16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

FACIAL VIDEO BASED RESPONSE REGISTRATION SYSTEM

Usman Saeed, Jean-Luc Dugelay

Institut Eurecom
2229 Route des Cretes, B.P. 193, 06904 Sophia Antipolis, France
phone: + (33) 04 93 00 81 00, fax: + (33) 04 93 00 82 00, email: saeed@eurecom.fr, jld@eurecom.fr
web: www.eurecom.fr

ABSTRACT

Today computers have become more accessible and easy to use for everyone, except the disabled. Though some progress has been made on this issue but still it has been focused on either a certain disability or is too expensive for real world scenarios. Major contributions have been made for people lacking fine motor skills and speech based interfaces, but what if they lack both. In this regard we have proposed an integrated video based system that enables the user to give commands by head gestures and enter text by lip-reading. Currently certain gestures and limited vocabulary is recognized by the system but this could be extended in the current framework.

1. INTRODUCTION

Over the past decades keyboard and mouse have been the prevalent interfaces for human computer communication, but they were designed with able-bodied individuals in mind. Unfortunately people who lack certain skills have been left out and tearing down this divide has been a challenging task. Initial efforts were made for people with fine motor disabilities by designing equipment that can substitute as pointing devices but they were specific to a certain disability and at times expensive. Next came speech based interfaces which enabled people to communicate by talking to the computer in a more natural way, but they have mostly been limited to text entry systems. A full fledge system that will allow a user with fine motor and speech disability to enter text and give commands is still far from reality.

In this paper we have proposed a human machine interface for the people with speech and fine motor impairment by using video input. Currently we have focused on two type of interfaces; first one for gesture recognition elaborated by a single choice question for which the user can respond by nodding of the head. The second interface for lip reading is illustrated by a multiple choice questions system where the user only articulates the lip motion of the digit of choice.

The novelty of our approach lies in proposing several image processing techniques that enable us to attain real-time (30f/s) detection of response. The yes/no detection is achieved by combing robustness of a holistic approach with the accuracy of a feature based technique. For digit recognition we have proposed a feature vector that is created by superimposing the outer lip contour of the video sequence. Using this single image as the feature vector reduces the computational cost of the classifier thus enabling us to attain results in real-time.

The rest of the paper is organized as follows: we explain the head gesture recognition system in section 2, then we detail our lip reading system in section 3 and finally we conclude this paper with remarks and future works in section 4.

2. HEAD GESTURE RECOGNITION

Head gesture recognition systems aspire to have a better understanding of subliminal head movements that are used by humans to complement interactions and conversations. These systems vary considerably in their application from complex sign language interpretation to simple nodding of head in agreement. They also carry additional advantage for people with disabilities or young children with limited capabilities.

As part of our project, we focused on a simple yet fast and robust head gesture recognition system to detect the response of users to Yes/No type question. We did not wish to be limited by using specialized equipment thus we have focused our efforts in using a standard webcam for vision based head gesture recognition.

2.1 State of Art

Head gesture recognition methods combine various computer vision algorithms for feature extraction, segmentation, detection, tracking and classification, so categorizing them based on distinct modules would be overly complicated. We thus propose to divide the current head gesture recognition systems into the following categories.

2.1.1 Holistic Approach

This category of techniques focuses on the head as a single entity and develops algorithms to track and analyze the motion of head for gesture recognition. The positive point of these techniques is that as head detection is the main objective, they are quite robust at detecting it. The main disadvantage is the accuracy in detecting small amounts of motion.

Systems introduced in [1, 2] have been based on color transforms to detect the facial skin color. In [3] the mobile contours have been first enhanced using pre-filtering and then transformed into log polar domain. [4] have build a mouse by tracking head pose using a multi-cues tracker combining,

color, templates etc. in layers so if one fails the other layer can compensate for it.

2.1.2 Local Feature Approach

These algorithms detect and track local facial features such as eyes. The advantage is accuracy in motion estimation but the drawback is that local features are generally much difficult and computationally expensive to detect. [5] have proposed a "between eye" feature that is selected by a circle frequency filter. [6] have based there gesture recognition on an infra-red camera with LEDs placed under the monitor to detect accurately the location of the pupil.

2.1.3 Hybrid Approach

The aim of these algorithms is to combine holistic and local feature based techniques. Thus in reality trying to find a compromise between robustness of holistic approaches and accuracy of local feature based techniques, but most of them end up being computationally expensive as they combine various different levels of detection and tracking.

[7] have reported a head gesture based cursor system that detects a heads using a statistical model of the skin color. Then heuristics were used to detect nostrils and tracked to detect head gestures. In [8] they have combined previous work that has been done in face detection and recognition, head pose estimation and facial gesture recognition to develop a mouse controlled by facial actions.

2.2 Proposed Method

The method proposed builds upon previously developed algorithms that are well accepted like Lucas Kanade for tracking. The specific requirements of our project dictate that the head gesture recognition algorithm should be robust to lighting and scale yet fast enough to maintain a frame rate of 30 f/s. On the other hand scenarios concerning occlusion and multiple heads in the scene have not been handled in the current implementation

2.2.1 Face Detection

The first module is the face detector, which is based on cascade of boosted classifiers proposed by [9]. Instead of working with direct pixel values this classifier works with a representation called "Integral Image", created using Haar-like features. The advantage of which is that they can be computed at any scale or location in constant time. The learning algorithm is based on AdaBoost, which can efficiently select a small number of critical visual features from a larger set, thus increasing performance considerably.

The classifier has been trained with facial feature data provided along the Intel OpenCV library [10]. The face detection (Refer figure 1.) using the above classifier is robust to scale and illumination but has two disadvantages, first although it can be considered fast as compared to other face detection systems but still it attains an average performance of 15 f/s. Secondly it is not as accurate as local feature trackers. Thus head detection was only carried out in the first frame and results passed on to the next module for local feature selection and tracking.

2.2.2 Feature Selection and Tracking

The next step involves the selection of prominent features (Refer figure 2.) within the region of the image where the face has been detected. We have applied the Harris corner and edge detector [11] to find such points. The Harris operator is based on the local auto-correlation function.

Tracking of these feature points is achieved by Lucas Kanade technique [12]. It uses the spatial intensity gradient of the images to guide in search for matching location, thus requiring much less comparisons with respect to algorithms that use a predefined search pattern or search exhaustively.

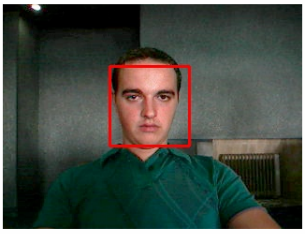

Figure 1: Detected Face

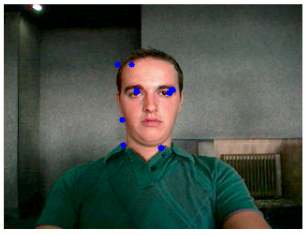

Figure 2: Tracking Points

2.2.3 Yes/No Decision

The final module analyzes the coordinate points provided by the tracking algorithm to take decision whether the gesture is a Yes or a No. First a centroid point is calculated from the tracked points, then the decision is taken based on the amount of horizontal or vertical motion of this centeroid. If the amount of vertical motion in the entire sequence is larger than the horizontal a yes decision is generated, similarly for No.

2.3 Experiments and Results

The development and testing was carried out on a basic 1.5 MHz laptop with 512 MB of RAM, without any specialized equipment. Video input of the frontal face was provided by a standard USB webcam with a resolution of 320X240 at 30 f/s.

Illumination and scale variability are the two main causes of errors in image processing algorithms, thus we have tried to replicate the scenarios most probable to occur in a real life situation. Although the amount of testing was limited to 10 people because of time concerns but due to the amount of variability introduced both in environmental conditions and subject characteristics (glasses/facial hair/sex), the tests are quite adequate.

2.3.1 Illumination Variability

As illumination variation is not dealt with explicitly in our algorithm, we defined three illumination scenarios (Refer figure 3.) to measure the effect of lighting change on our algorithms.

Figure 3: Light Variation

2.3.2 Scale Variability

The second important source of variability is scale, we have experimented with 3 different scale (Refer figure 4.) defined as the distance between the eyes in number of pixels. The 3 measures are S1: 15, S2: 20, S3: 30 pixels.

Figure 4: Scale Variation

2.3.3 Test Questions

The following five questions were selected due to the fact that they can be easily responded to by using head gestures of yes and no.

Q1. Are the instructions clear?
Q2. Are you male?
Q3. Are you female?
Q4. Are you a student of Computer Science?
Q5. Do you like chocolate?

2.3.4 Results

The system was tested with 10 people who were asked 5 questions each by varying the scale and lighting, we achieved a correct recognition rate of 92 % for the Yes/No gesture. The system did have some problem with detecting the face at large distances when illuminated from the side or in direct sunlight.

3. LIP READING

Lip-reading has been generally used to complement noisy audio signal for speech or speaker recognition but lately researchers have started to focus on using lip-reading for other applications like human computer interaction (HCI) and automatic indexing of television broadcasts.

Several situations could arise when we either do not have access to an audio capture device or the user is physically disabled, so for this project we have focused on providing a natural interface for a user to reply to a multiple choice question by using only the video signal. Currently the system can detect the motion of a users lip and recognize the first three digits of the decimal number system i.e. 1, 2, 3.

3.1 State of Art

Like most vision related problems, lip-reading can be subdivided into distinguishable units. The first is the localization of head / face, but here we shall not focus on this step as it has been described previously. Once the head / face is located one moves towards localization of mouth region and then detection, segmentation and extraction of lip features. Finally the extracted features are either integrated with audio features or are used independently for classification.

3.1.1 Mouth Localization

Within the detected face the localization of the mouth region can be considered trivial but has still received its due share of research. The foremost technique has been using the distinguishing qualities of lip color [13]. Mouth region has also been shown to consist of higher number of edges as compared to other facial features [14]. Finally some have used specialized equipment [15].

3.1.2 Lip Detection / Segmentation

Lip detection and segmentation is crucial for lip-reading and model based methods form the core set of techniques. Model based technique such as [16] have proposed the use of snakes for lip segmentation and [17] have built upon the previous method by proposing a "jumping snake". Active shape and appearance models proposed by [18], is a classical model based segmentation technique and can been generalize for lip detection and segmentation.

[19] have proposed lip detection based on point distribution model (PDM) of the face and [20] have reported the use of a 2D Gabor transform for using texture as basis for lip detection.

3.1.3 Feature Extraction

The features can be further classified as appearance based or shape based, in appearance based techniques the pixel values of the ROI are considered as the feature. Thus the feature vector can be obtained by concatenating the ROI pixel in grayscale [21], or color values [19]. The techniques that use optical flow as visual features [22] are also considered as appearance based.

Shape based techniques such as [19] employ a snake to estimate the lip contour, and then use a number of snake radial vectors as visual features, [23] use a lip template parameters instead.

3.1.4 Feature Classification

Hidden Markov models [14] are by far the most widely used classifiers in speech recognition but problems have mostly arisen due to the selection of model parameters. Neural Networks [24] on the other hand operate as black boxes and provide little information on the classification procedure. Efforts have also been made to integrate HMM and NN by [21].

3.2 Proposed Method

The algorithm proposed is a generalized lip reading system with restricted vocabulary. It proposes the use of a superimposed image of lip motion for the entire video sequence to be used as a feature vector for digit recognition. It gets the rough localization of the mouth region and then performs multiple classical image processing techniques to extract the outer lip contour and finally a support vector machine (SVM) is used to classify video as 1, 2, 3 digits from lip motion.

3.2.1 Lip Feature Extraction

The first step towards lip-reading is the detection and segmentation of lip, we propose the following sequence of steps for this purpose.

ROI Selection

The ROI around the lip is extracted first, using the anthropological standards and the head tracking algorithm defined before. All further processing is carried out in this restricted window.

Color Transform

The purpose of this step is to transform the color space so as to enhance the difference between the skin and lip. Due to the fact that lip color is not universal, so we did not wish to exclusively base our detection on color. Thus from the several color transform proposed in the literature we have selected the one proposed by [25]. This color transform is not based on a ratio of color components but only reduces the effect of the blue color which plays a subordinate role in

differentiating between skin and lip color. The color transform has been defined as

$$I = \frac{(2G - R - 0.5B)}{4}$$

Edge Detection

The next step is the extraction of the lip contours (Refer figure 5.), for this end we tested several traditional edge detection techniques like Canny, Sobel, Laplacian edge detectors and found Canny to suit our requirements adequately. The edge strength was controlled with Otsu's thresholding which minimizes the intra-class variance.

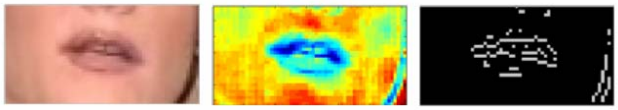

Figure 5: Colour transform and edge map of lip ROI.

Artifact Removal

As we are working in a ROI of the lip we have to carry out some clean up tasks to remove any extra artifacts. The objects in the ROI are first dilated with a disk shaped structuring element to improve the connectivity of the objects. Then any holes present in the objects are filled up. As it is presumed from the creation of the ROI that the lip is the largest object almost in the middle of the ROI all objects sharing a 4 connected pixel with the boundary and objects with small size are removed. Finally the object is eroded to its original size. To have a closed and natural looking lip contour we calculate the convex hull of this rough object and the perimeter of this convex hull is taken as the outer lip contour (Refer figure 6.).

Figure 6: Morphological processing of ROI.

Error recovery

Lip detection being an intricate problem is prone to errors, especially the lower lip edge. We faced two types of errors and propose appropriate error recovery techniques. The first type of error, which was observed more commonly, was caused when the lip was missed altogether and some other feature was selected, this error can easily be detected and corrected by applying feature value and locality constraints. The second type of error occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect and can only be partially corrected by a temporal smoothing filter.

Feature Vector

Once the outer lip contour has been detected for one lip image this procedure is repeated for all the images of the video, somewhat as in [26]. The lip boundary for all images is then superimposed (Refer figure 7.) to obtain the final feature vector. The pixel values of this image are used directly as the feature vector for digit recognition.

Figure 7: Lip detection sequence and superimposed image.

3.2.2 Digit Recognition

Classification of the feature vectors is performed by using a support vector machine (SVM); a supervised classification technique originally designed for a 2-class problem but now has been extended to multiple classes and can also perform regression. It first maps feature vectors into a higher-dimensional space using a kernel function, and then it builds an optimal linear discriminating function in this space. The solution is optimal because the margin between the separating hyperplanes and the nearest feature vectors is maximal.

3.3 Experiments and Results

The experiments were carried on a publicly available database [27]. Videos for 10 persons were randomly selected; each person had three repetitions of each digit (1, 2, 3). The dataset was then further divided into two sets; two third of the videos were used for training and the rest for testing.

Superimposed image of lip motion obtained from the image processing stage are now used as feature vectors for digit recognition. Classification is performed by using a 3 class SVM with RBF kernel, which can handle diverse type of data. The optimal parameters were selected using grid search method suggested by [28]. The following results were obtained when the parameters of SVM were tuned individually for each class and overall combined result.

	Class 1	Class 2	Class 3	Combined
Identification Rate	90 %	70 %	100 %	82 %

Table 1: Identification rate for digit recognition

4. CONCLUSIONS AND FUTURE WORKS

In this paper first we have introduced a real time and highly robust head gesture recognition system. It combines the robustness of a well accepted face detection algorithm with an accurate feature tracking algorithm to achieve a high level of speed, accuracy and robustness. Like all systems, our implementation does have its limitations which were partly enforced by the project definition. The first is that it cannot handle occlusion of the face and second is handling head gestures from multiple persons simultaneously in a given scene.

Secondly we have proposed an experimental digit recognition system with limited vocabulary. The novel approach in this system has been the use of a single image computed from a video sequence as a characteristic feature vector for recognition. This superimposed image greatly reduces the

complexity of the feature vector and enables the use of a much simpler and efficient classifier.
In the future our goal is to develop a real time head gesture and lip reading capable of understanding several complex head gestures with an enhanced vocabulary for lip reading. Other important contribution could be enhancing the interface by gaze estimation and personalizing modalities for specific user.

REFERENCES

[1] P. Lu, X. Huang, X. Zhu and Y. Wang, "Head Gesture Recognition Based on Bayesian Network," in Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, 2005, pp. 492.
[2] Pei Chi Ng and L.C. De Silva, "Head gestures recognition," in Proceedings of International Conference on Image Processing, 2001, vol.3, pp.266-269.
[3] A. Benoit, and A. Caplier, "Head nods analysis: interpretation of non verbal communication gestures," in Proceedings of International Conference on Image Processing, 2005, vol.3, pp. 425-8.
[4] K. Toyama, "Look, Ma--No Hands! Hands free cursor control with real-time 3D face tracking," in Proceedings of Workshop on Perceptual User Interface, 1998.
[5] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes"," in Proceedings of 4th International Conference on Automatic Face and Gesture Recognition, 2000, pp.40-45.
[6] A. Kapoor and R. Picard, "A real-time head nod and shake detector," in Proceedings of Workshop on Perspective User Interfaces, 2001.
[7] V. Chauhan and T. Morris, "Face and feature tracking for cursor control," in Proceedings of 12th Scandinavian Conference on Image Analysis, 2001.
[8] Pengyu Hong and T. Huang, "Natural Mouse-a novel human computer interface," in Proceedings of International Conference on Image Processing, 1999, vol.1, pp.653-656.
[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, 2001, vol.1, pp. 511-518.
[10] http://www.intel.com/technology/computing/opencv/
[11] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in Proceedings of 4th Alvey Vision Conference, 1988, pp.147-151.
[12] B.Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proceedings of DARPA Image Understanding Workshop, 1981, pp. 121–130.
[13] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Video based Sign Language Recognition," in Proc. IAPR Workshop Machine Vision Application, 2002, pp. 318-321.
[14] M. E. Hennecke, K. V. Prasad and D. G. Stork, "Automatic speech recognition system using acoustic and visual signals," In 29th Annual Conference on Signals, Systems and Computers, 1995, vol. 2, pp. 1214-1218.
[15] E. D. Petajan, *Automatic Lipreading to Enhance Speech Recognztion*, PhD thesis, University of IlIinois at Urbana-Champaign, 1984.
[16] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, pp. 321–331, 1988.
[17] N. Eveno, A. Caplier and P. Coulon, "Accurate and quasi-automatic lip tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, pp. 706 – 715, 2004.
[18] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models," in Proc. of European Conference on Computer Vision, 1998, pp. 484–498.
[19] C.L. Huang and Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," *J. Visual Comm. and Image Representation* , vol. 8, pp. 278-290, 1997.
[20] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," in Proc. Int'l Conf. Automatic Face and Gesture Recognition, 1998, pp. 454-459.
[21] C. Bregler, H. Hild, S. Manke and A. Waibel, "Improving connected letter recognition by lipreading," in Proc. International Conference on Acoustics, Speech and Signal Processing, 1993, pp. 557–560.
[22] K. Mase and A. Pentland, "Automatic lipreading by optical flow analysis." *Systems and Computers in Japan*, pp. 67–75, 1991.
[23] D. Chandramohan, and P. L. Silsbee, "A multiple deformable template approach for visual speech recognition," in Proc. International Conference on Spoken Language Processing, 1996, pp. 50–53.
[24] D. E. Rumelhart, J. L. McClelland, *Parallel Distributed Processing: Foundations and Psychological and Biological Models*, MIT Press, Cambridge, MA, 1986.
[25] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition," in Proceedings of the IAPR Workshop on Machine Vision Application, 2002, pp. 318–321.
[26] F. Matta and J-L. Dugelay, "TomoFaces: eigenfaces extended to videos of speakers," to appear in Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2008.
[27]ww.cmpe.boun.edu.tr/enterface07/outputs/final/p12docs.zip
[28] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," in Research Report of National Taiwan University, 2003.