MultiModal Information Fusion

Ling Guan
Ryerson Multimedia Laboratory & Centre for Interactive Multimedia Information Mining
Ryerson University, Toronto, Ontario Canada
lguan@ee.ryerson.ca
http://www.rml.ryerson.ca/

Acknowledgement

- The presenter would like to thank his former and current students, P. Muneesawang, Y. Wang, R. Zhang, A. Bulzacki, M. T. Ibrahim, N. Joshi and Z. Xie for their contributions to this research.
- Slides on fusion fundamentals provided by Prof. S-Y Kung of Princeton University are greatly appreciated.
- This research is supported by
  - The Canada Research Chair (CRC) Program,
  - Canada Foundation for Innovations (CFI),
  - The Ontario Innovation Trust (OIT), and
  - Ryerson University
Major Publications


Why Multimedia Multimodal Methodology? (revisit)

- Multimedia is a domain of multi-facets, e.g., audio, visual, text, graphics, etc.
- A central aspect of multimedia processing is the coherent integration of media from different sources or multimodalities.
- Easy to define each facet individually, but difficult to consider them as a combined identity
- Humans are natural and generic multimedia processing machines

Can we teach computers/machines to do the same (via fusion technologies)?
Potential Applications

- Human–Computer Interaction
- Learning Environments
- Consumer Relations
- Entertainment
- Digital Home, Domestic Helper
- Security/Surveillance
- Educational Software
- Computer Animation
- Call Centers

Source of Fusion for Classification

![Diagram showing the source of fusion for classification](Image here)
Feature (Data) level fusion

Direct Data (Feature) Level Fusion

Furthermore, let $S_i \in \mathbb{R}^p, i = 1, \cdots, N$ denote vectors comprising all the individual scores:

$$V_i = \begin{bmatrix} v_i^{(1)} \\ v_i^{(2)} \\ \vdots \\ v_i^{(q)} \end{bmatrix}, \quad i = 1, \cdots, N \quad (0.2)$$

Now, training data can be formed as the following input/teacher pairs

$$[V, T] = \{ [V_1, t_1], [V_2, t_2], \ldots, [V_N, t_N] \}$$

Prior knowledge can be incorporated into the fusion models by modifying

$$V_i = \begin{bmatrix} \mu v_i^{(1)} \\ \mu v_i^{(2)} \\ \vdots \\ \mu v_i^{(q)} \end{bmatrix}$$
Interaction level fusion

HMM (Face Model)
- forehead
- eyes
- nose
- mouth
- chin

Fused HMM
Decision (Score) level fusion

A Different interpretation

- Hierarchical Structure
- Each Sub-network \( E_i \) an expert system
- The decision module classifies the input vector as a particular class when

\[
Y_{\text{net}} = \arg \max_j y_j
\]
Score Fusion

1a Score Fusion (w/o supervision)
• Linear Score Fusion (confidence/prior knowledge)
• Nonlinear Score Fusion (ROC-based)

1b Score Fusion (via supervision)
• Linear Score Fusion (adaptive supervision)
• Nonlinear Score Fusion (adaptive supervision)

Score Fusion Architecture (Audio-Visual)

The scores are independently obtained, which are then combined.
• The lower layer contains local experts, each produces a local score based on a single modality
• The upper layer combines the score.
The most prevailing unsupervised approaches estimate the confidence based on prior knowledge or training data. Linear SVM (supervised) Fusion is an appealing alternative.

Nonlinear Adaptive Fusion (via supervision)
Data/Feature Fusion (early work)

- Simple and straightforward (Good)
- Curse of Dimensionality (Bad)
- Normalization issue

- Case study: Bimodal Human emotion recognition (also with a score fusion flavor)

Indicators of emotion

- Speech
- Facial expression
- Body language: highly dependent on personality, gender, age, etc
- Semantic meaning: two sentences could have the same lexical meaning but different emotional information

......
Objective

♦ To develop a generic language and cultural background independent system for recognition of human emotional state from audiovisual signals

Audio feature extraction

- Input speech
  - Preprocessing: Noise reduction, Leading and trailing edge elimination
  - Wavelet thresholding
- Hamming window
  - Windowing: 512 points, 50% overlap
  - MFCC
  - Formant
- Audio feature set
Visual feature extraction

Input Image Sequence → Key Frame Extraction → Maximum Speech Amplitude → Face Detection

Feature Mapping → Gabor Filter Bank

The recognition system - with Decision Fusion

Input video → Key frame extraction → Face detection → Gabor wavelet

Audio feature extraction: Pre-processing → Windowing → Prosodic → MFCC → Formant

Recognized emotion: AN → DI → FE → HA → SA → SU

Decision module: Corresponding Classifier → Decision module

Classification scheme: Visual feature extraction

Ryerson University – Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim
Modular Networks

Hierarchical Structure
- Each Sub-network $E_r$ is an expert system
- The decision module classifies the input vector as a particular class when
  $$Y_{\text{net}} = \arg \max y_j$$

Experimental results

Experiments were performed on 500 video samples from 8 subjects, speaking 6 languages.
- Six emotion labels: Anger, Disgust, Fear, Happiness, Sadness, and Surprise
- 360 samples (from six subjects) were used for training, and the rest 140 (from the remaining two subjects) for testing, there is no overlap between training and testing subjects

Ryerson University – Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim
Interaction Fusion

- In general, not straightforward
- Fusion takes place on intermediate results dynamically
- Case study:
  1. Speech Recognition
  2. Image Retrieval with Audio Information

Speech Recognition

Fusing information obtained from different processing methods

Two separate HMM based models:

- spectral features, missing data (MD),
- MFCC features.
- The Fused HMM model is used for the interaction level fusion.
Speech Recognition

<table>
<thead>
<tr>
<th></th>
<th>SNR 18dB</th>
<th>SNR 6dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>83.7</td>
<td>64.7</td>
</tr>
<tr>
<td>Conventional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC CMN</td>
<td>66.0</td>
<td>60.3</td>
</tr>
<tr>
<td>Conventional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral Features, MD</td>
<td>76.5</td>
<td>67.4</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC+MD</td>
<td>84.5</td>
<td>67.5</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC, CMN+MD</td>
<td>88.6</td>
<td>73.5</td>
</tr>
</tbody>
</table>

* CMN - Cepstral Mean Normalization

TABLE I. RECOGNITION RESULTS WITH TEST CORPUS + FACTORY NOISE

Image Retrieval with Audio Cues

General Idea

Audio information of a query image

Semantic class weighting

Weight propagation

Image matching using audio information

Visual information of a query image

Nearest neighbor retrieval

Bayesian fusion

Experimental Setup

- Database
  - 4400 images collected from Flickr
  - Featuring 44 kinds of animals
- Visual feature selection
- Audio feature selection
  - MFCC with frame length equal to 256.
Experimental Results

Precision as a function of the number of retrieval iterations.

Observations:
1. Major improvement is obtained within the first iterations.
2. Information fusion improve the performance and the audio relevance feedback further improves it.
Score Level Fusion

- Could be straightforward or involving more analysis.
- Rigid due to limit on information left

Case study:
1. Video Retrieval based on Audiovisual Cues
2. Video Retrieval

The scores are independently obtained, which are then combined.
**Visual**
- Adaptive Video Indexing (AVI)
  - Using visual templates
    
    $$f_1^{(h)} = \arg\min_{j \in \{0,1,\ldots,T_1-1\}} \left[ \frac{1}{N} \| h_j - \hat{h} \|^2 \right]$$
    
    $$f_2^{(h)} = \arg\min_{j \in \{0,1,\ldots,T_1-1\} - \{f_1^{(h)}\}} \left[ \frac{1}{N} \| h_j - \hat{h} \|^2 \right]$$

  - TF-IDF Model
    
    $$f_i[j] = \frac{\frac{f_i[j]}{\max\{f_i[j]\}} \times \log \frac{N_{ti}}{n[j]}}{N}$$

- Cosine Distance for Similarity Matching

**Visual Feature Representation**
- Laplacian Mixture Model (LMM) of wavelet coefficients of audio signal
  
  \[ p(w_i) = \alpha_1 p_1(w_i | b_1) + \alpha_2 p_2(w_i | b_2) \]
  
  \[ \alpha_1 + \alpha_2 = 1 \]

- Audio feature vector with model parameter (using EM estimator)
  
  \[ f_a = \{ [m_0, \sigma_0], \{ \alpha_{1,i}, b_{1,i}, b_{2,i} \} \}_{i=1,2,...,L-1} \]

- \( \{ \alpha_{1,i}, b_{1,i}, b_{2,i} \} \) are the model parameters obtained from the \( i \)-th high-frequency subband.
LMM vs GMM (2 COMPONENTS)

- Recognition rate obtained by the SVM based fusion model, using video database of 6,000 clips
- Five semantic concepts

<table>
<thead>
<tr>
<th>Type of concept</th>
<th>Accuracy (%)</th>
<th>False positive rate (%)</th>
<th>False negative rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love scene</td>
<td>90.97</td>
<td>8.91</td>
<td>19.70</td>
</tr>
<tr>
<td>Music video</td>
<td>91.03</td>
<td>9.03</td>
<td>0</td>
</tr>
<tr>
<td>Fighting</td>
<td>84.68</td>
<td>25.65</td>
<td>14.55</td>
</tr>
<tr>
<td>Ship crashing</td>
<td>91.81</td>
<td>7.54</td>
<td>26.87</td>
</tr>
<tr>
<td>Dancing party</td>
<td>99.68</td>
<td>0.30</td>
<td>2.08</td>
</tr>
<tr>
<td>Average</td>
<td>91.63</td>
<td>10.29</td>
<td>12.64</td>
</tr>
</tbody>
</table>

Video Classification Result
Multimodal Human Authentication

with
Signature, Iris and Fingerprint

Signature Recognition
Fingerprint Image Enhancement System Overview

Iris Segmentation System Overview
Fingerprint/Signature/Iris Fusion

Feature Extraction → Match Score → Score Fusion → Final Decision

Feature Extraction → Match Score

Feature Extraction → Match Score

Finger Network

Robot Applications

Face Tracking

Body Tracking

Emotion Recognition

Camera Tracking

Hand Gesture Recognition

Movement Recognition

Stereo Vision / Depth Calc.
Robot Application: Domestic Helper
- via emotion/intention recognition

1. Target people group:
   Elderly and disabled people at homes or community houses
2. Capable of simple gestures and body language
3. Capable of simple, and, sometime, incomplete verbal communications

Robot Application: Domestic Helper
- via emotion/intention recognition

1. Help the elderly and the disabled with their daily life.
2. Entertain the people they look after.
3. Call the nurse or emergency when in need.
Multimodal Fusion for Human Intention Recognition

- Hand gesture
- Body movement
- Audio cues
- Visual cues

Possible human intention

Challenges

- Will fusion help in a problem on hand?
- What is the best fusion model for a problem on hand?
  - Data/feature level,
  - Interaction level,
  - Score/Decision level,
  - Or multilevel.
- New data analysis and mining tools need to be developed to address the issues. Or the existing tools may be revisited.
Challenge 1: Does fusion help?

Sensor 1 (audio) → Recognition by audio → Fusion → Recognition by fusion → Recognition by video → Sensor 2 (visual)

Challenge 2: Fusion at which level?

Score (Decision) → Interaction → Score (Decision) → Data/Feature #1 → Interaction → Data/Feature #2
Information Entropy

- Entropy is a measure of the uncertainty associated with a random variable
- Uncertainty is useful information
- Entropy
  \[ H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \]
- Conditional entropy
  \[ H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \]

Multimodal source type

- **Conflict**
  \[ H(x), H(y) < H(x,y) \]
  joint uncertainty is larger
  negative relevance

- **Redundancy**
  \[ H(x,y) < H(x), H(y) \]
  joint uncertainty is smaller
  positive relevance

- **Complementary**
  \[ H(x) \approx H(x,y) \approx H(y) \]
  joint uncertainty unchanged
  weak relevance
Multimodal fusion method

- A feature has a conflict with other features
  - If the conflict is beyond threshold, eliminate conflict feature
  - Keep the total uncertainty at an acceptable level
    \[ H(X \mid Y) = \text{Max}(H(X)) \]

- Redundancy fusion (A ∩ B)
  - According to ranking of uncertainty value
- Complementary fusion (A ∪ B)

Mapping from entropy to weight

- Weights are the inverse of entropies
  - High entropy -> low confidence -> low weight
  - Feature corrupted -> maximum entropy -> zero weight
  - \( \sum w = 1 \)
    - \( H = 0, \text{then } w = 1 \)
    - \( H = H_{\text{max}}, \text{then } w = 0 \)
    - \( Hi = Hj, \text{then } wi = wj = 0.5 \)
Entropy vs. Correlation in Fusion

- Developed an entropy based method, Kernel Entropy Component Analysis (KECA)
- Compared with correlation based methods:
  - KPCA – Kernel Principal Component analysis.
  - KCCA – Kernel Canonical correlation analysis


• Experimental results

![Graphs showing comparison between KECA, KPCA, and KCCA on RML and eNTERFACE databases.](image-url)
Summary

- Fusion – coherent integration of multimedia multimodal information
- It is a natural process by human beings, but not straightforward for machines.
- It may be carried out at different information levels, but how to choose the right model?
- Several case studies are used to demonstrate the power of information fusion
- Multiple challenges are waiting to be addressed