



The audio-visual message - Structure and Emotion

Riccardo Leonardi
Università degli Studi di Brescia

Thanks to: Nicola Adami, Sergio Benini, Luca Canini, Ugo Ciraci




● Signals & Communications Lab - University of Brescia




Outline

- Structural dimension
 - Semantics versus structure
 - Elementary units
 - Shots
 - Shot-cut detection
 - Comparing shots (retrieval / clustering)
 - Logical story units
 - Extraction methods
 - Modeling
 - Using LSU for video skimming
 - Video scene retrieval: LSU comparison
- Emotional dimension
 - Inducing emotional cues in the viewer
 - Visualization of the emotional cues




● Signals & Communications Lab - University of Brescia




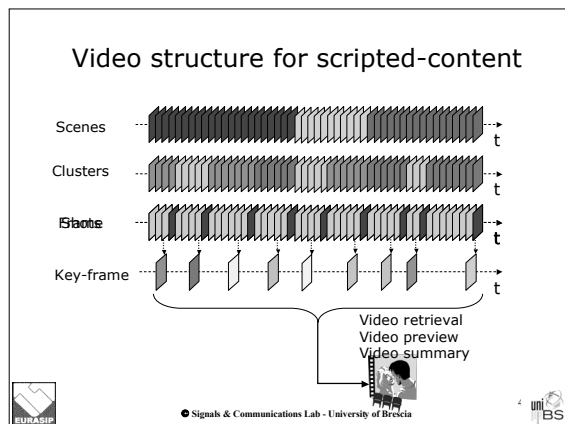
Segmentation versus Semantics

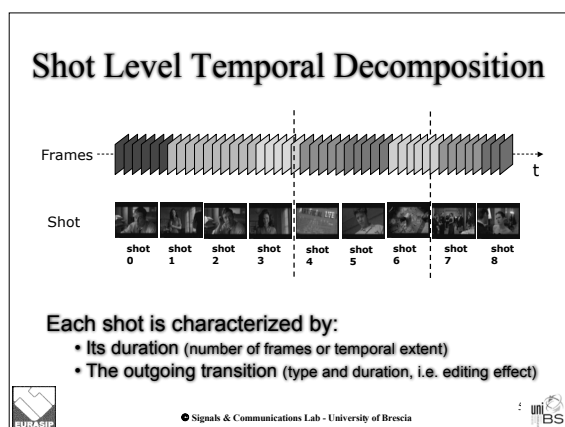
- Easy to associate a certain semantics to a segment of a partition of the domain of definition
 - 1D signals: time interval
 - 2D signals: region
 - 2D+t signals (video)
 - 2D decomposition: single object
 - 2D+t decomposition: moving object
 - t decomposition
 - > Single camera record (shot): action
 - > Consecutive set of camera records: interaction

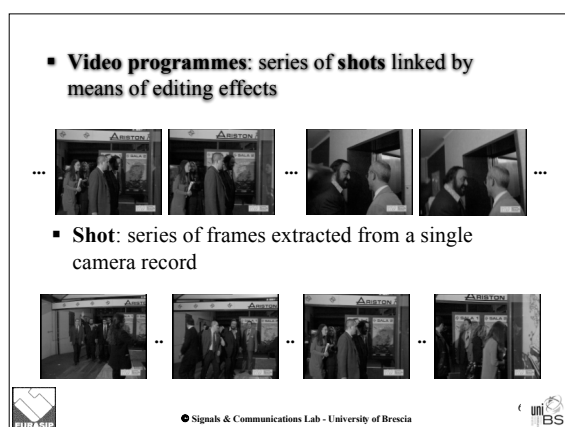


● Signals & Communications Lab - University of Brescia








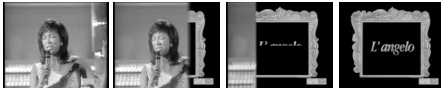


Editing effects: used to concatenate consecutive shots


Cut





Wipe





Dissolve



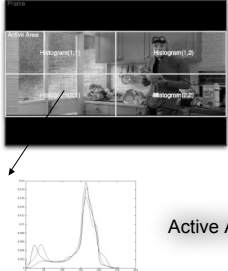

 Signals & Communications Lab - University of Brescia
 

Shot identification

- Cuts are typically detected by evaluating the distance between the Color Histograms of contiguous frames
- Dissolves are detected by statistical modeling
- Inter-frame difference and Motion information can be used to reject false positive
- Adaptive thresholds and active area detection can be used to improve detection performance




 Signals & Communications Lab - University of Brescia
 

Color Histogram



- Active area determination
- Arbitrary partitioning of the Active area into rectangles
- Selectable bin size and number of color components

Active Area can be dynamically estimated


 Signals & Communications Lab - University of Brescia
 

Dissolve detection

a frame contained in a dissolve (cross-fade) is obtained as a linear combination of the two concatenated shots

for
and

● Signals & Communications Lab - University of Brescia

Dissolve detection

- 2 shots participating in a dissolve can be considered as 2 realizations of 2 independent stochastic processes
- At each instant in time 2 independent RV's are linearly combined

I_{in}

I_{out}

p.d.f. of F_{in} image

p.d.f. of F_{out} image

● Signals & Communications Lab - University of Brescia

Dissolve detection

- the histogram of a frame involved in a dissolve is obtained as the convolution of the histograms associated to the frames of the 2 shots H_{in} and H_{out} involved in the dissolve

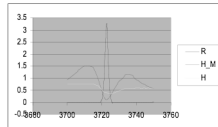
Central frame of dissolve

p.d.f. of central image of dissolve

H_{in} estimated by convolution

● Signals & Communications Lab - University of Brescia

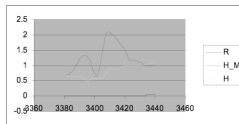
Statistical independence assumption



● Signals & Communications Lab - University of Brescia



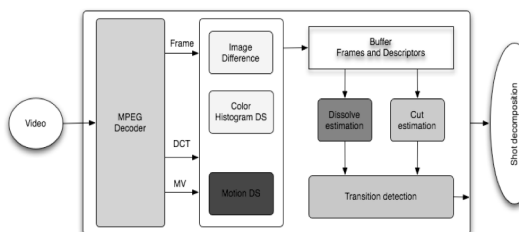
Behavior with statistical dependency



● Signals & Communications Lab - University of Brescia




Shot detection modules



● Signals & Communications Lab - University of Brescia




Demo




● Signals & Communications Lab - University of Brescia

14




Query by Example



● Signals & Communications Lab - University of Brescia


17



Motion Activity in Video


• **"High" activity scenes include**


- Goals in football
- Car chasing



• **"Low" activity scenes include**


- Talking heads
- Interviews






● Signals & Communications Lab - University of Brescia



18



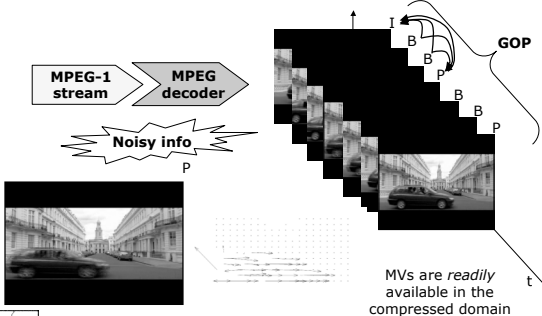
Motion-based video retrieval



- Motion activity
- Motion direction
- Motion spatial localization



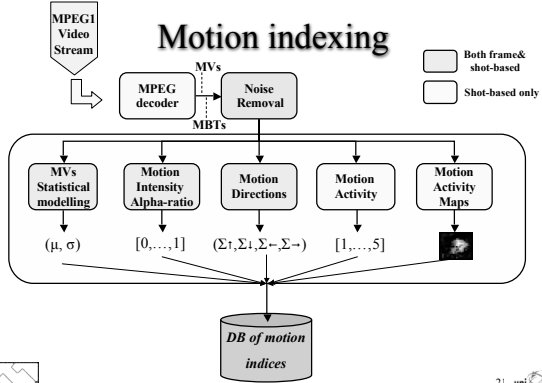

 Signals & Communications Lab - University of Brescia
 



Mpeg Motion Vectors (MVs)




 Signals & Communications Lab - University of Brescia
 

Motion indexing






 Signals & Communications Lab - University of Brescia
 

Motion Directions

- Direction of activity



- While a video shot may have several objects with differing activity, we can often identify one or more *dominant* directions


 Signals & Communications Lab - University of Brescia
 



Motion Activity Maps (MAM)

- Spatio/temporal distribution of activity

- MAM is an image that accumulates motion activity on the image grids along time axis
- The intensity of MAM pixel is the numeric integral of the motion vector field along time



 Signals & Communications Lab - University of Brescia
 

Motion-based distance on shots














 Signals & Communications Lab - University of Brescia
 

“Anchor man” retrieval

Query shot 11



Low motion


Rank #1	shot 351
	  
Rank #2	shot 330
	  
Rank #3	shot 324
	  

● Signals & Communications Lab - University of Brescia













2: uni BS

“Motorbike” retrieval

Query shot 435



High motion


Rank #1	shot 171
	  
Rank #2	shot 443
	  
Rank #3	shot 434
	  

● Signals & Communications Lab - University of Brescia













2: uni BS

“Jeep” retrieval

Query shot 441



High motion


Rank #1	shot 433
	  
Rank #2	shot 178
	  
Rank #3	shot 184
	  

● Signals & Communications Lab - University of Brescia


2: uni BS

“Football game” retrieval


Query shot 165




High motion




Rank #1 shot 144



Rank #2 shot 170




Rank #3 shot 156




● Signals & Communications Lab - University of Brescia 21 uni BS

“Moving legs” retrieval


Query shot 309




High motion




Rank #1 shot 312



Rank #2 shot 321



Rank #3 shot 205



● Signals & Communications Lab - University of Brescia 21 uni BS

Investigating the video structure

- Temporal segmentation of video content as:
 - a base for video information retrieval
 - a mean to summarize information
- Several semantic hierarchies
 - Shot decomposition
 - Scene decomposition
 - Programme decomposition
 - ...

● Signals & Communications Lab - University of Brescia 31 uni BS

LSU segmentation - definition

LSU - Logical Story Unit (*best approximation of a "scene"*)
 A sequence of contiguous and interconnected shots sharing a common semantic thread (i.e. time /location)

The diagram illustrates the segmentation of a video stream into Logical Story Units (LSUs). A horizontal timeline labeled 't' shows a sequence of shots from 'shot 0' to 'shot 8'. A dashed vertical line separates the first four shots (shot 0 to shot 3) from the remaining shots (shot 4 to shot 8). The first group of shots is labeled 'LSU'.

Signals & Communications Lab - University of Brescia

Decoding and Key-frame selection

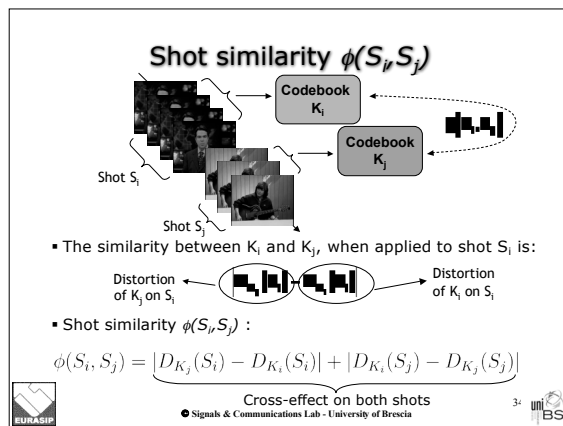
The diagram shows the process of decoding and key-frame selection. An 'MPEG-1 stream' is processed by an 'MPEG decoder' to produce a sequence of frames. 'Shot segmentation information' is used to identify key-frames, such as 'Key-Frame Shot 0' and 'Key-Frame Shot 1'. The frames are arranged in a staircase pattern along a time axis 't', with 'Shot 0' and 'Shot 1' labeled.

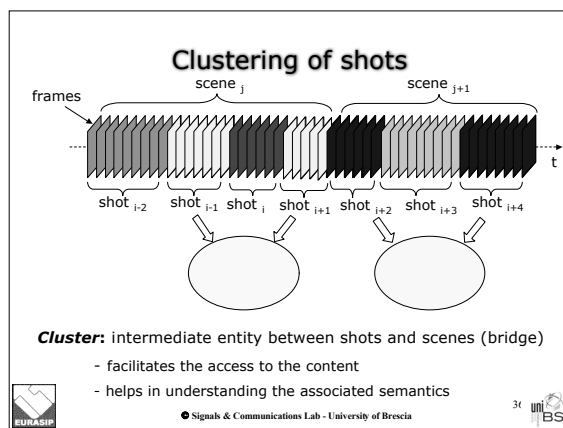
Signals & Communications Lab - University of Brescia

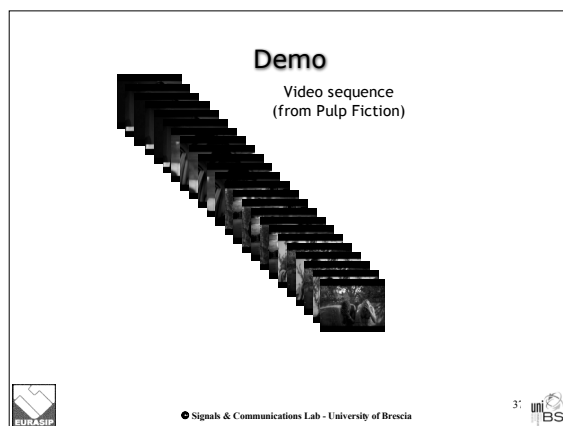
Vector Quantization process

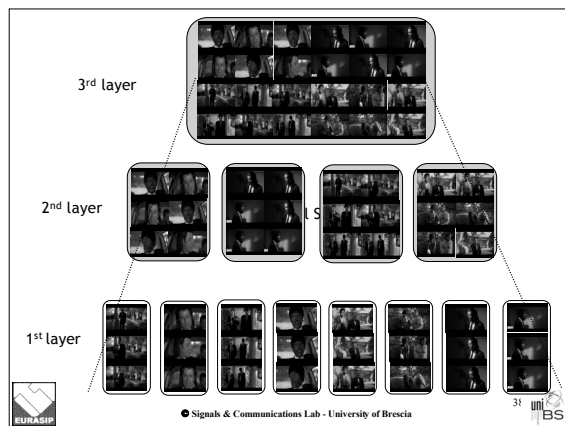
The diagram illustrates the Vector Quantization process. A video frame (352x288) is subsampled into a 176x144 block. The number of vectors is 1584 (blocks 4x4). The process involves a search for the best match in a 'Codebook K_j ' using a 'Hypercube (48 dimensions)'. The dimension of each vector is 48 (4x4 pixel, 3LUV each). The process starts at 'START' and ends at 'STOP'.

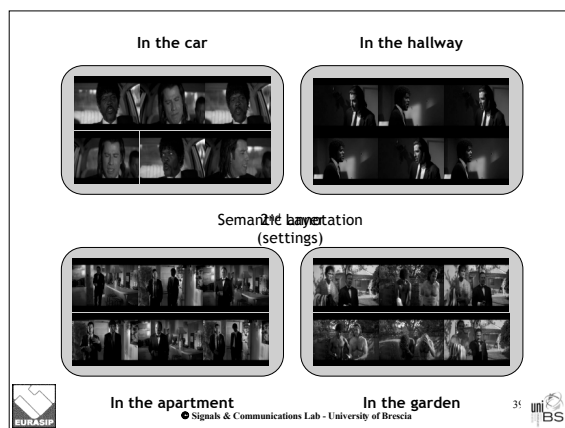
Signals & Communications Lab - University of Brescia

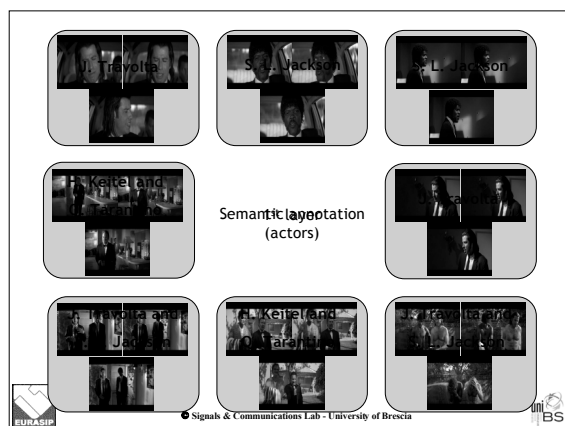








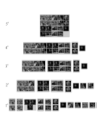




LSU segmentation


i. *Cluster validity analysis*

- to determine the best hierarchical level



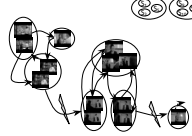
ii. *Temporal constraint analysis on clusters*


- to separate very long distance recurring shots



iii. *Scene Transition Graph*


- to identify LSU boundaries



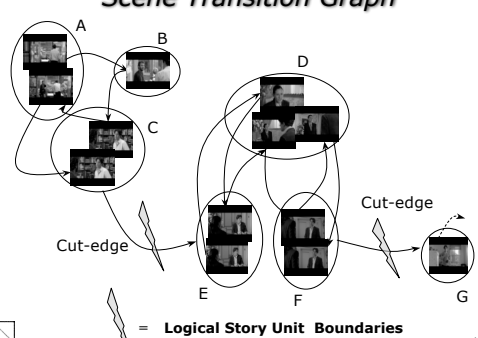


● Signals & Communications Lab - University of Brescia


41



Scene Transition Graph




= **Logical Story Unit Boundaries**



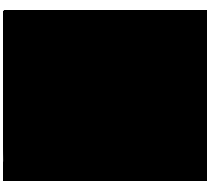
● Signals & Communications Lab - University of Brescia


41



Investigating the video structure


- Segmentation into Logical Story Units (LSU): an example





● Signals & Communications Lab - University of Brescia

41



Digital media summarisation

EURASIP

Signals & Communications Lab - University of Brescia

41 UniBS

Video structure

Scene

Group

Frame

Key-frame

video preview

EURASIP

Signals & Communications Lab - University of Brescia

41 UniBS

What is a skim?

Video

Skim

EURASIP

Signals & Communications Lab - University of Brescia

41 UniBS

Why skims?

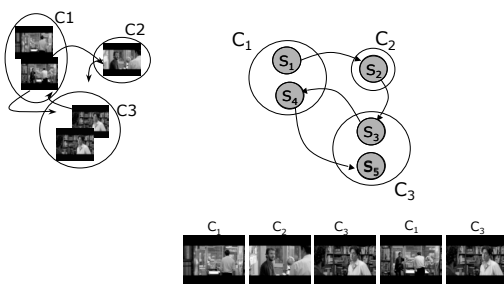
- *Video skims can be used?*
 - When the user focuses on a specific content
- *Professional trailers available only for movies*
 - For other content a skimmed version can be helpful
- *Trailers are art objects by themselves...but*
 - Are they really useful for a clear insight of content?



● Signals & Communications Lab - University of Brescia



Analysis of shot patterns

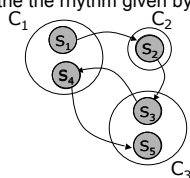


● Signals & Communications Lab - University of Brescia



Levels of information in a scene

1. Content information
 - Information associated to the concept represented (e.g., woman in red)
2. Pace information
 - Information related to the rhythm given by the transitions



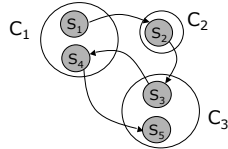
● Signals & Communications Lab - University of Brescia



Levels of information in a scene

3. Structure information

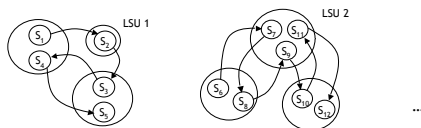
- Related to the cardinality N of the concepts in the story
AND the pattern S in which they are observed



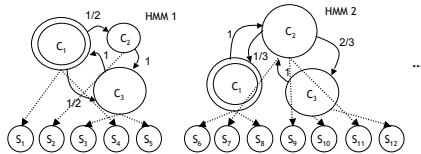
Signals & Communications Lab - University of Brescia



From scenes...



- ... to Hidden Markov Models (HMM)



Signals & Communications Lab - University of Brescia



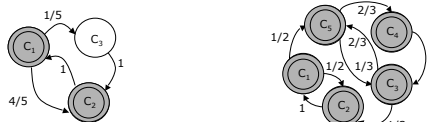
Examples of HMM

HMM of a dialogue

- After sequence $C_1 C_2 C_1$ it is likely that concept C_2 appears again

HMM of a dynamic scene

- It is difficult to predict the next concept



Signals & Communications Lab - University of Brescia



Criteria for skimming

■ *Content informativeness*

- more “informative” video shots should be preferred with respect to less informative;

• *Representativeness & Coverage*

- each scene should participate to the skim proportionally to its duration in the original video;

• *Structure informativeness*

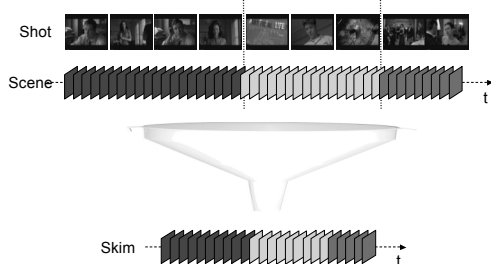
- information conveyed by shot patterns inside stories (e.g. dialogues,...) should be preserved;



● Signals & Communications Lab - University of Brescia



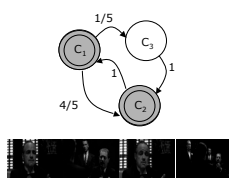
Coverage and representativeness



● Signals & Communications Lab - University of Brescia



Structure informativeness



Skim ratio = 50%



● Signals & Communications Lab - University of Brescia



Content informativeness

▪ Which shots to select?

- Salient features!
 - saliency depends on the video genre and the user
- Two possible options
 - Dynamic shots (action movies, etc.)
 - Presence of human faces (drama, comedy, etc.)



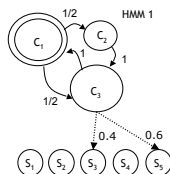
• Signals & Communications Lab - University of Brescia



Salient features and observations

▪ To privilege informative segments, salient shots are assigned higher probability of observation

- Probability $\sigma_j(k)$ is given by



• Signals & Communications Lab - University of Brescia



Example: Matrix

▪ *Movie with mixed content*

- Dialogues
- Dynamics scenes (chasings, etc.)

▪ Two provided skims


- 10% of first 30 minutes
 - A) Emphasis on motion dynamics
 - B) Emphasis on the presence of faces



• Signals & Communications Lab - University of Brescia




Demo




A
Fs=motion

B
Fs=faces

51






● Signals & Communications Lab - University of Brescia

User tests


Automatic Skim Generation Using Hidden Markov Models
 USER TEST - Step 7 of 10


(Already done with this skim? If yes, please go to the [NEXT STEP](#))

This is a 3min-length preview generated from the first 30min of a blockbuster movie.
 Please watch it and then click the button below to fill out the questionnaire.



61







● Signals & Communications Lab - University of Brescia

Test set

- **10 blockbuster movies:** introductory part (30 min)
 - 3 skims generated with 3 different methods (A, B, C)
 - Skimming ratio 10% (~3 min)
- **A:** selects video segments randomly
- **B:** our algorithm
 - Drama: emphasis on face presence
 - Action movies: motion dynamics
- **C:** manually generated by a cinema lover and editing system expert

61





● Signals & Communications Lab - University of Brescia

Usefulness and quality / Informativeness

Automatic Skim Generation Using Hidden Markov Models
 USER TEST - Step 7 of 10

This is a question sheet about the preview you've just watched.
 Please answer the questions and click 'Submit' to send your data and go to the next step.
 (Legend: 1=min, ..., 5=MAX)

A - Usefulness/quality

A1) Is the preview useful in understanding the genre of the movie it's generated from?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

A2) Is the preview able to give you an idea of the movie's atmosphere?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5



A3) Is the preview able to give you an idea of the narration pace?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

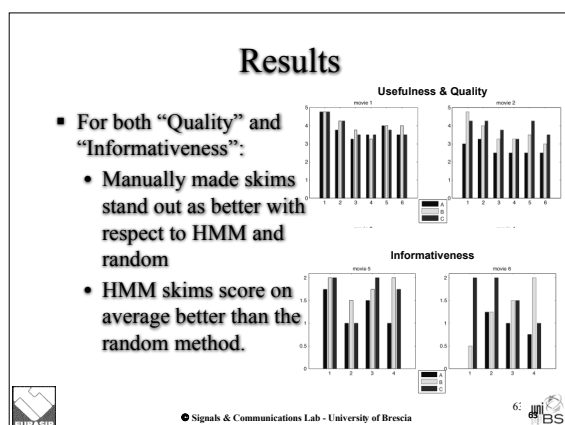
A4) Is the preview able to give you an idea about the characters involved?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

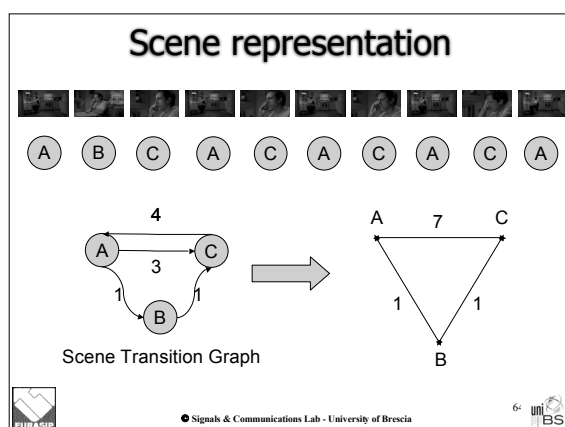
A5) Do you think this preview can be useful in understanding if you will like the movie?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

A6) Please give the preview a global score:
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

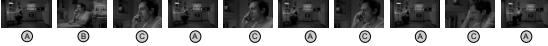
A7) If you've any comment about the preview, write it down in the box below.

  Signals & Communications Lab - University of Brescia





Topological + content features




Associate to each node (A,B,C) topological information:


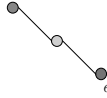
- In-degree, out-degree
- Betweenness
- Path distances

	in	out	bw	pd
A				
B				
C				

In-degree, out-degree



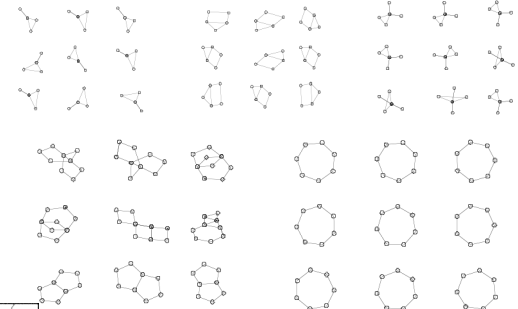
Betweenness

6: uni BS

● Signals & Communications Lab - University of Brescia

Results



● Signals & Communications Lab - University of Brescia

6: uni BS

Ground truth scenes

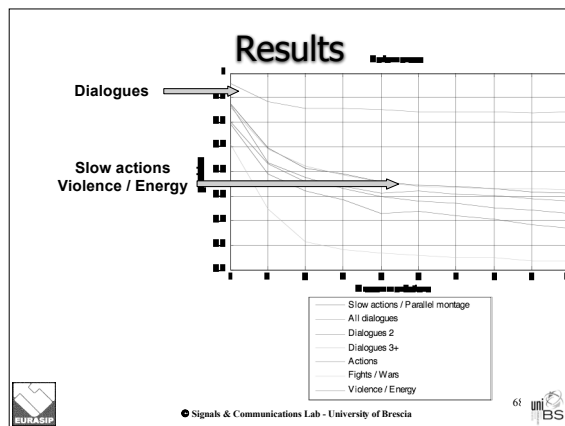
- Collection of 180 movies;
- About 35000 scenes automatically generated as LSUs;
- Selection of 800 scenes manually tagged by several *conceptual categories* chosen thinking to common situations of interest;

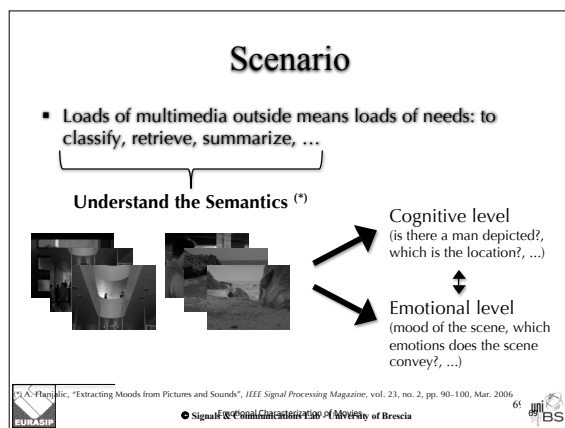
Categories:

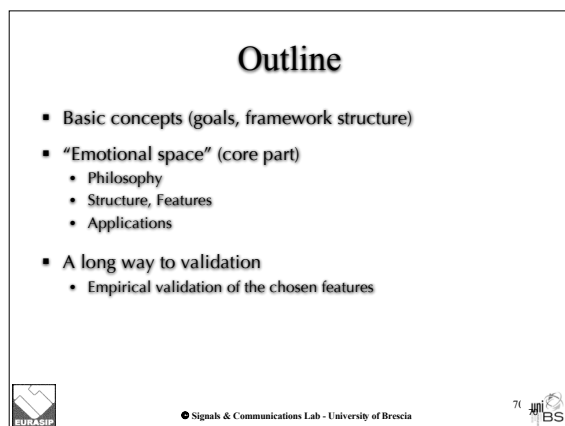
- *Dialogues*
- *Slow actions / Parallel montage*
- *Actions*
- *Fights*
- *Wars*

● Signals & Communications Lab - University of Brescia

6: uni BS







Basic concepts

- Long term goal: to describe the emotional identity of a feature movie

A product in which almost every element (actions, the shooting style, ...) is thought to convey a message

The mixture of feelings the director intended to communicate by using the movie in its complexity

- To do that, a Framework is under construction
 - Core part of the system: the Emotional space ^(*)
 - Several plug-in modules...

(*) Camini, Berini, Migliorati and Leonardi, "Emotional Identity of Movies", In Proc. of ICIP, Cairo, Egypt, Nov. 2009

● Signals & Communications Lab - University of Brescia

7:

EMOTIONAL SPACE

● Signals & Communications Lab - University of Brescia

7:

ES: overview

- 3D (extendible) space whose dimensions are related to features eligible to have an emotional impact

- The trajectory evolution over time provides a strong description of the multimedia content

● Signals & Communications Lab - University of Brescia

7:

ES: overview

- Consider a movie as a piece of art and inspect concepts used to study design objects (*)

"Design" concept	Object Description
Natural	Does it convey <i>passion</i> or <i>reflection</i> ?
Temporal	Turned to the <i>past</i> or projected toward the
Energetic	Is it seductive and <i>energetic</i> or flat and <i>basic</i> ?

↓ Our Transposition

High level concept	Dichotomic adjectives	Low level feature
Natural	Warm / cold	Illuminant of the scene
Temporal	Dynamic / slow	Motion dynamics
Energetic	Energetic / minimal	Sound energy

(*) C. T. Canelli, "Trini diagram: imaging emotional identity 3D positioning tool," In Proc. of SPIE, Dec. 1999.

Signals & Communications Lab - University of Brescia

71

ES: overview

Signals & Communications Lab - University of Brescia

71

Natural dimension

- Illuminant**
 - It's used by directors to underline the mood of a scene
 - It's suitable for the natural axis
 - natural (passionate/reflective) ~ warm/cold
 - colours can be classified into warm/cold (*)
 - It is estimated by using a "white patch" algorithm (**)
- Then, it is mapped on the natural axis**
 - Axis is 1D, color is 3D. How to deal with that?

The elements of color, J. Wiley & Sons, 1970

Vahde-Weijer et al., "Edge-Based Color Constancy," IEEE Transactions on Image Processing, vol. 16, no. 32, pp. 2207-2214, Sep. 2007.

Signals & Communications Lab - University of Brescia

71

25

Natural dim: colour perception

illuminant

warm/cold

natural dimension

● Signals & Communications Lab - University of Brescia

Natural dimension: white patch

illuminant

warm/cold

natural dimension

- Simple but effective and fast
 - Hyp.1: there is at least a "truly white" area (*)
 - Hyp.2: the illumination is quite uniform
 - Algorithm
 - o individuate area with maximum reflectance $s(\lambda)$
 - o its colour is a good estimation of the illuminant
 - o procedure frame by frame, averaged over the shot

(*) truly white area = an area which would be white under "white light" ($s(\lambda)=1$)
[in the picture, zones with the sea foam are good candidates]

● Signals & Communications Lab - University of Brescia

Natural dimension: mapping

illuminant

warm/cold

natural dimension

- Black Body Radiation: a suitable starting point
- Illuminant colour mapped on the axis interval which is at minimum Euclidean distance from it

● Signals & Communications Lab - University of Brescia

Temporal dimension

- Motion dynamics
 - Used by directors to stress the identity of a scene
 - o ex: speed and dynamism VS calm and tranquility
 - o achieved with shot type/length, amount of motion...
 - Defined as a combination of
 - o shot length (sequences of short shots - high narration pace)
 - o motion activity (MPEG7 descriptor) ^(*)

(*) S. Jeannin et al., "MPEG-7 Visual Motion Descriptors",
IEEE Trans. on Circuits and Systems for Video Technology, vol. 11, no. 6, Jan. 2001

Signals & Communications Lab - University of Brescia

Temporal dimension

- For each shot, motion dynamics:

related to shot pace related to motion activity
(short shot → high value) (std dev of motion vector modules)

T = value on the temporal axis
 l_{shot} = shot length
 l_{avg} = average shot length computed on a large movie database^(*)
 k = scale coefficient

(*) Cinemetric Database - www.cinemetrics.lv/database

Signals & Communications Lab - University of Brescia

Energetic dimension

- Audio dynamics
 - Scene semantics is usually stressed thanks to audio
 - o ex: gentle for romantic moments VS aggressive for actions
 - o audio: speech + soundtrack + ...
 - Audio energy is the selected low-level feature
 - o why such a simple feature?
 - o lots of efforts have been done with good results
 - o ...but no previous works in the group on audio topics

ideas are welcome ☺

Signals & Communications Lab - University of Brescia

Energetic dimension

audio energy

energetic/minimal

energetic dimension

- For each shot, audio energy:
 - Down-sampling at 8 kHz single-channel
 - Average log-energy of samples > threshold
 - highlight brief and intense events (thunders, screams, ...)

● Signals & Communications Lab - University of Brescia

81

Drawing movies

Movie trajectory

- each shot → point in the space
- trajectory = 3D cubic spline of points
- many shots → too complex representation

↓

3D solid

- summarizes trajectory characteristics
- solid parameters:
 - colour: average illuminant colour
 - shape: related to trajectory's smoothness
 - centroid: average shot positions
 - dimensions: std dev of points over the three axis

● Signals & Communications Lab - University of Brescia

81

EMOTIONAL SPACE APPLICATIONS

● Signals & Communications Lab - University of Brescia

81

Overview

- The emotional space:
 - Represents the movie as a trajectory thanks to features which are involved in the conveyance of the emotional message
 - The representation with solids provides a compact description
- Ideas for possible applications:
 - Retrieval/Recommendation by using distance between emotional identities (solids)
 - Content mash-up by using trajectory comparison
 - Content clusterization with respect to different space zones
 - ...



Signals & Communications Lab - University of Brescia



Retrieval: measure of distance

- Distance D between solids (S_A, S_B)
 - Distance between centroids (\bar{S}_A, \bar{S}_B)
 - Difference in dimensions ($|S_A - S_B|$)
 - w_i for normalization, w_i weights the terms



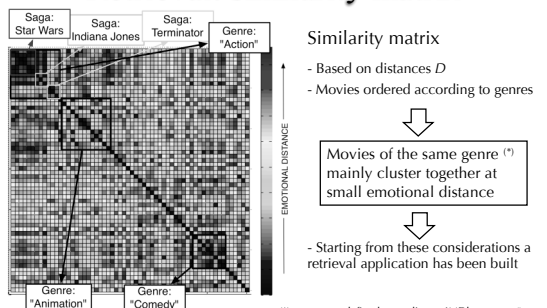
- D could be useful for retrieving emotionally similar movies from a database
- In first instance, movie genres are considered as a good (even if raw) emotional categorization



Signals & Communications Lab - University of Brescia



Retrieval: similarity matrix



(*) genres are defined according to IMDb (www.imdb.com)

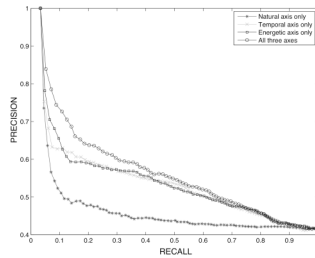


Signals & Communications Lab - University of Brescia



Retrieval: performance

- Retrieval of movies whose 3D solids are at low emotional distance from the query solid



Signals & Communications Lab - University of Brescia



LONG WAY TO VALIDATION



Signals & Communications Lab - University of Brescia



Validation: motivations

- So far
 - Features that are likely to be involved in the emotional characterization of a movie have been used
 - These choices are supported by several considerations
 - Well known cinematographic techniques
 - Common sense of our cultural background
 - Our ideas ☺
- It's worth trying to do a robust validation of our choices
 - By doing tests to check if "perceived \leftrightarrow conveyed"
 - Currently we are doing some tests; others are planned



Signals & Communications Lab - University of Brescia



