A Non-invasive Multi-sensor Capturing System for Human Physiological and Behavioral Responses Analysis

Senya Polikovsky, Maria Alejandra Quiros-Ramirez, Takehisa Onisawa, Yoshinari Kameda, and Yuichi Ohta

Graduate School of System and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8573 Japan
senya@image.iit.tsukuba.ac.jp

Abstract. We present a new noninvasive multi-sensor capturing system for recording video, sound and motion data. The characteristic of the system is its 1msec. order accuracy hardware level synchronization among all the sensors as well as automatic extraction of variety of ground truth from the data. The proposed system enables the analysis of the correlation between variety of psychophysiological model (modalities), such as facial expression, body temperature changes, gaze analysis etc... . Following benchmarks driven framework principles, the data captured by our system is used to establish benchmarks for evaluation of the algorithms involved in the automatic emotions recognition process.

Keywords: sensor-fusion, synchronization, benchmarks.

1 Introduction

Automatic recognition of emotions has been actively studied in the last decade [7]. Although strong benchmark environment is necessary for the development of this field it is usually neglected. In this work we present a new noninvasive multi-sensors capturing system for collecting video, sound and motion data that allows the creation of benchmarks. The recorded multi-sensor data enables the analysis of correlation between diverse psychophysiological models, such as facial expression, body temperature changes, gaze analysis, and voice.

Psychophysiological models are usually studied independently and fusion between new sensing technologies is barely utilized. There are four reasons for limited use of sensor fusion: first, there is an absence of a single off-the-shelf system that integrates a variety of modern sensors and simplifies their manipulation. Second, having several recording devices brings up the challenge of their synchronization. Synchronization is a key point in order to analyze emotional changes in time and relation between different clues in representing emotions. Third, the overflow of the recorded video and other data makes its management and analysis difficult. Finally, in order for the recorded data to be used as a benchmark for tracking and classification algorithm development, a variety of

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 63–70, 2013.

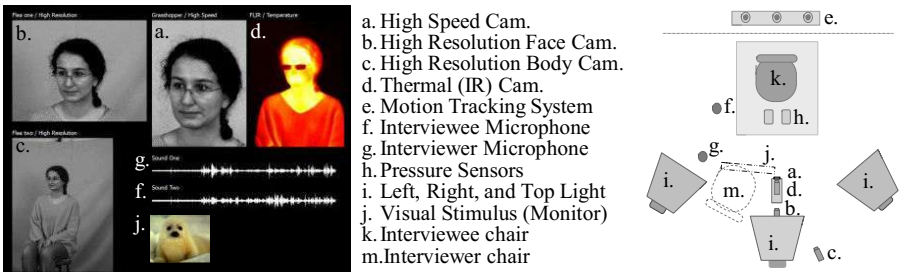

Fig. 1. Left: Visualization of the captured data during Computer-to-Human interview by our system. Right: Diagram of the sensors location during data collection.

ground truth information should be annotated which represents a challenging and time consuming task.

The capturing system we present contains high speed, high resolution and thermal cameras, eyes and 3D marker tracking devices, microphones and pressure sensors. The system provides a simple control over sensors and ensures 1msec order accuracy hardware level synchronization among all the sensors. The design of the system allows relatively simple extensibility of additional sensors. In addition, we introduce sensors for speeding-up the annotation process by segmenting points of interest in recorded data as well as extracting ground truth information such as position of the body parts and gaze direction. These characteristics are required for the creation of benchmarks. The system was used to record a cross-cultural database containing 36 subjects, presented at [11]. Figure 1, left side shows an example of signals captured by the system and right side is a diagram of the sensors location during the recording.

Based on our benchmark driven framework [2] used for development of emotion sensing systems, the presented system corresponds to the "Data Capturing" step of the framework. The data captured by the system is used to create the benchmarks of the remaining steps of the framework. This topic will be explained in detail in the section 3.

There is few research focused on the design of capturing systems. The research presented in [5] is an example of some guidelines for designing a capturing system with special emphasis on synchronization between video, sound and eye tracking sensors. In this paper we present a system design that implements a broader range of sensors utilizing a different design approach.

Due to the variety of terms in the field, we chose *psychophysiological model* to refer to similar terms such as behavioral clue or modality.

2 Multi-sensors Capturing System

This section introduces the design of our multi-modal capturing system. The design of a capturing system for emotional sensing purpose consists of the following steps: 1) Definition of the scenarios in which the system will be used.

2) Choice of the psychophysiological models that will be analysed. 3) Selection of system sensors based on measurement accuracy, synchronization capabilities, sensor hardware and software interfaces, as well as the price. 4) Choice of hardware configuration for system computers, operation system, development environment and communication protocols in the system. During the design process the steps go through a number of iterations until the final configurations are established. From our experience due to the large number of factors and limited guidelines on capturing system design, the process is more empirical than methodological. Therefore an introduction of a variety of capturing system designs is necessary for advancing the field since it will save the development time that it requires to build the system from scratch.

Next we describe the scenarios and the selected psychophysiology models, then we define the requirements of the capturing system and present the system design. Finally, we introduce the system sensors and synchronization scheme.

2.1 Scenario and Psychophysiology Models

In this stage we are focusing our interest on human-to-human and human-to-computer indoor sitting interview scenario. This scenario allows to control environmental factors such as lighting conditions, room temperature and space background. The use of the chair limits the movement of the interviewee and dictates sensors location.

As for the psychophysiology models, based on collaborations with a police negotiation unit as well as a psychologist [3] from Arizona University, we have identified the five most promising psychophysiology models for behavior analysis from a technological and psychological point of view. These models are used to differentiate between normal and abnormal behavior, detection of deception and stress.

1. Micro-expressions - Ekman at al. showed that facial expressions are the most important behavioral source for lie indication and danger demeanor detection [1]. Micro-expressions appear with low muscular intensity, which makes it impossible to analyse using standard speed cameras. Thus, a high speed camera is required [4].
2. Facial Feature Area Temperate - Pavlidis at al [6]. demonstrated the correlation of increased blood perfusion in the orbital muscles and stress levels for human beings. It has also been suggested that this periorbital perfusion can be quantified through processing thermal video captured by thermal (infrared) camera.
3. Eyes analysis - Pupil dilation as well as gaze direction [10] can also be used for stress, interest and drug use detection. The newest eye tracking systems provide high accuracy analysis.
4. Body Language - Analysis of head, shoulders and hands movement can be used for deception detection. Body language has been used for decades by psychologists for human behavior analysis [3].
5. Voice Stress Analysis - Used to recognize stress responses that are present in human voice, when a person suffers psychological stress [8].

By combining these five approaches that rely on different sources of information, we increase trustfulness and robustness of the final analysis. In addition, the accurate synchronization between the measurements related to each one of the models provides the ability to analyse the timing correlation between them. The synchronization of 1msec order was selected as a trade-off between sufficient accuracy (that allows to combine EMG signals in the future) and the cost and complexity of the system.

2.2 Capturing System Requirements and Design

The system was designed based on the following list of requirements: 1) all sensors are controlled using a simple interface from a single computer, 2) all sensors have the ability to be synchronized with 1msec order accuracy, 3) simple integration of new sensors is allowed, 4) video data is captured with no compression, allowing analysis of the influence of compression in the future, 6) detection of missing frames in video sequence is supported, 7) automatic extraction of stimulus timing to speed-up the segmentation of the recorded signals, 8) automatic extraction of ground truth by off-the-shelf devices, 9)system is capable of recording large amount of information into the HDDs, 10) system is transportable.

Figure 2 introduces the capturing system design, due to the limitation of the space the overview of the design is the caption of the figure. For more details on the design we encourage the reader to contact the authors.

2.3 Sensors

The sensors of the system can be separated in two groups: the first group contains the sensors that will be used in real-time implementation; the second group contains support sensors that are aimed to extract reliable ground truth measurements. Some of the sensors belong to both groups (see Table 1). The sensors from the first group are: high-speed camera for analysis of facial micro-expressions, infrared camera measuring temperature changes, high-resolution camera for body analysis and speaker interaction, microphones capturing voice, pressure sensors for capturing body weight changes and rapid legs motion. The second group contains motion capturing system for automatically detect precise head location and orientation as well as shoulder level. The MCS markers are attached on the backside in such way that they cannot be seen by front cameras. Eye tracking system for automatic extraction of gaze and pupil dilatation. Photosensor for capturing precise timing of the visual stimulus presented on the screen.

A photosensor with response similar to human eyes is attached to the computer screen to detect the exact timing of visual stimulus presented on the screen during human computer experiments. Having the exact timing of the stimulus allows the automatic segmentation of the recorded data in order to extract the important sections and reduce the amount of final data.

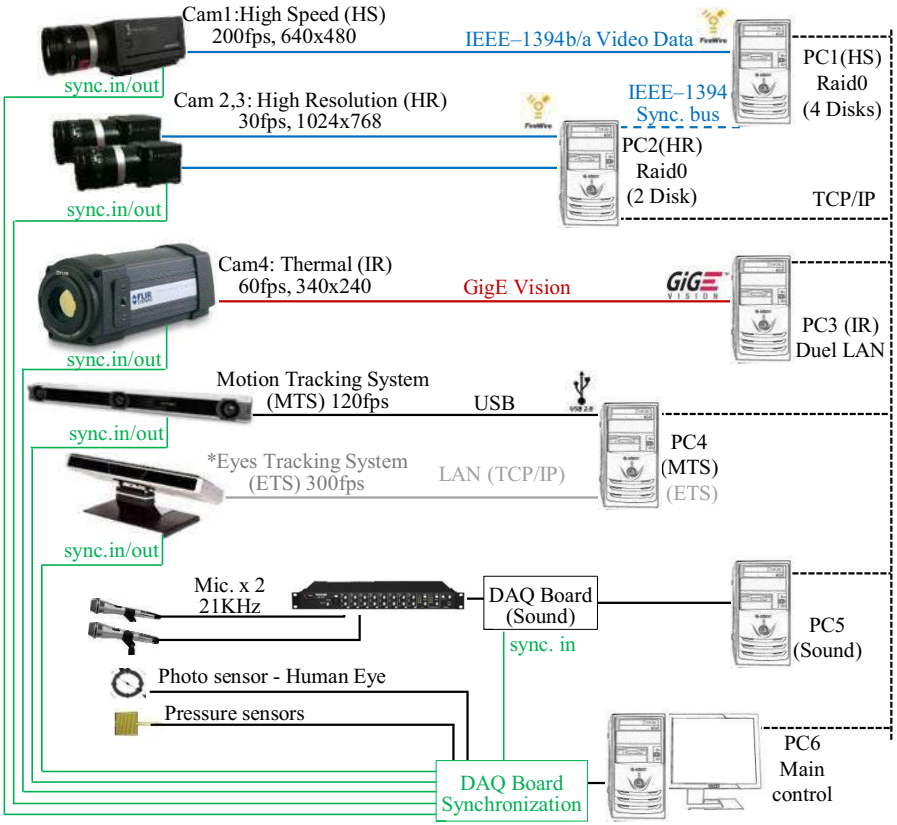

Fig. 2. Capturing system design. The input sensors are: high speed, thermal, and two high resolution cameras, motion and eyes tracking systems, two microphones, pressure sensors, and a photosensors with a response similar to the one of human eyes. The corresponding cameras' data transfer interfaces can be seen in the figure. Due to 1) incapability of number of cameras' drivers to be install on the same computer, 2) reuse of already exiting equipment and 3) a large capacity of the recorded data, separate computer is dedicated for each on of the sensor.

The system consists of six computers connected to the same network, PC1 and PC2 contains RAID0 hard disk setup for high speed stream data recording. Special attention should be given to configuration of the RAID hardware to meet the required writing speed. For HS and HR cameras, the use of HPwv8400 workstation with onboard RAID card provided the sufficient writing speed. The rest of the computers are standard configuration PC. The main computer controls the simultaneous recording process of all the sensors through MailSlot interprocess communication. Computers that receive data from 1394 protocol have an additional synchronization 1394 bus (dash blue line). PointGrey provides synchronization software on top of 1394 bus, however it supports only on WinXP OS, therefore WinXP and Win7 are used as OS in the system. All the sensors are wired with synchronization in/out cable (green line) connected to data acquisition board (DAQ), additional synchronization signals are supplied by a sound recording system. Photo and pressure sensors and sampled using the same DAQ, this way all the digital and analog signals are recorded in the same time line.

* At this time Eyes Tracking System was not used during the data collection.

Table 1. System sensors parameters

Sensor	fps	Data type	Sync	role
Hi-speed (Grasshopper, PTG)	200	640x480, RGB	GPIO	on-line
High-resolution (Flea2, PTG)	30	1024x768, RGB	GPIO	on-line/support
Thermal (A325, FLIR)	60	320x240, 16 bits	GPIO*	on-line/support
Motion Tracking System (Trio)	120	3D points location	GPIO**	support
Eyes Tracking System (TobiiTX300)	300	Gaze direction	GPIO	on-line/support
Microphones x2	21(KHz)	1D	Acq.Card	on-line/support
Photodiode	2(KHz)	1D	Acq.Card	on-line/support
Pressure sensors X4	2(KHz)	1D	Acq.Card	on-line/support

PTG - Pointgrey, * - No Shutter out, ** - Start / stop control,
Tobii - sync. option coming soon

2.4 Synchronization

In this section we describe our synchronization strategy and alignment of the recorded data. Current off-the-shelf capturing products are not capable to ensure high accuracy synchronization between cameras based on different interfaces such as FireWire (IEEE 1394) and GigE Vision and Camera Link. One possible solution is to use an external trigger; however high-speed middle range price cameras lack accuracy in external trigger mode or not support it as with thermal cameras. The high-end price of such cameras can reach order of 100k$. In our strategy the main computer, that controls the operation of all the sensors, starts the recording of all sensors asynchronous. Next, after insuring that all the sensors

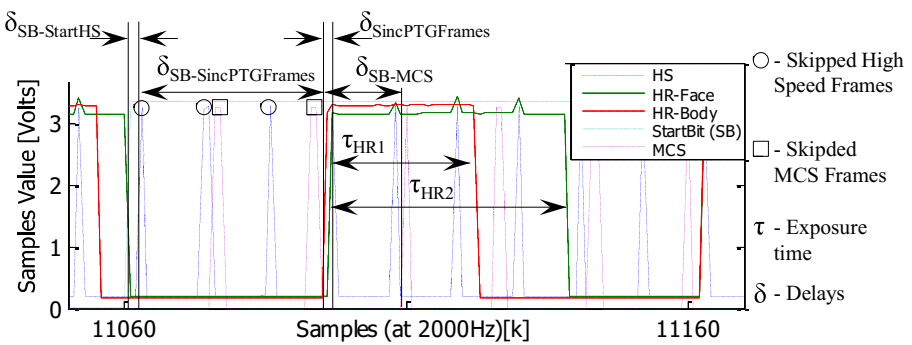

Fig. 3. Cameras strobe signals captured by DAQ that are used to extract the time delay between the sensors. The delays are calculated in correspond to "start bit" (SB) (dash green). $\delta_{SB\ \ FirstHS}$ delay between SB and start frame of high speed camera. $\delta_{SB\ \ SincPTGFrames}$ - SB and the first time all the frames of PTG cameras are sync. $\delta_{SincPTGFrames}$ - PTG cameras sync. error (500msec.). $\delta_{SB\ \ MCS}$ - SB and motion capturing system frame. ○ and □ - indicates the HS and MCS frames that should be skipped during the alinement process. τ represents the cameras exposure time.