Basic Concept for fusion of sound and video

A. Störmer, D. Arsić, and G. Rigoll
Institute for Human-Machine Communication
TUM (Munich University of Technology)
Germany
Reminder: TUM's Basic Concept for behavior recognition

- Bottom Up Strategy
- LLF – Low Level Features
- LLA – Low Level Activity
  - Sound will probably inserted here
    - “ready“ modules of sound recognition
- HLA – High Level Activity
  - Adapt scenario descriptions to sound events

Beside this, directional information provided by sound will probably be useful in WP3, bayesian occupancy filter.
LLF - Low Level Features for sound

- MFCC
- magnitude / loudness
- ....

Better leave this to the experts...

... we suggest a late fusion based on the results of „sound event detectors“
LLA - Low Level Activities

Late fusion: Include the detectable sound events to the list of available Low Level activities.

• Classification results of sound
  • detected sound events
    • gunshot
    • screaming
    • talking
    • ...
• A list of recognizable classes/ events is needed
Reminder: HLA – High Level Activity

• put LLA in context to each other and to environment
  – environment map with knowledge about context information (e.g. position of ATM/ entrance areas/ desk, restricted area ..)

• formulate scenarios (grammar, state machine)
  – experience with HMMs, states modelled by Bayesian Nets (but could also be anything else (Petri Nets, Rules))

• Concepts and structures developed for
  – Approaching a stationary object/person
  – Person walking/standing next to each other
  – Person following other person
  – Two persons approaching each other
  – operating ATM
  – loitering in the region of ATM
  – approaching ATM while being operated
  – running away from ATM
Insert detectable sound events into the grammar

Example: persons having a small talk

- approaching another person
- standing AND talking
- leaving the person

distances, movement
distances, movement, sound
distances, movement

IMAGE
IMAGE & SOUND
IMAGE

bayesian net?