Audiovisual Event Detection Towards Scene Understanding

C. Canton-Ferrer, T. Butko, C. Segura, X. Giró, C. Nadeu, J. Hernando, and J.R. Casas
Technical University of Catalonia, Barcelona - Spain

1. Goals

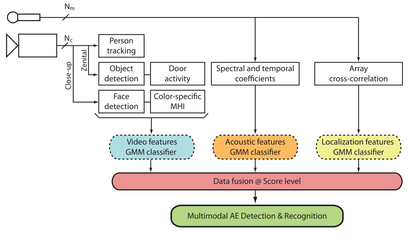
- Detect and recognize events from a **multimodal** source of information
- The events to be recognized are usually produced in a room scenario:
  - Laughing, coughing, keyboard typing, clapping, door slam, yawning, phone ringing, paper wrapping, etc.
- These events have been usually detected and recognized using acoustic information solely but they have a visual counterpart that can be exploited to recognize them
- Three sources of information are employed: **acoustic features**, **acoustic localization**, **video features**

Acoustic Event	Audio	Localization	Video
Applause	✔	✔	✔
Cup clinking	✔	✔	✘
Chair movement	✘	✘	✔
Coughing	✔	✔	✔
Door slamming	✔	✘	✔
Key jingling	✔	✔	✘
Door knocking	✔	✔	✘
Keyboard typing	✘	✘	✔
Phone ringing	✔	✔	✘
Paper wrapping	✘	✘	✔
Footsteps	✘	✘	✔

There is room for multimodal fusion and modality compensation

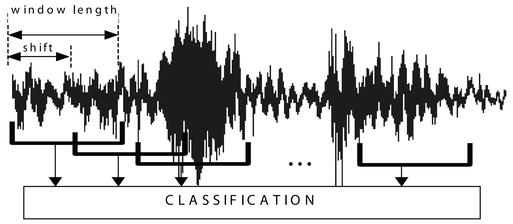
2. System Flowchart

- The data analysis is processed following a pipeline flow



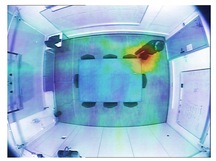
3. Monomodal Event Detection

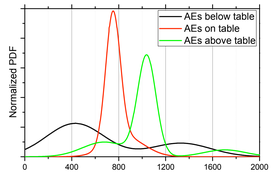
Acoustic Features



- **Features employed (spectro-temporal):** 16 frequency-filtered log filter-bank energies with their derivatives (plus its temporal evolution).
- **Classification method:** GMM-based classifier (5 Gaussians). All events are trained using acoustic features and determine the baseline performance algorithm.

Acoustic Localization



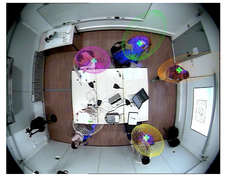


AE localization pdf's

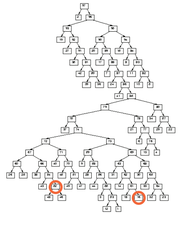
- **Features employed:** Spatial 3D localization of the acoustic source employing the SRP-PHAT localization method, based on computing time delays among microphone pairs.
- **Classification method:** Based on defining a meta-classes grouping those events according to the *xy*-position in the analysis scenario (i.e. door slam) and to their *z*-position (i.e *footsteps* or *clapping*).

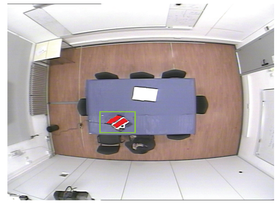
Video Features

- **Person tracking (position+velocity):** This information is useful to recognize events like *footsteps*. Information about height changes is employed to detect *chair moving* (a sudden change in height).

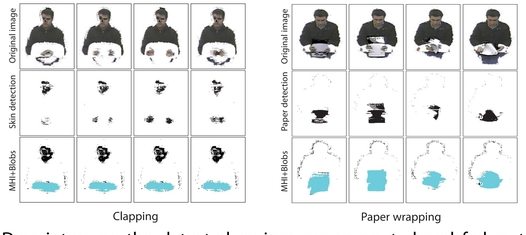


- **Object detection:** Detection of specific objects in the room may allow recognizing activities related to them. A particularly challenging event such as is *keyboard typing* benefitted from a laptop detector.





- **Color-specific MHI:** Motion History Image and Energy (MHI and MEI) can be tailored to describe motion of a specific color.



Descriptors on the detected regions are computed and fed as the input to the video based classification system.

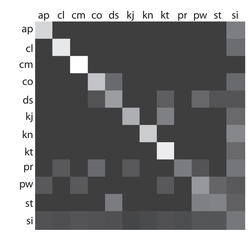
- **Face detetion:** Combining information from color-specific MHI with the position of the face gives a useful feature to detect events like *coughing* or *phone ringing* (since the hand motion ends close to the face).
- **Position-specific descriptors:** Some activities like *door slam* are well localized, therefore, visual activity close to the doors of the room may give a hint to detect this event.
- **Classification method:** A series of GMM-based classifiers are defined for every event using the features provided by the video systems. The output is given as a vector with the probability of every event.

4. Multimodal decision fusion

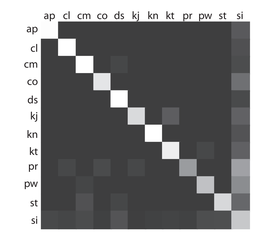
- All modalities are synchronized and normalized
- Two schemes are discussed in this paper: Weighted Mean Average (WAM) and Fuzzy Integral (FI). Some remarks can be drawn:
  - WAM is based on a trained linear combination of all information sources but does not account for crossed dependencies
  - FI is presented as a more efficient alternative to WAM

5. Results

- Experiments conducted over the recorded dataset, showed that some events can be better recognized when using a multimodal approach. Particularly:
  - Footsteps: 244% improvement
  - Paper wrapping: 15% improvement
  - Overall improvement: 7.5%
- Confusion matrices showed this effect:



Baseline AED



A+L+V FI-AED

6. Contributions & Conclusions

- A dataset has been designed to evaluate the performance of the proposed techniques. Description:
  - 5 cameras at a resolution of 765x576 pixels, at 25 fps
  - 6 T-shaped 4-microphone clusters, at 44 kHz
  - Calibration and synchronization fulfilled

The employed dataset is available for research purposes. Ask the authors.

- Exploiting the visual counterpart of events that have been historically considered as "acoustic", leads to an improvement of their recognition
- Acoustic localization also provides discriminative features to recognize meta-classes of events (below/on/over table, near/far to door, etc.)
- Future work involves exploring more sophisticated information fusion schemes