ATTENTIVE TOYS

Ismail Haritaoglu[1], *Alex Cozzi*[1], *David Koons*[1], *Myron Flickner*[1], *Dmtry Zotkin*[2], *Yaser Yacoob*[2]

[1]IBM Almaden Research, San Jose, CA 95120, USA
[2]Computer Vision Laboratory University of Maryland, College Park, MD 20742, USA

ABSTRACT

We describe an attentive system that pay attention to people so they can attend to people's needs using visual and audio sensors. It integrates real-time video and audio processing techniques to detect and track multiple people in the scene, and speech recognition and eye contact to develop an communication interface with them in a natural way that human uses. We implemented it as a visually interactive toy robot (VTOYS) and demonstrated successfully to many people in different age class to explore new ways on human-machine interactions and interfaces on how people interact with machines when machines have some ability to perceive around him.

1. INTRODUCTION

Human cognition depends on highly developed abilities to perceive, integrate, and interpret visual, auditory information. Without a doubt, machines would be much more powerful if they had even a small fraction of the perceptual ability of humans. Adding such perceptual abilities to machines would enable machines and humans to work together more as partners. People uses eyes, ears, voice, gesture, facial expressions while communicating the each other. These are natural ways people use to communicate with other people, however, they cannot use these natural modalities to communicate with machines with current input devices. Machines are currently visually blind and deaf. They cannot recognize the presence of humans, identify their communicative messages, or react to these messages. Traditionally, autonomous machines are designed to operate as independently and remotely as possible from humans interaction. However, recent applications, such as, entertainment, education, are driving the development of machines that can interact and cooperate with people, and play a part in their daily lives.

We have been developing attentive devices which use non-obtrusive sensing technology, such as video cameras and microphones, to support the natural communication modalities of humans, such as, facial expression, body posture, gesture, gaze direction, and voice and to identify and observe a user's actions, and to extract key information. These cues are analyzed to determine the user's physical, emotional, or informational state, and needs which in turn can be used to help make the user more productive by performing expected actions or by providing expected information. We implemented video, audio algorithms and integrated into an attentive toy robot (VTOY) which uses natural ways that human uses to perceive environment, and react to the people in natural way, such as, voice, eye contact, emotions, and expressive behaviors which play an important role in human communication, such as being happy or angry when somebody start playing or ignore him. We have integrated IBM Via Voice speech recognition engine with Vtoys so that people can use voice to communicate with Vtoys and Vtoys response them with speech.

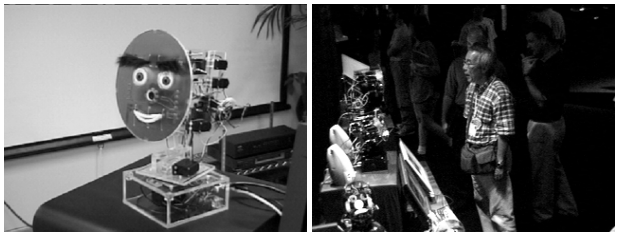

Figure 1: IBM's attentive toys

Recently, there has been efforts to make a human-like robot or machines. Honda's humanoid human size walking robot [3] is a good example of showing how robotics can achieved difficult task, such as walking up/down on the stairs. Meanwhile, while the efforts on the robotics site has been progressed, researchers try to focus on more sensor, social and behavior aspect of building smart devices that can interact with people in similar way people interacts with each other. Perhaps, the most similar work to VTOYS is MIT's the Sociable Machines Project [4] which develops an expressive robot called Kismet that engages people in natural and expressive face-to-face interaction. Sony's IBO toy is another example to show with little perceived information can make difference in human-machine interaction.

In this paper we described video and audio processing techniques used in VTOYS for enabling robots to better perceive and understand human actions and mechanically react to them by executing a cartoon-like character. Figure 2 shows the system diagram of VTOYs. There are three main units: sensor processing unit, behavior/attention unit and motor-servo unit. Sensor unit has two main modules: video and audio processing units. Video processing unit which consists of silhouette and motion based people detection, eye/gaze detection, and face expression tracking modules which allows Vtoys detect where people are and how many of them in the scene, whether they do eye contact, their face expressions. Audio processing unit which consists of acoustic source detection and speech recognition modules allows vtoys receive voice inputs from people and determine location of sound source. Features which has been detected in sensor unit are processed in the attention/behavior unit which decides what VTOY is going to do, such as, where it directs the attention, how it start engagement with people, showing emotions/facial expression to people. Attention/behavior unit communicate with motor/servo unit to control the robotics part of VTOYs via RS232 interface. All algorithms are running near frame rate that vtoy can response to people without delay. Our experiments shows that people notices and any delay

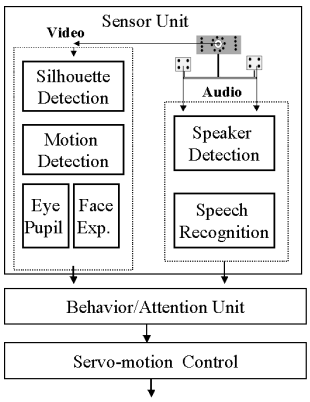

Figure 2: System diagram of IBM's Attentive Toys

bigger than 100ms. Currently, Vtoys has 30 fps video processing speed and 10 estimation per frame speed for acoustic processing on a single 866 Mhz dual Pentium III system.

2. SYSTEM OVERVIEW

The VTOY robot is a prototype of a future toy which has a total of 12 servos resulting in 11 1/2 degrees of freedom(Figure 1). The eyes constructed from ping-pong ball each have two degrees, azimuth and elevation, of rotational freedom. VTOY has infrared illuminators and an infrared camera in his face, two array of microphones on left and right side, and speakers which allows him to simulate "eye" and "ear" and having the following capabilities:

- **Detection of the presence of humans, their gaze and facial actions.** It has a video camera that is employed to recover visual information about humans in its vicinity. It monitors its space searching for human presence. It then locates face features of a person and determines their rough gaze.
- **Controlled neck and eye movements and generation of facial deformations such as mouth and eyebrows deformations.** It determines the gaze of the human and rotates its eyes and neck to follow the movement and gaze directions. It also uses controllers to flex its deformable facial features, thus conveying cartoon-like expressions.
- **Acoustic based speaker localization and speech recognition and synthesis for verbal commands and communication** It has capability to determine location of people using simple array of microphones system when people are out of its sight, shifting its attention to those people by moving its neck. In addition, It has speech recognition ability to understand simple command that human ask and uses speakers to be able to speak with people.
- **A small set of behaviors designed to attract the attention of a people** The basic capabilities of VTOY are augmented by a set of simple behaviors designed to attract a viewer. It deforms its facial features during speech to invite the user to look at it. It mimics, exaggerates and smiles at user facial deformations, thereby arising the curiosity of the user.

Figure 3: Silhouette based people detection: Statistically modeled background scene (top-left), detected foreground regions (bottom-left), final head and torso segmentation (bottom-right), eye detection (top-right)

3. VIDEO AND AUDIO BASED PEOPLE DETECTION AND TRACKING

Eyes and ears are the important part of human to perceive information. We are using a camera and an array of microphones to simulate the eye and the ear in Vtoy. Camera based solutions allows us to determine the location of each person when they are in the field of the camera by combining static shape, dynamic motion, and eye-contact information. Acoustic based solution allows us to determine to location of people when they are speaking even they are not in the field of view of the camera by computing the 3D location of single sound source using multiple microphone.

3.1. Silhouette-based People Detection

Vtoys detects objects through a statistical background subtraction process when it needs to detect people without neck motion. We used a model of background variation that is computationally efficient. We assume each pixel intensity distribution is *bimodal*. The background scene is then modeled by representing each pixel by three values; its minimum $m(x)$ and maximum $n(x)$ intensity values and the maximum intensity difference $d(x)$ between consecutive frames observed during this training period. Foreground objects are segmented from the background in each frame of the video sequence by a four stage process: thresholding, noise cleaning, morphological filtering and object detection. Each pixel is first classified as either a background or a foreground pixel using the background model (detail information can be found in [2]). Then Vtoys simply classified those objects as individual person, group of people or "other" on the basis of static size and shape properties. If an object was classified as a person, then It segmented the shape into body parts (head, torso), built appearance models for the torso and head, and tracked the person in the camera's field of view. If an object contains a group of people, Vtoys attempts to find out many people comprise that a foreground object corresponding to a collection of people and segment the foreground object into its constituent individuals.

Intuitively, what types of information might be used to count

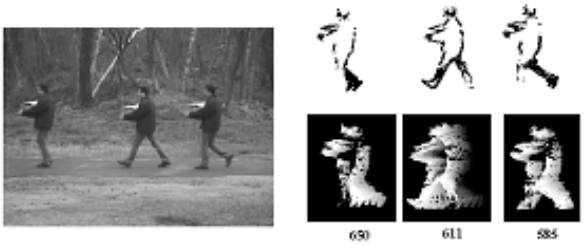

Figure 4: motion-detection results and their motion templates

people and segment groups? **Local shape information:** For example, by analyzing the boundary of the object, we might find pieces that look like heads, etc. In fact, Vtoys operates by first attempting to locate heads based on local shape approximations to the object boundary. **Global shape information and constraints:** Vtoys employs a global shape constraint derived from the requirement that the head be aligned with the axis of the torso. In particular, by projecting the object region on an axis perpendicular to the assumed torso axis, one should observe a peak in the projection in the vicinity of a head. **Appearance information.** So, for example, a hypothesized head could be verified by matching the texture of the region to a prototypical face (as in face detection) assuming the face was visible from the camera. Vtoys verify detected the head location using pupil location.

People are detected based on two types of shape analysis of binary foreground regions. First the results of a local corner detection algorithm are analyzed to find possible head locations on the boundary. Next, a region-based shape analysis is conducted. A vertical projection histogram of each silhouette is constructed to find the vertical boundaries between people. During tracking, a dynamic template for each head detected is generated and updated. A second-order motion model, which combines robust techniques for region tracking and matching of silhouette edges with recursive least square estimation, is used to predict the location of the head and the person in subsequent frames. A normalized distance region segmentation is applied to the foreground region to obtain a distance map used to segment the silhouette into regions representing individual people. Distance values from each pixel to each potential person are normalized to obtain the normalized distance map for each silhouette; then, each pixel is assigned to a person according to this map. Finally, an appearance model is generated and updated for each person during tracking so that the person can be identified after occlusion. Figure 3 shows and example of people detection results based on silhouette analysis.

3.2. Motion-based People Detection

The other information used to detect people is motion when people move in the scene. This allows detection of people while vtoys is in active motion with no background scene information available. During active tracking, the background model will not be immediately available when Vtoy's neck is positioned at an angle different from initial neck positions. It uses a motion cues to detect and track people until the background model parameters are computed for the camera current field of view. People are detected by motion-history templates which are computed by subtracting three consecutive images and thresholding the difference images to determine pixel that may be moving. Each pixel is first classified

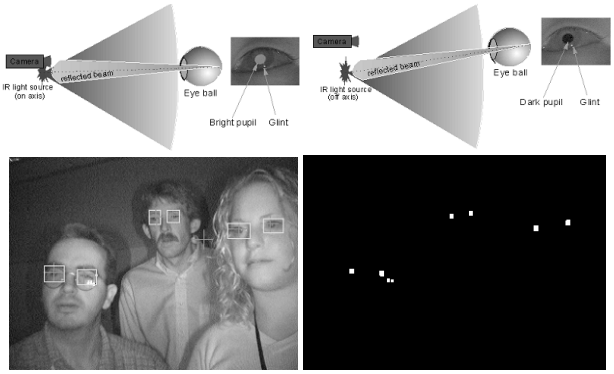

Figure 5: Illustration of pupil detection techniques use in Vtoy: On-axis illumination (top-left), off-axis illumination (top-right), detected pupil location (bottom)

as foreground or background pixel, Then A fast binary connected component operator is applied to foreground pixel to detect the foreground region. Vtoys employs a second order motion model for the biggest foreground object to predict its location in subsequent frames. The active tracking of a walking person with its motion history and motion detection templates are shown in Figure 4. Vtoys gives high priority to motion-based based people detection results in order to response back to people by shifting his attention to where the sudden motion occurs.

3.3. Eye Detection and Face Expression Tracking.

We integrated fast eye/pupil detection techniques along with face expression tracking to Vtoys in order to determine eye contact, face expressions which play an important role in human communication. We developed a fast, robust, and low cost pupil detection technique that uses two infra red (IR) time multiplexed light sources, synchronized with the camera frame rate [5]. One light source is placed very close to the camera's optical axis, and the second source is placed off-axis. The pupil appears bright in the camera image during on-axis illumination (similar to the red eye effect from flash photography), and dark when illumination is off-axis. Our experiments using a real-time implementation of the system show that this technique is very robust, and able to detect pupils using wide field of view low cost cameras under different illumination conditions, even for people with glasses (Figure 5. By using simple and fast image subtraction methods, eye regions are detected and verified with silhouette and motion based results to detect whether or not people are looking at vtoys and keep eye contact. Once two eyes for each people in the scene are detected, vtoys attempt to compute head orientation and locate other face features(eye brow, mouth), based on pupil location. In Figure 6, detected face features are shown while the head is in different orientation. The face features are used to track facial expression of people that allows vtoys determine the emotion of people based on their facial expression.

Facial expression tracking has until recently been considered as an off-line problem. One of the challenging problem in our project was computing the facial expression of people in real-time so that Vtoys can sense the emotion of people, or mimick them. We integrated basic techniques for motion analysis of human facial deformations [6]. A system that simultaneously tracks the im-

Figure 6: Face feature detection based on eye location when the head is in different orientation

age motion of a human face and measures the nonrigid motion of face regions such as the eyes, eyebrows and mouth was developed [6]. The system was applied initially to the problem of recognition of the so-called six universal facial expressions. The system employs robust estimation techniques to fit parametric motion models to image regions - a planar model for the rigid body motion of the head/face in an image sequence, and extended affine models for the nonrigid motion of face features such as the eyebrows and mouths. Figure 7 show instances while vtoy was tracking user's facial expression and mimicking it.

3.4. Acoustic Speaker Detection

The audio detection and localization uses algorithms based on the signal power and time difference of sound arrivals (TDOAs) to perform signal detection and localization. The audio detection and localization module is builtwith 8 microphones around the PowerDAQ 16 channel data acquisition board with total frequency of up to 1.25 MHz. We used 8 channel at 22.5 kHz sampling rate each. Each channel connected to a preamplifier whose output is fed to data acquisition board. The microphones are divided into two subarray of 4 microphones. Forming two "ears" of the system. Microphones in each array are arranged in a square pattern and ywo sub array the left and right side of the Vtoys.

Time differences of arrival time between microphones are determined by computing generalized cross correlation between signals arriving at the microphones of each subarray and obtaining the peaks. Computed TDOAs are used to determine the azimuth and elevation of the source. The solutions obtained separately for the two microphone arrays, and the bearing angles computed intersected to find the source location in 3D. The coordinate of the source are transform in to the vtoys coordinate frame and vtoys look angle are computed. To prevent neck motions dues to false source detection, first acoustic signal checked to ensure there is a sufficient power in the speech frequency bands. After that, additional filtering is performed based on consistence of last computed source location in small time frame and speech length and that gives additional advantage to prevent jerky neck motion for short sounds possible noise. Detailed information on how TDAOs computed are found in [1].

4. CONCLUSION AND FUTURE WORK

We describe an attentive system that pay attention to people so they can attend to people's needs, such as, where you are, what you see, say, do, talk, using visual and audio sensors. It integrates real-time video and audio processing techniques to detect and track multiple people in the scene, and speech recognition and eye contact to communicate with them in a natural way that human uses. We

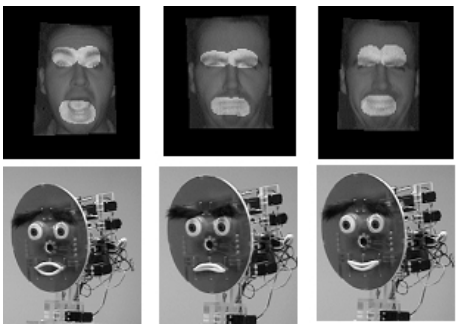

Figure 7: Vtoys has ability to track and mimick facial expressions

implemented it as a visually interactive toy robot (VTOYS) and demonstrated successfully to more than 5000 people in different age class to explore new ways on human-machine interactions and interfaces on how people interact with machines when machines have some ability to perceive around him. It allowed us to test interactive strategies between human and attentive devices. Since this research area is little explored, we are learning and focusing about interaction scenarios in which people are comfortable, as well as define the future capabilities that attentive machines needs to acquire. Developing real-time system and algorithms for tracking and understanding human movement at coarse and fine levels was one of the challenging problem in our system. We are working on add new video sensors to allow us to track people in 3D using stereo and color information , and acoustic beam-forming algorithm to improve the speech signal quality for better speech recognition results. The rapid embedding of computers in the human environment can be made more human friendly by adding visually interactive capabilities both at the level of perceptual understanding of human body and facial movements, and the human reaction to robotic emulation of human traits.

5. REFERENCES

[1] D. Zotkin, R. Duraiswami, Ismail Haritaoglu, T. Otsuka and L. Davis, A real-time audio-video front-end for multimedia applications. The Journal of the Acoustical Society of America – October 1999 – Volume 106, Issue 4, p. 2271

[2] I. Haritaoglu, David Harwood, Larry Davis W4:Areal time system for detection and tracking of people and monitoring their activities IEEE Transaction on Pattern Analysis and Machine intelligence, August, 2000.

[3] Honda Humanoid Robot Project http://world.honda.com/robot/

[4] Breazeal, C Sociable Machines: Expressive Social Exchange Between Humans and Robots. Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT.

[5] C. Morimoto, D. Koons, A. Amir and M. Flickner, Pupil Detection and Tracking Using Multiple Light Sources accepted to Image and Vision Computing Journal (IVC), special issue on Advances in Facial Image Analysis and Recognition Technology, 2000

[6] M.J. Black and Y.Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. Int. Journal of Computer Vision, 25(1), 1997, 23-48.