A MULTIPLE CAMERA APPROACH TO FACIAL GESTURE RECOGNITION FOR CHILDREN WITH SEVERE SPASTIC QUADRIPLEGIA

Brian Leung and Tom Chau
Institute of Biomaterials and Biomedical Engineering, University of Toronto

ABSTRACT

This paper presents the theoretical framework for a new approach to computer vision-based facial gesture recognition that accommodates the physiological conditions of children with severe spastic quadriplegic cerebral palsy (CP). Clinical observation suggests that these children can exploit one or more facial gestures (e.g. tongue protrusions) to operate a facial gesture access modality with adequate proficiency. The proposed approach uses independent input video data from multiple cameras observing from different viewpoints, in order to maximize the detection of intentional facial gestures in the presence of spastic head movements common to children with severe CP. Also, this paper outlines a case series methodology for further developing and evaluating the proposed algorithm and briefly discusses preliminary image processing issues.

INTRODUCTION

It has been reported in literature on education that interactive learning is crucial to a child's cognition and communication skills development [1]. Two examples of interactive learning are children playing with toys and children accessing the computer for online learning exercises. In particular, today's elementary education programs balance passive instruction and active learning so that children may begin their intellectual and creative development through diversified learning perspectives and experiences at an early age [2]. Access to interactive learning, even as simple as being able to play, is indispensable for receiving the full benefits of modern education.

Children with severe cerebral palsy (CP) are at a disadvantage relative to typically developing children with regards to opportunities for interactive learning, because they lack a dependable method of using the mechanical (switches and buttons) interface implemented in most toys and computer peripherals. In most cases of CP, the motor disorder is characterized by frequent extraneous gross movements at the limbs and the head. These involuntary movements make targeting mechanical switches a very difficult task, especially in severe cases of CP measuring at levels 4-5 in the Gross Motor Function Classification Scale (GMFCS) [3]. The lack of dependable access has been identified a primary risk factor for higher incidences of developmental and learning problems in children with severe CP [4].

Children with severe spastic quadriplegic cerebral palsy (severe *spastic quadriplegia*) are affected by muscle spasticity (abnormal hyperactivity, rigidity, and spasm) at all four limbs and the neck. Furthermore, these children are usually non-verbal secondary to spastic quadriplegia. Therefore, mechanical or voice switches typically are not feasible to realize dependable access.

Some children with spastic quadriplegia may be able to exploit facial gestures, such as eye blinks and tongue thrusts, to reliably operate a computer vision-based facial gesture recognition system. Such system would detect facial gestures by image processing of the video input from a camera placed at a distance in front of the user. A valid detection then triggers transient activation of a logical "high" voltage signal to mimic the output of a switch. A representative example is BlinkLink by Grauman *et al.* [5], where thresholding on the cross-correlation of an opened-eye template image discerns closed eyes from opened eyes, and thus eye blinks. An alternative approach to realizing access from computer vision-based facial gesture recognition is by movement tracking of the head or a face feature to control a computer mouse cursor [6, 7]. However, spastic head movements limit the feasibility for these children to use the head or face as a control.

For the same reason as above, single camera systems have limited utility for children with severe spastic quadriplegia. In the aforementioned examples of computer vision systems, having a frontal view of the face is important for facial gesture recognition. The extraneous head movements complicate the user task of facing the camera squarely and steadily for proper facial gesture recognition. A possible alternative is to have multiple cameras available for facial gesture recognition from multiple independent viewpoints. Theoretically, this solution is feasible because the extra cameras can detect gestures that are not

properly presented in the original (single) camera's field of view.

THE MULTIPLE CAMERA APPROACH

Assume that multiple cameras are set up at different positions with their fields of view directed at a user's face. The proposed treatment of multiple cameras is to consider each camera as an independent input channel of the facial gesture recognition algorithm. The video data from all the cameras are not coordinated, calibrated, or dependent on each other. This approach is analogous to having several single camera systems performing gesture recognition from different perspectives of the same gesture source. Figure 1 shows the example of a multiple camera setup with three cameras.

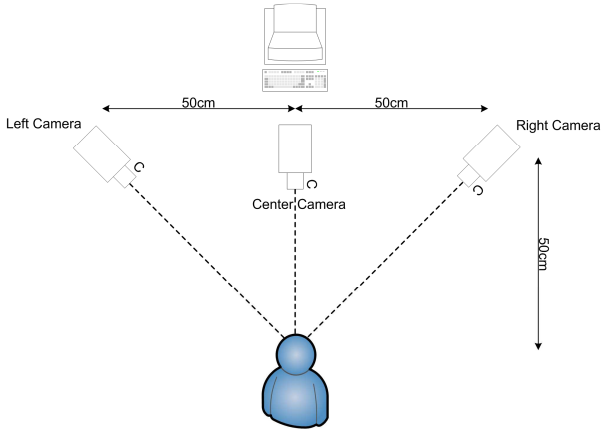

Figure 1: A three-camera setup relative to a user.

This proposed treatment of multiple cameras has the potential of overcoming the aforementioned issues of single camera in the facial gesture recognition for children with severe spastic quadriplegia by:

- increasing the likelihood that the frontal view of the face is present in at least one camera for proper gesture recognition, and
- allowing the user to focus more on producing the facial gestures and worry less about targeting a specific camera.

The motivation for investigating a multiple camera approach is supported by empirical evidence, which shows the extra video perspectives can potentially increase coverage of the frontal view for proper facial gesture recognition. In 18 minutes of pilot video footage recorded from a child with severe spastic quadriplegia and with the setup in Figure 1, the frontal view of the child's face was lost to the center camera (i.e. rotated to the left or right by 30° or more) for an approximate total of 6 minutes. More importantly, during the episodes of extraneous head movements the left or right camera did capture a better frontal view than the center camera (see Figure 2). Thus, in these episodes the side cameras would have been more suitable for facial gesture recognition.

Figure 2: An example frame of the pilot video footage recorded from the left, center, and right camera respectively. Here, the frontal view of the face is best in the right camera and so the facial gesture of tongue protrusion should be best detected from the right camera.

Algorithm Architecture

Figure 3 shows the algorithm architecture for realizing the proposed multiple camera facial gesture recognition. Each of the *N* cameras is driven by an instance of the single camera algorithm to realize single camera facial gesture recognition. This follows directly from the treatment of multiple cameras as the fusion of gesture recognition from several independent single camera systems. The function of the multiple camera algorithm is to produce a single gesture recognition decision by fusion of gesture recognition results from all *N* cameras.

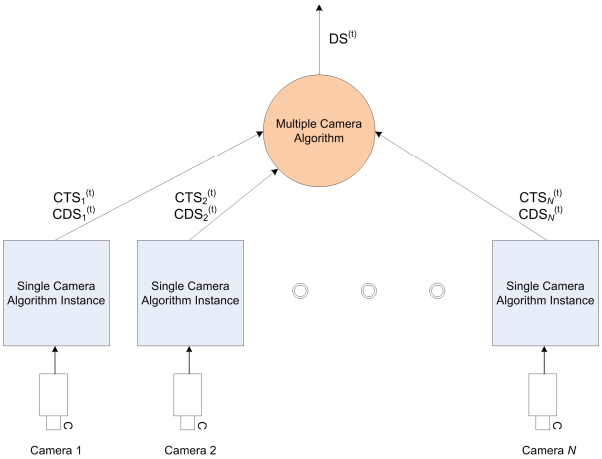

Figure 3: Algorithm architecture of the multiple camera facial gesture recognition.

Requirements of the Single Camera Algorithm

The functional requirements of the single camera algorithm are that it processes video data on a frame-by-frame basis and that it produces two scores on each frame: the camera tracking score (CTS) and the

camera detection score (CDS). $CTS_i^{(t)}$, a value in the real interval [0,1], measures how well the frontal view of the user's face is presented to the i^{th} camera at frame t. The frontal view is lost if the CTS is less than some empirical threshold value T_r. Similarly, $CDS_i^{(t)}$, also a value in [0,1], measures how well the facial gesture is presented to the i^{th} camera at frame t. A binary gesture recognition decision is made by thresholding the CDS. The i^{th} camera is said to have detected a valid facial gesture at frame t if $CDS_i^{(t)}$ is greater than some empirical threshold value T_d.

Because the gesture recognition decision of the multiple camera algorithm depends on the decisions from the single camera algorithm instances, it is expected that the more robust the single camera algorithm is, the more robust the multiple camera algorithm will become.

Multiple Camera Algorithm

A suitable multiple camera algorithm should only incorporate the gesture recognition decisions from those cameras that have not lost the frontal view of the face, i.e. those cameras with CTS greater than the threshold T_r. Moreover, the algorithm should favour the recognition decisions from cameras that have a better frontal view, i.e. those cameras with larger CTS values.

As a pre-processing step, the tracking score ($TS_i^{(t)}$) for the i^{th} camera at frame t is calculated by running average of the F most recent CTS_i values from frame t, where F is a empirically determined number. The purpose of this running average process is to attenuate any noise resulting from sudden and inconsequential fluctuations in the frame-by-frame CTS values.

A basic multiple camera algorithm is to forward the gesture recognition decision from the best camera as the multiple camera gesture recognition decision, i.e. the multiple camera detection score at frame t is

$$\mathrm{DS}^{(t)} = \begin{cases} 0, & \text{if } \mathrm{TS}_i^{(t)} < T_r \text{ for all } i = 1,\ldots,N \\ \mathrm{CDS}_j^{(t)}, & \text{where } j = \arg\max_i \left\{\mathrm{TS}_i^{(t)}\right\}_{i=1}^{N} \end{cases}$$

and $DS^{(t)}$ is a real value from [0,1].

Another candidate algorithm is to set the multiple camera detection score as a weighted average of the detection scores from all cameras that have not lost the frontal view of the face. Define the index set C = {c $\in$ {1, … , N} | $TS_c^{(t)} \geq T_r$} that contains the indices of all cameras that have not lost the frontal view. Then, the multiple camera detection score at frame t is

$$\mathrm{DS}^{(t)} = \frac{\sum_{c \in C} \mathrm{TS}_c^{(t)} \times \mathrm{CDS}_c^{(t)}}{\sum_{c \in C} \mathrm{TS}_c^{(t)}}$$

The multiple camera gesture recognition decision is similar to the binary recognition decision used in the single camera algorithm. The multiple camera system is said to have detected a valid facial gesture at frame t if $DS^{(t)}$ is greater than the threshold T_d.

CASE SERIES METHODOLOGY

The proposed multiple camera algorithm is to be further developed, implemented, and evaluated within the context of three independent case studies with three children at Bloorview Kids Rehab. Each case study consists of one to two 20-minute experiment sessions per week with the case participant. A minimum of 15 sessions is to be completed by each participant. All three case studies are scheduled to start in March 2008 and finish by the end of May 2008.

Target Population

The child participants are of ages 4–11, have severe spastic quadriplegia (GMFCS levels 4–5), show the ability to reliably produce at least one facial gesture, and can make yes/no decisions.

Instrumentation

Figure 4 shows the instrumentation to be used in the experiment sessions. The prototyping platform consists of three Logitech Quickcam Pro 5000 web cameras and a laptop computer with an Intel Core 2 Duo 2.16 GHz processor and 2 GB of RAM. The web cameras record at a resolution of 320 x 240 pixels and a frame rate of 15 Hz. The laptop executes the multiple camera algorithm. Upon gesture detection, the laptop delivers a transient logical "high" voltage signal to an adapted computer mouse of the classroom computer. The transient voltage signal triggers a left button click on the adapted mouse. This single switch action allows the participants to operate a single switch program on the classroom computer. The user and system setup follows the one shown in Figure 1.

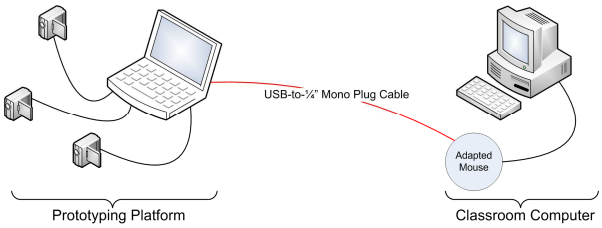

Figure 4: Instrumentation for design and testing the multiple camera facial gesture recognition system.

Experiment Sessions

For each candidate multiple camera algorithm, the participants will be asked to complete one round of a single switch game and one round of word matching exercise. During each activity, which requires about 25 switch actions each to complete, the program running on the classroom computer tallies the number of true positives (N_{TP}), false positives (N_{FP}), true negatives (N_{TN}), and false negatives (N_{FN}). These values are defined as follows:

- N_{TP}: the number of intentional gestures successfully detected.
- N_{FP}: the number of instances a switch action is triggered in the absence of intentional gesture.
- N_{TN}: the number of instances the gesture recognition system is inhibited under the intentional absence of gesture.
- N_{FN}: the number of intentional gestures missed by the gesture recognition system.

Performance Metrics

The multiple camera algorithms will be evaluated according to their positive and negative predictivities. The definitions of positive predictive value (PPV) and negative predictive value (NPV) are

$$\text{PPV} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \text{ and } \text{NPV} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FN}}}$$

respectively. Moreover, logs of the CTS and CDS values will be maintained to determine the proportions of gesture detections that are made from the center camera and from the side cameras.

PRELIMINARY IMAGE PROCESSING

The first step of the single camera algorithm is to locate the user's face from the current frame. Image segmentation based on skin colour thresholding is used to identify the skin regions. Skin colour thresholding is chosen due to its low computational complexity. The difficulty with this method is that walls and wood furniture in the classroom can sometimes be incorrectly identified as skin regions.

It is possible to exploit the spastic head movements of a child with spastic quadriplegia to improve the image segmentation. Because the user's head is frequently moving but the background is still, background subtraction can be used to inhibit the number of incorrect skin regions. A running Gaussian average background subtraction [8] is suitable for this purpose in terms of its low computational complexity and ability to identify background pixels. Figure 5 shows an example of the image segmentation and the effects of background subtraction.

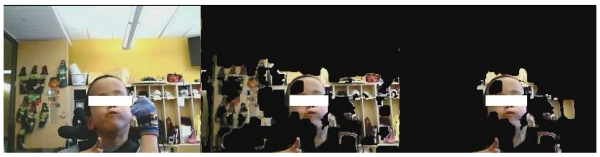

Figure 5: An image processing example. Left: Original frame. Middle: Frame after skin colour segmentation. Right: Segmentation with background subtraction.

CONCLUSION

This paper established the theoretical framework for a multiple camera approach to facial gesture recognition. The paper also outlined the case series methodology and briefly discussed some preliminary image processing issues. The case studies are on schedule to finish by the end of May 2008.

ACKNOWLEDGEMENTS

This research is supported by NSERC, Barbara and Frank Milligan Graduate Fellowship (University of Toronto), and Bloorview Research Institute.

REFERENCES

[1] E. Segers and L. Verhoeven, "Multimedia support of early literacy learning," *Computers and Education*, vol. 39, no. 3, pp. 207–221, 2002.

[2] M. H. Cooney, P. Gupton, and M. O'Laughlin, "Blurring the lines of play and work to create blended classroom learning experiences," *Early Childhood Education Journal*, vol. 27, no. 3, pp. 165–171, 2000.

[3] R. Palisano *et al*., "Development and reliability of a system to classify gross motor function in children with cerebral palsy," *Developmental Medicine and Child Neurology*, vol. 39, no. 4, pp. 214–223, 1997.

[4] C. Hunt and P. Curtis, "My child has cerebral palsy…," *A guide book for families of Bloorview Kids Rehab*.

[5] K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. Bradski, "Communication via eye blinks and eyebrow raises: video-based human-computer interfaces," *Universal Access in the Information Society*, vol. 2, no. 4, pp. 359–373, 2003.

[6] M. Betke, J. Gips, and P. Fleming, "The Camera Mouse: visual tracking of body features to provide computer access for people with severe disabilities," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 1, pp. 1–10, 2002.

[7] R. Kjeldsen, "Improvements in vision-based pointer control", in *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 189–196, 2006.

[8] C. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.