MultiModal Information Fusion

Ling Guan
Ryerson Multimedia Laboratory & Centre for Interactive Multimedia Information Mining
Ryerson University, Toronto, Ontario Canada
lguan@ee.ryerson.ca
https://www.ryerson.ca/multimedia-research-laboratory/

Acknowledgement

- The presenter would like to thank his former and current students, P. Muneesawang, Y. Wang, R. Zhang, M.T. Ibrahim, L. Gao, C. Liang and N. El Madany for their contributions to this research.
- Slides on fusion fundamentals provided by Prof. S-Y Kung of Princeton University are greatly appreciated.
- This presentation is supported by
  - The Canada Research Chair (CRC) Program,
  - Canada Foundation for Innovations (CFI),
  - The Ontario Research Fund (ORF), and
  - Ryerson University
Relevant 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
Feature (Data) level fusion

Direct Data (Feature) Level Fusion

Furthermore, let \( V_i \in \mathbb{R}^d, i = 1, \ldots, N \) denote vectors comprising all the individual features:

\[
V_i = \begin{bmatrix}
V_i^{(1)} \\
V_i^{(2)} \\
\vdots \\
V_i^{(q)}
\end{bmatrix}, \quad i = 1, \ldots, N
\]  

(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:

\[
V_i = \begin{bmatrix}
\nu^{(1)} V_i^{(1)} \\
\nu^{(2)} V_i^{(2)} \\
\vdots \\
\nu^{(q)} V_i^{(q)}
\end{bmatrix}
\]

Ryerson University - Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim
Representation level fusion

HMM (Face Model)

- forehead
- eyes
- nose
- mouth
- chin

Fused HMM

Bayesian network

Information fusion through inference

Audio stream

Video stream

Feature extraction

Audio feature vector

Video feature vector

Ryerson University – Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim
**Decision (Score) level fusion**

- **Audio stream**
- **Video stream**

- **Classifier** SVM, HMM etc
- **Confidence score, Likelihood etc**
- **Weighted average**

---

**Modular Networks (Decision Level)**

- **Hierarchical Structure**
- **Each Sub-network** $E_i$, an expert system
- **The decision module** classifies the input vector as a particular class when

$$Y_{net} = \arg\max y_j$$
Score Fusion Architecture (Audio-Visual)

- The lower layer contains local experts, each produces a local score based on a single modality.
- The upper layer combines the score.

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.

Linear Fusion

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)

(Kernel, SVM)

Score 1

Score 2

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

Case study:
1. Bimodal Human emotion recognition (also with a score level fusion flavor)
2. 3D human action recognition
3. Fusion by feature mapping
Bimodal Human emotion recognition

— Also with a score level 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

Mehrabian’s Communication Rule

Silent messages. Wadsworth, Belmont, CA, 1971

- 7%-38%-55% Rule
  - Derived from experiments dealing with communications of feelings and attitudes
  - 7%: Verbal (Words, What you say)
  - 38%: Vocal (Tone of voice, How you say them)
  - 55%: Visual (Gaze, Facial expression, Body language)

Audio feature extraction

Preprocessing
- Noise reduction
- Leading and trailing edge elimination
- Wavelet thresholding

Hamming window
- Prosodic
- MFCC
- Formant

Audio feature set

Visual feature extraction

Input Image Sequence

Key Frame Extraction
- Maximum Speech Amplitude
- Face Detection

Visual Features

Feature Mapping

Gabor Filter Bank
The recognition system
- with Decision Fusion

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

Input speech
- Pre-processing
- Windowing
- MFCC
- Formant

Audio feature extraction

Visual feature extraction

Feature selection

Decision module

Corresponding classifier

Recognition emotion

Prosodic

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

Modular Networks (Decision Level Fusion)

Human emotion recognition

- Hierarchical Structure
- Each Sub-network $E_r$ an expert system
- The decision module classifies the input vector as a particular class when

$$Y_{net} = \arg \max y_j$$
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.

3D Human Action Recognition

3D Human Action Recognition

3D Action Recognition Representations

- Depth based Approaches
- Skeleton based Approaches
- RGB (Color) based Approaches
- Fusion based Approaches

The Recognition Framework

Training $x_1$

- Training Data
- Depth
- Features
- Depth Feature Extraction
- Sequences’ Labels

Testing $y$

- Testing Data
- Depth
- Features
- Depth Feature Extraction
- Collaborative Representation Learning for Dictionary $D$

Compute $\arg\min f$
Experiment on SBU interaction dataset

- A collection of two-person interactive activities
- Composed of 8 interactions performed by 21 subjects, total 265 sequences (6822 frames).
- Following 5-fold cross-validation, randomly split the dataset into 5 folds of 4-5 two-actor sets each.

Subject01--Subject07

Kicking  Punching  Hugging  ShakingHands

Results -- SBU Interaction dataset

- The energy-based segmentation is better than time-guided segmentation
- Heterogeneous Features Fusion based on CCA-serial method is better than CCA-sum method
- The best recognition accuracy of the proposed: 95.39%
- Better than the 6 deep learning based methods:
  - Structured Model [21],
  - ST-LSTM + Trust Gates [22],
  - Hierarchical RNN [23],
  - LSTM + co-occurrence [23],
  - SkeletonNet (Skeleton + CNN) [27],
  - Global Context-Aware + LSTM [26],

Comparison of existing methods on SBU-REISET interaction dataset:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body-Pose Feature + Linear SVM [4]</td>
<td>87.6</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Body-Pose Feature + MLP [4]</td>
<td>91.1</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Poselet Mining [18]</td>
<td>86.9</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Deep Model [21]</td>
<td>93.4</td>
<td>RGB D</td>
</tr>
<tr>
<td>ST-LSTM + Trust Gates [22]</td>
<td>93.3</td>
<td>Skeleton</td>
</tr>
<tr>
<td>LSTM + Cooccurrence + Dropout [23]</td>
<td>90.41</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Hierarchical RNN [23]</td>
<td>90.55</td>
<td>Skeleton</td>
</tr>
<tr>
<td>SkeletonNet (Skeleton + CNN) [27]</td>
<td>93.47</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Global Context-Aware + LSTM [26]</td>
<td>94.1</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Ours (Time-guided)</td>
<td>90.0</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Ours (Time-guided Depth)</td>
<td>91.08</td>
<td>Depth</td>
</tr>
<tr>
<td>Ours (Concurrent Fusion by CCA-sum)</td>
<td>90.77</td>
<td>Both</td>
</tr>
<tr>
<td>Ours (Energy-guided)</td>
<td>94.56</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Ours (Energy-guided Depth)</td>
<td>87.69</td>
<td>Depth</td>
</tr>
<tr>
<td>Ours (Energy-guided Concurrent Fusion by CCA-sum)</td>
<td>92.31</td>
<td>Both</td>
</tr>
<tr>
<td>Ours (Energy-guided, Fusion by CCA-serial)</td>
<td>95.29</td>
<td>Both</td>
</tr>
</tbody>
</table>
Fusion by Feature Mapping


Knowledge Discovery (revisit)
Knowledge Discovery - 2 (revisit)

In the schematic diagram on the previous slide, knowledge discovery includes:

- Feature generation block: generate features/descriptors by
  - Identify the key-points
  - Generate features at the key-points
  - Classical methods incorporating prior knowledge (hand crafted features)
  - Deep leaning structure such as CNN (hand crafted architecture)

- The feature mapping block: map the features into more effective representation by
  - Feature selection
  - Explicit mapping by Statistical Machine Learning (SML), or Implicit mapping by FFNN, normally including pooling, a statistical processing step

- Feature generation and mapping are two different, but complementary processing steps. Both are critically important in information discovery.

SML methods are solidly rooted in mathematics, and the analysis procedure could be clearly and convincingly presented.

Feature Mapping (1)

- The Purpose:
  - Explicitly or implicitly reorganize the information for best possible analysis and recognition performance

- Advantages
  - Provide more complete and discriminatory description of the intrinsic characteristics of different patterns

- Major Challenges
  - How to extract the discriminatory description of the intrinsic characteristics from the data?
  - How to design a mapping strategy that can effectively utilize the complementary information presented in different datasets?
When multiple features/modalities are involved, it is in essence multi-modal information fusion [L1]. When only one feature set is involved, it is simplified to information transformation. The following presentation focuses on mapping by LMCCA, DMCCA for the purpose of multimodal information fusion. The common characteristic: generic or feature independent.

**Feature Mapping (2)**

- When multiple features/modalities are involved, it is in essence multi-modal information fusion [L1].
- When only one feature set is involved, it is simplified to information transformation.
- The following presentation focuses on mapping by LMCCA, DMCCA for the purpose of multimodal information fusion.
- The common characteristic: generic or feature independent.

---

**Discriminative Multiple Canonical Correlation Analysis (DMCCA) (1)**

- In DMCCA,
  - the correlation among feature sets derived from multi-modal, multi-feature, or multiple channels is taken as the metric of the similarity;
  - the within-class correlation and the between-class correlation are considered jointly, leading to a more discriminant space.
DMCCA (2)

A key characteristics of DMCCA (analytically verified):

- Capable of extracting the discriminatory representation
- The number of projected dimension ($d$) corresponding to the optimal recognition accuracy is smaller than or equals to the number of classes ($c$) being studied.
- Or mathematically: $d \leq c$
- The property can be graphically illustrated accurately.

Experimental Results (1)

- We conduct experiments on different applications
  - Face Recognition (ORL database) — [http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html](http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html)
Experimental Results(2)

- **Experimental Settings**

  Table. 1 Experimental settings on different database

<table>
<thead>
<tr>
<th>Name of database</th>
<th>Total samples</th>
<th>Training samples</th>
<th>Testing samples</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>3000</td>
<td>1500</td>
<td>1500</td>
<td>10</td>
</tr>
<tr>
<td>ORL</td>
<td>400 (All samples)</td>
<td>200</td>
<td>200</td>
<td>40</td>
</tr>
<tr>
<td>Caltech</td>
<td>31 to 800 images/classes</td>
<td>3030</td>
<td>6084</td>
<td>101</td>
</tr>
</tbody>
</table>

Experimental Results(3)

- **Hand-crafted Feature Extraction**

  - Handwritten digit recognition
    - 24-dimensional: the mean of the digit images transformed by the Gabor filters.
    - 24-dimensional: the standard deviation of the digit images transformed by the Gabor filters.
    - 36-dimensional: Zernike moment features.
  
  - Face recognition
    - 36-dimensional: The histogram of oriented gradient (HOG) feature.
    - 33-dimensional: The local binary Patterns (LBP) feature.
    - 48-dimensional: Gabor transformation feature with the mean and standard deviation of the face images transformed by each filter.
Experimental Results(4)

Table 2: The performance with different methods on MNIST
(Training samples 1500; Testing samples 1500)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial Fusion [L6]</td>
<td>70.33%</td>
</tr>
<tr>
<td>CCA [L7]</td>
<td>74.60%</td>
</tr>
<tr>
<td>GCCA [L8]</td>
<td>75.53%</td>
</tr>
<tr>
<td>MCCA [L9]</td>
<td>72.73%</td>
</tr>
<tr>
<td>CNN [L10]</td>
<td>76.40%</td>
</tr>
<tr>
<td>PCANet [L11]</td>
<td>79.20%</td>
</tr>
<tr>
<td>DMCCA</td>
<td>82.60%</td>
</tr>
</tbody>
</table>

Graphically Selecting Optimal Projection in DMCCA (MNIST)

J(η) in handwritten digit recognition on MNIST Database
(peak at 9 < 10)
Experimental Results (5)

- Table 3: The performance with different methods on ORL

<table>
<thead>
<tr>
<th>Methods</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial fusion [L6]</td>
<td>77.5%</td>
</tr>
<tr>
<td>CCA[L7]</td>
<td>94.5%</td>
</tr>
<tr>
<td>GCCA[L8]</td>
<td>95.5%</td>
</tr>
<tr>
<td>MCCA[L9]</td>
<td>94.5%</td>
</tr>
<tr>
<td>Discriminative Sparse Representation (DSR)[L12]</td>
<td>94.5%</td>
</tr>
<tr>
<td>Collaborative representation classification (CRC)[L13]</td>
<td>88.5%</td>
</tr>
<tr>
<td>l1-regularized least squares (L1LS)[L14]</td>
<td>92.5%</td>
</tr>
<tr>
<td>Fast L1-Minimization Algorithms (FLMA)[L15]</td>
<td>90.0%</td>
</tr>
<tr>
<td>CNN[L10]</td>
<td>76.0%</td>
</tr>
<tr>
<td>PCANet[L11]</td>
<td>92.0%</td>
</tr>
<tr>
<td>DMCCA</td>
<td>98.5%</td>
</tr>
</tbody>
</table>

Graphically Selecting Optimal Projection in DMCCA (ORL)

$J(\eta)$ in face recognition on ORL Database (peak at 27 < 40)
Experimental Results (6)

- Deep NN based Feature Extraction

- Caltech 101 dataset - AlexNet

![AlexNet structure]

Experimental Results (7)

- The parameters of fc6, fc7 and fc8 and recognition results are given as follows:
  - fc6: 4096 fully connected layer. Recognition rate: 77.84
  - fc7: 4096 fully connected layer. Recognition rate: 77.65
  - fc8: 1000 fully connected layer. Recognition rate: 73.31
### Experimental Results (8)

- Table 4 Comparison with AlexNet on Caltech 101

<table>
<thead>
<tr>
<th>Methods</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet fc6</td>
<td>77.80%</td>
</tr>
<tr>
<td>AlexNet fc7</td>
<td>77.65%</td>
</tr>
<tr>
<td>AlexNet fc8</td>
<td>77.31%</td>
</tr>
<tr>
<td>LMCCA [2018]</td>
<td>83.68%</td>
</tr>
<tr>
<td>LMCCA (new)</td>
<td>87.21%</td>
</tr>
<tr>
<td>DMCCA [2018]</td>
<td>89.38%</td>
</tr>
<tr>
<td>DMCCA+KECA [unpublished]</td>
<td>90.61%</td>
</tr>
</tbody>
</table>

### Experimental Results (9)

- Table 5 Comparison with different methods on Caltech 101

<table>
<thead>
<tr>
<th>Methods</th>
<th>Year</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. Du et al. [L16]</td>
<td>2017</td>
<td>78.60%</td>
</tr>
<tr>
<td>L. Mansourian et al. [L17]</td>
<td>2017</td>
<td>75.37%</td>
</tr>
<tr>
<td>P. Tang et al. [L18]</td>
<td>2017</td>
<td>82.45%</td>
</tr>
<tr>
<td>G. Lin et al. [L19]</td>
<td>2017</td>
<td>78.83%</td>
</tr>
<tr>
<td>W. Xiong et al. [L20]</td>
<td>2017</td>
<td>75.90%</td>
</tr>
<tr>
<td>S. Kim et al. [L21]</td>
<td>2017</td>
<td>83.00%</td>
</tr>
<tr>
<td>W. Yu et al. [L22]</td>
<td>2018</td>
<td>77.90%</td>
</tr>
<tr>
<td>L. Sheng et al. [23]</td>
<td>2018</td>
<td>74.78%</td>
</tr>
<tr>
<td>DMCCA [L2]</td>
<td>2018</td>
<td>89.38%</td>
</tr>
</tbody>
</table>
Summary

1. Transformation based feature coding can improve the quality of visual features, both hand crafted and those obtained by deep NN.
2. Optimal coding by LMCCA, DMCCA and DMCCA+KECA have been analytically derived and experimentally verified.

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 Feature Representation

- Visual
  - Adaptive Video Indexing (AVI)
    - Using visual templates
      
      \[
      I^{(h)}_1 = \arg \min_{j \in \{0, 1, \ldots, T_h - 1\}} \left( \frac{1}{N} \| h_j - \hat{h}^i \|^2 \right)
      \]
      
      \[
      I^{(h)}_2 = \arg \min_{j \in \{0, 1, \ldots, T_h - 1\}} \left( \frac{1}{N} \| h_j - \hat{h}^i \|^2 \right)
      \]

  - TFxIDF Model
    
    \[
    f_r[j] = \frac{fr[j]}{\max\{fr[j]\}} \times \log \frac{N_m}{n[j]}
    \]

- 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)

\[ f_a = [ (m_0, \sigma_0), [\alpha_{i_1}, b_{i_1}, b_{i_2}]] \quad i = 1, 2, ..., L - 1 \]

- \( \{\alpha_{i_1}, b_{i_1}, b_{i_2}\} \) 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

Ryerson University - Multimedia Research Laboratory
Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, Yun Tie, Adrian Bulzacki, Muhammad Talal Ibrahim
Multimodal Human Authentication

with

Signature, Iris and Fingerprint

Signature Recognition

(a) (b) (c)
Fingerprint Image Enhancement System Overview

Iris Segmentation System Overview
Fingerprint/Signature/Iris Fusion

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

Robot Applications

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

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,
  - Representation level,
  - Score/Decision level,
  - Or multilevel.
- New data analysis and information 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

Sensor 2 (visual) → Recognition by video

Fusion

→ Recognition by fusion

Challenge 2: Fusion at which level?

Data/Feature #1 → Representation → Score (Decision) → Data/Feature #2 → Representation → Score (Decision)
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
    \[ H(X | Y) = \text{Max}(H(X)) \]

- Redundancy fusion \((A \cap B)\)
  - according to ranking of uncertainty value
- Complementary fusion \((A \cup B)\)

Mapping from entropy to weight

- Weights – reversely proportional to 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)
- Fused entropy based method with Discriminative Analysis (KECA-DMCCA)
- Compared with correlation based methods:
  - KPCA – Kernel Principal Component analysis.
  - KCCA – Kernel Canonical correlation analysis


**Experimental results**

- KECA
- KPCA
- KCCA

- RML database
- eINTERFACE database
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