T-61.6030 Multimedia Retrieval

Multimodal content-based video retrieval

Dušan Sovilj

11.04.2008.

Outline

- Chapter 10 of 'Multimedia retrieval' by Blanken
- Introduction
- Processing audio signal
- Fusion of audio and video signal
- Superimposed text
- Summary

Introduction

- Detection of important events (hightlights) in Formula1 races
- Fusion of evidence from different modalities:
  - audio
  - video
- Using superimposed text in video signal for powerfull querying

Introduction

- Audio
  - detection of excited speech
- Video
  - capturing high level concepts (specific highlights) based on low-level features
- Superimposed text
  - using domain specific features for complex queries

Processing audio signal

- Audio signal from the car race is complex and ambiguous
- Filter out unneccesary noise, car engines, crowd, etc. leaving final audio with only speech
- Goal:
  - find segments of excited speech in filtered signal
  - recognize domain specific keywords
    - Formula1 - pit-stop, crash, passing

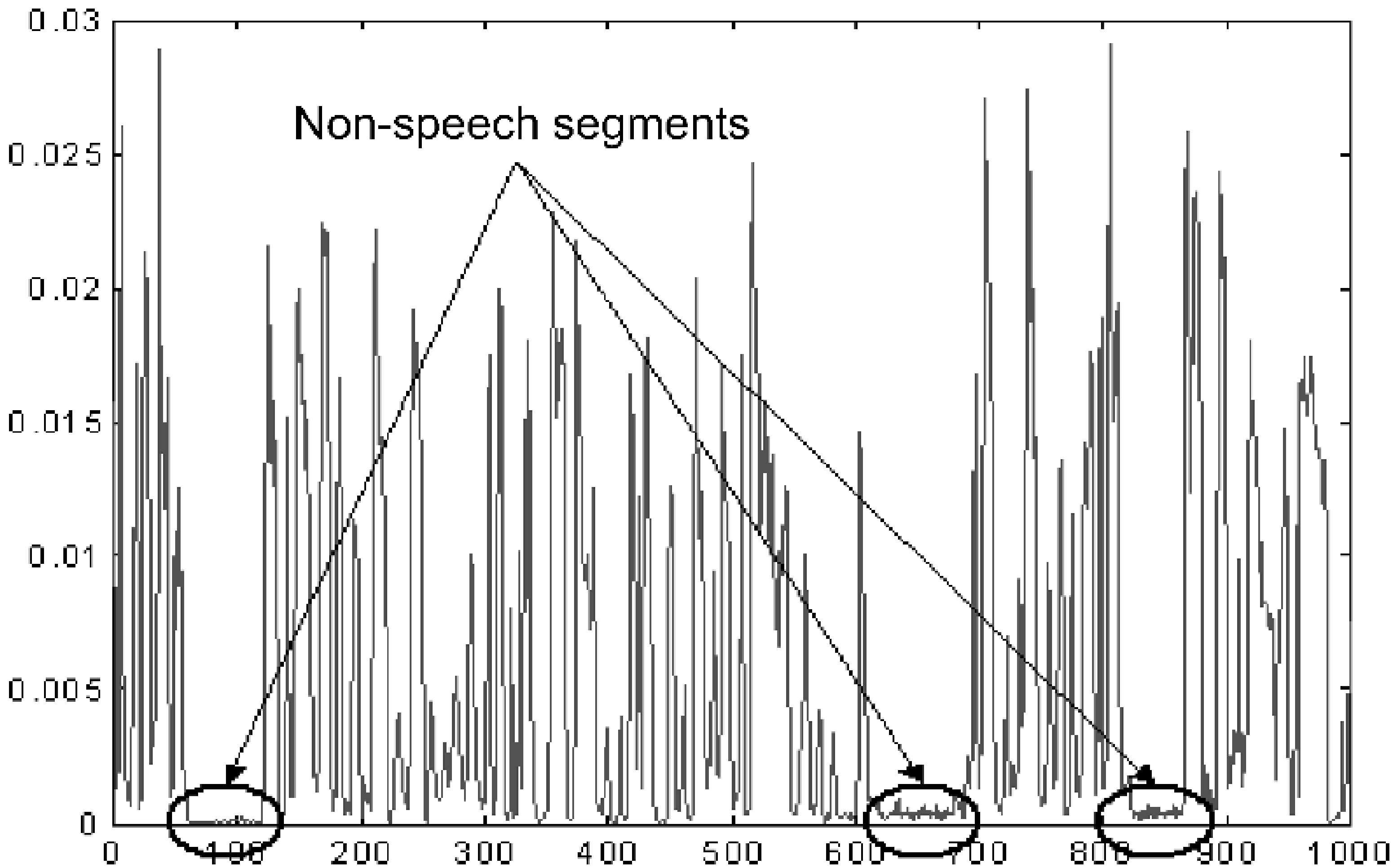
Processing audio signal

- Detection of speech segments based on following low level features
  - Short time energy (STE)
  - Mel-Frequency Cepstral Coefficients
  - Pitch
  - Pause rate
- Divide audio singal into suitable resolution:
  - frames (10ms) and segments (100ms)

Processing audio signal

- Idea:
  - For detecting speech segments use Short-Time Energy and MFCC
  - Pitch and pause rate are responsible for detection of excited speech
- All features are used together (not in steps)

Speech sequence detecion



Short-term energy calculations for 1000 audio frames (source: Blanken)

Keyword spotting

- Focus on recognizer for limited number of words as it gives less false alarms then general recognizer
- 30 in case of Formula1 racing
- Based on finite state grammar

Detection of excited speech

- Choice of features and calculations
- Model used:
  - Bayesian network (BN)
  - Dynamic Bayesian network (DBN)
- Influence of network structure
- For DBN:
  - influence of temporal dependencies

Choice of features

- For splitted signal in segments, derive many features:
  - keywords (f_1)
  - pause rate (f_2)
  - average value of STE (f_3)
  - dynamic range of STE (f_4)
  - average MFCC (f_7)
  - ...

Bayesian network

- Selected features are used as input to this probabilistic framework, and are considered as evidence nodes (observed variables)
- Dependencies between these variables are captured with hidden nodes (stochastic variables)

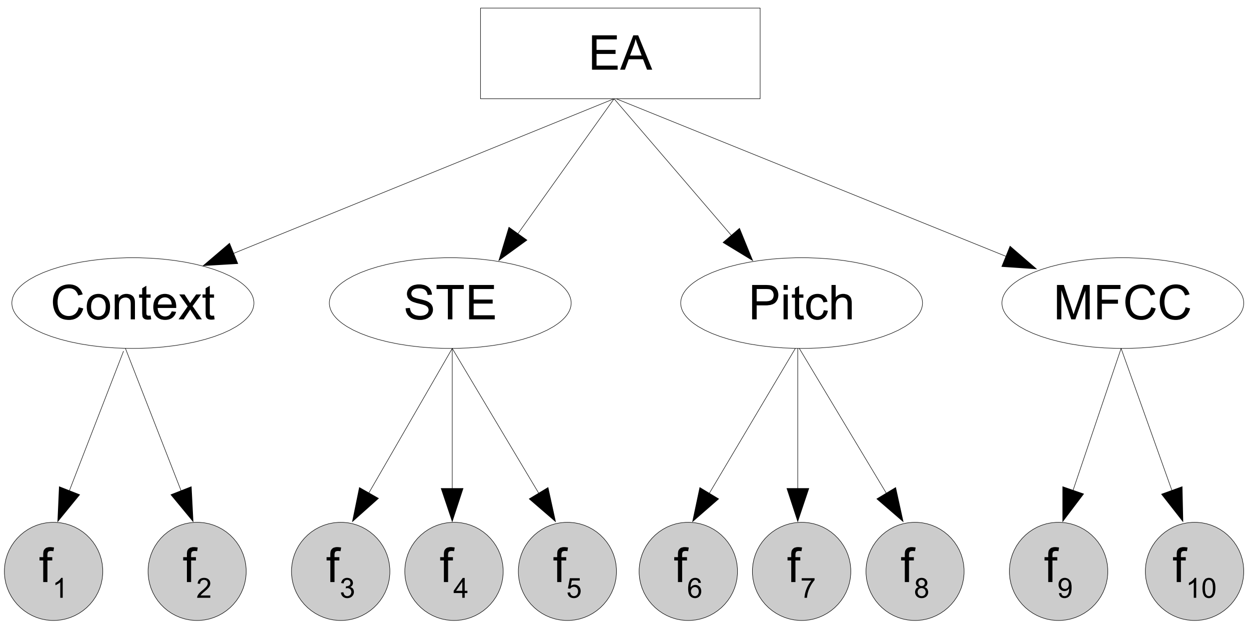
Dynamic Bayesian network

- Dynamic BN can deal with time aspect
- Stochastic variables may depend on observed (stochastic) variables from previous time segment
- Satisfies first order Markov property
- In Bayesian network dependencies from segments are not allowed. Only within one time segment and between observed and stochastic variables.

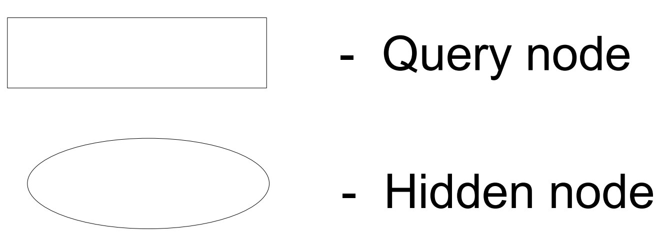
(Dynamic) Bayesian network

- Conditional probabilities are learned from the training set for both types of networks
- EM is used to find these conditional probabilities

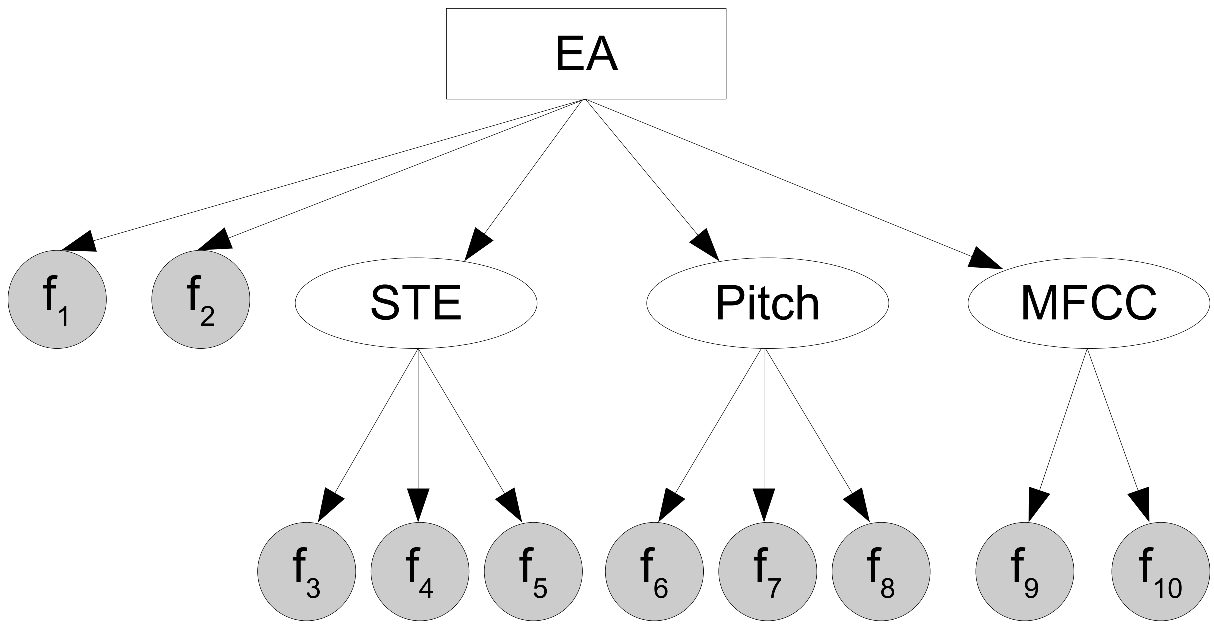
Influence of network structure



First solution for connections between nodes in BN (source: Blanken)

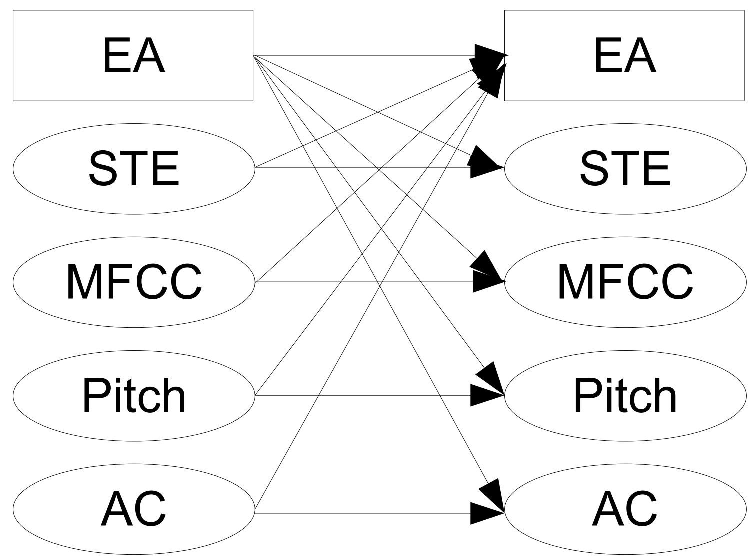


Influence of network structure



Second solution for connections between nodes in BN (source: Blanken)

Influence of network structure



Temporal dependencies between hidden nodes for DBN (Source: Blanken)

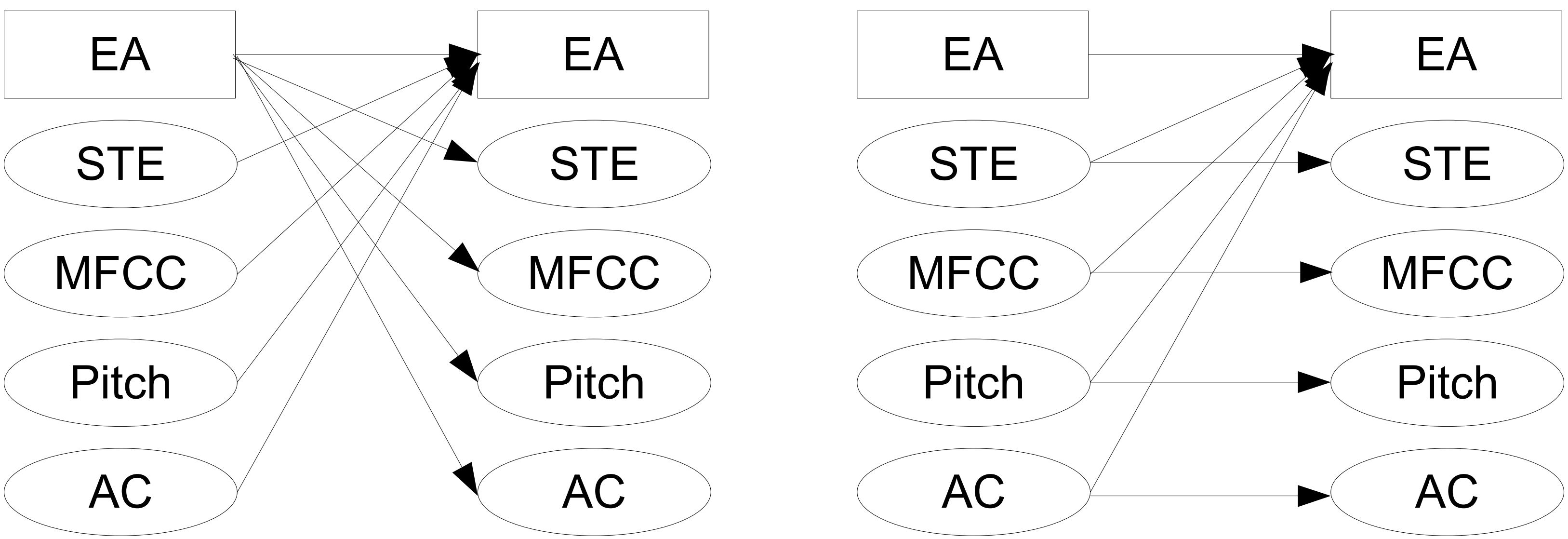
Results

- Training set: 300s from audio signal, in 3000 samples of 100ms
- For DBN same 300s into 12 segments of 25s
- Processing output values for BNs

Network structure	BN (type 1)	BN (type 2)	DBN (type 1)
Precision	60%	54%	85%
Recall	66%	61%	81%

Influence of temporal dependencies

- Specifying different temporal dependencies between stochastic nodes gives different results



Another way of specifying temporal dependencies (source: Blanken)

Results

- Fully connected (parametrized) dependencies gives best results and its performance is shown on other 2 races for evaluation

Race	German GP	Belgian GP	USA GP
Precision	85%	77%	76%
Recall	81%	79%	81%

Analyzing image stream

- Some highlights can be missed if using only audio signal
- Using low-level video features (color-histogram, dominant color, shape moments) to further improve results
- Focus on concept and domain of racing to extract specific events: passing, start of race, fly-outs
- Problems
  - replay sections
  - shot division

Motion

- Motion information is based on block-matching or optical flow techniques
  - low-level and extracted from motion vectors
  - high-level (camera motion - zooming, panning, tilting)
- Optical flow from motion vectors formed from pixel colors used in DBN

Shot segmentation

- Sequences of more or less same content based frames
- Color histogram difference (N - number of colors in frame)

$$HD(H_t, H_{t-1}) = \sum_{i=1}^{N} \frac{(H_t(i) - H_{t-1}(i))^2}{H_t(i)}$$

- Setting appropriate threshold for HD gives good results

Replay detection

- Detecting word 'Replay' in image frames
  - easy, but differs for every race
- Digital Video Effects (DVE)
  - special sequences marking start and end of replay sections
  - must be learned for every race
- One easy and fast way
  - simple RGB color change detection on central part of image frame (feature f_{12})

Image 'features' 1

- Start of race
  - defined by amount of motion and red lights on semaphore
  - detection of motion based on pixel color difference for 3 colors: red, green and blue (f_{13})
  - measuring the amount of red light with filtering the image for red color (f_{14})

Image 'features' 2

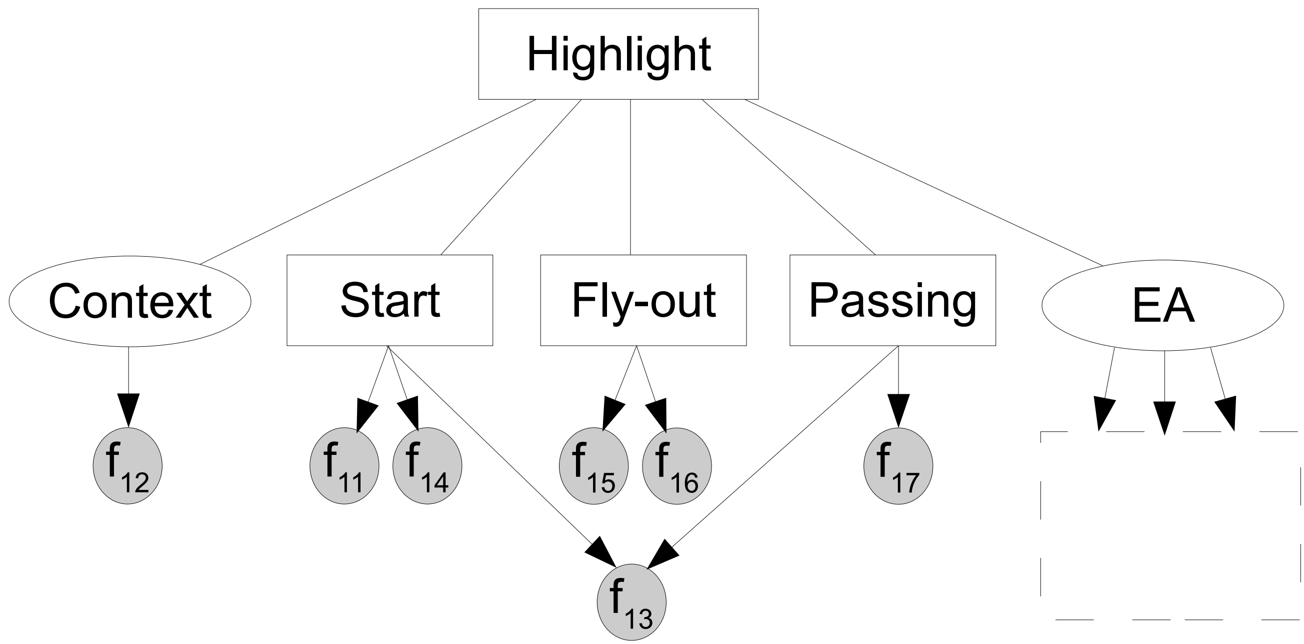
- Passing of another car
  - detecion based on motion histograms from consecutive still images
  - pixel-color difference, same as for start (f_{13}) and amount of motion (f_{17})

Image 'featues' 3

- Fly-outs
  - accompnied with lof of dust and sand
  - find out if there is dust or sand in images, looking for dominant colors
  - first find out the dominant color based on several still images
  - for actual evidence use filtered RGB images and calculate the amount of dominant colors
    - sand f_{15}
    - dust f_{16}

Highlight detection

- Training on 6 sequences of 50 seconds of 1 race



Audio-visual DBN for one time slice (source: Blanken)

Results

	Audio/video DBN	Ger.GP	Bel.GP	USA GP
Highlights	Precision	84%	43%	73%
	Recall	86%	53%	76%
Start	Precision	83%	100%	100%
	Recall	100%	67%	50%
Fly-out	Precision	64%	100%	-
	Recall	78%	36%	-
Passing	Precision	79%	28%	
	Recall	50%	31%	

Superimposed Text (ST)

- Text imposed in video signal for better understanding of its content
- Differs from scene text (billboards, text on vehicles...)
- Process
  - detection of superimposed text - regions
  - refinement of detected text
  - recognition

(ST) Detection

- Superimposed text has certain spatial properties
  - specific width/height
  - duration of appearance
- There properties are of course domain specific, and will have different position, font, etc.
- Text region
  - same text on the same position in image over several frames, in other words
  - horizontal rectangular structure of clustered sharp edges
  - use horizontal differential filtering

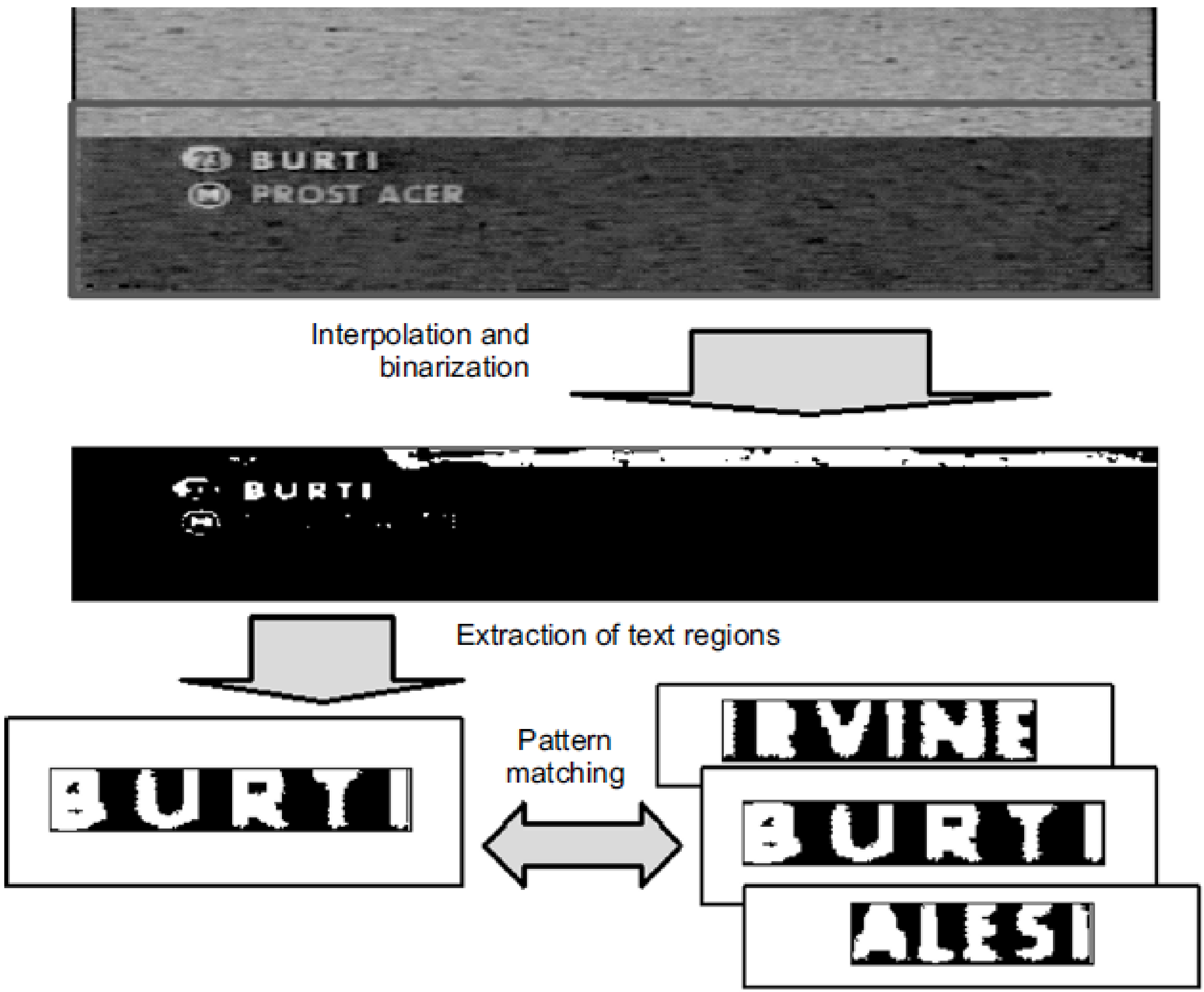
(ST) Detection

- Exploiting domain of video signal
  - in Formula1 races these text regions are in the bottom of the image and shaded (season 2000)
  - check color features of bottom pixels
- Further
  - not many words appear in text that are displayed in the same font

(ST) Refinement

- Binarization on text region to make text character stand out more based on intesities
- Then filter and interpolate
  - clean, clear and big characters
- Use word recognition phase, insted of character recognition - faster
- Words are defined as sequence of characters close to each other (pixel distance)

Example of ST extraction



source: Blanken

(ST) Recognition

- Based on pattern recognition
  1. extract reference patterns for each word (name of driver, team, etc)
  2. split words into categories having different word length
  3. matching - use pixel difference metric

$$PD = \sum_{(x,y)} I_{ref}(x,y)\, I_{extr}(x,y)$$

- Select a pattern with largest pixel difference and above specified threshold

Integrated querying

- Combining highlight detection with DBNs and pattern matching of superimposed text, possible queries are:
  - 'Driver A takes first position'
  - 'Driver B flying out on 10th lap'
  - 'Driver C in pit-stop'

- The results for these queires are obtained as highlights where those video sequences are marked having certain imposed text in them

Summary

- Automatic derivation of high level video content based on raw video data using (Dynamic) Bayesian networks
- Experiments on Formula1 races with audio, video and superimposed text
- Influence of networks' structure and temporal dependencies for DBN's
- Use of superimposed text llows powerfull queries for extraction of highlights in video signal

References

- Blanken et al., *Multimedia Retrieval,* Springer, 2007