Multimedia Analysis:
Marriage of Signal Processing and Machine Learning

_Tsuhan Chen_ 陈祖翰
Professor, Carnegie Mellon University
tsuhan@cmu.edu
A 10-Year Journey...


MMSP Workshops

International Conf on Multimedia (ICME)

IEEE Transactions on Multimedia, March 1999~

SPS Distinguished Lecturer 2007
Multimedia Analysis: Bits vs. Content

Reconstruct

It’s Bayesian in machine learning, a.k.a., prior…

Number of all possible 16×12 images = $2^{16×12×8}$

$>> 30×60×60×24×365×\text{human history}×\text{world population}$

$>> \text{number of all possible face images}$
Thoughts

“The most compelling shapes are those near to our hearts: people’s faces, a gracefully moving body, a natural scene with rustling leaves and flowing water. Evolution has tuned us to these sights…”

[Lengyel, 1998]

Multimedia analysis... more than processing bits...

it’s all about the content ...

it’s signal processing + machine learning

[Chen, 2006]
Content-Based Information Retrieval

Many Interesting Applications...
Logos
[Leung&Chen ICME’02]

Hand-Drawn Query

Retrieved Trademarks
Hand-Drawn Sketches
[Leung&Chen ICME’03]
3D Objects [Zhang&Chen ACM MM'01]
3D Protein Structures [Chen&Chen ICIP’02]
Content-Based Information Retrieval (CBIR)

- Multimedia Database
- Feature Extraction
- Low-Level Feature Space
- Indexing
- Indexed Feature Database
- Similarity Measure
- Retrieval Results
High-Level Semantics

- e.g. hierarchical attribute tree

**BIG CHALLENGE:** How to bridge the gap between low-level features and high-level semantics?
Possible Solutions

Multimedia Database

Feature Extraction

Query

Low-level Feature space

Feature Extraction

Indexed Feature Database

Hidden Annotation

Indexing

Retrieval Results

Relevance Feedback

Similarity Measure

$f_1$

$f_2$

$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$
Semantic Information

- Hidden annotation
  - Object \#n has Attribute \#k → explicit

- Relevance feedback
  - Objects \#m and \#n are (not) similar → implicit

Q: How to represent and propagate semantic information?

Q: How to use explicit/implicit semantic information to improve retrieval?
## Semantic Information as Probabilities

<table>
<thead>
<tr>
<th>Object 1</th>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>Attribute 3</th>
<th>…</th>
<th>Attribute K</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_{11}</td>
<td>p_{12}</td>
<td>p_{13}</td>
<td>…</td>
<td></td>
<td>p_{1K}</td>
</tr>
<tr>
<td>p_{21}</td>
<td>p_{22}</td>
<td>p_{23}</td>
<td>…</td>
<td></td>
<td>p_{2K}</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>p_{N1}</td>
<td>p_{N2}</td>
<td>p_{N3}</td>
<td>…</td>
<td></td>
<td>p_{NK}</td>
</tr>
</tbody>
</table>

$p_{nk}$: Attribute Probabilities
Annotate one object...

<table>
<thead>
<tr>
<th></th>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>Attribute 3</th>
<th>...</th>
<th>Attribute K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When an object is annotated, $p_{nk}$ is set to 0/1

Q: How to propagate?  A: Based on low-level features
Semantic Propagation

“Biased Kernel Regression”
Semantic Propagation (cont.)
Semantic Propagation (cont.)

<table>
<thead>
<tr>
<th></th>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>Attribute 3</th>
<th>...</th>
<th>Attribute K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>$p_{13}$</td>
<td>...</td>
<td>$p_{1K}$</td>
</tr>
<tr>
<td>Object 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Object $N$</td>
<td>$p_{N1}$</td>
<td>$p_{N2}$</td>
<td>$p_{N3}$</td>
<td>...</td>
<td>$p_{NK}$</td>
</tr>
</tbody>
</table>

Q: Which to annotate next?
Active Learning

- Choose the most **uncertain** object to annotate
  - Uncertainty determined by the entropy of attribute probabilities
- “Selective sampling”
  - May want to consider density in feature space too
Recap...

- Maintain attribute probabilities of each object
- Set an attribute probability to 1/0 when annotated
- Propagate probabilities to non-annotated objects
- Choose the most uncertain object in the database to annotate next
  - Use probabilities to estimate uncertainty
- Use probabilities to measure semantic distance...
Result [Zhang&Chen T-MM’02]

3D Objects (1750 total)
Relevance Feedback

Relevance feedback
- Ask for user’s feedback during the retrieval
  - “Object #i is (not) similar to the query”
  - “Objects #m and #n are (not) similar”
- Implicit semantic information

Use feedback to improve retrieval
- Way 1: Move the query point
- Way 2: Weigh the features
- Way 3: “Warp” the feature space
An Example

Retrieved Results (inside circle)

Query
An Example

Retrieved Results (inside circle)

Query

Positive feedback
Negative feedback
Move the Query Point

New Retrieved Results (inside circle)

New Query

- Positive feedback
- Negative feedback
Feature Weighting

New Retrieved Results (inside ellips)

Query

- Positive feedback
- Negative feedback
Feature Space Warping

Before Warping

After Warping

Retrieved Results (inside circle)

New Retrieved Results (inside circle)
Feature Space Warping

This is also semantic propagation!!!
Experiment Result [Bang & Chen ICIP’02]

Performance Improvement

3D Objects (1750 total)
Semantic Propagation is the Key

- Without semantic propagation, hidden annotation and relevance feedback are not very useful

- With enough relevance feedback, can we accomplish information retrieval without low-level features at all?
Pushing Content to Extreme

Content-Free Information Retrieval...
With enough relevance feedback, retrieval is based more and more on feedback, less and less on features.

In the extreme case, retrieval based on feedback only:
- Retrieval based on user history

E.g., Amazon.com
Example -- How CFIR Works

<table>
<thead>
<tr>
<th>User History</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="User 1" /></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><img src="image2.png" alt="User 2" /></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><img src="image3.png" alt="User 3" /></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>
**Example -- How CFIR Works**

<table>
<thead>
<tr>
<th>User History</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="User 1" /></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><img src="image" alt="User 2" /></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><img src="image" alt="User 3" /></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

and ![User 4](image) are more similar than and ![User 5](image).
CBIR vs. CFIR

Will user $U$ like image $X$?

Two different approaches:

- Look at what $U$ likes
  - Characterize images \(\rightarrow\) Content-based IR
- Look at which users like $X$
  - Characterize users \(\rightarrow\) Content-free IR
Bayesian Framework

Pair-wise conditional probability matrix

\[ \tilde{P}(x_i = 1 \mid x_j = 1) \quad \forall i, j \]

Retrieval based on

\[
P(x_i = 1 \mid x_{j_1} = 1, \ldots, x_{j_F} = 1) \propto \prod_{k=1}^{F} P(x_i = 1 \mid x_{j_k} = 1) \frac{P(x_i = 1)^{F-1}}{P(x_i = 1)^{F-1}}
\]
Experiment Results

CBIR Methods

CFIR Methods

Sample Images
Content without User Feedback

Extracting content from nothing...
Unsupervised Image Categorization

[Caltech face + background dataset]
Unsupervised Image Categorization

[UIUC car dataset]
“Bag of Words” Representation

DoG interest point detector + SIFT descriptor [Lowe]

Vector Quantization

Codebook

Count

Words
Graphical Model

$z$ : topic
$w$ : word

$P(z, w) = P(z)P(w|z)$

**Topic Appearance**
Need to handle background...

Solution: Add a hidden layer → PLSA
Probabilistic Latent Semantic Analysis (PLSA)

- Hofmann 01, Monay and Gatica-Perez 04, Sivic et al. 05, Quelhas et al. 05
- Can model complex scenes

Inference:
- Categorization $P(z|d)$
- Segmentation $P(z|d, w)$

$d$: image
$z$: topic
$w$: word

$$P(d, z, w) = P(d)P(z|d)P(w|z)$$

Document Characteristic
Topic Appearance

- Maximum likelihood estimation using EM algorithm
Problem with PLSA

- $P(z|d, w) \geq 0.5$
- $P(z|d, w) < 0.5$
“Bag of Words” is the problem...

- As long as the parts are present, the exact position does not matter too much

Picasso, 1943

Not so for general objects!

Dali, 1936
Enforcing Clustering

- A number of $S = 10$ fixed spatial distributions

\[ \mu \sum_s \]

\[ \begin{array}{c}
\text{position} \\
\text{appearance}
\end{array} \]

$S_1$ to $S_{10}$ $S_9$
Enforcing Clustering: Semantic-Shift

\[ p(d, z, w, x) = P(d)P(z|d)P(w|z)p(x|z, d) \]

\( d \) : image
\( z \) : topic
\( w \) : word appearance
\( x \) : word position

[Source: Liu & Chen CVPR’06]
Representing Location Semantics

- Assume single foreground object
- Location semantics $p(x | z, d)$
  - Foreground:
    $$p(x | z_{FG}, d_i) \equiv \mathcal{N}(x | \mu_i, \Sigma_i)$$
  - Background:
    Complement of foreground distribution
Learning in Semantic-Shift

\[ P(z_k|d_i, w_j, x_p^d) \propto p(x_p^d|z_k, d_i)P(z_k|d_i)P(w_j|z_k) \]

posterior

\[ \mathcal{N}(x|\mu_i, \Sigma_i) \]

Document Characteristic

\[ P(z_k|d_i) \propto \sum_j \sum_p n_{ijp}P(z_k|d_i, w_j, x_p^d) \]

Topic Appearance

\[ P(w_j|z_k) \propto \sum_i \sum_p n_{ijp}P(z_k|d_i, w_j, x_p^d) \]

\[ n_{ijp} = n(d_i, w_j, x_p) : \text{bag of word position pairs} \]

Carnegie Mellon
## Learning in Semantic-Shift

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x</td>
<td>z, d)$</td>
</tr>
<tr>
<td>$P(z</td>
<td>d, w, x)$</td>
</tr>
<tr>
<td>$P(w</td>
<td>z)$</td>
</tr>
</tbody>
</table>

This is why “semantic-shift”

Learning all 3 terms simultaneously…

Completely unsupervised…
Results

[PLSA PLSA Semantic Shift Semantic Shift]

0.7 $\rightarrow$ 0.85

0.9 $\rightarrow$ 0.98

[Liu&Chen CVPR’06]
Future Work

- Intra-object modeling
- Video: spatial-temporal modeling
- Training the codewords in the loop
- Multiple objects
Conclusions

- Machine learning can bridge the gap between low-level features (bits) and high-level semantics (content)
- Hidden annotation and relevance feedback can help; semantic propagation is the key
  - Active learning
- “Content-free” information retrieval is possible
  - Bayesian framework
- Content extraction without user feedback is possible
  - Unsupervised learning; graphical models
Afterthoughts...

- **Feng-Shui (风水)**
  - Ancient Chinese room arrangement technique

- **Way 1 (low-level):**
  - Write down all the rules
  - Too many and do not generalize

- **Way 2 (high-level):**
  - Imagine how a dragon would move through the room to arrange it in a livable manner
  - Intuitive and creative
  - Done by some Feng-Shui masters
Advanced Multimedia Processing Lab

Please visit us at:

http://amp.ece.cmu.edu