SEE WHAT YOU HEAR AND HEAR WHAT YOU SEE
AUDIO VIDEO TRACKING

D'Arca E., Hopgood J., Robertson N. - Heriot Watt University, Edinburgh University
{ed88, n.m.robertson}@hw.ac.uk, James.Hopgood@ed.ac.uk

Abstract

Surveillance in urban environments relies almost exclusively on visual information. Video information is spatially very informative but there are important outstanding research problems: tracking, for example, can be difficult when large occlusions occur. Since anomalous events often have a distinct audio component, for example, gunshot or car crash, ***the aim of this project is to investigate how to combine concurrent audio and visual information at the signal level in order to determine whether a significant event has occurred.***

Aims and Novelty

- To realize a system that tracks objects and recognizes and differentiates dynamic events in video scenes like a human operator.
- To try to reproduce the ability of the human brain to recognizes events relying on multimodal information, especially the audio and the video ones

Applications

An interesting and general application of the Audio Video (AV) data fusion could be a multi modal system that is able of continuously inferring the **presence of humans**and to **recognize their behaviours** as well as to **remove the ones that become inactive** in the analysed scenario.

Challenges

- Audio and Video should fairly contribute to describe the recorded scenes
- A lot of data to be processed
- Noisy signals (audio reverberations, occlusions and AV background noise)

Fusing Audio and Video:Background

The core of the systems will be definitely the **fusion module**. In fact all the computational efforts should be convey in this in order to obtain a meaningful association of the two different data kinds.
Tracking: The work that has been done so far is used to combining data in a unique vector in all the situations in which targets have to be localized over time.
Recognition: Instead correspondences between the two signals are exploited mostly to detect and recognize events or to discover anomalies.

Modeling the Fusion
In general two main approaches have been followed in the literature:

- AV features vectors are concatenated and used as a unique vector and processed in single modality [1, 2]
- Each signal is processed independently from the other one and the fusion (performed on a based common feature such as synchrony f.e.) is the final step[3, 4]

AV Tracking

Involves:

- Estimation of the arrival angle of the audio
- Video detection
- Signals filtering and smoothing
- Fusion
- Joint state estimate

AV Recognition

Involves:

- Understanding of the audio signal
- Scene interpretation
- Signals filtering and smoothing
- Fusion
- Joint state estimate

Proposed Approach

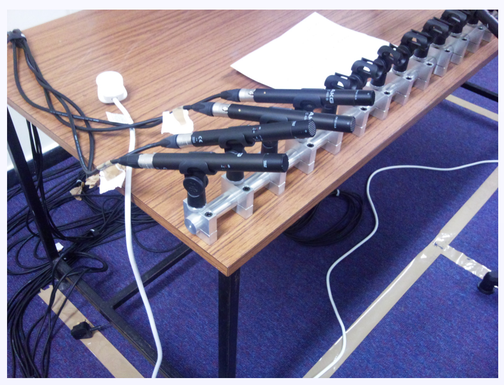
Our grand aim is to realize a new fusion method employing the two signals assuring that one cue will not overwhelm the other. Besides we want to join the signals at the ground level (i.e. the signal level) in order to take advantage of the overall information without the risk of discarding some useful data.
Specifically we will focus on **tracking people and show that the coupled information overcome reverberations (audio) or occlusions (video)issues quantifying how much.** Then we will use this information to **predict in video the place where the event may be occurring or vice versa to steer the microphones towards the last.**

Experiments Setup

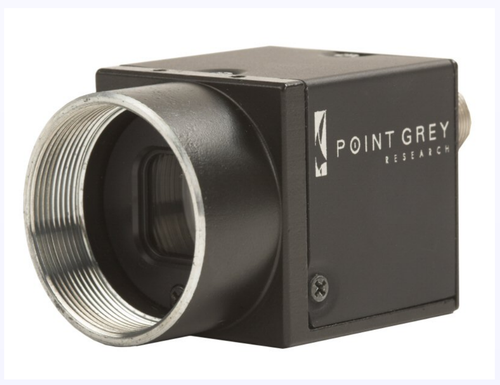
The first experiment for investigating our problem will be performed in an indoor environment. We will working off-line starting with a simple scenario i.e. a man who is walking and speaking continuously in the room. At first *we want to understand how video data could help audio to recover the position of the target information as this is usually a difficult task due to sound noise but mostly reverberations*.This all will be done assuring a correct calibration and synchronization between cameras and microphones. Our system will be composed by up to 4 CCD cameras and 4 pairs of cardioid microphones as well as a PC (CPU: Intel Core2 Duo @ 2.66 GHz). The parameters of the cameras and the positions of the microphone in the room will be calculated in advance. The layout of the experimental environment and a picture of a similar tracking context are shown in Figure b) and c) respectively.

Tools		
DETECTIONS	SENSOR NAMES AND SPECS	TRACKING ALGORITHM
Video	Flea2 - CCD camera max res 1392x1032@15FPS	mean shift
Audio	AKG C480B - cardioid pattern with ULN amplifier	GCC+PHAT+EKF

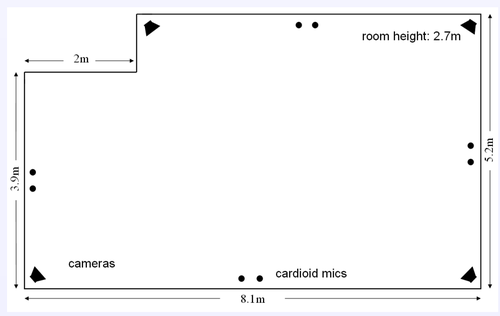
a) Equipment Table



b) AKG C480 microphones



c) FL2-14S3M/C camera



d) Room Layout

e) A Similar Scenario

Objectives

The long term objective of our research is to explore new solutions of **fusing audio and video** signals in order to obtain an enhancement framework which **outperforms single modality systems**.In particular we want to:

- combine the two signals to predict how likely is to appear a source in the video/audio field;
- aid one type of two weak signal to avoid a track loss exploiting the complementarity of the other one.

Tracking Goals

- How can video enhance the performance of an audio based tracker
- How can audio enhance the performance of a video based tracker

Event Detection Goals

- How different sounds can help a classifier to distinguish between different actions
- How the visual information could be used to disambiguate between similar sounds

References

[1] Gatica-Perez, D. and Lathoud, G. and Odobez, J.-M. and McCowan, I.,'Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings',in *Audio, Speech, and Language Processing, IEEE Transactions on*,2007

[2] Checka, N. and Wilson, K.W. and Siracusa, M.R. and Darrell, T.,'Multiple person and speaker activity tracking with a particle filter', in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004

[3] Cristani, M. and Bicego, M. and Murino,V.,'Audio-Visual Event Recognition in Surveillance Video Sequences'in *Multimedia, IEEE Transactions on*,2007

[4] Barzelay, Z. and Schechner, Y.Y.,'Harmony in Motion', in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007