RegimVid Semantic Indexing System at TrecVid 2010

Speaker:
Dr. George Quénot

On behalf of:
Nizar Elleuch – Mohamed Zarka – Issam Feki – Dr. Anis Ben Ammar – Prof. Adel M. Alimi

November 15, 2010
Content

1. System Overview
   - RegimVid Overview
   - Visual Features Extraction
   - Audio Features Extraction
   - Multimodal Fuzzy Fusion

2. Experiments

3. Conclusion And Future Works
1 System Overview
   • RegimVid Overview
   • Visual Features Extraction
   • Audio Features Extraction
   • Multimodal Fuzzy Fusion

2 Experiments

3 Conclusion And Future Works
Content

1 System Overview
   - RegimVid Overview
   - Visual Features Extraction
   - Audio Features Extraction
   - Multimodal Fuzzy Fusion

2 Experiments

3 Conclusion And Future Works
RegimVid Overview

System Overview
Experiments
Conclusion And Future Works

RegimVid Overview
Visual Features Extraction
Audio Features Extraction
Multimodal Fuzzy Fusion

Slide :  3 / 24  RegimVid at TrecVid2010
RegimVid Overview

Slide: 3 / 24  |  RegimVid at TrecVid2010
The **RegimVid** indexing system provides an automatic analysis of video contents by using frame description based on low-level features.
RegimVid Indexing Sub-System

1. The system extracts the low-level features for each modality of the video shot.
2. The system represents contents for labeling them, later, by basing on score detection via classification process.
3. The predicted score are merged to obtain multimodal fusion.
RegimVid Runs in TrecVid2010

Participation in the Semantic Indexing Task (SIN)
RegimVid Runs in TrecVid2010

**Regim\textsubscript{4}**  A visual modality analysis orientated towards an automatic categorization of video contents to create relevance relationships between low-level descriptions and semantic contents according to a user point of view.

**Regim\textsubscript{5}**  A Multimodal fuzzy fusion using positive rules extracted from LSCOM Ontology. The fusion process employs a deduction reasoning engine.

**Regim\textsubscript{6}**  A Multimodal fuzzy fusion using positive and negative rules extracted from LSCOM Ontology.
Visual Features Extraction Approach

Initial unlabeled images

Collaborative Soft annotation

Annotated database

SVM Classification

Concept detection

TP P PP

IACC.1 key-frame data

Interest keypoints detection

Local feature extraction

Elementary codebook

Bag of Pseudo-Sentences
Visual Features Extraction Approach

Aggregate the training data at three relevance levels or classes, namely "highly relevant" (TP), "relevant" (P) and "somewhat relevant" (PP).
Visual Features Extraction Approach

Aggregate the training data at three relevance levels or classes, namely "highly relevant" (TP), "relevant" (P) and "somewhat relevant" (PP).
Visual Features Extraction Approach

**Interest keypoints detection**

The main idea is to exploit a detector based on luminance and variation of the orientation of edge.

**Step 1:** Use a pyramid 4 scales 8 orientations for each image of a concept

**Step 2:** To detect the edge with CANNY method

**Step 3:**
- To detect the discontinuity of the orientation of edge
- To detect the homogeneous areas (luminance)

**Step 4:** Detect points of interest
Visual Features Extraction Approach

Local feature extraction

we use several visual descriptors of different modalities (color, texture and shape) as Color Histogram, Co-occurrence Texture, Gabor, ....

After extracting the visual features, we proceed to the early fusion step.

Elementary codebook

One of the most important constraints of discrete visual codebook generation is in the uniform distribution of visual words over the continuous high-dimensional feature space.

- to generate a codebook of prototype vectors from the above features, we utilize the SOM-based clustering

- after the learning process of the SOM map, we grouped the similar units by using of partitive clustering using K-means.
Visual Features Extraction Approach

**Local feature extraction**

we use several visual descriptors of different modalities (color, texture and shape) as Color Histogram, Co-occurrence Texture, Gabor, ....

After extracting the visual features, we proceed to the early fusion step.

**Elementary codebook**

One of the most important constraints of discrete visual codebook generation is in the uniform distribution of visual words over the continuous high-dimensional feature space.

- to generate a codebook of prototype vectors from the above features, we utilize the SOM-based clustering
- after the learning process of the SOM map, we grouped the similar units by using of partitive clustering using K-means.
Bag of Pseudo-Sentences

- We are interested in the spatial distribution of key-points to enhance the classification process and concepts categorization.

- To generate these pseudo-sentences, we used only two stages of spatial clustering based on the Relative Euclidean Distance (RED) calculated between each visual elementary word in each image.

- The size of the obtained codebook allows having more discriminative models, but also a need for memory, storage, and computing time to train a classifier much more important. Therefore, we perform a refinement step to reduce the size of the obtained pseudo-sentences codebook.

- The refinement process is likened to a problem of optimization of the pseudo-sentences construction. To resolve this problem, two steps are considered: the analysis of syntax and the occurrence of all constructed pseudo-sentences, and the subdivision of pseudo-sentences having a low occurrence.
Visual Features Extraction Approach

SVM Classification (1/2)
- use the LIBSVM implementation
- we use Platt’s method that produces probabilistic output using a sigmoid function.
  - The first considers the examples annotated “highly relevant” as positive examples and the other represents the negative ones.
  - The second merges the two classes ”highly relevant” and ”relevant” in a positive class and others are considered as negative examples.
  - The third consider the examples of ”highly relevant”, ”relevant” and ”irrelevant” as positive examples, and examples of ”neutral” and ”irrelevant” as negative examples.
Once the three classifiers are learnt with probabilistic SVM, we merge the three outputs by calculating the weighted average to obtain the final model using this formula:

\[ C = \alpha \ast C_{tp} + \beta \ast C_{tp+p} + \gamma \ast C_{tp+p+pp} \]
A complete three modules process, acting dependently:

1. Pre-processing
2. Acoustic sources separation
3. Training and classification

Block diagram of sound concepts extraction and classification
Audio Feature Extraction

A complete three modules process, acting dependently:

1. Pre processing
2. Acoustic sources separation
3. Training and classification
The audio stream is segmented into clips that are 3 seconds long with 1 second overlapping with the previous ones.

2. STE: Short Time Energy Feature

3. A merge module of no silence segments remaining runs to the preparation to a new segmentation.

4. Segmentation is well oriented to the detection of speech and music classes of the audio stream obtained.
1. The audio stream is segmented into clips that are 3 seconds long with 1 second overlapping with the previous ones.

2. STE : Short Time Energy Feature

3. A merge module of no silence segments remaining runs to the preparation to a new segmentation.

4. Segmentation is well oriented to the detection of speech and music classes of the audio stream obtained.
Acoustic sources separation module

Step 1: No silence segments are separated into speech and non-speech segments by two features: LSTER (Low Short Time Energy Ratio) and SF (Spectrum Flux).

Step 2: No speech segments are classified into music and environmental sound, by a BP (Band Periodicity feature).
Acoustic sources separation module

**Step 1**: No silence segments are separated into speech and non-speech segments by two features: LSTER (Low Short Time Energy Ratio) and SF (Spectrum Flux).

**Step 2**: No speech segments are classified into music and environmental sound, by a BP (Band Periodicity feature).
Learning and classification concepts module

1 - Concepts introduce and MFCC extraction

- Labeling user sets audio concepts for identification
- Audio samples of each concept are introduced by a cepstral description MFCC (Mel Frequency Cepstral Coefficient)

2 - SVM for classification

A support vector machine (SVM) is a two-class classifier constructed from sums of a kernel function $K(.,.)$,

$$f(x) = \sum_{i=0}^{N} \alpha_i y_i K(x,x_i) + b$$

$x$ is the vector needed to classify and $x_i$ are support vectors obtained from the training sets by an optimization process, $y_i$ is either 1 or -1 depending on the corresponding support vector belongs to class 0 or class 1.
Learning and classification concepts module

1 - Concepts introduce and MFCC extraction
- Labeling user sets audio concepts for identification
- Audio samples of each concept are introduced by a cepstral description MFCC (Mel Frequency Cepstral Coefficient)

2 - SVM for classification
A support vector machine (SVM) is a two-class classifier constructed from sums of a kernel function $K(\cdot, \cdot)$,

$$f(x) = \sum_{i=0}^{N} \alpha_i y_i K(x, x_i) + b$$

$x$ is the vector needed to classify and $x_i$ are support vectors obtained from the training sets by an optimization process, $y_i$ is either 1 or -1 depending on the corresponding support vector belongs to class 0 or class 1.
Multimodal Fusion Approach

- Fuse visual and audio concepts
- Fusion process $\neq$ aggregate concepts

Why we fuse?
- To generate coherent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Approach
- The fusion system is based on three different levels of the JDL/DFS Data Fusion Model:
  - level 1: Object refinement (dealing with conflicting situations)
  - level 2: Situation refinement (enrich semantic interpretation)
  - level 4: Fusion Process control
- The fusion process uses:
  - A fuzzy deduction reasoning engine (using LSCOM Ontology)
  - A fuzzy abduction reasoning engine
Multimodal Fusion Approach

- Fuse visual and audio concepts
- Fusion process $\neq$ aggregate concepts

Why we fuse?
- To generate coherent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Approach

The fusion system is based on three different levels of the JDL/DFS Data Fusion Model:

- Level 1: Object refinement (Dealing with conflicting situations)
- Level 2: Situation refinement (enrich semantic interpretation)
- Level 4: Fusion Process control

The fusion process uses:

- A fuzzy deduction reasoning engine (Using LSCOM ontology)
- A fuzzy abduction reasoning engine
Multimodal Fusion Approach

- Fuse visual and audio concepts
- Fusion process \neq aggregate concepts

Why we fuse?
- To generate coherent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Approach

- The fusion System is based on three different levels of the JDL/DFS Data Fusion Model:
  - level 1: Object refinement (Dealing with conflicting situations)
  - level 2: Situation refinement (enrich semantic interpretation)
  - level 4: Fusion Process control

- The fusion process uses:
  - A fuzzy deduction reasoning engine (Using LSCOM Ontology)
  - A fuzzy abduction reasoning engine
Level 1: Object Refinement

- This level deals with mixed unimodal semantic interpretations.
- As input, every concept has a list of indexed video content sorted by their descending pertinent ranks.
- These ranks are fuzzified.

Let \( r \) be the rank of a concept for a video content, and \( R \) is the highest rank of the same concept for all video contents. We seek for a fuzzified rank called \( r_N \) as follow:

\[
r_N = \left( \frac{(\epsilon - 1)}{(R - 1)} \right) \times (R - r) + 1
\]

Where \( \epsilon \) is a positive integer.
Level 1: Object Refinement

- This level deals with mixed unimodal semantic interpretations.
- As input, every concept has a list of indexed video content sorted by their descending pertinent ranks.
- These ranks are fuzzified.

Let $r$ be the rank of a concept for a video content, and $R$ is the highest rank of the same concept for all video contents. We seek for a fuzzified rank called $r_N$ as follow:

$$r_N = \left( \frac{(\epsilon-1)}{(R-1)} \right) \ast (R - r) + 1$$

Where $\epsilon$ is a positive integer.
Level 2: Situation Refinement

The purpose of this level is to look for new concepts by analysing available interpretations.
Level 2: Situation Refinement

The purpose of this level is to look for new concepts by analysing available interpretations.
Level 2: Situation Refinement

The purpose of this level is to look for new concepts by analysing available interpretations.

Deduction Engine

Abduction Engine
The table below shows concept detection improvement, given by our multimodal fusion system vs. the unimodal visual analysis system, in terms of indexed shots number.

<table>
<thead>
<tr>
<th>TV10 Concept ID</th>
<th>TV10 Concept Name</th>
<th>REGIM_4</th>
<th>REGIM_5 and REGIM_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Animal</td>
<td>627</td>
<td>737</td>
</tr>
<tr>
<td>12</td>
<td>Bicycles</td>
<td>55</td>
<td>249</td>
</tr>
<tr>
<td>15</td>
<td>Boat_Ship</td>
<td>177</td>
<td>246</td>
</tr>
<tr>
<td>21</td>
<td>Car</td>
<td>565</td>
<td>599</td>
</tr>
<tr>
<td>50</td>
<td>Face</td>
<td>1800</td>
<td>1925</td>
</tr>
<tr>
<td>51</td>
<td>Female_Person</td>
<td>1501</td>
<td>1874</td>
</tr>
<tr>
<td>67</td>
<td>Indoor</td>
<td>336</td>
<td>972</td>
</tr>
<tr>
<td>75</td>
<td>Male_Person</td>
<td>1883</td>
<td>2407</td>
</tr>
<tr>
<td>87</td>
<td>Outdoor</td>
<td>383</td>
<td>4636</td>
</tr>
<tr>
<td>90</td>
<td>Person</td>
<td>1998</td>
<td>9672</td>
</tr>
<tr>
<td>91</td>
<td>Plant</td>
<td>323</td>
<td>527</td>
</tr>
<tr>
<td>93</td>
<td>Politicians</td>
<td>391</td>
<td>418</td>
</tr>
<tr>
<td>108</td>
<td>Sky</td>
<td>845</td>
<td>845</td>
</tr>
<tr>
<td>111</td>
<td>Sports</td>
<td>1111</td>
<td>1277</td>
</tr>
<tr>
<td>125</td>
<td>Vegetation</td>
<td>1909</td>
<td>1909</td>
</tr>
<tr>
<td>126</td>
<td>Vehicle</td>
<td>728</td>
<td>1165</td>
</tr>
</tbody>
</table>
REGIM_4, REGIM_5 and REGIM_6 Results (2/3)

REGIM_4 Precision

Run score (dot) versus median (---) versus best (box) by feature.
REGIM_5 Precision

Run score (dot) versus median (---) versus best (box) by feature number.
REGIM_6 Precision

Inferred average precision versus Feature number

Run score (dot) versus median (---) versus best (box) by feature
The table below shows the precision at number of shot of each runs in our system.

It demonstrates the effectiveness of the multimodal fuzzy fusion system indexing.

<table>
<thead>
<tr>
<th>n Shot</th>
<th>Precision REGIM_4</th>
<th>Precision REGIM_5</th>
<th>Precision REGIM_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.630</td>
<td>0.630</td>
<td>0.630</td>
</tr>
<tr>
<td>100</td>
<td>0.536</td>
<td>0.528</td>
<td>0.527</td>
</tr>
<tr>
<td>1000</td>
<td>0.181</td>
<td>0.193</td>
<td>0.194</td>
</tr>
<tr>
<td>2000</td>
<td>0.094</td>
<td>0.102</td>
<td>0.102</td>
</tr>
</tbody>
</table>
Conclusion

- Preliminary experiments and obtained results are presented.
- The main direction for the **REGIMVid** enhancement is the multi-modal video indexing.
- Actually, the different video modalities indexing (visual and audio) are collectively performed.

Future Works

- We plan to incorporate motion information to detect concepts involving activities more effectively.
- **REGIMVid** Toolbox functionalities will be enhanced by complementary tools as personalization and visualization.
Conclusion

- Preliminary experiments and obtained results are presented
- The main direction for the REGIMvid enhancement is the multi modal video indexing.
- Actually, the different video modalities indexing (visual and audio) are collectively performed

Future Works

- We plan to incorporate motion information to detect concepts involving activities more effectively.
- REGIMVid Toolbox functionalities will be enhanced by complementary tools as personalization and visualization.
Thanks For Your Attention