Introduction to Multimedia Semantics

Dr. Gerald Friedland
International Computer Science Institute
Berkeley, CA
fractor@icsi.berkeley.edu
Introduction to Semantic Multimedia Analysis

This presentation includes material from slides by:
Dr. Shu–Ching Chen
Professor, School of Computing and Information Sciences Florida International University, Miami, Florida, USA
Overview

- Introduction and Examples
- Visual Representation
- Acoustic Content Representation
- Modeling Techniques
- Multimodal Integration
- Concluding Remarks
Multimedia Analytics

- Context
- Understanding
  - Semantic Computing
- Machine Learning
  - Artificial Intelligence
- Filtering
  - Signal Processing
- Features
  - Images
  - Audio
  - Video
  - Text
Part I

Definitions and Examples
Multimedia: Definition

Entry:
  multimedia

Function:
  noun plural but singular or plural in construction

Date:
  1950

A technique (as the combining of sound, video, and text) for expressing ideas (as in communication, entertainment, or art) in which several media are employed; also: something (as software) using or facilitating such a technique.

(Merriam-Webster online dictionary)
Multimedia Content Extraction

Automatic analysis of the content (semantics) contained in data directly encoded for human perception (audio, images, video, touch) and its associated meta data (natural text, computer-encoded data).
Typical Problems

• Which videos contain person X
• Is this upload copyright infringement?
• Who does what when and where in this scene?
• What are the speakers, scenes, objects, narrative themes in this video collection?
Semantic Analysis of Multimedia Data

• enables automatic logical inference on perceptually encoded data

• enables more “natural” interaction with the computer: “do what the user means”

• assists in the creation of content: “do what the author means”
Image Retrieval

Image Retrieval: Surfing Shoes
Image Retrieval: Surfing Shoes
Image Retrieval: Surfing Shoes

ShoeSurfer by Implicit Interfaces

Displaying: Best matches for this shoe

Quick-view this shoe at Zappos.com.

Badgley Mischka: Shiloh $548.95
Sizes in stock: Unknown
Image Retrieval: Surfing Shoes
Examples: Cut & Paste in Images

Speaker Diarization: Who Spoke When?

Audio Track:

Segmentation:

Clustering:

Analyzing Meetings

Copyright Detection

Cruxle Copyright Detector Demo

1 copyright violations identified. Please find those videos in the right panel. Do you want to take them down now?
Part II

Visual Content Representation
- RGB (Red-Green-Blue) encoding three bytes enabling $(2^8)^3 = 2^{24}$ colors.

Color Spaces

- CMY(K): cyan, magenta, yellow, (and black)
- YCbCr
- HSV (hue, saturation, and value)
- CIE L*a*b*, CIE L*u*v*
From Image to Video

(a) Motion Estimation (ME)
- Matched block: has the smallest MAD
- Current pixel
- Block: size N
- Position: (x, y)

(b) Motion Compensation (MC):
- When two reference frames are available
- Linear interpolation

Search area

I: reference
P: dynamic frame
I: reference

Motion vector: (i, j)

(x, y)

(x-L, y-L)

(x+L, y+L)
“Natural” Segmentation

Video

Scene

Shot

Camera View 1

Camera View 2

Camera View n

Frame

1 2 3 4

5 6 7 ...
Typical Video Features

Low-level features:

- Color features: color dominant, color histogram, color moment, etc.
- Texture features: structural features, statistical features
- Shape features: edge detectors, boundary-based, region-based, etc.
Typical Video Features

**Color moments**

- First order: mean \( \mu_i = \frac{1}{N} \sum_{j=1}^{N} f_{ij} \)

- Second order: variance \( \sigma_i = \left( \frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^2 \right)^{1/2} \)

- Third order: skewness \( s_i = \left( \frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^3 \right)^{1/3} \)

- Forth order: kurtosis \( k_i = \left( \frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^4 \right)^{1/4} \)
Typical Video Features

Texture features

- **Co-occurrence matrix**
  \[ P_d[i, j] = n_{ij} \]

- **Energy**
  \[ C(k, n) = \sum_i \sum_j P_d[i, j]^2 \]

- **Contrast**
  \[ C(k, n) = \sum_i \sum_j (i - j)^k P_d[i, j]^n \]

- **Entropy**
  \[ C_e = -\sum_i \sum_j P_d[i, j] \ln P_d[i, j] \]

- **Homogeneity**
  \[ C_h = \sum_i \sum_j \frac{P_d[i, j]}{1 + |i - j|} \]

- **Correlation**
  \[ C_c = \frac{\sum_i \sum_j [ij P_d[i, j]] - \mu_i \mu_j}{\sigma_i \sigma_j} \]
Typical Video Features

Edge Detectors
- Sobel,
- Prewitt,
- Canny,
- ... and others
Typical Video Features

Medium-level features

- Difference frames, optical flow
- Region detection: shapes patterns, skin-color, textures
- Scale-Invariant Feature Transform (SIFT)
Typical Video Features

High-level features:

- Face detection: number of faces, location of face, etc.
- Categories: indoor and outdoor, play and non-play, etc.
- Metadata: GPS coordinates, compression rate, time
Part III

Audio Content Representation
A Sound Wave...

Analog wave:

Digitized wave:

Typical formats:
- 8000Hz, 8-bit log. companded (µ-law): telephone
- 16000Hz, 16-bit linear: speech (Skype)
- 44100Hz, 16-bit linear, stereo: Compact Disk
- 48000Hz, 32-bit linear, stereo: Digital Audio Tape
Typical Audio Features

Amplitude space:
- Energy (usually rms)
- Zero-Crossing Rate
- Thresholding
- “delta” and “delta-delta”s
- Entropy

\[ x_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}. \]
Typical Audio Features

Frequency space:
- Pitch
- Voicedness/Unvoicedness (HNR)
- Long-Term Average Spectrum (LTAS)
- Formants F1...F5
- Speaking rate
Typical Audio Features

A Speech Signal:

Pitch:

Intensity:

Formants:

HNR:
Other spaces:
- Linear-Prediction Coefficients (LPC)
- Mel-Frequency-Scaled Coefficients (MFCC): MFCC12, MFCC19, MFCCxx+delta+deltadelta
- PLP (Perceptual Linear Prediction)
- RASTA, RASTA-PLP, MSG
MFCC: Idea

Audio Signal

Pre-emphasis

Windowing

FFT

Mel-Scale Filterbank

Log-Scale

DCT

MFCC

power cepstrum of signal = \left| F \left\{ \log\left( \left| F \{\text{the signal}\}\right|^2 \right) \right\} \right|^2
MFCC: Mel Scale
MFCC: Result

Input audio

FFT Spectrum

Mel Spectrum

Mel Cepstrum
Part IV

Modeling Techniques
Development of a Content Analysis Algorithm
Modeling Techniques

Unsupervised Techniques:
- K-Means, X-Means
- PCA

Supervised Techniques:
- Gaussian Mixtures
- Neuronal Networks
- Support-Vector machines
- Hidden-Markov Models
K-Means

Algorithm Outline (Expectation Maximization)

Choose k initial means $\mu_i$ at random
loop
  for all samples $x_j$:
    assign membership of each element to a mean (closest mean)
  for all means $\mu_i$
    calculate a new $\mu_i$ by averaging all values $x_j$ that were assigned members
until means $\mu_i$ are not updated significantly anymore
Gaussian Mixtures
Training of Mixture Models

Goal: Find $a_i$ for

$$f_X(x) = \sum_{i=1}^{n} a_i f_Y(x; \theta_i).$$

Expectation:

$$y_{i,j} = \frac{a_i f_Y(x_j; \theta_i)}{f_X(x_j)}.$$

Maximization:

$$a_i = \frac{1}{N} \sum_{j=1}^{N} y_{i,j}$$

$$\mu_i = \frac{\sum_j y_{i,j} x_j}{\sum_j y_{i,j}}.$$
Part V

Multimodal Integration
Multimodal Integration

- ... is a field of cognitive psychology.
- Before 1960: Unimodal approach
- Initial results in the 1960’s, recently hyped again (2003+)
Multimodal Integration

Human psychology suggests:

- Uncertainty in sensory domains results in increased dependency of multisensory integration (Alais & Burr 2004)
- Multiple sensory inputs increase the speed of the output (Hershenson 1962)
Multimodal Integration

In computer science:

- How to create systems that benefit from multimodal integration in similar ways the brain does, i.e. they are
  - more accurate, robust, and/or faster than unimodal state of the art and/or
  - offer qualitative improvements over unimodal approaches
Some signal is observed and reduced...

...to the essentials relevant to the problem, ...

...statistical models are used to compute a score (e.g. probabilities) for the given observations, ...

... so that a decision function can decide on the classification.
Feature-Level Integration

Features are integrated before the model layer using a function ‘+’.

For example concatenation: n-dimensional vector ‘+’ m-dimensional vector = n+m-dimensional vector
Model–Level Integration

Output scores are integrated using a function ‘+’.

For example weighted combined log-likelihoods.
Decision-Level Integration

Output decision are fused using a function ‘+’.

For example majority voting.

WARNING: Meta-data fusion in general is a difficult research problem.
Remarks

• Signal-level integration is unlikely because of intractable data dimensionality.

• Multi-Level integration is also possible.

• In reality, a classification algorithm is more complicated than this scheme (e.g. feedback loops)

• The integration function ‘+’ may also be learned automatically.
Example: “3D from Audio”

Part VI

Concluding Remarks
Most Important Lesson

Multimedia content analysis is hard.

Therefore every possible cue should be considered for a solution, including context and user presumptions.
Computing methods become semantic computing methods when they are indistinguishable from understanding to the user.

In other words, when the computer does “what the user means” with the minimum communication possible.
Outlook

Coming up: Hands-on Session