Multimodal Signal Processing, Saliency and Summarization

Petros Maragos, Alexandros Potamianos, Athanasia Zlatintsi and Petros Koutras

slides: http://cognimuse.cs.ntua.gr/icassp17

Tutorial at IEEE International Conference on Acoustics, Speech and Signal Processing 2017, New Orleans, USA, March 5, 2017
Tutorial Outline


2. Visual Processing and Saliency: P. Koutras

3. Audio processing and Saliency: A. Zlatintsi

4. Text processing and Saliency: A. Potamianos

5. Multimodal Video Summarization: All
Part 1: Multimodal Signal Processing, Audio-Visual Perception and Fusion
Part 2: Visual Processing and Saliency

Visual Saliency Models

Eyes Fixation Prediction

Framewise Saliency

Spatio-Temporal Processing

Visual Saliency Models

- Input Image
- Multiscale
- Low-level features
- Features
- Training
- Outputs
- Other
- Motion, density, etc.
- Image features, etc.

Eyes Fixation Prediction

- Original RGB Frames
- Luminance SDE
- Color Stream SDE
- Framewise Saliency

Framewise Average

Discrete Spatial Frequency U
Discrete Spatial Frequency V

Spatio-Temporal Dominant Energy

Tutorial: Multimodal Signal Processing, Saliency and Summarization
Part 3: Audio Processing and Saliency

---

Modulation Features and Saliency

Input Audio

Feature Extraction

Fusion

Audio Saliency

Summarization Algorithm

Audio Events

Audio Markers

Audio Segmentation

Audio Summary

Multiband Teager Energies

Saliency Curve

Audio annotated salient segments
Summary x2: Included segments
Part 4: Text Processing and Saliency

From word frequencies to semantic networks and beyond …

e.g., semantic-affective mapping
Part 5: Multimodal Video Summarization

Monomodal Saliencies

Multimodal Saliency & Movie Summarization Algorithm

Emotion Annotation

Tutorial: Multimodal Signal Processing, Saliency and Summarization
Part 1
Multimodal Signal Processing, Audio-Visual Perception and Fusion

Petros Maragos

Tutorial at IEEE International Conference on Acoustics, Speech and Signal Processing 2017, New Orleans, USA, March 5, 2017
Part 1: Outline

- Applications – Motivations of A-V signal processing
- A-V Perception
- Bayesian Formulation of Perception & Fusion Models
- Application: Audio-Visual Speech Recognition
- Application: Multimodal Gesture Recognition in HRI
Applications - Motivations
Human versus Computer Multimodal Processing

- Nature is abundant with multimodal stimuli.
- Digital technology creates a rapid explosion of multimedia data.
- Humans perceive world multimodally in a seemingly effortless way, although the brain dedicates vast resources to these tasks.
- Computer techniques still lag humans in understanding complex multisensory scenes and performing high-level cognitive tasks.

Limitations: inborn (e.g. data complexity, voluminous, multimodality, multiple temporal rates, asynchrony), inadequate approaches (e.g. monomodal-biased), non-optimal fusion.

Research Goal: develop truly multimodal approaches that integrate several modalities toward improving robustness and performance for anthropo-centric multimedia understanding.
### Multimedia Data Challenges

- **Data are Voluminous:**
  - 24 hrs of TV = 430 Gb = 2,160,000 still (frame) images
  - WWW: 300-hr videos are uploaded on YouTube per minute.
  - 300 millions images are uploaded on FaceBook per day.
  - Kinect sensor: 250 MB/sec (uncompressed RGB)

- **Data are Dynamic**
  - Temporal video, Website updating, News quickly get obsolete

- **Different Temporal Rates**
  - Video: 25-30 frames/second
  - Audio: 44000 sound samples/sec,
  - Speech: 100 feature-frames/sec, 4 syllables/sec

- **Cross-Media asynchrony**
  - image and audio scene boundaries are different
A fundamental phenomenon in speech perception (McGurk & MacDonald)

Improving Automatic Speech Recognition (ASR) systems performance in adverse acoustical conditions:

- Noise, Interferences
Audio-Visual Recovery of Vocal Tract Geometry

Applications:
- Speech Mimics
- Articulatory ASR
- Speech Tutoring
- Phonetics

Multimodal HRI: Applications and Challenges

assistive robotics

education, entertainment

Challenges

- Speech: distance from microphones, noisy acoustic scenes, variabilities
- Visual recognition: noisy backgrounds, motion, variabilities
- Multimodal fusion: incorporation of multiple sensors, integration issues
- Elderly users
Multimodal Saliency & Movie Summarization

COGNIMUSE: Multimodal Signal and Event Processing In Perception and Cognition

website: http://cognimuse.cs.ntua.gr/
Audio-Visual Perception and Fusion

Perception: the sensory-based inference about the world state
Multicue or Multimodal Perception Research

- **McGurk effect**: Hearing Lips and Seeing Voices [McGurk & MacDonald 1976]

- **Modeling Depth Cue Combination using Modified Weak Fusion** [Landy et al. 1995]
  - scene depth reconstruction from multiple cues: motion, stereo, texture and shading.

- **Intramodal Versus Intermodal Fusion of Sensory Information** [Hillis et al. 2002]
  - shape surface perception: intramodal (stereopsis & texture), intermodal (vision & haptics)

- **Integration of Visual and Auditory Information for Spatial Localization**
  - Ventriloquism effect
  - Enhance selective listening by illusory mislocation of speech sounds due to lip-reading [Driver 1996]
  - Visual capture [Battaglia et al. 2003]
  - Unifying multisensory signals across time and space [Wallace et al. 2004]

- **AudioVisual Gestalts** [Monaci & Vandergheynst 2006]
  - temporal proximity between audiovisual events using Helmholtz principle

- **Temporal Segmentation of Videos into Perceptual Events by Humans** [Zacks et al. 2001]
  - humans watching short videos of daily activities while acquiring brain images with fMRI

- **Temporal Perception of Multimodal Stimuli** [Vatakis and Spence 2006]
McGurk effect example

- [ba – audio] + [ga – visual] → [da]  (fusion)

- [ga – audio] + [ba – visual] → [gabga, bagba, baga, gaba]  (combination)

Speech perception seems to also take into consideration the visual information. Audio-only theories of speech are inadequate to explain the above phenomena.

Audiovisual presentations of speech create fusion or combination of modalities.

One possible explanation: *a human attempts to find common or close information in both modalities and achieve a unifying percept.*
Attention

- **Feature-integration theory of attention** [Treisman and Gelade, CogPsy 1980]:
  - “Features are registered early, automatically, and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention.
  - This theory of attention suggests that attention must be directed serially to each stimulus in a display whenever conjunctions of more than one separable feature are needed to characterize or distinguish the possible objects presented.”

- **Orienting of Attention** [Posner, QJEP 1980]:
  - Focus of attention shifts to a location in order to enhance processing of relevant information while ignoring irrelevant sensory inputs.
  - **Spotlight Model**: focus visual attention to an area by using a cue (a briefly presented dot at location of target) which triggers “formation of a spotlight” and reduces RT to identify target. Cues are *exogenous* (low-level, outside generated) or *endogenous* (high-level, inside generated).
  - Overt / Covert orienting (with / without eye movements): “Covert orientation can be measured with same precision as overt shifts in eye position.”

- **Interplay between Attention and Multisensory Integration**: [Talsma et al., Trends CogSci 2010]: “Stimulus-driven, bottom-up mechanisms induced by crossmodal interactions can automatically capture attention towards multisensory events, particularly when competition to focus elsewhere is relatively low. Conversely, top-down attention can facilitate the integration of multisensory inputs and lead to a spread of attention across sensory modalities.”
Perceptual Aspects of Multisensory Processing

**Multisensory Integration**: unisensory auditory and visual signals are combined forming a new, unified audiovisual percept.

**Goal**: Perceiving Synchronous and Unified Multisensory Events

**Principles**: Multisensory integration is governed by the following rules:

- **Spatial rule**,
- **Temporal rule**,
- **Modality Appropriateness**:
  - Visual dominance of spatial tasks.
  - Audition is dominant for temporal tasks.
- **Inverse effectiveness law**:
  - In multisensory neurons, multimodal stimuli occurring in close space-time proximity evoke supra-additive responses. The less effective monomodal stimuli are in generating a neuronal response, the greater relative percentage of multisensory enhancement.
  - Is this the case for behavior? Recent experiments indicate that inverse effectiveness accounts for some behavioral data.

**Synchrony and Semantics** are two factors that appear to favor the binding of multisensory stimuli, yielding a coherent unified percept. Strong binding, in turn, leads to higher stream asynchrony tolerance.

[ E. Tsilionis and A. Vatakis, “Multisensory Binding: Is the contribution of synchrony and semantic congruency obligatory?”, COBS 2016.]
Computational audiovisual saliency model

- Combining audio and visual saliency models by proper fusion
- Validated via behavioral experiments, such as pip & pop:

[ A. Tsiami, A. Katsamanis, P. Maragos and A. Vatakis, ICASSP 2016.]
Bayesian Formulation of Perception

\[ P(S|D) = \frac{P(D|S)P(S)}{P(D)} \]

S: configuration of auditory and/or visual scene of world
D: mono/multi-modal data or features.

\( P(S) \): Prior Distribution, \( P(D/S) \): Likelihood, \( P(D) \): Evidence

\( P(S/D) \): Posterior conditional distribution

S \( \rightarrow \) D: World-to-Signal mapping

Perception is an ill-posed inverse problem

\[ \hat{S}_{MAP} = \arg \max_S P(D|S)P(S) \]
Strong Fusion: Bayesian formulation

Audio-Visual Data $(A, V)$

Likelihood $P_{AV}(A, V | S)$

World State $S$

Prior $P_{AV}(S)$

Posterior $P_{AV}(A, V | S)P_{AV}(S)$

Estimate $\hat{S}_{AV}$

\[
P_{AV}(S|D_A, D_V) = \frac{P_{AV}(D_A, D_V | S)P_{AV}(S)}{P_{AV}(D_A, D_V)}
\]

[ Clark & Yuille 1990 ]
Weak Fusion: Bayesian formulation

If the two single monomodal MAP estimates are close, their fusion is weighted average

\[ \hat{S}_{AV} = \frac{w_a \hat{S}_A + w_v \hat{S}_V}{w_a + w_v} \]

[Yuille & Bulthoff, 1996]
Models for Multimodal Data Integration

Levels of Integration:
- *Early* integration
- *Intermediate* integration
- *Late* integration

Time dimension:
- *Static*: CCA- Canonical Correlation Analysis: e.g. “cocktail-party effect”
  - Max Mutual Information
  - SVMs- Support Vector Machines: kernel combination

- *Dynamic*: HMMs (Hidden Markov Models)
  - DBNs (Dynamic Bayesian Nets)
  - DNNs (Deep Neural Nets)
Multi-stream Weights for Audio-Visual Fusion

\[ B(S|D_A, D_V) = \left[ P_A(D_A|S') \right]^{q_1} \left[ P_V(D_V|S) \right]^{q_2} \frac{P(S)}{P(D)} \]

- Intermediate case between weak and strong fusion
- Select exponents \( q_1, q_2 \) for aural and visual streams
- Work in the LogProb domain \( \rightarrow \) Weighted Linear combination
Multi-Stream HMM Topologies for Audio-Visual (A-)Synchrony

Synchronous HMMs
Synchrony at each state

Two-Stream HMMs
Phone-synchronous
State-asynchronous

Product-HMMs: Controlled synchronization freedom

Parallel-HMMs for Sign Recognition


[ C. Vogler & D. Metaxas, CVIU 2001 ]

[ S. Theodorakis, A. Katsamanis & P. Maragos, ICASSP 2009 ]
Synchronous Multi-Stream HMMs

\[ p(y^{(1)}, y^{(2)} \mid x) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1})p(y_t^{(1)} \mid x_t)p(y_t^{(2)} \mid x_t) \]
Asynchronous Multi-Stream HMMs

\[ p(y^{(1)}, y^{(2)} | x^{(1)}, x^{(2)}) = p(x_0^{(1)}, x_0^{(2)}) \]

\[ \prod_{t=1}^{T} p(x_t^{(1)}, x_t^{(2)} | x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(y_t^{(1)}, y_t^{(2)} | x_t^{(1)}, x_t^{(2)}) \]

[ Fig. Credit: G. Gravier ]
DBNs: Coupled HMMs

\[ p(y^{(1)}, y^{(2)} | x^{(1)}, x^{(2)}) = p(x_0^{(1)})p(x_0^{(2)}). \]

\[
T \prod_{t=1}^{T} p(x_t^{(1)} | x_{t-1}^{(1)}, x_{t-1}^{(2)})p(x_t^{(2)} | x_{t-1}^{(1)}, x_{t-1}^{(2)})p(y_t^{(1)} | x_t^{(1)})p(y_t^{(2)} | x_t^{(2)})
\]

[Fig. Credit: G. Gravier]

[A. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", EURASIP J. ASP 2002]
DBNs: Factorial HMMs

\[ p(y^{(1)}, y^{(2)} | x^{(1)}, x^{(2)}) = p(x_0^{(1)})p(x_0^{(2)}) \cdot \prod_{t=1}^{T} p(x_t^{(1)} | x_{t-1}^{(1)})p(x_t^{(2)} | x_{t-1}^{(2)})p(y_t^{(1)} | x_t^{(1)}, x_t^{(2)})p(y_t^{(2)} | x_t^{(1)}, x_t^{(2)}) \]

[Fig. Credit: G. Gravier]

Multimodal Hypothesis Rescoring + Segmental Parallel Fusion

Bayesian Co-Boosting for Multimodal Gesture Recognition

\[ H(x_i) \]

\( x_i \): training instance; \( w_{i,t} \): training instance \( x_i \)'s weight at the \( t \)-th iteration; \( h_{t,v}(x_i) \): weak classifier learnt from modality \( v \) at the \( t \)-th iteration; \( H(x_i) \): final strong classifier.

Two-Stream CNN

- RGB
- Optical Flow

Fusion after conv4 layer

- single network tower

Fusion at two layers (after conv5 and after fc8)

- both network towers are kept
- one as a hybrid spatiotemporal net
- one as a purely spatial network

Audio-Visual Speech Recognition

Main reference:


General References:

Speech: Multi-faceted phenomenon
CONSONANTS.

X Glottis closed, (catch.)
I " narrow, (voice.)
O " open, (aspirate.)
0 Super- Glottal Passage contracted, (whisper.)
| Soft Palate depressed, (nasal.)
C Back of Tongue, (contracting oral passage.)
☐ Front of do. (do.)
☐ Point of do. (do.)
☐ Lips, (do.)

VOWELS.

I Back of Tongue high.
I Back and Front do. do.
I Front do. do.

[The dotted lines show the 'high, 'mid,' and 'low' positions of the tongue, as subsequently explained.]
Audio Feature Extraction

- MFCCs
- Formants
- LPC Analysis
- Symmetric/Anti-symmetric Polynomials
- LSFs

Tutorial: Multimodal Signal Processing, Saliency and Summarization
Visual Feature Extraction:
Active Appearance Modeling of Visible Articulators

- Active Appearance Models for face modelling
- Shape & Texture related articulatory information
- Features: AAM Fitting (nonlinear least squares problem)
- Real-Time, marker-less facial visual feature extraction

$$\begin{align*}
\text{Active Appearance Models} &= \text{Shape} + p_1 + \text{Texture} + p_2 + \cdots \\
\text{Visible Articulators} &= \text{Face} + \lambda_1 + \text{Texture} + \lambda_2 + \cdots
\end{align*}$$
Example: Face Analysis and Tracking Using AAM

- Generative models like AAM allow us to qualitatively evaluate the output of the visual front-end,

shape tracking
reconstructed face

original
Measurement Noise and Adaptive Fusion

Conventional View: Features are directly observable

\[ p(c \mid x_{1:S}) \propto p(c) \prod_{s=1}^{S} p(x_s \mid c) \]

Our View: We can only measure noise-corrupt features

\[ p(c \mid y_{1:S}) \propto p(c) \prod_{s=1}^{S} \int p(x_s \mid c) p(y_s \mid x_s) \, dx_s \]

\[ p(c \mid y_{1:S}) \propto p(c) \prod_{s=1}^{S} \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} \mathcal{N}(y_s ; \mu_{s,c,m} + \mu_{e,s}, \Sigma_{s,c,m} + \Sigma_{e,s}) \]
Demo: Fusion by Uncertainty Compensation

- Classification decision boundary w. increasing uncertainty
  - Two 1D streams (y1 and y2-streams), 2 classes
AV-ASR Evaluation on CUAVE Database
Audio-Visual Recognition

Audio–Visual Multistream–HMM ASR Results

Average Absolute Improvement due to Visual information
AV-W-UC vs. A-UC
28.7 %

- Weights and Uncertainty Compensation
- Hybrid Fusion Scheme
Asynchrony Modeling with Product-HMMs

Average absolute improvement due to modeling with Product-HMM vs. Multistream-HMM

1.2 %
A Real-Time AV-ASR Prototype

Image Acquisition
- Firewire color camera, 640x480 @25 fps

(Re)initialization

Face detector
- Adaboost-based, @5 fps

Face tracking & feature extraction
- Real-time AAM fitting algorithms

GPU-accelerated processing
- OpenGL implementation

HMM-based backend

System Overview

Transcription

Training Tools
- one
- two
- three
- four
- six
- five

Recogniser
Audio-Visual Speech Recognition Demo
(WACC: AV=89%, A=74% at 5 dB SNR babble noise)
Audio-Visual Gesture Recognition and Human-Robot Interaction
Multimodal Gesture Signals from Kinect-0 Sensor

RGB Video & Audio

Depth
(vieniqui - *come here*)

Skeleton
(vieniqui - *come here*)

User Mask
(vieniqui - *come here*)

ChaLearn corpus

Multimodal Hypothesis Rescoring + Segmental Parallel Fusion

Audio-Visual Fusion & Recognition

- Audio and visual modalities for A-V gesture word sequence.
- Ground truth transcriptions (“REF”) and decoding results for audio and 3 different A-V fusion schemes.
- Results in top rank of ChaLearn (ACM 2013 Gesture Challenge – 50 teams - 22 users x 20 gesture phrases x 20 repeats).

[V. Pitsikalis, A. Katsamanis, S. Theodorakis & P. Maragos, JMLR 2015]
Audio-Gestural command recognition: Overview of our multimodal interface

MOBOT robotic platform


[Image of the MOBOT robotic platform with connected MEMS linear array, Kinect RGB-D camera, and other components. Diagram shows the flow of auditory and visual data, leading to N-best hypotheses and scores, multimodal late fusion, and the best AV hypothesis.]
Multimodal fusion: Complementarity of visual and audio modalities

- Similar audio, distinguishable gesture
- Distinguishable audio, similar gesture

“Come Here”  “Come Near”  “Turn right”  “Park”
Multimodal gesture classification results

- Leave-one-out experiments (Mobot-I.6a data: 8p,8g)
- Unimodal: audio (A) and visual (V)
- Multimodal (AV): N-best list rescoring

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p11</th>
<th>p12</th>
<th>p13</th>
<th>p4</th>
<th>p7</th>
<th>p8</th>
<th>p9</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>96.87</td>
<td>81.81</td>
<td>79.16</td>
<td>87.50</td>
<td>78.26</td>
<td>79.16</td>
<td>90.32</td>
<td>80.00</td>
<td>84.13</td>
</tr>
<tr>
<td>V</td>
<td>40.62</td>
<td>63.63</td>
<td>70.83</td>
<td>34.37</td>
<td>65.21</td>
<td>50.00</td>
<td>64.51</td>
<td>66.66</td>
<td>56.98</td>
</tr>
<tr>
<td>AV</td>
<td>87.50</td>
<td>90.90</td>
<td>95.83</td>
<td>84.37</td>
<td>100.0</td>
<td>79.16</td>
<td>96.77</td>
<td>86.66</td>
<td>90.15</td>
</tr>
</tbody>
</table>

**Multimodal confusability graph**
Part 1: Conclusions

- Audio-Visual Fusion → Better Results (ASR, Gesture, Saliency).
- More Big Data → Needs for summarization (not enough time for humans to see the videos). Not only data compression or dimensionality reduction for storage or fast access.
- More Data → Big Databases → Better training algorithms (Training processes work better if we have significant amounts of training data).
- Multimodal Data (audio, visual, depth, text):
  - Need for advanced signal processing algorithms for each modality (different nature of each modality).
  - Signal modalities or dimensions are complementary (i.e. microphones arrays enhance audio signal for distant ASR, audio-visual integration and fusion for speech/gesture understanding, video summarization).
Collaborators

Evangelopoulos, Georgios
Iosif, Elias
Kardaris, Nikos
Katsamanis, Nasos
Panagiotaropoulos, Georgia
Papandreou, George
Pavlakos, Georgios
Potamianos, Gerasimos
Pitsikalis, Vassilis
Theodorakis, Stavros
Tsiami, Antigoni
Rapantzikos, Kostantinos
Rodomagoulakis, Isidoros

Research Projects / Sponsors

**COGNIMUSE:**  [http://cognimuse.cs.ntua.gr/](http://cognimuse.cs.ntua.gr/)

**MOBOT:**  [http://mobot-project.eu/](http://mobot-project.eu/)

**I-SUPPORT:**  [http://www.i-support-project.eu/](http://www.i-support-project.eu/)

**BabyRobot:**  [http://www.babyrobot.eu/](http://www.babyrobot.eu/)