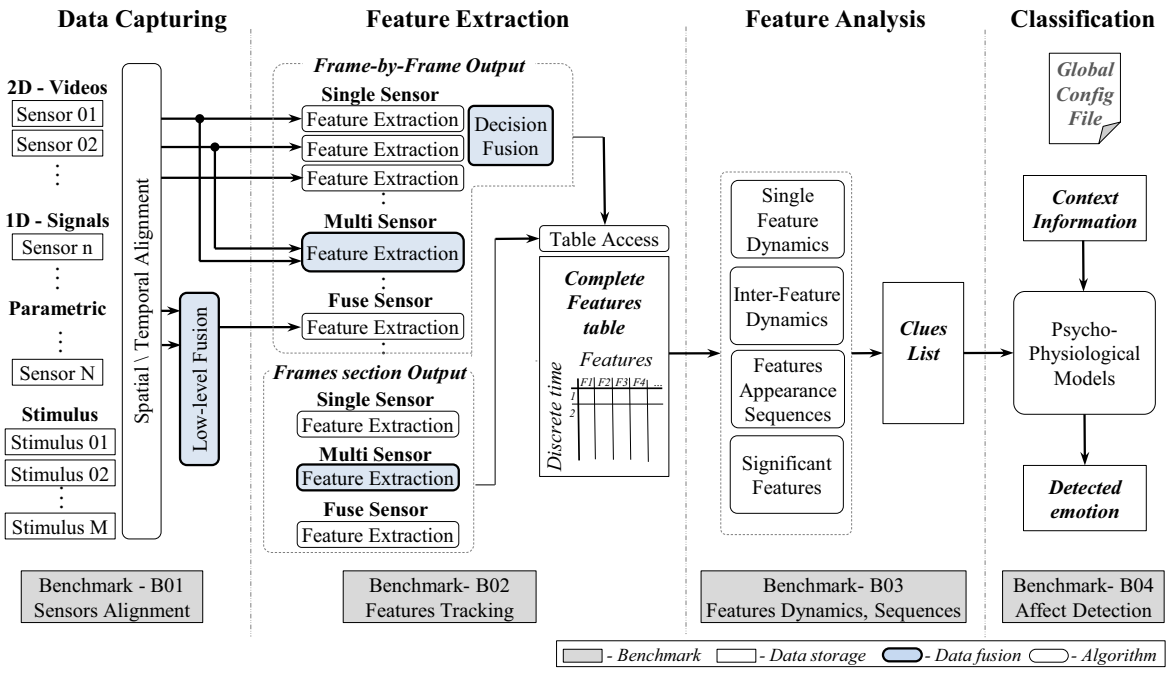

Fig. 1. Benchmark driven development framework flowchart. The framework consist of four benchmarkable steps: 1) *Data Capturing* - sensor manipulation and raw data retrieval; 2) *Feature Extraction* - data fusion and feature detection / tracking; 3) *Feature Analysis* - adaptation between extracted features and clues; 4) *Classification* - application of psychophysiological models. Algorithms communication is implemented through feature table located in shared memory. All steps support high configurability via global configuration file. 'Feature' refers to the measurements made using raw sensory data, while the term 'clue' corresponds to the input data of the psychophysiological models. 'Context' refers to variables of the dynamics of the interaction and subject characteristics like appearance, culture, age, etc.

on individual algorithms we introduced several ground-truth annotation methods with multilayer approach. For example, facial feature motion defined by FACS as well as by facial point's 3D location. The multi-annotation of ground-truth gives the ability of deeper analysis of algorithm performance.

In addition, our framework uses the concepts of "features" and "clues". 'Feature' refers to the measurements made using raw sensory data, while the term 'clue' corresponds to the input data of the psycho-physiological models for emotion classification. Transition from features to clues is done in the feature analysis step that will be described below.

Next we describe the four benchmarkable steps in detail:

1) *Data capturing step* - It corresponds to the retrieval and preprocessing of raw data. The functionalities handled by this step are: sensors operation, data retrieving, data preprocessing, spatial alignment (3D relative location of cameras and 3D scanners in the working space), temporal alignment (sensor synchronization), and low-level sensor data fusion. The outputs of this step are three types of preprocessed and aligned signals: 1) data from single sensor, 2) low-level fused data from multiple sensors, and 3) timing of audio, video and gesture stimulus. An example of fused data is the combination of an infrared and a visual frame into a single frame, each pixel including both RGB and temperature channels.

   Benchmark B01 - used for evaluation of the algorithms and tools employed for spatial alignment, multimodal sensors synchronization and sensor data fusion. This benchmark contains definitions of testing procedures and data.

2) *Feature Extraction step* - In this step a variety of feature detection and tracking algorithms can be applied on input data for multiple features extraction, such as algorithms for face tracking, facial expressions detection, temperature changes, pupils dilation, etc. Each algorithm is classified based on its inputs and outputs. There are three types of inputs (single sensor, multi-sensor and fused sensor) and two types of output (frame-by-frame tracking results and results corresponding to group-of-frames). Such classification simplifies interchangeability and comparison of algorithms of similar classes.
   The output of all algorithms is stored in shared memory, organized as a features/frames table. This approach provides an easy sharing of information across all algorithms. It also provides a complete overview of all tracking results across time, useful for debugging and data visualization.

   Benchmark B02 - provides a platform for evaluation of features extraction algorithms. The benchmark consists of multi-sensor data as inputs and mechanism for comparing the algorithms outputs to the ground-truth.

The systematic organization of the ground-truth allows comparison of algorithms even if their inputs or outputs are not identical.

The following example will illustrate the comparison of three algorithms for facial expression detection: First, an algorithm based on AAM calculates the facial expression, face location and orientation simultaneously. A second algorithm based on facial feature point tracking, which provides the location of every feature point; expression detection based on analysis of the location of the points being a separate task. The third algorithm is based on spatial-temporal descriptors applied on the video of the face; here face tracking is a separate task. Direct comparison of these algorithms is difficult.

However, our benchmark provides programmable interface to construct a variety of comparison procedures. The interface simplifies the control over multi-sensor video signals as inputs and the access to ground-truth parameters such as location and orientation of the head, facial skin area, facial point location, and FACS annotations for every frame.

3) *Feature analysis step* - This step performs the role of adapter between the extracted features and the clues used during the classification step. There are four main analysis types: 1) single feature dynamics detection, such as speed and acceleration of facial features displacement, 2) timing relationship between features, for example delay between the face expression and temperature changes, 3) detection of specific features sequences; for example appearance of multiple apexes during facial micro-expressions (happen commonly before deep breath-in), 4) definition of significant features that will be used without any additional analysis as a model input.

   Benchmark B03 - provides a platform for benchmarking features analysis algorithms testing the quality of the extracted clues. The input is a collection of different features tables with the option to add noise to them and the ground-truth is the expected clues list.

4) *Classification step* - In this last step classification algorithms based on psychophysiological models are applied. The inputs of the models are a clues list and a context information such as age group, visual racial appearance, cultural background, evaluation of the mood, etc. The output is the final evaluation of the emotional state.

   Benchmark B04 - provides a psychophysiological models development and evaluation platform. The input is a variety of clues lists, and the ground-truth is emotional tags using several emotion description approaches. At this time, the classification step does not include detailed procedures for development of psychophysiological models. However, the framework provides a simple access to an extensive list of emotional clues available from data analysis and this benchmark. This way the development of the psychophysiological models becomes independent from other more technical steps.

A single configuration file is used for configuring all steps. This provides simple control on execution flow, fusion process and diverse algorithms parameters.

The current implementation has been developed in Matlab with GPGPU to speed up the execution time. Matlab is a rapid and flexible development environment with variety of open source libraries. The presented framework is not particularly designed to support the development of real-time systems, but rather the development of feature extraction algorithms and psychophysiological models. Future work will include implementation of these algorithms and models into real time systems.

III. Discussion

The presented development framework for multimodal emotional sensing supporting system is organized into clear benchmarkable steps, allowing 1) evaluation of various algorithms, 2) measurement of the significance of each of the steps in the framework, and 3) improvement of the collaboration between psychologists and technical researchers by separating psychophysiological models development from technical steps.

We are currently tackling the development of an intuitive user interface (UI) for ESSS that presents data to the interviewee in a way that optimizes decision making and minimizes distraction from the interview.

We are also finishing the annotation of a first phase of multi-sensor hostile detection database, by using this data we create benchmarks based on the presented framework. The database and benchmarks will be publicly available soon.

Through this work, our objective is to provide a framework that can serve as the ground for the creation of standard benchmarks for ESSS and their modules.

References

[1] J. F. Nunamaker Jr., D. C. Derrick, A. C. Elkins, J. K. Burgoon, and M. W. Patton, "Embodied conversational agent based kiosk for automated interviewing," *Journal of Management Information Systems*, vol. 28, pp. 17–48, 2011.

[2] P.Petta, C. Pelachaud, and R. Cowie, *Emotion-oriented systems: the HUMAINE handbook*. Springer, 2011.

[3] B. Dumas, D. Lalanne, and S. Oviatt, "Human machine interaction," D. Lalanne and J. Kohlas, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. Multimodal Interfaces: A Survey of Principles, Models and Frameworks, pp. 3–26.

[4] B. Dumas, D. Lalanne, and R. Ingold, "Hephaistk: a toolkit for rapid prototyping of multimodal interfaces," in *Proceedings of the 2009 international conference on Multimodal interfaces*, ser. ICMI-MLMI '09. New York, NY, USA: ACM, 2009, pp. 231–232.

[5] J. Shen and M. Pantic, "A Software Framework for Multimodal Human-Computer Interaction Systems," vol. 203143, no. October, pp. 2038–2045, 2009.

[6] J.-Y. L. Lawson, J. Vanderdonckt, and B. Macq, "Rapid Prototyping of Multimodal Interactive Applications Based on Off-The-Shelf Heterogeneous Components," in *User Interface Software and Technology*, Oct. 2008, pp. 41–42.

[7] H. O. North, "Imagery library for intelligent detection systems (i-lids)," aug 2011. [Online]. Available: http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/

[8] D. North, "Introducing behaviour driven development," jan 2012. [Online]. Available: http://dannorth.net/introducing-bdd/