Audiovisual Data Fusion for Successive Speakers Tracking

Quentin Labourey[1,2], Olivier Aycard[1], Denis Pellerin[2] and Michele Rombaut[2]
[1]*LIG, Grenoble, France*
[2]*GIPSA-lab, Grenoble, France*

Keywords: Audiovisual Data Fusion, Skin Detection, Sound Source Tracking, Talking Face Tracking.

Abstract: In this paper, a human speaker tracking method on audio and video data is presented. It is applied to conversation tracking with a robot. Audiovisual data fusion is performed in a two-steps process. Detection is performed independently on each modality: face detection based on skin color on video data and sound source localization based on the time delay of arrival on audio data. The results of those detection processes are then fused thanks to an adaptation of bayesian filter to detect the speaker. The robot is able to detect the face of the talking person and to detect a new speaker in a conversation.

1 INTRODUCTION

Over the last decades, robotics has taken a growing importance, in our society and imagination as well as in Science. Until a few years ago, "robots" often only amounted to sensors placed in the infrastructure. Those types of sensors can make people feel ill-at-ease because they are intrusive. Companions robots that share the environment with humans are a way to remedy the problem. They are equipped with sensors and are able to perceive the behavior of people and interact with them. There are various kinds of sensors, such as video and audio sensors. However the data acquired from those sensors is most of the time uncertain, noisy or partial. That is why multi-sensor data fusion is now an extremely prolific field of research, as it allows to obtain more complete information from partial data in a robust way.

This work presents a method based on audiovisual data fusion for tracking the positions of the successive speakers of a human conversation with a robot equipped with visual and audio sensors. The robot must be able to detect the faces around him and choose which one is the speaker. If the speaker is not in the visual scene, he must be able to track his position.

As the audio and video modalities are very different, a late fusion seems adequate for the present work (Snoek, 2005), where detection is performed on each modalities and a final decision is taken as to where the speaker is by fusing the detection results. It implies that this work is based on three main axis:

FACE DETECTION ON VISUAL DATA: The literature on face detection is quite abundant. In their survey of the state-of-the-art of face detection, Zhang (Zhang and Zhang, 2010) show that most of the recent works on face detection are feature-based and appearance-based methods, where a classifier is trained either with direct positive examples of faces or with visual features extracted from faces (patterns, edges, etc...). Although a lot of work with different methods exist, such as Support Vector Machines (Osuna et al., 1997) , Neural Networks (Vaillant et al., 1994), one of the principle contribution of the last decades on face contributions is Viola & Jones algorithm (Viola and Jones, 2004).

AUDIO SOUND SOURCE LOCALIZATION: Various methods exist to localize a sound source thanks to microphone array. Brandstein (Brandstein and Silverman, 1997) divides those methods into 3 broad parts: steered-beamformer based methods, which are probabilistic methods (Valin et al., 2004) and are most of the time quite time-costly, high-resolution spectral-estimation based methods, which are based upon a spatiospectral correlation matrix, and time-difference of arrival methods (Gustafsson and Gunnarsson, 2003), which use the delay of arrival between microphones and are widely spread on various complexity levels.

FUSION OF THE RESULTS FROM BOTH MODALITIES: Various types of fusion methods exist in different fields, a large part of them being based either on classifiers and machine learning (Rao and Trivedi, 2008), or on estimation of the state of objects (Nguyen and Choi, 2010). Hospedales (Hospedales and Vijayakumar, 2008) worked on a similar problem with unsupervised learning methods.

The method of the present work is a 2-steps pro-

Labourey Q., Aycard O., Pellerin D. and Rombaut M..
Audiovisual Data Fusion for Successive Speakers Tracking.
DOI: 10.5220/0004852506960701
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 696-701
ISBN: 978-989-758-003-1

cess (figure 1): In the perception step, faces are detected on visual data thanks to a skin detection algorithm based on naive bayesian classifiers while sound source localization is performed on audio data thanks to the cross-correlation method. In the decision step, the results of the perception step are fused thanks to an adaptation of bayesian filter. Finally the choice of the position of the speaker is performed by the robot.

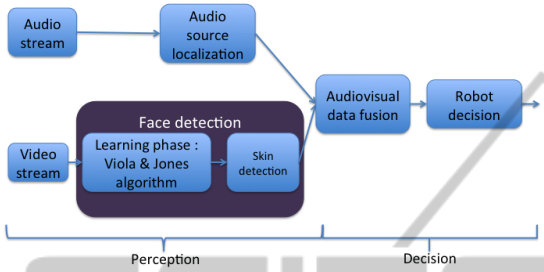

Figure 1: Speaker detection: the two steps of perception and decision.

This method provides with a real-time and robust way to keep track of the speaker in a conversation, while being adapted to low-quality data (independent from morphological patterns) and low-quality microprocessors (only one scanning of each frame).

This article is elaborated as follows: Section 2 describes the features of the robot used in this work, section 3 details the method for the detection on audio and video modalities independently and section 4 describes the decision process thanks to the audiovisual data fusion method. Finally, results are presented and discussed in section 5.

2 EXPERIMENTAL PLAFTFORM

This section details the specifications of the robot chosen to illustrate the speaker tracking method and the features on which this work has been performed.

Reeti® (figure 2) is a "humanoid" robot produced by Robopec (www.reeti.fr/), in France. It is equipped with low-cost sensors for perception (two cameras in the eyes and a stereophonic microphone in the base), and with servo-motors for action (turning the head and face expressions).

The two cameras can be accessed simultaneously and produce a 640x480 pixels color RGB video, at 10 frames per seconds. Only one camera is used instead of steroscopic vision to be faster with a less complex algorithm.The sound is acquired in stereo, with the two channel inputs at 15 centimeters apart.

An important constraint is that the field of audio localization covers the whole angular field in front of

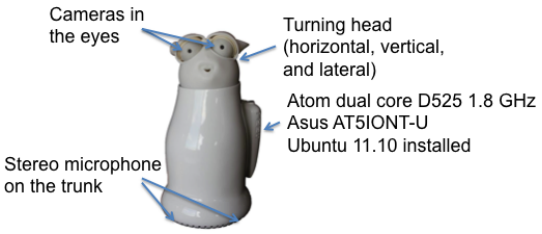

Figure 2: Reeti® the robot.

the robot from 0 to 180 degrees, whereas the field of visual detection covers only about 50 degrees.

The global objective here is to track the angular position α of the current speaker in a horizontal plane with respect to the robot (figure 3).

3 PERCEPTION STEP

This section details our algorithms for face detection from video data and sound source localization from audio data.

3.1 Visual Data: Face and Skin Detection

The goal of the video processing is the frame-by-frame detection of the angular positions of every face present in the visual scene with respect to the robot, i.e. the process should give out a number N of angles $[\alpha_1, \cdots, \alpha_N]$, each of them being the estimated angular position of a candidate face with respect to the robot in a 2D horizontal plan.

Some particular cases can happen such as occlusion (the faces can be fully or partially hidden by an object like a hand, handkerchief, other person...), position changes (the faces may move during a conversation and their angle and position change), lighting changes (e.g. a person standing in front of a window moves, etc...).

Viola & Jones algorithm (Viola and Jones, 2004) performs very well on separate images with a high rate of detection and a low rate of false alarms. However it is shown that the particular cases cited above related to the dynamicity of human conversation make the exclusive use of this algorithm unadapted to this work. Besides, the needed algorithm must be able to process a video stream in real-time, on a low-quality microprocessor. In such context, Viola & Jones algorithm cannot be used online and for each frame for tracking as it is too time-costly, although it can be used to obtain examples of face pixels.

This is why skin detection (Chai and Ngan, 1998) is used, based on naive bayesian classifiers (Schnei-

derman and Kanade, 1998) on the 3 video color channels. The proposed algorithm relies on three steps:

ONLINE CREATION OF THE TRAINING DATASET: The training set is composed of skin (positive samples) and non-skin pixels (negative samples). To create the actual training set, the Viola & Jones algorithm is performed on each frame until some faces are detected. The pixels that are detected as part of faces are positive examples while those that are not are negative examples. If necessary, to increase the number of positive samples, it is possible to perform Viola & Jones on several successive frames.

TRAINING OF THE NAIVE BAYESIAN CLASSIFIERS: 3 classifiers are trained, one for each color channel. Let R be the event "The considered pixel has a red value of R", with $R \in [0;255]$ and C be the event "The considered pixel is a positive sample". From the training set, the aim is to learn the probability distributions $P(R|C)$ and $P(R|\neg C)$. This is done by computing the histograms of the positive and negative samples.

Let $H_{C,R}(i)$, $i \in [0;255]$, be the histogram computed on all the positive samples of the training set on the red channel, and $H_{\neg C,R}(i)$ be the histogram computed on all the negative samples of the training set on the red channel. Then the probability distributions $P(R|C)$ and $P(R|\neg C)$ are computed as follows:

$$P(R|C) = \frac{H_{C,R}(R)}{\sum_{i=0}^{255} H_{C,R}(i)}, P(R|\neg C) = \frac{H_{\neg C,R}(R)}{\sum_{i=0}^{255} H_{\neg C,R}(i)} \quad (1)$$

A similar reasoning is performed to obtain the probability distributions on the blue channel $P(B|C)$ and $P(B|\neg C)$, and on the green channel $P(G|C)$ and $P(G|\neg C)$.

CLASSIFICATION OF INCOMING PIXELS: The classification is made with the assumption that the three color channels are independent. A score between 0 and 1 is given to each pixel. The higher the score obtained is, the higher the chance that this pixel is skin:

$$Score = \frac{Score_{skin}}{Score_{skin} + Score_{\neg skin}} \quad (2)$$

where

$$\begin{aligned} Score_{skin} &= \eta_1 \cdot P(r|C) \cdot P(g|C) \cdot P(b|C) \\ Score_{\neg skin} &= \eta_2 \cdot P(r|\neg C) \cdot P(g|\neg C) \cdot P(b|\neg C) \end{aligned} \quad (3)$$

with η_1 and η_2 as normalization parameters, and r,g,b respectively the red, green and blue values of the pixel.

Each incoming frame is processed with the naive bayesian classifiers. At the end of the classification, 0 or more skin regions are detected. Each detected skin region with a number of pixels under an arbitrary number of pixels (decided depending on the minimal size of head to detect) is removed: it enables to get rid of noise in the detection and to avoid keeping track of very small regions of the images for the detection.

3.2 Video Angular Conversion

At the end of the frame processing, the centers of gravity of the N detected skin regions are obtained, with $N \geq 0$.

The goal of the angular conversion is to transform the x-value of each center of gravity into an angle between 0 and 180 degrees in a horizontal plane in the robot coordinate system. The robot reference, defined as the center of the head of the robot is used as the origin of the coordinate system.

In this work, it is considered that the camera reference is the same as the robot reference (figure 3).

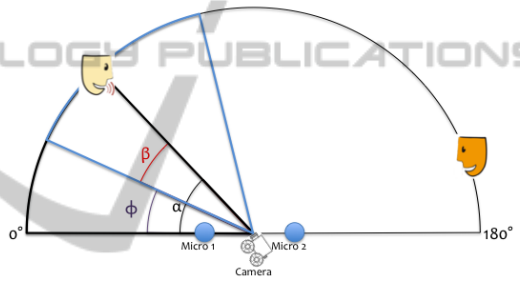

Figure 3: Angular definition: view from above

The angular position of the center of gravity with respect to the robot is computed as follows:

$$\alpha = \beta + \phi \quad (4)$$

- ϕ is the angular position of the camera obtained by linear conversion: 1 step between two camera positions is equivalent to 1.27 degrees.
- β is the angular position of the pixel in the image, obtained by linear conversion: 1 degree contains 13 pixels.

The skin detection process thus gives out a number N of angles $[\alpha_1, \cdots, \alpha_N]$ with respect to the robot at which candidates for faces are detected.

3.3 Audio Data: Sound Source Localization

The goal of the audio data processing is to produce a unique angle δ corresponding to the estimated angular position of the current speaker with respect to the robot.

The hypotheses on the framework for sound source

localization are: one person is talking at a time, in a silent environment (the noise is inherent to the microphones and the robot's motors and fan). For this work, the detection does not have to be extremely precise as it will be fused with the video data. Because only two microphones are used and real-time processing is required, it seemed thus logical to choose a state-of-the-art cross-correlation method (C.C.M.) thanks to the Time Difference of Arrival (T.D.O.A.) (Gustafsson and Gunnarsson, 2003) method.

4 DECISION STEP: AUDIOVISUAL DATA FUSION

The perception step is performed on visual & audio data at a period of 100 ms. The set of N faces extracted at time t, called the set of video observations is noted $O_t^v = (\alpha_t^1, \cdots, \alpha_t^N)$, with $N \geq 0$. In the same manner, the unique audio observation at time t is noted δ_t. All the observations lie within a $[0;180]$ range.

From those data, a decision has to be taken as to where the speaker in the conversation is located. Those observations are uncertain, and there is high variety of possible visual and audio configurations. That is why an adaptation of a bayesian filter is used (Thrun et al., 2005) with a differentiation of cases similar to a model rupture. The bayesian filter fits with the problem as it is a method to predict/estimate the state of a system equipped with sensors from the observations made by those sensors.

In the present case, the state S_t of the system at time t is the angular location of the speaker. As the angular measures are uncertain, the angular field is divided into 18 possible states, each of them being an angular region 10 degrees wide (figure 4). The set of states is noted $\mathbb{S} = (S^0, \cdots, S^{17})$.

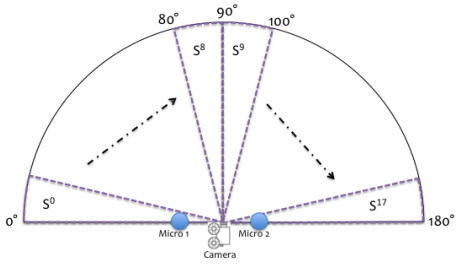

Figure 4: State repartition: 10 degrees for each state.

As the visual field is narrower than the field of audio localization, the position of the head has an impact on the model. There are 93 possible head positions $(Po^1, \cdots, Po^{93})$. The position of the head at time t is noted Po_t

In the present section, the field of visual detection will indifferently be referred to as field of vision (F.O.V.) or field of visual detection.

4.1 Dynamic Model

The dynamic model corresponds to the model of the natural evolution of the conversation, i.e. the probability distribution $P(S_t|S_{t-1})$. Because the speaker in a conversation can change at any time, it is difficult to estimate the position at time t+1 of a speaker when it is known at time t.

In first approximation, the evolution model is built just to track the current speaker: it is considered that from a time t to a time t+1, or in 100 ms interval, the same person is still talking, even though this person might have moved in between. The probability distribution $P(S_t|S_{t-1})$ is thus a discrete "gaussian-like" form around S_{t-1}. Figure 5 is an example for $S_{t-1} = S^6$.

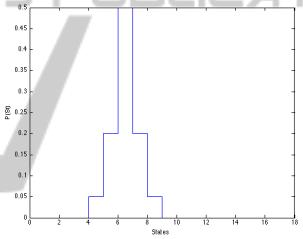

Figure 5: Dynamic model: $P(S_t|S^6)$.

The case where the speaker changes between t and t+1 is not handled by this dynamic model.

This is considered thanks to a simplified version of interacting multiple models (IMM) (Farmer et al., 2002): when the information from both sensors or only the audio sensor is in complete conflict with the previous estimate, the information coming from the previous estimate is nullified and a uniform distribution is used. This case is detailed a bit more later on in this section.

4.2 Audio Sensor Model

The sensor model represents the trust put into the sensor. Let $P(\delta_t|S_t)$ be the audio sensor model. The audio measure is not highly precise. The probability of δ_t knowing S_t should be a gaussian which mean-value is the mean angular value of the angular interval S_t and which variance σ_δ is high. Figure 6 shows an example of possible audio sensor model with $\sigma_\delta = 20$ for $S_t = S^7$. In this case the mean-value of the gaussian is 75.

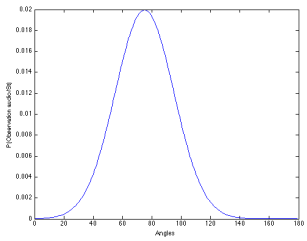

Figure 6: Dynamic model: $P(\delta_t|S^7)$, $\sigma_\delta = 20$.

4.3 Video Sensor Model

As the F.O.V. of the camera is not 180 wide, the observations depend on Po_t, and the sensor model must be computed at each iteration: The video sensor model is in fact $P(\alpha_t|S_t, Po_t)$. 2 separate cases must be reviewed for the video sensor model: if the speaker is in the visual scene then the probability of detecting a face is centered on the state in which the speaker is. If he is not, the probability of detecting a face is uniform.

$$P(\alpha_t|S_t, Po_t) = \begin{cases} U(\alpha_t), \text{ if } S_t \cap F.O.V. = \emptyset \\ \mathcal{N}(\mu_\alpha, \sigma_\alpha) \text{ if } S_t \cap F.O.V. \neq \emptyset \end{cases} \quad (5)$$

where U is the uniform distribution over α_t, $\mathcal{N}(\mu_\alpha \sigma_\alpha)$ is a normal distribution of mean value μ_α equal to the mean value of S and of variance $\sigma_\alpha = 10$.

4.4 Estimation of $\mathbf{P}(S_t|O_t^v, \delta_t, Po_t)$

Once the dynamic model and the video models are defined, the estimate of S_t can be computed. Two main cases have to be differentiated here:

THE SPEAKER IS IN THE VISUAL SCENE: If all the present faces are well detected, one or more video observations are available. There is a choice to be made as to what visual observation is the best. This is done by taking the video observation the closest to the audio observation. The video observation α_t used for fusion is chosen as follows:

$$\alpha_t = \underset{\alpha_t^i \in O_t^v}{\arg\min}(|\alpha_t^i - \delta_t|) \quad (6)$$

The fusion process is then the state-of-the-art prediction/estimation process.

The prediction step is:

$$P(S_t|\alpha_{0:t-1}, \delta_{0:t-1}, Po_{1:t}) = \sum_{S_{t-1}} (P(S_t|S_{t-1}) \cdot P(S_{t-1}|\alpha_{1:t-1}, \delta_{0:t-1}, Po_{1:t-1})) \quad (7)$$

The estimation step is computed with the hypothesis that the video and the audio sources are independent:

$$P(S_t|\alpha_t, \delta_{0:t}, Po_{1:t}) = \eta \cdot P(\delta_t|S_t) \cdot P(\alpha_t|S_t) \cdot P(S_t|\delta_{1:t-1}, \alpha_{0:t-1}, Po_{1:t}) \quad (8)$$

with η a normalization factor.

Between the iteration t and the iteration t+1, if the speaker has changed, $P(\delta_t|S_t)$ and $P(\alpha_t|S_t)$ will overweight $P(S_{t-1}|\alpha_{1:t-1}, \delta_{0:t-1}, Po_{1:t-1})$ and modify the model in a few iterations, corresponding to the delay.

Once the estimation is over, the probability distribution of the speaker location is known. The most likely face video observation is chosen. Let ML be the maximum of the location probability distribution.

$$ML = \max_{S \in \mathbb{S}}(P(S_t|\alpha_{0:t}, \delta_{0:t}, Po_{1:t})) \quad (9)$$

The chosen video observation O^{Final} is the closest observation to ML.

THE SPEAKER IS NOT IN THE VISUAL FIELD: The model has to change. A test has to be made on the audio: if the audio observation is outside the video, it is a model rupture, then the weight of the previous estimate and the video observations are nullified and only the audio observation is considered: the final observation is δ_t.

5 RESULTS

Some results obtained from data extracted from the robot can be observed on figure 7 and an example of full video is accessible on the internet at the following URL: https://www.youtube.com/channel/UCl4mcfTKzB8orarl-SC7qgg.

It is difficult to evaluate the method as there are a lot of parameters that are linked to the experimental platform. However, some global remarks can be made:

- The full perception-decision chain is performing in real-time.
- The particular cases described in section 3.1 are fixed by the skin detection.
- Even if the audio detection is rough, when fused with the face detection it allows us to easily differentiate between two persons side by side as long as the space between their faces is bigger than the resolution of skin detection, so that two video observations are made.

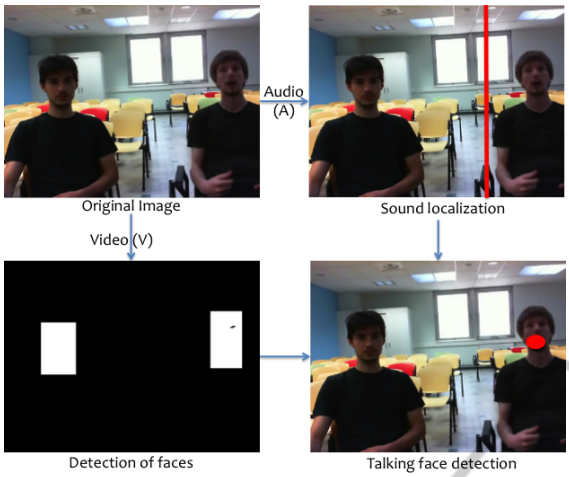

Figure 7: Results of audiovisual data fusion for a conversation involving 2 speakers.

- The time of response when the speaker change is no more than a few frames (around 4) before the detection of the new speaker (figure 8).
- In the different videos that were tested the speaker is always detected after some time

Figure 8: Detection of a new speaker among 3 persons after a small delay.

6 CONCLUSIONS

This work focuses on three main issues: produce detection methods on low-quality visual and audio data from low-cost sensors, elaborate a robust audiovisual data fusion method adapted to the situation, and make the method follow a real-time constraint.

A full processing chain has been elaborated: it relies on the independent processing of the audio stream (sound source localization) and the video stream (face detection), fused in a late fusion process to create the decision. One of the main advantages of this processing chain is that each of its links can be modified and upgraded in future works. There are perspectives to improve this processing chain: the robot should be able to differentiate two speakers speaking at the same time at this stage, and the detection process can be improved by dealing with the bias introduced in the coordinate system. A lot of improvement could come from the inclusion of new sensors such as depth sensors or laser sensors to discriminate between region of interests to be explored directly instead of exploring the whole frame at each time. Improvement can also be made from using motion in the video, instead of frame-by-frame processing.

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

REFERENCES

Brandstein, M. S. and Silverman, H. F. (1997). A practical methodology for speech source localization with microphone arrays. *Computer Speech & Language*.

Chai, D. and Ngan, K. (1998). Locating facial region of a head-and-shoulders color image. In *Automatic Face and Gesture Recognition. Proc.*

Farmer, M. E., Hsu, R., and Jain, A. K. (2002). Interacting multiple model kalman filters for robust high speed human motion tracking. *ICPR'02*.

Gustafsson, F. and Gunnarsson, F. (2003). Positioning using time-difference of arrival measurements. In *ICASSP'03*.

Hospedales, T. M. and Vijayakumar, S. (2008). Structure inference for bayesian multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Nguyen, Q. and Choi, J. (2010). Audio-visual data fusion for tracking the direction of multiple speakers. In *Control Automation and Systems*.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *CVPR'97*.

Rao, B. D. and Trivedi, M. M. (2008). Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition. *In Acoustics, Speech, and Signal Processing, 2008.*

Schneiderman, H. and Kanade, T. (1998). Probabilistic modeling of local appearance and spatial relationships for object recognition. In *CVPR'98*.

Snoek, C. G. M. (2005). Early versus late fusion in semantic video analysis. *ACM Multimedia*.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press.

Vaillant, R., Monrocq, C., and Le Cun, Y. (1994). Original approach for the localisation of objects in images. *Vision, Image and Signal Processing, IEEE Proc.*

Valin, J.-M., Michaud, F., Hadjou, B., and Rouat, J. (2004). Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach. In *ICRA'04*.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *IJCV'04*.

Zhang, C. and Zhang, Z. (2010). A survey of recent advances in face detection.