Look to Talk: Embedded Face Localization for Activating Speech Recognition

X. Kong[1], J. F. Guitarte Perez[2] and K. Lukas[2]

[1] *Munich University of Technology, Email: Xiangpeng.Kong@mytum.de*
[2] *Siemens AG, Munich, Email: Lukas@siemens.com, Jesus.Guitarte@siemens.com*

Abstract

In this paper we present a highly efficient solution to activate speech recognition on embedded platforms and to enhance its performance by utilizing visual information. The idea of 'Look to Talk' employs a face localization algorithm which detects the face when the user is close enough to an appliance and looks at it. This substitute of the traditional 'Push to Talk' system makes the usage more convenient and user interface friendlier, as no manual interaction is required. After the face localization, an embedded lip finding and tracking algorithm is applied, which detects the movement of the lips and activates the speech recognition. This solution improves the recognition results significantly and avoids interferences from the surrounding, especially in multi-user and acoustically noisy environments. The face localization algorithms are dedicated to low resource platforms and can therefore be embedded in various appliances like ticket vending machines or household equipment.

Introduction

There are many scenarios where speech recognition is applied using a "Push to Talk". The user has to activate the system by pushing a button when he wants to interact with the machine. But it is desirable that the user can speak directly to the machine without making extra actions. This is especially interesting in vendor machine scenarios, e.g. coffee machines. If the user has to press a button in order to activate the speech recognition (Push to Talk to avoid high false alarm rates), he would rather directly press the button to get the coffee. The problem of avoiding Push to Talk has been profusely studied in speech recognition and solutions such like word spotting have been proposed.

In this article we present a solution that gets the attention of a machine by using visual information rather than audio information, which we call "Look-to-Talk". The use of visual information for voice activity detection was already proposed in [1]. We have developed an application that will only be active when the user is looking at the machine and his face is close enough to the machine; therefore it is assumed that the user wants to use it. This is also the normal situation in natural human conversations. Furthermore, the activation of the speech recognition will not occur immediately when the speaker is in front of the device, but until visual speech activities (lip movements) are detected. This would be a Visual-Voice-Activity-Detector (V-VAD). This feature makes our system advantageous against other activity detectors such like movement sensors that will detect the presence (movement of the speaker) but not when he speaks.

Lip Finding and Tracking

The first task that must be solved is the automatic detection and tracking of the mouth region. Our algorithm [2] is made up of two different functions: lip finding and lip tracking. Lip finding is applied when no previous information of the lip position is available. This happens in the first frame of a sequence or whenever the lips in the previous frame have not been correctly located. Lip finding is based on a geometric model of the face. Structures of pixels are evaluated in order to know if their relative positions match a simplified prior model of the face (see figure 1.a). In particular, this model only takes into account the relationship between the locations of the eyebrow(s) and the mouth.

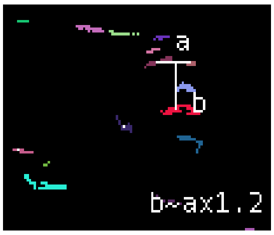

Fig. 1.a: Lip Finding

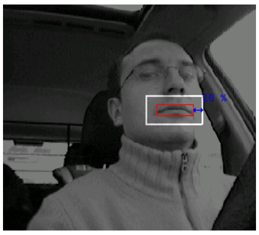

Fig. 1.b: Lip Tracking

Lip tracking proceeds when knowledge of the lip position is available in the previous frame. In this case we rely on the hypothesis that the position of the lips will not change very much between adjacent frames. Lips will be searched in an area that is 10% larger than the region where they were located in the previous frame, see figure 1.b. Furthermore, lip tracking is more reliable and requires less resource than lip finding. The algorithm requires 4 MHz (2,7% of CPU load), providing a correct lip detection and tracking in more than 94% of the frames. For the tests an emulation of an ARM920T with 150 MHz and 16 Kbyte bi-directional cache has been used (external memory access speed is 150 nsec for non sequential and 10 nsec for sequential access).

Lip Movements: ASM

In previous paragraph we have shown how the system can automatically know whether there is a speaker in front of the camera or not, now we will show how to use the lip movement information. For the design of this system, the shape information of the lips is extracted by using Active Shape Models (ASM) [3].

In ASM a priori knowledge of the plausible mouth deformations is learnt in a training process. A set of points (landmarks) must be consistently located in the mouth contours of the training set. Rigid transformation dependencies are first removed by using Procustes Analysis. Then Principal Component Analysis (PCA) is applied on the aligned points. PCA computes the main variation modes of the points. This allows the deformations to be described only by a small set of parameters:

The eigenvectors ϕ_i (where *i* refers to each variation mode) and the corresponding eigenvalues λ_i of the landmark coordinates covariance matrix *S* are computed and sorted so that $\lambda_i > \lambda_{i+1}$. If Φ_i contains the *t* eigenvectors corresponding to the *t* largest eigenvalues, a set of points describing the mouth contour *x* can be approximated by:

$$x = \bar{x} + \Phi \cdot b \qquad (1)$$

where $\Phi = (\phi_1 \mid \phi_2 \mid \ldots \mid \phi_t ,)$ and *b* is a *t*-dimensional vector given by:

$$b = \Phi^T \cdot (x - \bar{x}) \qquad (2)$$

The *b* coefficients will describe the different variations of the mouth with respect to its mean value which can be seen in figure 2.

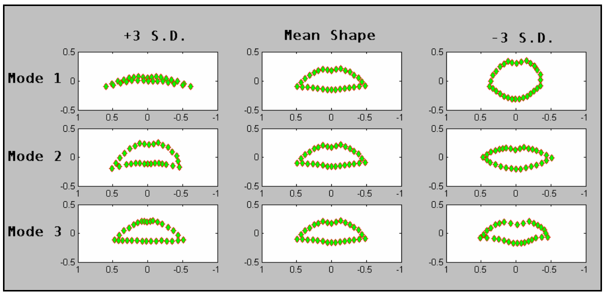

Fig. 2: The first three ASM lip models are shown; in the middle column the mean mouth is presented and the other columns represent the three most significant shape variation modes for ± 3 standard deviation.

The first coefficient (weighting of Mode 1) is very adequate for detecting voice activity, as it provides information of the mouth opening and it is independent on the scaling, rotation and translation. This implies that when the speaker is moving his head but not speaking (there is no lip movement), the system will not produce false alarms, which would probably be generated if just a movement sensor was used. A reconstruction of the mouth using only this first coefficient ϕ_1 can be seen in figure 3.

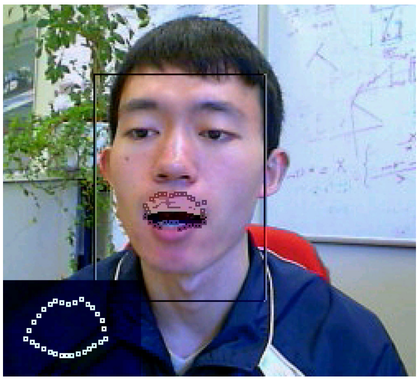

Fig. 3: Mouth reconstruction using 1st ASM Mode

For every frame *n*, the value *M(n)* will be set to 1 (active) when the first coefficient ϕ_1 changes more than a threshold Th_{move}, which is obtained empirically.

$$M(n) = \begin{cases} 1 & if\,(\phi_1(n) - \phi_1(n-1) > Th_{move} \\ 0 & if\,(\phi_1(n) - \phi_1(n-1) \le Th_{move} \end{cases} \qquad (3)$$

Finally, we are going to assume that the user in front of the device has spoken to the system during the recognition frames *[n-T,n]* when the value *M* was active for a number of frames higher than a threshold. In this way the system robustness against short time lip tracking errors will be improved.

$$A(n) = \begin{cases} 1 & \sum_{j=0}^{T} M(n-j) > Th_{time} \\ 0 & \sum_{j=0}^{T} M(n-j) \le Th_{time} \end{cases} \qquad (4)$$

The microphone is going to be activated when the lip finding and tracking algorithm finds the user's lips. Speech recognition is going to provide a recognition result. This will be a hypothesis associated with the period of time the microphone is activated, frames *[n-T,n]*. But the system will assume the hypothesis as correct only if the activation signal *A(n)* is 1 at the end of the recognition.

Conclusion

The proposed system solves the problem of Push to Talk, as for activation the user only has to approach to the device. Face and lip localization algorithms dedicated to embedded platforms enable an automatic activation of the speech interaction. Furthermore they reduce the insertions of words said by other users, as the recognition result will only be considered as correct when lip movements were detected during recognition. Several applications can take profit of our embedded visual activity detection, for example vendor machines, industrial appliances or house hold equipment.

Acknowledgement

The authors wish to express special thanks to E. Lleida, A. Frangi, F. Althoff, S. Hoch, V. Ramesh and N. Paragios for their good advises and interesting discussions.

References

[1] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, and C. Jutten, "An Analysis of Visual Speech Information Applied to Voice Activity Detection", Proc. ICASSP, vol 1, pp. 601-604, 2006

[2] J. F. Guitarte, K. Lukas, A. F. Frangi, "Low Resource Lip Finding and Tracking Algorithm for Embedded Devices," Proc. Eurospeech, vol. 3, pp. 2253-2256, 2003.

[3] J. Luettin, N. A. Thacker, "Speech reading using probabilistic models," Computer Vision and Image Understanding, vol. 65, pp. 163-178, 1997.