Audio-Visual Data Fusion using a Particle Filter in the Application of Face Recognition

Michael Steer, Ayoub Al-Hamadi, Bernd Michaelis
Institute for Electronics, Signal Processing and Communications (IESK)
Otto-von-Guericke-University Magdeburg, Germany
{Michael.Steer, Ayoub.Al-Hamadi}@ovgu.de

Abstract

This paper describes a methodology by which audio and visual data about a scene can be fused in a meaningful manner in order to locate a speaker in a scene. This fusion is implemented within a Particle Filter such that a single speaker can be identified in the presence of multiple visual observations. The advantages of this fusion are that weak sensory data from either modality can be reinforced and the presence of noise can be reduced.

1. Previous Work

The idea of fusing audio and video data together in the application of tracking is not new. Various other implementations have been attempted. J. Vermaak et al describe a method [5] in which a standard contour tracking algorithm consisting of an edge detector and a Particle filter is used in conjunction with a Time Difference on Arrival (TDOA) calculation to deduce the speaker location from auditory data received by a pair of microphones. The audio data is used for initialization and the video for localization in an attempt to utilize the strengths of each modality. This method enhances the existing visual tracking successfully and can detect speaker ‘ping-pong’. The current implementation is however not a real time solution. Huiyu Zhou et al.[6] use a histogram matching based technique for image based detection and a TDOA algorithm once again for audio localization. The audio undergoes pre-processing to remove noise and a Ricatti Kalman filter is used to further reduce spurious detections before both audio and video observations are passed to a Weighted Probabilistic Data Association (WPDA) filter for fusion and tracking. This paper describes an attempt to combine these ideas into a solution that achieves data fusion in real time to provide useful segmentation as a pre-processing step for subsequent tracking tasks.

2. Strategy Overview

The overall strategy for the application involves taking observation vectors created from both audio and visual localization routines as inputs to a particle filter. The physical configuration of the sensors is comprised of two microphones placed equidistant from a central camera on a horizontal plane. Observations are taken from both the audio and visual sensors at each time step. Steps are taken to make the observations as accurate as possible, but in the event of multiple candidate matches for speaker position in either modality, no effort is made to select the correct one. Instead all observations are collated and combined with the fusion algorithm within the particle filter.

3. Audio strategy Overview

Audio observations take the form of an angle describing the direction to a sound source from a point centered on the camera position. This angle is calculated by a Time Difference On Arrival (TDOA) [3] algorithm which exploits the fact that sound arriving at each microphone will be delayed according to its direction of origin. To find this time delay a cross correlation function is used. For two signals $S_1$ and $S_2$ with one signal delayed by time $\omega$ the cross correlation is given as:

$$GCC(\omega) = \sum_{n=-\infty}^{\infty} S_1[n]S_2[n-\omega]$$ (1)
Figure 1. Example of how audio buffer is split into frames. The sample delay $d$ can be used together with the sample rate to determine the time difference.

A fixed search window of $N$ is applied to the time delays, yielding:

$$GCC(\omega) = \begin{cases} 
\sum_{n=-1}^{N-\omega} S_1[n]S_2[n+\omega], & 0 \leq \omega \leq N - 1 \\
\sum_{n=1}^{N+\omega} S_1[n]S_2[n-\omega], & 1 - N \leq \omega \leq 0 
\end{cases}$$

Due to the nature of audio sensory information, multiple detections are possible for a single source. An algorithm exploiting the properties of Signal to Noise (SNR) and Zero Crossing Rate (ZCR) [1] for each audio frame is used to distinguish between speech and non-speech. Only frames identified as containing speech are marked for further processing. In order to account for reverberation a Generalised Cross Correlation function using a Phase Transform (GCC-PHAT) as introduced by Knapp and Carter and Brandstein and Silverman [4] is used. This results in a more robust version of the correlation function. The composition of frames is chosen as to exploit the physical characteristics of the sensory arrangement.

4. Video Strategy Overview

When a set of video observations are required, a frame is taken from the camera and processed through a face detection algorithm based on a skin colour model [1]. A logical AND operation is performed on the output of three color model filters representing RGB, HSV and CrCb colour spaces. These colour spaces all define boundaries of pre-learnt skin colour pixel values. This technique gives a balance between the efficiency of lower level feature extraction and the robustness of more computationally expensive methods. The resulting skin colour regions are then examined for contours which may describe faces in the scene. Morphological methods are then used to discard contours which do not fit facial characteristics of size and shape. As a final step, bounding boxes matching the facial segmentation in terms of size and rotation are generated and output.

5. Particle Filter Implementation

The particle filter is an implementation of the CONDENSATION algorithm [2], the use of which is motivated primarily due to its ability to deal with substantial observational clutter. This is required as the multimodality of the observation model can lead to numerous false positives with respect to possible speaker positions. Sample parameters are chosen to reflect a bounding box describing the current state of a speaker in the scene. A sample set $X$ of size $N$ at time $t$ is defined as:

$$X^t = \{x_0, x_n, \ldots, x_N\}$$

Parameters of each sample are

$$x(\gamma) \in \{\text{xpos}, \text{ypos}, \text{width}, \text{height}, \text{angle}\}$$

The set of observations at time $t$ as

$$Z^t = Z_A^t \cup Z_V^t$$

$$Z_A^t = \{a_0, \ldots, a_M\}$$

$$Z_V^t = \{v_0, \ldots, v_K\}$$

In this case $M$ is the dimensionality of each audio observation and $K$ is that of the video. Observations take the form

$$a(\gamma) \in \{\text{xpos}\}$$

$$v(\gamma) \in \{\text{xpos}, \text{ypos}, \text{width}, \text{height}, \text{angle}\}$$

Each sample is held within the filter with an associated weight representing the confidence the filter has in the
samples ability to describe the correct speaker position with respect to the observations. It is the initialization and propagation of these weights which determine the output of the filter.

5.1. Initialization

The initial step is to generate the particle set which begins by obtaining a set of initialization observations. The 5 parameters for each sample $x$ are generated by sampling from a multi-modal Gaussian with each mode being an audio or video observation value for that component. The distributions for each parameter $\gamma$ can be described as:

$$p(x | Z , \gamma = 0) = \sum_{m=0}^{M} \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} e^{-\frac{(x[\gamma] - Z_m^x[\gamma])^2}{2\sigma_\gamma^2}}$$

$$+ \sum_{k=0}^{K} \frac{1}{\sqrt{2\pi\phi_\gamma^2}} e^{-\frac{(x[\gamma] - Z_k^v[\gamma])^2}{2\phi_\gamma^2}}$$

$$p(x | Z , \gamma > 0) = \sum_{k=0}^{K} \frac{1}{\sqrt{2\pi\phi_\gamma^2}} e^{-\frac{(x[\gamma] - Z_k^v[\gamma])^2}{2\phi_\gamma^2}}$$

(9)

5.2. Selection

In this step, particles are selected from the distribution by means of factored sampling. This is achieved by selecting a random index into an array containing the cumulative probabilities of the particle set. This leads to particles with weights corresponding or in close proximity to peaks in the probability distribution being sampled many times, whereas samples at low points in the distribution are discarded. The probability distribution itself is the set of weights assigned to each particle in the particles set the previous time step during the measurement step given by (9).

5.3. Prediction

In this step, the state of each sample is altered to reflect an underlying temporal behavioral model. As we are modeling human behavior it is difficult to derive a deterministic model to capture movement in the scene. However the properties of the auto regressive model make it effective at capturing the behavior of many naturally occurring phenomena so this is used to drive the prediction step. The first order autoregressive model

![Figure 2. Generalization of sample set weighting. The top diagram shows a set of observations with the green boxes being extracted from the video and the blue from audio. The common sample parameter between the modes is the 'x' value so the given sample set yields a particle density with respect to the x-axis shown in the second diagram. The bottom diagram shows the overall effect of the partitioning on the particle density over the (x,y) domain with the darker areas showing a higher particle density.](image-url)
Figure 3. Example output showing skin colour segmentation (top left), GCC output (bottom left), probability densities obtained from the observations (top right). The output (bottom right) shows two video observations in red, multiple audio detections in blue and the output of the filter in green.

yields the following relationship between the current value of a particle \( x_t \) and the predicted particle \( x_{t+1} \)

\[
x_{t+1} - \mu_\gamma = \theta_\gamma (x_t - \mu_\gamma) + \sigma_\gamma \beta \tag{10}
\]

Where \( \mu_\gamma \) and \( \sigma_\gamma \) are the mean and standard deviations and \( \sigma_\gamma \) a scaling factor for the parameter number \( \gamma \) and \( \beta \) is normalized Gaussian noise. The standard deviation and scaling factors are chosen such as to capture differences in behavior between the different parameters. This is model specific and must be obtained by a result of experimentation however in general it can be said that the values are set to reflect the fact that the \( x \) and \( y \) parameters have a higher variance than the width, height and angle.

5.4. Measurement

In this step, a set of observations is obtained and an observational density over the 5 dimensional state space is calculated. Each sample is then assigned a weight according to its particular position in state space relative to observational density. The 5 dimensional state space is described by the same mixture of Gaussians defined in (9). In order to correctly weight a particle with respect to the observation density it is necessary to compute a normalized sum of each of its component parameters.

\[
p(x | Z) = \frac{1}{5} \sum_{\gamma=0}^{4} p(x_\gamma | Z) \tag{11}
\]

6. Results

Results show that the described technique is good enough to track a speaker in a scene and distinguish between multiple speakers. Performance is very good, real-time on a standard desktop workstation. Fig 3 shows successful fusion of multiple observations to arrive at the correct speaker position. The audio observations require further refinement before they can be considered as strong a discriminator as the video observations. Therefore an initial step for further work is to increase the microphone count so as to produce more accurate measurements and add a second dimension other than a single horizontal measurement.

Acknowledgement

This work is supported by Foerschungspraemie (BMBF, FKZ: 03FPB00213), Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by DFG, and BMBF Bernstein-Group (FKZ: 01GQ0702).

References