Sfax University
National Engineering School of Sfax

REsearch Group on Intelligent Machines
Groupe de Recherche sur les Machines Intelligentes

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

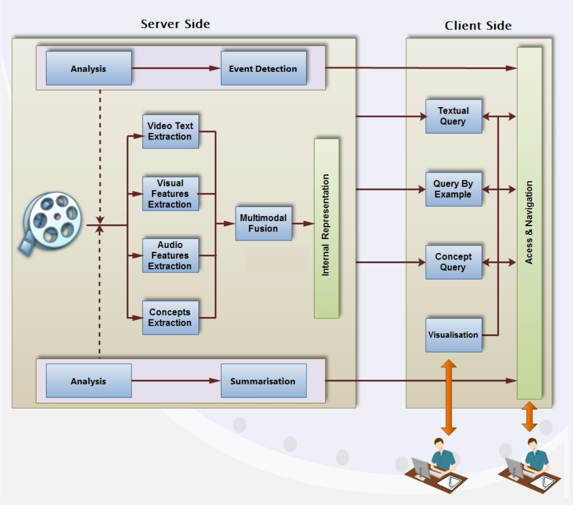
Content

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

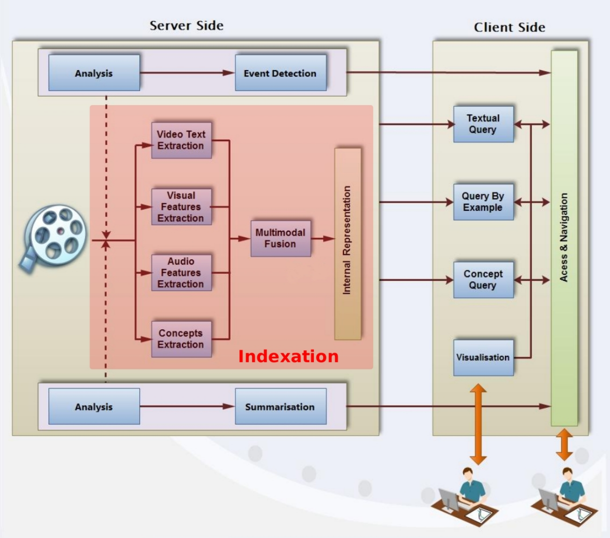
Content

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

RegimVid Overview

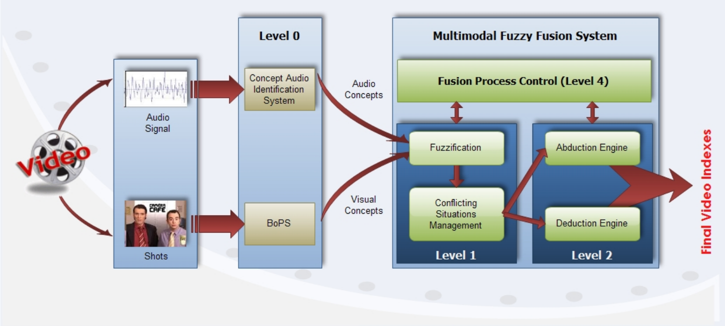


RegimVid Overview

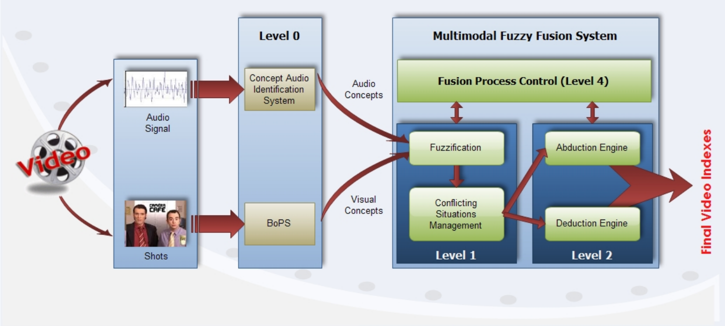


RegimVid Indexing Sub-System

The **RegimVid** indexing system provides an automatic analysis of video contents by using frame description based on low-level features.



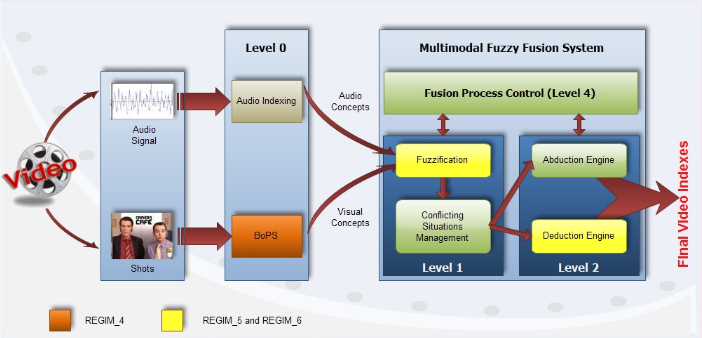
RegimVid Indexing Sub-System



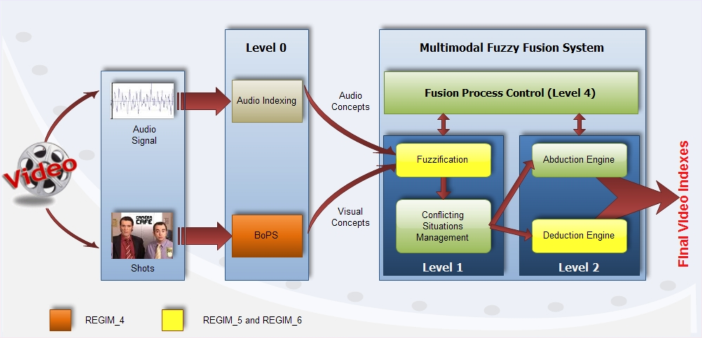
1. The system extracts the low-level features for each modality of the video shot
2. The system represents contents for labeling them, later, by basing on score detection via classification process.
3. The predicted score are merged to obtain multimodal fusion.

RegimVid Runs in TrecVid2010

Particpation in the Semantic Indexing Task (SIN)



RegimVid Runs in TrecVid2010

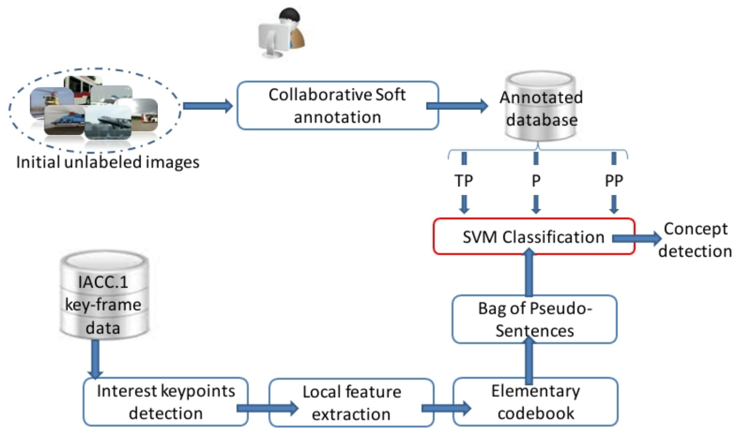


Regim$_4$ A visual modality analysis orientated towards an automatic categorization of video contents to create relevance relationships between low-level descriptions and semantic contents according to a user point of view

Regim$_5$ A Multimodal fuzzy fusion using positive rules extracted from LSCOM Ontology. The fusion process employs a deduction reasoning engine

Regim$_6$ A Multimodal fuzzy fusion using positive and negative rules extracted from LSCOM Ontology.

Visual Features Extraction Approach

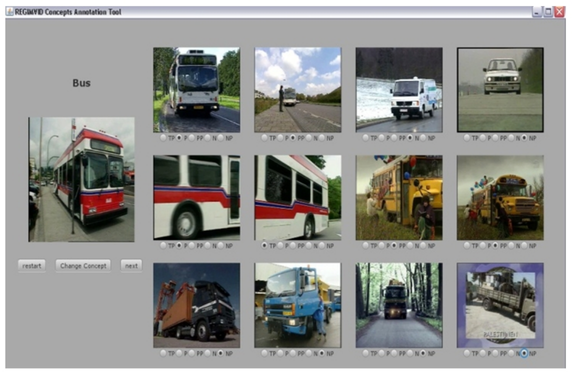


Visual Features Extraction Approach

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.

Visual Features Extraction Approach

Bag of Pseudo-Sentences

- We interested in 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 the memory, storage and the 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.

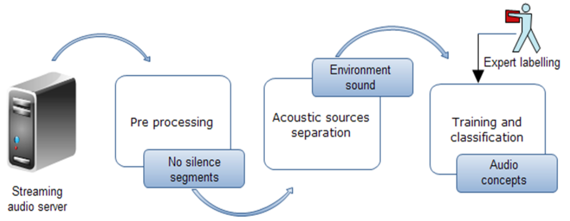
Visual Features Extraction Approach

SVM Classification (2/2)

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 :

$$C = \alpha * C_{tp} + \beta * C_{tp+p} + \gamma * C_{tp+p+pp}$$

Audio Feature Extraction

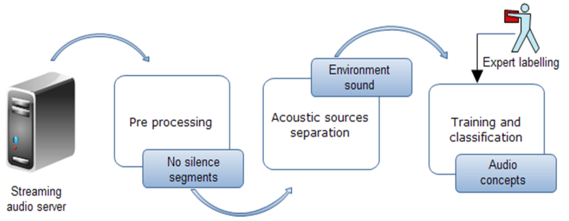


Block diagram of sound concepts extraction and classification

A complete three modules process, acting dependently :

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

Audio Feature Extraction

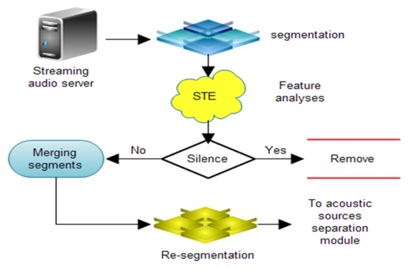


Block diagram of sound concepts extraction and classification

A complete three modules process, acting dependently :

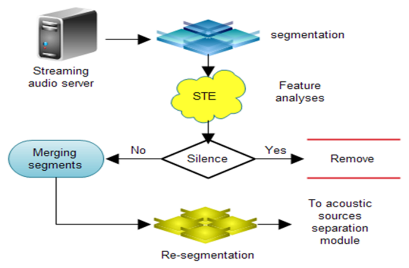
1. Pre processing
2. Acoustic sources separation
3. Training and classification

Pre processing module



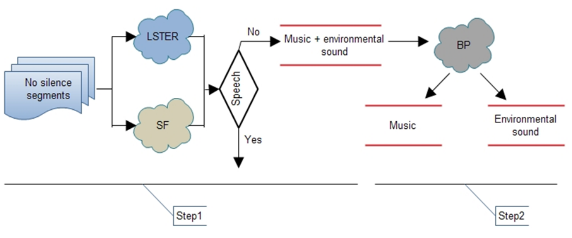
1. The audio stream is segmented into clips that are 3 seconds long with 1 second overlapping with the previous ones.
2. STE : Short Time Energy Feature
3. A merge module of no silence segments remaining runs to the preparation to a new segmentation
4. segmentation is well oriented to the detection of speech and music classes of the audio stream obtained.

Pre processing module



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.

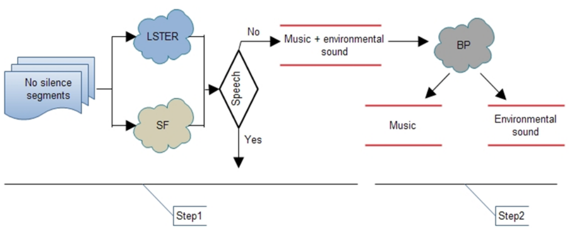
Acoustic sources separation module



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 Coeificient)

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 Coeificient)

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.

Multimodal Fusion Approach

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

Why we fuse ?

- To Generate cohrent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Aproach

- The fusion System is based on three different levels of the JDL/DFS Data Fusion Model :
  - level 1 : Object refinement (Dealing with conflincting situations)
  - level 2 : Situation refinement (enrich semantic interpretation)
  - level 4 : Fusion Process control
- The fusion process uses :
  - A fuzzy deduction reasoning engine (Unsing 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 cohrent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Aproach

- The fusion System is based on three different levels of the JDL/DFS Data Fusion Model :
  - level 1 : Object refinement (Dealing with conflincting 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 cohrent semantic interpretation
- To look for further concepts
- To enrich the semantic interpretation

Fusion Aproach

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

Fusion : Object Refinement

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 :

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

Where ϵ is a postive integer.

Fusion : Object Refinement

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 :

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

Where ϵ is a postive integer.

Fusion : Situation Refinement

Level2 : Situation Refinement

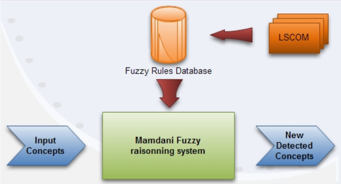
The purpose of this level is to look for new concepts by analysing available interpretations

Fusion : Situation Refinement

Level2 : Situation Refinement

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

Deduction Engine

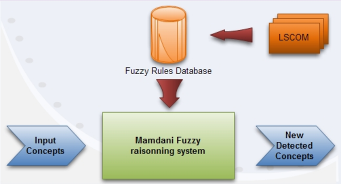


Fusion : Situation Refinement

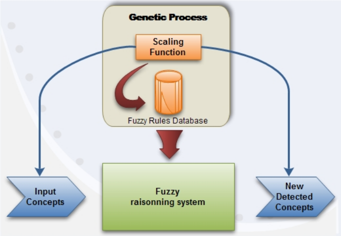
Level2 : Situation Refinement

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

Deduction Engine



Abduction Engine



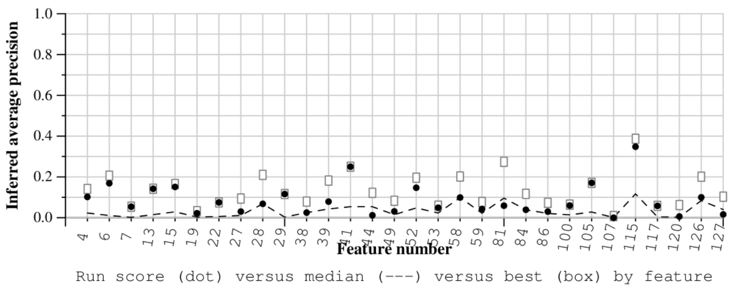
REGIM_4, REGIM_5 and REGIM_6 Results (1/3)

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.

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

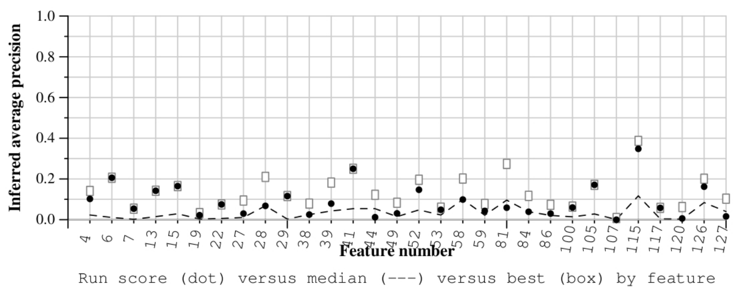
REGIM_4, REGIM_5 and REGIM_6 Results (2/3)

REGIM_4 Precision



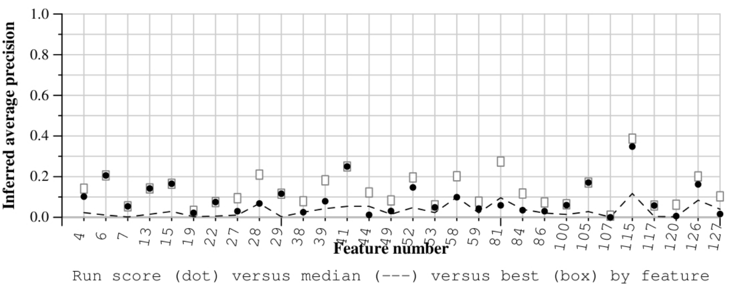
REGIM_4, REGIM_5 and REGIM_6 Results (2/3)

REGIM_5 Precision



REGIM_4, REGIM_5 and REGIM_6 Results (2/3)

REGIM_6 Precision



REGIM_4, REGIM_5 and REGIM_6 Results (3/3)

- 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.

n Shot	Precision REGIM_4	Precision REGIM_5	Precision REGIM_6
10	0.630	0.630	0.630
100	0.536	0.528	0.527
1000	0.181	0.193	0.194
2000	0.094	0.102	0.102

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