Hough transform-based mouth localization for Audio Visual Speech Recognition

Gabriele Fanelli, Juergen Gall, Luc Van Gool
Overview

- Introduction
- Method
- Results
- Conclusions
AVSR - overview

Video Signal

Audio Signal

Visual Features Extraction

Audio-Visual Fusion

Recognition

“Hello”
AVSR - overview

Video Signal

Audio Signal

Visual Features Extraction

Audio-Visual Fusion

Recognition

“Hello”
AVSR – our approach

- Face Tracking
- Eye Detection
- Eye Tracking
- Mouth Localization
- Visual Features Extraction
- Audio Features Extraction
- Audio-Visual Fusion
- Recognition

"Hello"
Motivation

Facial feature points detectors can fail because of:

- Partial occlusions
- Facial hair
- Large mouth deformations
Eye detection based on isophote curvature

- Isophotes - curves of equal image intensity
- Each pixel votes for a radius, i.e., an eye center, with:
  - Length = reciprocal of the curvature
  - Direction = gradient
  - Orientation = bright to dark
- Votes from points with high curvedness are given higher weights
- The peak in the accumulator image is chosen as the eye center

Ref. Accurate Eye Center Location and Tracking Using Isophote Curvature, R. Valenti and T. Gevers, CVPR08
Eye tracking

- Eye detection fails when the iris is not visible, e.g., blinks
- Kalman filters are employed to track the eyes' positions
- Located eyes tell us mouth's scale and rotation
- Mouth localization can focus on a scaled and rotated region of interest
HT-based mouth localization

- Hough transform-based methods are robust to large appearance and shape variations, partial occlusions
- Implicit Shape Model as random forest
- Position and appearance of a patch are learned and cast votes for the object's center
- Votes are summed up into a Hough image, peak is the mouth

Ref. Class-specific Hough Forest for Object Detection, J.Gall and V. Lempitsky, CVPR09
Audio-Visual Speech Recognition

- Audio-visual fusion and recognition is achieved by MSHMM, each modality being modeled as a Gaussian mixture
- We extract DCT features from the stream of normalized mouth images
- Audio features are mel-frequency cepstral coefficients
- First and second order derivatives are added in both modalities
Tests - scale and orientation

- BioID database
- 1521 images of 23 individuals
- Uncontrolled illumination
- Resolution 384x288
- Facial hair, talking people, glasses, eyes closed
- Ground truth locations available for eyes and other facial features
Tests – mouth localization

- Two tests:
  - Full pipeline
  - Eye locations given

- Compared to Vukadinovic and Pantic's point detector, ICSM05
Results - successes
Results - failures

Mainly caused by wrong face and eyes detections
AVSR Experiments - CUAVE DB

- 36 speakers
- Digits 0-9 in American English
- Good image conditions
- Faces are frontal

- DB split in 6 sets, train on 5, test on the 6th
- Audio corrupted by white noise at several SNR levels
- Mouth detector trained on the full BioID database
Tests – Video-Only Speech Recognition

![Graph showing word recognition rate vs. number of visual features. The graph compares automatic vs. manual recognition.]
Tests – AVSR

We acknowledge Dr. Mihai Gurban (EPFL) for providing source code for feature extraction and audio-visual speech recognition.
Questions?