Integrating Concept Ontology for Semantic Video Classification

Jianping Fan
Department of Computer Science
UNC-Charlotte, USA
Outline of presentation

- Motivation
- New Video Representation Framework
- Automatic Salient Object Detection
- Semantic Medical Concept Modeling
- Classifier Training and Feature Selection
- Semantic Video Classification
- Conclusions
1. Research Motivation

**Challenges:** Content Representation and Concept Interpretation

Most existing CBVR systems describe multimedia like six blind men!
1. Research Motivation

Semantic Medical Concepts

Semantic Gap

Low-Level Multimodal Signals

We focus on a medical education video domain!
2. Problem Definition

- Quality of features
- Video content representation
- Semantic video concept modeling
- Classifier training and feature selection
- Semantic video classification
- Knowledge Visualization for Video Access
3. Video Content Representation

- Video content representation
  - Video Shots
    - Easy to implement
    - Features may not be representative
  - Semantic Video Objects
    - Difficult to implement
    - Features are representative

How to take both advantages?
3. Video Content Representation

- **Semantic Medical Concepts**
- **Principal Video Shots**
- **Low-Level Multimodal Signals**

Semantic Gap

Gap 1

Gap 2
3. Video Content Representation

Definition of Principal Video Shot

- Video Shots
- Salient Objects of Interesting

Principal Video Shot

Physical Video Shot

Concept-Sensitive Salient Objects
3. Video Content Representation

- **Salient Object Definition**

![Diagram of Taxonomy of Medical Video Components]

- **Medical Video Stream**
- **Video Stream**
  - Surgery Objects
  - Medical Equipment
  - Video Text
- **Audio Stream**
  - Medical Environment
  - Human Voice
  - Noise
  - Music

**Taxonomy of Medical Video Components**
3. Video Content Representation

**Salient object definition:** semantic-sensitive, feature-invariant objects in the domain of interest
3. Video Content Representation

- Automatic Salient Object Detection
3. Video Content Representation

- Automatic Salient Object Detection

Support Vector Machine is used for binary region classification!
### 3. Video Content Representation

- **Automatic Salient Object Detection**

<table>
<thead>
<tr>
<th>Salient Objects</th>
<th>Human Face</th>
<th>Lecture Slide</th>
<th>Blood Regions</th>
<th>Background Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>90.3%</td>
<td>92.4%</td>
<td>90.2%</td>
<td>94.6%</td>
</tr>
<tr>
<td>Recall</td>
<td>87.5%</td>
<td>89.6%</td>
<td>86.7%</td>
<td>81.2%</td>
</tr>
<tr>
<td>Salient Objects</td>
<td>Human Speech</td>
<td>Human Skin</td>
<td>Lecture Sketch</td>
<td>Blue Cloth</td>
</tr>
<tr>
<td>Precision</td>
<td>89.8%</td>
<td>96.3%</td>
<td>93.3%</td>
<td>96.7%</td>
</tr>
<tr>
<td>Recall</td>
<td>82.6%</td>
<td>95.4%</td>
<td>88.5%</td>
<td>95.8%</td>
</tr>
</tbody>
</table>
3. Video Content Representation
3. Video Content Representation
### 3. Video Content Representation

<table>
<thead>
<tr>
<th>Text</th>
<th>Face</th>
<th>Hair</th>
<th>Inside</th>
<th>Skin</th>
<th>Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Text Example" /></td>
<td><img src="image2.png" alt="Face Example" /></td>
<td><img src="image3.png" alt="Hair Example" /></td>
<td><img src="image4.png" alt="Inside Example" /></td>
<td><img src="image5.png" alt="Skin Example" /></td>
<td><img src="image6.png" alt="Blood Example" /></td>
</tr>
</tbody>
</table>
| Graft Survival at 3-Years:  
  - Living Donor Kidneys: 69.7%  
  - Cadaveric Donor Kidneys: 65.7% | | | | | |
3. Video Content Representation

Original Video Clips → Salient Object Detection → Salient Object → Object Volume Extraction → Confident Object Volume
3. Video Content Representation

Original Frame → Homogeneous Regions

Automatic Image Segmentation

Image Region Classification & Aggregation

Temporal Statistic

Principal Video Shot → Salient Object
3. Video Content Representation

- Global Visual Features for Video Shot Representation;

- Local Visual Features for Salient Object Representation
3. Video Content Representation

- **Global Visual Features for Video Shot Representation**

  - **Color**
    - HSV color histogram, dominant color, ...

  - **Texture**
    - Edge histogram, wavelet coefficients, Tamura features, ...

  - **Motion**
    - Directional motion histogram, Camera motion, ...

  - **Other features**
3. Video Content Representation

- **Local Visual Features for Salient Object Representation**

- **Color**
  - HSV color histogram,
  - dominant color, ...

- **Texture**
  - Edge histogram,
  - Tamura, ....

- **Shape**
  - Rectangular box,
  - moments, ..... 

- **Motion**
  - Trajectory, motion
  - histogram, ...

- **Confidence Map**
4. Semantic Video Concept Modeling

Semantic Video Concept 1

Semantic Video Concept i

Semantic Video Concept Nc

PVS Class 1

PSV Class j

PSV Class S

Low-Level Auditory Signal

Low-Level Vision Signal

Low-Level Image-Textual Signal
4. Semantic Video Concept Modeling
4. Semantic Video Concept Modeling
4. Semantic Video Concept Modeling

How can we formulate this composition relationship (concept modeling)?
4. Semantic Video Concept Modeling

How to model the data distributions?
4. Semantic Video Concept Modeling

Atomic Video Concept Modeling

\[
P(X, C_j, \Theta_{c_j}) = \sum_{l=1}^{\kappa_j} P(X \mid C_j, S_l, \theta_l) \overline{\omega}_l
\]
Hierarchical Video Concept Modeling

4. Semantic Video Concept Modeling

Semantic Video Concept at the second level:

\[
P(X, C_j, \Theta_{c_j}) = \sum_{j=1}^{\kappa_i} \omega_{c_j} \sum_{l=1}^{\kappa_j} P(X | C_j, S_l, \theta_l) \omega_l
\]

Semantic Video Concept at the nth level:

\[
P(X, C_j, \Theta_{c_j}) = \sum_{m=1}^{\kappa_m} \omega_{c_m} \ldots \sum_{j=1}^{\kappa_i} \omega_{c_j} \sum_{l=1}^{\kappa_j} P(X | C_j, S_l, \theta_l) \omega_l
\]
5. Video Concept Learning for Classification

- Model Selection & Parameter Estimation

\[
P(X, C^j, \Theta_{c^j}) = \sum_{l=1}^{\kappa_j} P(X | C^j, S_l, \theta_l) \omega_l
\]

We need techniques for estimating:

Model Structure \( \mathcal{K}_j \)

Model Parameters \( \Theta_{c_j} = \{\theta_l, \omega_l | l = 1, \ldots, \kappa_j\} \)
5. Video Concept Learning for Classification

- Adaptive EM Algorithm

(a) Start from a large $K$

(b) Perform automatic merging, splitting, & elimination

© Incorporating negative samples for concept learning
Adaptive EM Algorithm

Overlapped mixture components from same concept model: *merging*
5. Video Concept Learning for Classification

Adaptive EM Algorithm

Elongated mixture components from same concept model: *splitting*
5. Video Concept Learning for Classification

Adaptive EM Algorithm

Tailed mixture components from different concept models: *splitting & elimination*
5. Video Concept Learning for Classification

- **Merging:**
  \[ J_m(i, k, \theta_{ik}) = JS(C_j, \theta_{ik}) + \varphi JS(C_j, \theta_i, \theta_k) \]

- **Splitting:**
  \[ J_s(i, m, \theta_i) = \frac{\varphi JS(C_j, C_h, \theta_i, \theta_m)}{JS(C_j, \theta_i)} \]

- **Elimination:**
  \[ J_e(i, \theta_i) = \frac{\varphi}{JS(C_j, \theta_i)} \]
5. Video Concept Learning for Classification

- **Normalization**

\[ \sum_{i=1}^{\kappa_j} J_e(i, \theta_i) + \sum_{i=1}^{\kappa_j} \sum_{k=i+1}^{\kappa_j} J_m(i, k, \theta_{ik}) + \sum_{i=1}^{\kappa_j} \sum_{m=i+1}^{\kappa_h} J_s(i, m, \theta_i) = 1 \]

- **Acceptance Probability**

\[ P_{accept} = \min \left( \exp \left[ - \frac{L(C_j, \Theta_1) - L(C_j, \Theta_2)}{\tau} \right], 1 \right) \]
Advantages

(a) Performing model selection and parameter estimation jointly in a single algorithm!

(b) Negative samples are incorporated for discriminative learning of finite mixture models with higher classification accuracy!

© New approach for hierarchical classifier training by performing automatic merging, splitting, and elimination of components!
6. Learning with Unlabeled Samples

- Unlabeled Samples for Concept Learning

\[ \Psi(X_i, C_j, t) = \varphi_a(X_i, C_j, t) \varphi_b(X_i, C_j, t) \]

\[ \gamma_i = \begin{cases} 1, & e^{\gamma_i} - e^{-\gamma_i} \\ e^{\gamma_i} + e^{-\gamma_i} & \end{cases} \]
6. Learning with Unlabeled Samples

- Unlabeled samples partitioning

- Certain unlabeled samples

- Informative unlabeled samples

- Outliers & New Concepts
6. Learning with Unlabeled Samples

- Unlabeled Samples for Concept Learning

**Confidence Score:**

\[ \Psi(X_l, C_j, t) = \frac{\phi_a(X_l, C_j, t) \phi_b(X_l, C_j, t)}{\text{Confidence Score}} \]
6. Learning with Unlabeled Samples

- Unlabeled Samples for Concept Learning

**Penalty Term:**

\[
\gamma_l = \begin{cases} 
1, & \text{Certain Unlabeled Samples} \\
\frac{e^{y_l} - e^{-y_l}}{e^{y_l} + e^{-y_l}}, & \text{Uncertain Unlabeled Samples}
\end{cases}
\]

\[
y_l = |\Psi(X_l, C_j, t+1) - \Psi(X_l, C_j, t)|
\]
6. Learning with Unlabeled Samples

- **Birth of New Mixture Components**

\[
P(X, C_j, \Theta_{C_j}) = (1 - \omega_{\kappa_j+1}) \sum_{l=1}^{\kappa_j} P(X | C_j, S_l, \theta_l) + \omega_{\kappa_j+1} P(X | C_j, S_{\kappa_j+1}, \theta_{\kappa_j+1})
\]

*Birth* operation is added to capture **unknown image context classes** that are interpreted by **informative unlabeled samples**.
7. Hierarchical Video Concept Learning

Second-Level Semantic Video Concept

Atomic Video Concept 1

Atomic Video Concept i

Atomic Video Concept $K_i$

$P(X, C_j, \Theta_{c_j}) = \sum_{j=1}^{\kappa_i} \omega_{c_j} \sum_{l=1}^{\kappa_j} P(X | C_j, S_l, \theta_l) \omega_l$
7. Hierarchical Video Concept Learning

- **Model Integration by combining mixture components from all its children nodes**

\[
P(X, C_i, \Theta_{c_i}) = \sum_{l=1}^{\kappa} P(X | C_j, S_l, \theta_l) \overline{\omega}_l
\]

where the initial model structure is:

\[
\kappa = \sum_{j=1}^{\kappa_i} \kappa_j
\]
7. Hierarchical Video Concept Learning

- **Elimination**

\[ P(X, C_i, \Theta_{c_i}) = \frac{1}{1 - \Theta} \sum_{l=1}^{\kappa-1} P(X | C_j, S_l, \theta_l) \omega_l \]

- **Merging**

\[ P(X, C_i, \Theta_{c_i}) = \sum_{l=1}^{\kappa-2} P(X | C_j, S_l, \theta_l) \omega_l + P(X | C_j, S_{mn}, \theta_{mn}) \omega_{mn} \]
7. Hierarchical Video Concept Learning

Splitting

\[ P(X, C_i, \Theta_{c_i}) = \sum_{l=1}^{\kappa-1} P(X \mid C_j, S_l, \theta_l) \bar{\omega}_l + P(X \mid C_j, S_r, \theta_r) \bar{\omega}_r + P(X \mid C_j, S_t, \theta_t) \bar{\omega}_t \]

Conclusion:

Using class distribution for high-level concept learning can achieve better performance than using their prediction outputs!
8. Semantic Video Classification

Bayesian Rule for Video Classification:

\[ P(C_i \mid X, \Theta_{c_i}) = \frac{P(X \mid C_i, S_l, \theta_l)\omega_l}{\sum_{l=1}^{\kappa} P(X \mid C_i, S_l, \theta_l)\omega_l} \]
8. Semantic Video Classification
8. Semantic Video Classification
8. Semantic Video Classification
9. Benchmark and Algorithm Evaluation

- Precision
  \[ \rho = \frac{\nu}{\nu + \gamma} \]

- Missing-recall
  \[ \bar{\rho} = \frac{\nu}{\nu + \tau} \]
9. Algorithm Evaluation

- Convergence of Adaptive EM Algorithm

![Graph showing convergence of adaptive EM algorithm](image)
9. Algorithm Evaluation

- Convergence of Adaptive EM Algorithm
9. Algorithm Evaluation

- Effectiveness of Adaptive EM Algorithm
9. Algorithm Evaluation

- Hierarchical Concept Learning

(a) More concept levels may reduce hypothesis variance!

(b) More concept levels may increase risk of transmission of errors among levels!
9. Algorithm Evaluation

- Training with Unlabeled Samples
## 9. Benchmark

- Classification accuracy vs. representation

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Class Lecture</th>
<th>Trauma Surgery</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Salient Objects</strong></td>
<td>81.6%, 82.1%</td>
<td>82.3%, 81.9%</td>
<td>81.7%, 89.3%</td>
</tr>
<tr>
<td><strong>Video Shots</strong></td>
<td>64.9%, 68.5%</td>
<td>64.3%, 65.7%</td>
<td>61.1%, 59.7%</td>
</tr>
<tr>
<td><strong>Concepts</strong></td>
<td><strong>Gastrointestinal Surgery</strong></td>
<td><strong>Dialog</strong></td>
<td><strong>Burn Surgery</strong></td>
</tr>
<tr>
<td><strong>Salient Objects</strong></td>
<td>85.1%, 89.3%</td>
<td>79.2%, 78.6%</td>
<td>81.4%, 85.2%</td>
</tr>
<tr>
<td><strong>Video Shots</strong></td>
<td>71.5%, 69.3%</td>
<td>67.5%, 68.2%</td>
<td>64.8%, 69.1%</td>
</tr>
</tbody>
</table>
9. Benchmark

- Our hierarchical approach
- Flat approach

- Tricuspid surgery
- Stomachic surgery
- Colon surgery
- Leg & Body surgery
- Face surgery
- Cardivascular surgery
- Endoluminal vascular surgery
- Gastroinestinal surgery
- Pancreatic surgery
- Orthopadic surgery
- Burn surgery
- Vascular surgery
- Endocrine surgery
- Traumatic surgery

0.4 0.5 0.6 0.7 0.8 0.9 1
10. Video Skimming
11. Video Privacy Protection
11. Video Privacy Protection
12. Semantic Video Concept Modeling via Support Vector Machines
12. Semantic Video Concept Modeling via Support Vector Machines

- Positive support vectors
- Negative support vectors
- Positive samples
- Negative samples
- Margin
12. Semantic Video Concept Modeling via Support Vector Machines

\[
\begin{align*}
\min_{\omega, b, \xi} & \quad \frac{1}{2} \omega \cdot \omega + C \sum_{l=1}^{N} \xi_l \\
\text{subject to} & \quad G_j(X_l) \cdot (\omega \cdot \Phi(X_l) + b) \\
& \quad \xi_l \geq 0, \quad l = 1, \ldots, N.
\end{align*}
\]

\[
h(x) = \sum_{i=1}^{M} y_i \lambda_i \kappa(x_i, x) + c \sum_{i=1}^{M} \xi_i
\]
13. SVM Classifier Training

- Support Vector Machine may be one solution, but...

- **Training cost** is $O(M^3)$, $M$ is sample size

- **Curse of Dimensionality** — large-scale training samples are needed for reliable SVM classifier training! High-dimensional feature space!

- **Kernel functions** should be able to characterize the underlying geometric properties of image data! Heterogeneous feature space!
If RBF kernel is used, we need techniques for estimating the SVM parameters: $\gamma, C$.
13. SVM Classifier Training

- Incremental SVM Classifier Training & Convergence
  ---How to incorporate informative unlabeled samples for classifier training?

Batch-Based Approach

Incremental Approach
13. SVM Classifier Training

- Unlabeled Samples for Classifier Training

- Weak SVM Classifier
  - Unlabeled Samples
  - Certain Unlabeled Samples
  - Informative Unlabeled Samples
  - Uncertain Unlabeled Samples

- Incremental SVM Classifier Training
  - New Concepts & Outliers
14. Feature Partition and Selection

- **Feature Heterogeneity** – Different feature types
- **RBF kernel** cannot effectively characterize the underlying geometric properties of image data in such heterogeneous feature space

- More existing feature selection techniques can only work on **homogeneous feature space**! **filtering, wrapping, boosting**

  Feature correlation is ignored!
14. Feature Partition and Selection

**Semantic Video Concept**

**High-Dimensional Visual Features**

- Prior knowledge
- PCA
- Feature Selection by selecting the most relevant classifiers!

$$H(x) = \sum_{i=1}^{n} \alpha_i h_i(x)$$

Boost on both training samples and features!
Advantages

-- The dimensions for each feature subset are relative low, thus less training samples are needed --- speed up SVM classifier training

-- Feature selection in heterogeneous space via classifier selection -- Hierarchical feature selection

-- The geometric property of image data in each feature subset can be characterized effectively by using RBF kernels.
15. Benchmark

- More is Less: Optimal Feature Subset
16. Hierarchical Video Database Summarization & Visualization
16. Hierarchical Video Database Summarization & Visualization

- Concept-Oriented Video Summarization

- General Surgery
  - Trauma Surgery
    - Face Surgery
    - Leg Surgery
  - Skin Surgery
    - Establishment of Pneumoperitoneum
  - Eye Surgery
    - Stomachic Surgery
    - Colon Surgery
17. News Video Analysis & Visualization

- Problems of keyword-based query formulation
  - Requirements for the system
    - Must be able to detect most underlying video semantics
      - Current semantic classification technique supports at most hundreds of video concepts
    - Must be able to translate visual semantics to keywords
      - Many visual concepts are difficult to translate to language
  - Requirements for the users
    - Must have clear idea of what is needed
    - Must be able to translate information needs to keywords

More General Video Domain
17. News Video Analysis & Visualization

- Index video clips via information
- Represent information via visualization
17. News Video Analysis & Visualization

Quantify the interestingness via frequency
- Higher frequency implies higher interestingness?
  - “George Bush”
- Proposed solution: unpredictability
17. News Video Analysis & Visualization
17. News Video Analysis & Visualization
18. Who work on these topics?

1. CMU Informedia Project
2. Columbia University Advent project
3. University of Amsterdam
4. IBM Research Center
5. Google
6. Yahoo
7. MSN