Smart multimodal interfaces for human-computer interaction on train stations

L.J.M.Rothkrantz, J.C. Wojdel, D. Datcu

Abstract: *This paper describes some approaches to processing audio and video signals with support for speech recognition. Speech recognition and lip-reading are currently very popular research topics in the scientific community. They deal with the audio and the video signals and their interactions with each other. The goal of this paper is to show how simple methods can be used for both modalities. Namely, we describe a method for finding the position of the mouth, which is a very important step in audio-visual detection. At the end of this paper we describe some experiments with the proposed system and we compare different techniques.*

Key words: *Human Computer Interaction, Computer Vision, Multimodal Communication, Automatic Speech Recognition.*

INTRODUCTION

The project Creating Robustness in Multimodal Interaction (CRIMI) aims at improving the speech recognition and human-computer interaction in noisy environments (such as train stations). A preliminary processing stage of the current project is the detection of a speaking person who sits or stands in front of the camera. This information is further used by the system in order to track the person's mouth. As soon as some information about the mouth's position is known, more detailed processing techniques are aimed at performing lip-reading or speech recognition. By defining a specific hierarchy for the development of such a complex system, it allows for efficiently focusing on the important tasks such as recognizing speech and on the communication with the user.

Related work

The topic of multimodal interfaces is attracting more and more attention in the research community. The focus is especially on the combination of automatic speech recognition with other modalities, either to improve recognition, as in the work done on audio-visual speech recognition by IBM [1] [3] or to provide a natural human-computer interface as in the German SmartKom project [4]. At Delft University of Technology there is a project running on multimodal interfaces. Systems have now been developed for several modalities. An automatic speech recognizer has been built [6] using Hidden Markov Toolkit (HTK) and an automatic lip-reader has been developed [5]. In this paper we report about the interaction between these two modalities and present the speech recognizer and an integrated audio-visual recognizer. Experiments with these systems were conducted under noisy conditions to show that by including the visual modality in a speech recognizer leads to more noise robustness.

Model

The aim of the CRIMI project is to obtain better performance of the human computer interaction in multi-modal processing system. That means not only that the multiple modalities are to be used in the prototype, but also that each one of them must contribute somehow to the robustness of the system. A prototype of the system has been designed and implemented on an information kiosk at a railway station. In the early stages of the project planning it has been decided that the role of the video processing will be to be aware, to some extent, of the presence of the user in front of the information booth and his/her involvement in the communication. It was assumed that it is possible to correlate the video signal with the audio signal so that to assess the direction in which the person is located while he/she speaks.

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

A schema of the whole project is shown in figure 1. The left part of the figure represents the current project about localization of persons and detection of speech activities. The noisy speech input is processed at first independently from the visual information. Later, the visual cues are used to enhance the signal even further before passing it to the speech recognizer.

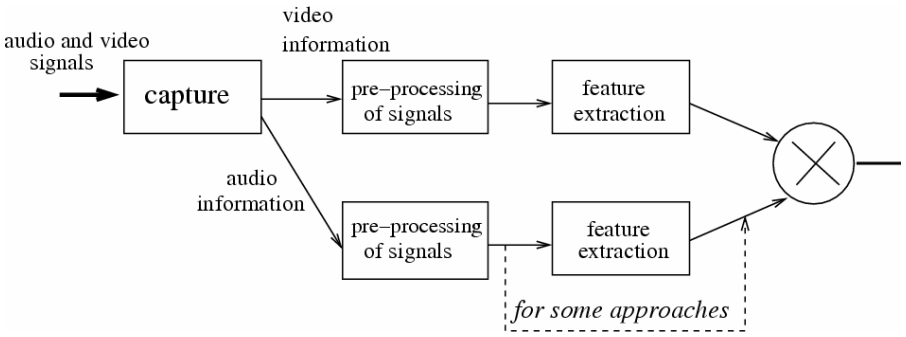

Fig. 1 The basic chain of lip-reading

Locating the face

There is a great number of face location techniques described in the literature. The number of those that can be easily implemented as real-time systems is however rather small. For the current research, two such techniques that are in our opinion the easiest and at the same time provide a good prospect of the robustness, have been considered.

- **Skin colour based locator**

The idea behind the skin colour based face locator is as follows. At first, it may be assumed that the users of the system will be clothed and so the only visible parts of the human skin would be the face and the hands. If the skin colour can be located in the image, then a fairly narrow set of possible face locations is available. Fortunately, the human skin colour has some characteristic properties that are sustained in a range of illumination conditions and across the wide range of skin complexities. It has been shown that using appropriate colour representation it is possible to filter out the visible human skin regions in the image with pretty good accuracy [9].

- **Lip-selective image filtering**

As a first step in video processing, we have to locate the mouth in the image somehow. Fortunately for us, in a video containing only the face, the lips have a distinct colouring that allows us to find them without the need of complicated object recognition techniques. We used hue based filtering for lip selective filtering. It incorporates a parabolic shaped filter that cuts off the colours that fall outside a given interval (formula 1).

$$F_{hue}(h) = \begin{cases} 1-(h-h_0)^2 / w^2, \mid h-h_0 \mid \leq w \\ 0, \mid h-h_0 \mid > w \end{cases} \quad (1)$$

The filter is defined by the center of the interval h_0 and its half width w. Filtereng the image starts with calculating the hue components of a given pixel and then passing it through the F_{hue} function. In order to select lips in the image the value of h_0 must be around zero (red colour) and w about 0.25.

- **IR eye tracking**

Skin colour based face tracking has its own advantages, but it does not work in all cases. Following some research [7], we have found that colour based tracking does not work well at the extremes of both illumination conditions and complexity ranges. The main reason consisted in the occurrence of confusion correlated with the very dark skin colours, given various changes in the illumination. In both cases the problem lays in the fact that with decreasing luminosity of the pixel, the inevitable noise creates growing disturbances of the chromacity. Therefore a method that could be independent of those factors is also strongly needed. One of such methods is near infra red (IR) eye tracking. Using a relatively simple hardware, the scene can be observed in IR part of the spectra. Further more, thanks to some characteristics, the human eye can be detected in the image. By applying special techniques, it is even possible to track the eyes in a video sequence. Because of the way the human eyes are constructed, they will always seem dark in the IR (with the exception of bright retina reflection). An example of simple thresholding operation revels that the eyes can be easily tracked in such image (figure 2).

Fig. 2 IR image thresholded for eye tracking

The literature describes two possible methods for eye tracking using IR: bright pupil method and dark pupil method. The first of them concentrated on the fact the retina is a highly reflective surface in this part of the light spectrum. Therefore the eyes should be visible as bright spots in the image. This approach proves to work well in close range situations when both the camera and the light source are well focused on the eye and situated close to it. However, the case can't be guaranteed for the situations the system aims to deal with. In case of the physical setup, the retinal reflection can be clearly observed only if subject looks straight to the camera. This however will never be the case as the camera must be placed outside the user's area of interest (i.e. the screen).

The second method (dark pupil) works well in long distance setups and it is well suited to fit the project's requirements. It disregards the retinal reflection which can be avoided if the light source is moved sufficiently far form the optical axis of the camera.

In order to capture the near IR imagery, the available hardware had to be modified to some extent. Fortunately, most of the light sensitive elements in modern cameras are highly IR sensitive. In typical consumer electronics this is actually a problem that must be amended with the use of high pass filter that blocks the IR rays from reaching the element. As an example, the figure 3 shows the construction of optical element of the Philips ToUCam Pro. At the end of the lens there is a hot-mirror filter that blocks the IR part of the spectrum. This filter can be easily removed (warranty avoiding procedure) so that the CCD element receives also the IR part of the light. In order to make the camera IR-only it was necessary to replace it with the low-pass filter that can be easily made from non-exposed and fully processed diapositive film.

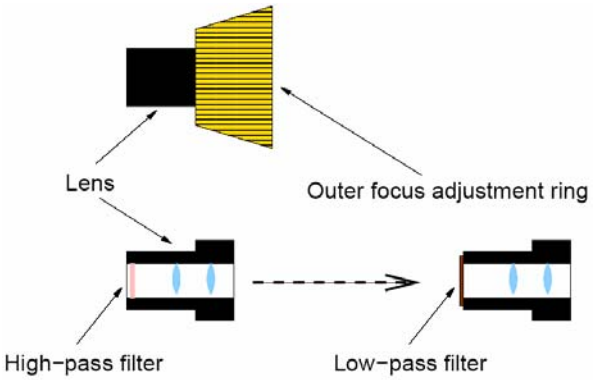

Fig. 3 Lens element in Philips ToUCam Pro web-cam

With the optics modified according to the above instructions, capturing the IR imagery became possible. One more requirement needed for the eye tracking regarded the location of the IR flashlight to be close to the camera. IR flashlights can be bought separately, but in our case it was also easy to build one from scratch.

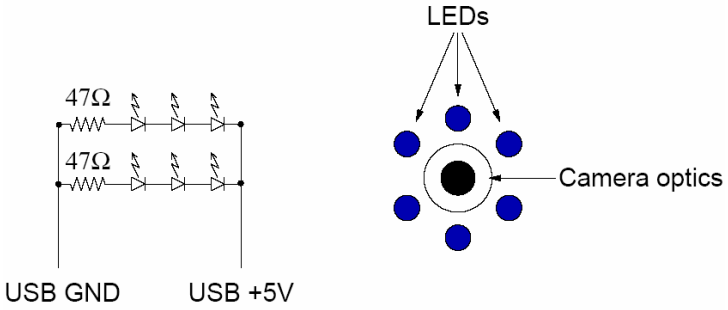

Fig. 4 Electrical circuitry and placement for LEDs' for IR camera

The web-cam that was used for the recordings could also be used as a power source for the array of LEDs that constituted the light (figure 4). The prototypes of such instalment were built using Philips ToUCam Pro and Philips Vesta Fun cameras with the electronics placed on a multifunctional PCB (figure 5). After some experiments [8] we concluded that even though the camera for this part of the system did not have to be of top quality, the recordings obtained from ToUCam Pro were satisfactory.

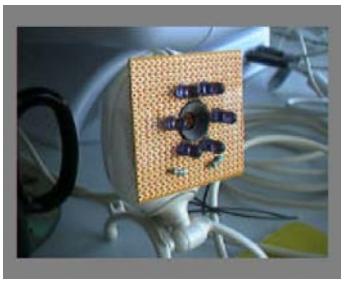

Fig. 5 Modified Philips ToUCam Pro web cam

Experimental setup

A set of audio and video recordings has been taken on a railway station in Eindhoven. During this time a total of 12 people performed some simple tasks using prototype booth (figure 6) that was setup in two different places on the station. The setup consisted of three independent parts:

- interactive application for assessing audibility of natural and synthesized speech
- audio recording hardware capturing the sound from four independent microphones

- video recording hardware with one full-colour camera and one near infra red (IR) web cam

In this paper we describe the preliminary findings concerning the video material gathered during those experiments.

The cameras used for the experimental setup were a Sony DV camcorder placed on top of the screen and a Philips web cam modified to capture IR images. A prior overview of the recordings taken form the experimental setup made by using IR camera was done so as to assess the relevancy of data for the goal of the analysis. For that some random recording samples were selected and rated according to a set of seven different criteria. These criteria are as follows:

Person in front: the user is in the front of the camera. The person's body might be partially occluded, barely visible, etc. but it is somewhere there and most probably is using the system.

Chin/mouth/eyes/top-of-head visible: the user's specific facial areas are visible

Self occlusion: the person occludes the view in some way. This might relate to the cases the user's hand is stretched towards the screen or the head is rotated so that the face is no longer visible.

Other occlusions: occlusions that cannot be identified as inflicted by parts of the user's body. The category relate to occlusions provoked by other people around and also to the case when the person moves outside of the view in the horizontal direction.

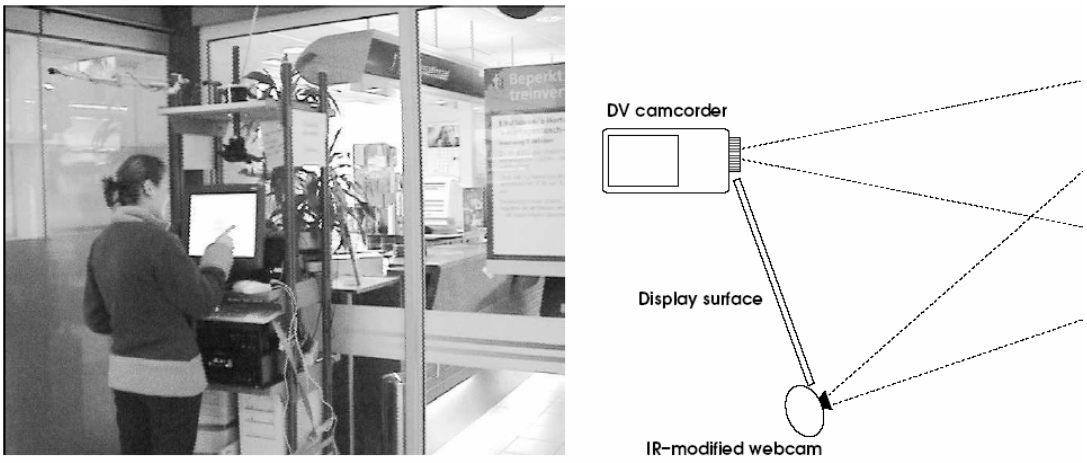

Fig. 6 The experimental setup with (a) the overview of the rack and (b) the schematics of camera placement

Experimental results

The results of manual labelling done on the data are shown in table 1. The percentage values are given per data set and per relevant part. The term *relevant part* of one dataset refers to the situation a person is in front of the booth. The table contains also derivatives of the gathered statistics. Namely, the term *Tractable* relates to the amount of images where both mouth and eyes of the user are clearly visible. The case allows for person tracking and processing of lip-movements. The term *Special case* refers to the occlusions triggered by specific arrangements of the experiment setup. Such cases are not assumed to exist beyond the experimental environment (i.e. occluding face by the AV-synchronization device, occluding the whole view by a hand of the operator standing behind the screen and leaning towards it).

It can be seen that the amount of tractable cases is only about half of the recorded interaction. That means that useful information is available only half of the running time though the video processing is highly robust. One additional aspect of the recordings is

relevant: one of the respondents was an elderly woman who had the tendency to lean on the screen coming almost to the touch. This behaviour rendered the whole video sequence unusable (she wasn't the only elderly person in the recorded data, yet the only one behaving like that).

Table 1: Statistics on the whole dataset

	In relevant data	In all data
Person in front		58.2%
Tractable	50.5%	29.3%
Chin visible	57.0%	33.1%
Mouth visible	73.8%	42.9%
Eyes visible	79.4%	46.2%
Top-of-head	52.3%	30.4%
Self occlusion	13.1%	7.6%
Other occlusions		23.9%
Special cases		7.1%

CONCLUSIONS

There are several important facts that can be concluded from the analysis of the data so far. The information coming from the IR camera will only yield useful information in at most 60 % of the time. Therefore only in those cases the audio processing may benefit from the visual information. Self occlusion happens only in about 10% of the time, so putting the camera below the screen makes a lot of sense. The only reason a person can't be tracked is because it is out of the field of view, not because it occludes the view. Therefore using a wider angle objective for the camera may be a straight way for improving the robustness of the tracking.

REFERENCES

[1] C. P. Neti, G. Leuttin, J. I. Mattews, H. Glotin, D.Vergyri, Audio-Visual Speech Recognition, IBM T.J. Watson Research Center, Summer Workshop, Final Report, 2000.

[2] K. van Turnhout, Audibility of synthetic speech in the train station environment. Experimental set-up internal CRIMI report, April 4, 2002.

[3] A. Verma, T. Faruquie, C. Neti, S. Basu, A. Senior, Late integration in audio-vidual continuous speech recognition, Automatic Speech Recognition and Understanding, 1999.

[4] W. Wahlster, N. Reithinger, A. Blocker, SmartKom: Multimodal Communication with a Life-Like Character, proceeding of Eurospeech, Scandinavia, 2001.

[5] P.Wiggers, L.J.M. Rothkrantz, Integration of Speech Recognition and Automatic Lip-Reading, Springer: Lecture Notes in Artificial Intelligence, vol. LNAI 2448, In: Proceedings of TSD 2002.

[6] P. Wiggers, J. Wojdel, L. Rothkrantz, A speech recognizer for the Dutch Language, In: Proceedings of Euromedia, Modena, Italy, 2002.

[7] J. Wojdel, L. Rothkrantz, Using Aerial and Generic Features in Automatic Lip-reading, Proceedings of Eurospeech, Scandinavia, 2001.

[8] J. C. Wojdel, Preliminary recording requirements for gathering video data, internal CRIMI report, May 17, 2002.

[9] J. C. Wojdel, Automatic lipreading in the Dutch language, PhD thesis Delft University of Technology, ISBN 83-89003-62-7, 2003.

ABOUT THE AUTHOR

Assoc. Prof. L. J. M. Rothkrantz, Department of Man-Machine Interaction, Delft University of Technology, Phone: +31 15 2787504, E-mail: L.J.M.Rothkrantz@ewi.tudelft.nl.