Real-Time Extraction of Video Objects:
Algorithms and applications

Aishy Amer

Concordia UNIVERSITY

Montréal - Québec - Canada

amer@ece.concordia.ca

Real-Time Extraction of Video Objects:
Algorithms and applications

Aishy Amer

Concordia UNIVERSITY

Montréal - Québec - Canada

amer@ece.concordia.ca

Outline

① Introduction

---

② Extraction of video objects: a motivation

③ Our framework: an overview

④ Related work

⑤ A modular framework
➜ Video enhancement: enhanced images
➜ Video analysis: moving objects
➜ Video interpretation: high-level content

⑥ Applications

⑦ Conclusion and outlook

---

⑧ Video demonstration

Introduction

➜ Continuous evolution of Data Processing:
from numbers, text, audio, graphic, image, video..

Introduction

➜ Continuous evolution of Data Processing:
from numbers, text, audio, graphic, image, video..

➜ Video communication & processing:

- **1895** First public motion picture presentation (France)

- **1920s** First experimental TV broadcasting (NY 1927, MTL 1930)

- **1950s** Introduction of color TV

- **1980s** Digital TV Studios, Digital Signal Processing in TV receivers

- **1990s** Coding standards, DCam, DTV, DVD, Internet video..

- **2000s** Video phone? video email? interactive TV?
'smart' cameras? largely automated video surveillance?..

Introduction

➜ Continuous evolution of Data Processing:
   from numbers, text, audio, graphic, image, video..

➜ Video communication & processing:

- **1895** First public motion picture presentation (France)
- **1920s** First experimental TV broadcasting (NY 1927, MTL 1930)
- **1950s** Introduction of color TV
- **1980s** Digital TV Studios, Digital Signal Processing in TV receivers
- **1990s** Coding standards, DCam, DTV, DVD, Internet video..
- **2000s** Video phone? video email? interactive TV?
  'smart' cameras? largely automated video surveillance?..

⇒ Ever-increasing amount of *raw unstructured* video

Video processing

➜ Video processing deals with computational frameworks to

- enhance   video data
- extract   useful video information
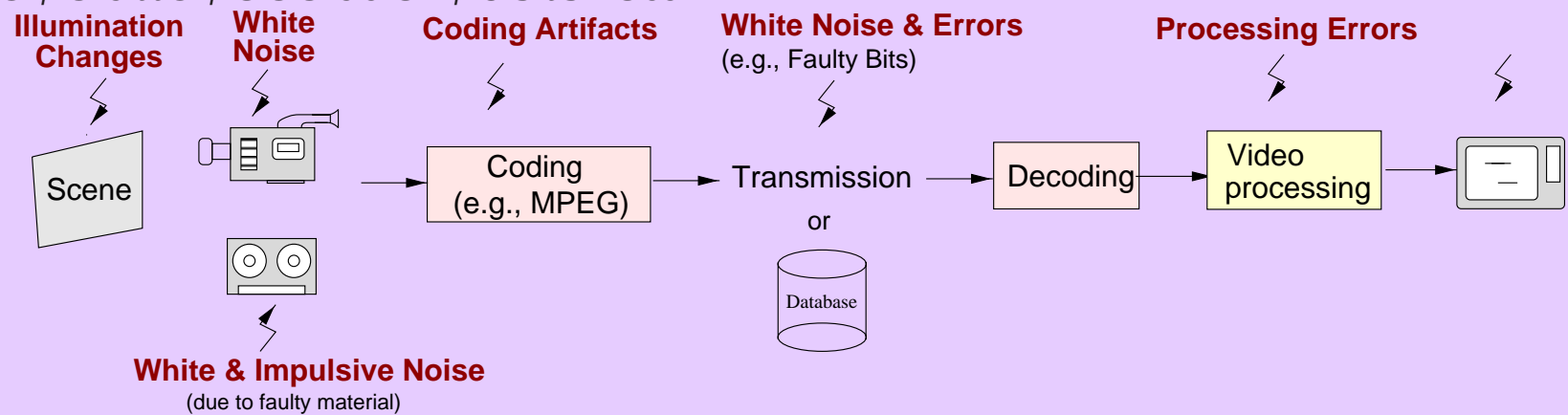- represent (structure)   raw unstructured video

⇒ Future development strongly depends on efficient processing and representation of video
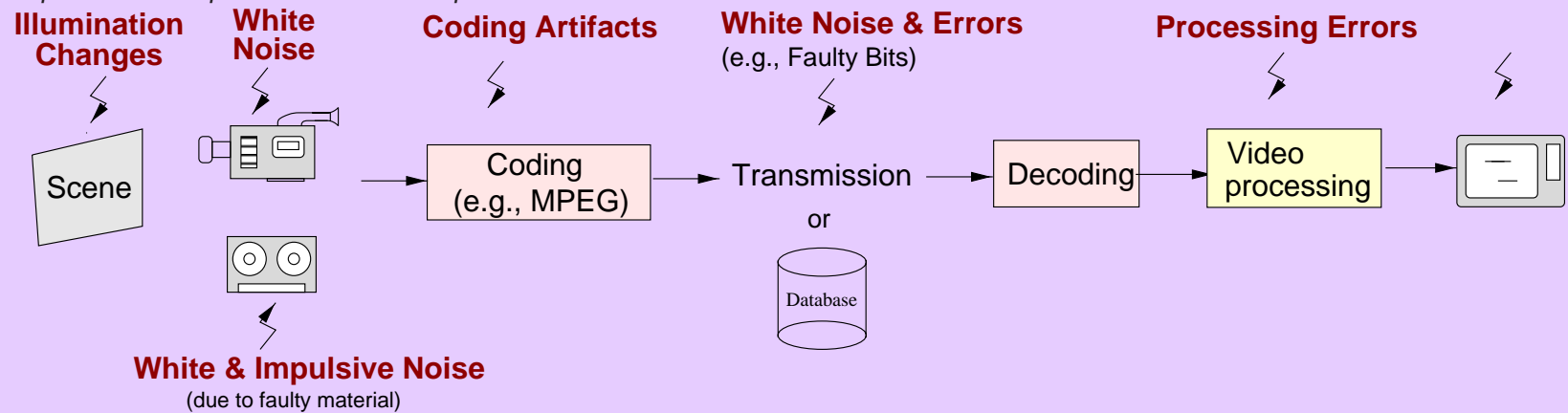
Video processing: difficulties

➜ Data processing theorem: processing destroys information

⇒ measure & extract only part of the real world

➜ Jitter, clutter, occlusion, data loss..

Illumination Changes   **White Noise**   **Coding Artifacts**   **White Noise & Errors** (e.g., Faulty Bits)   **Processing Errors**

Scene → Coding (e.g., MPEG) → Transmission or Database → Decoding → Video processing →

White & Impulsive Noise (due to faulty material)

Video processing: difficulties

→ Data processing theorem: processing destroys information

⇒ measure & extract only part of the real world

→ Jitter, clutter, occlusion, data loss..

Illumination Changes

White Noise

Coding Artifacts

White Noise & Errors (e.g., Faulty Bits)

Processing Errors

Scene → Coding (e.g., MPEG) → Transmission or Database → Decoding → Video processing →

White & Impulsive Noise (due to faulty material)

→ Real-time constraints

→ Performance evaluation

  • No objective measures to date

  • When are tasks completed? Integrated in end-to-end systems?

Video processing: research directions

➜ Continuous evolution: from pixel to block to object-based

Video processing: research directions

➜ Continuous evolution: from pixel to block to object-based

➜ Advances in technology & applications <mark>intensify</mark>

- • User-oriented design: what users need? how people find content?
- • Very low bit-rate content-based video compression
- • 'Intelligent' video representation

  - ○ Extraction of visually meaningful objects
  - ○ High-level interpretation of low-level video features

⇒ Extraction of visually meaningful content increasingly required

Outline

① Introduction √

---

② Extraction of video objects: a motivation

③ Our framework: an overview

④ Related work

⑤ A modular framework
- ➜ Video enhancement: enhanced images
- ➜ Video analysis: moving objects
- ➜ Video interpretation: high-level content

⑥ Applications

⑦ Conclusion and outlook

---

⑧ Video demonstration

Extraction of meaningful objects: Motivation

➜ Content-oriented video applications:

Transfer raw *un*structured video to structured data:

separate, select, & describe video content

Extraction of meaningful objects: Motivation

➜ Content-oriented video applications:

Transfer raw *un*structured video to structured data:

separate, select, & describe video content

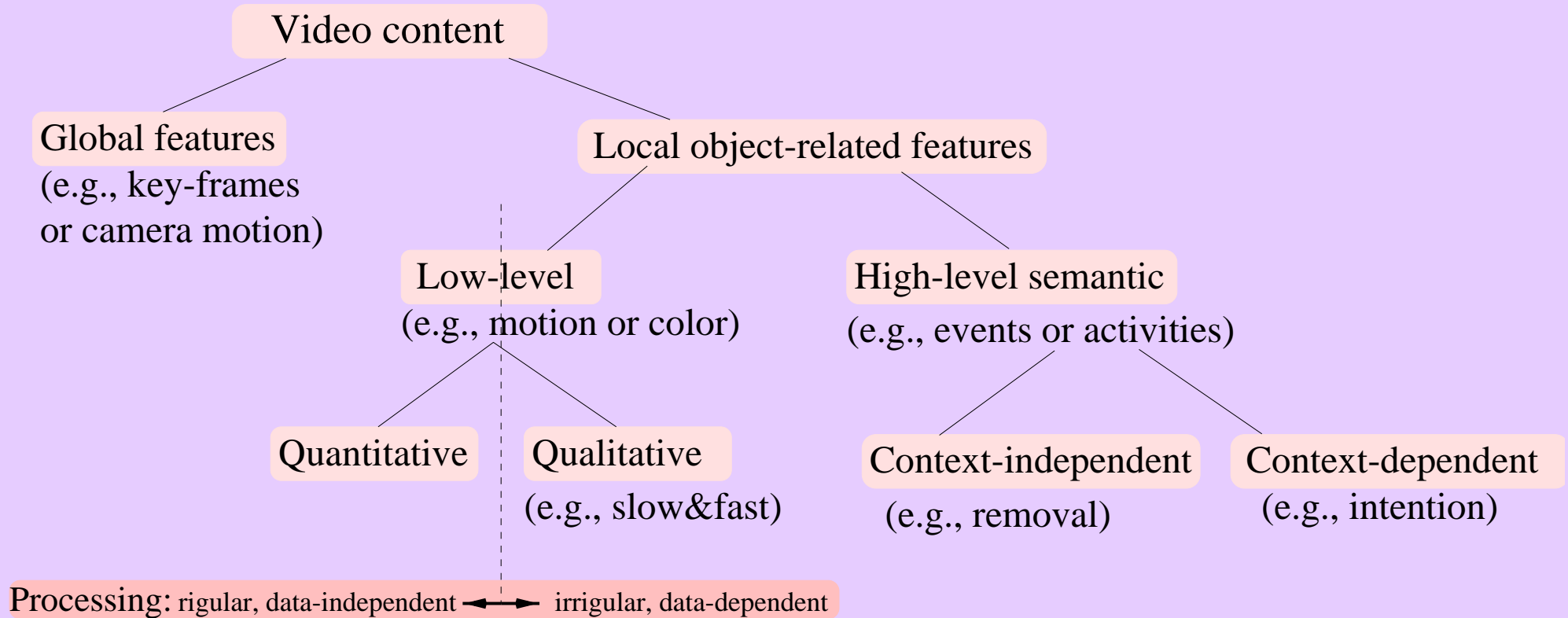➜ To date: manual content selection & description (e.g., film, surveillance)

costly , time consuming , & subjective

Motivation: Units of a video sequence

Video

Stories

Scenes

Shots

Objects & Meaning
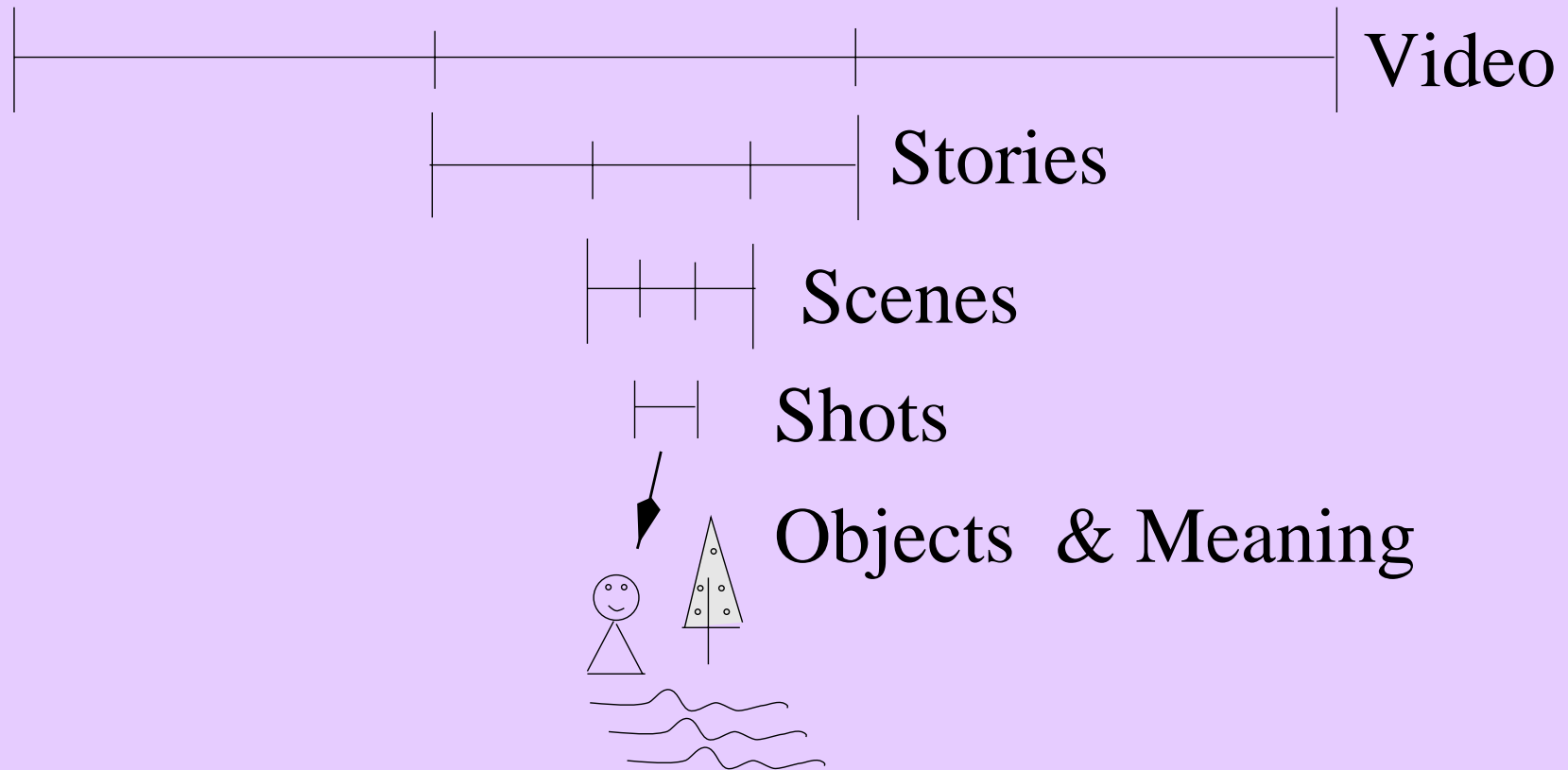
➜ Key: how to effectively describe video shots

Motivation: Shot content

Video content
- Global features (e.g., key-frames or camera motion)
- Local object-related features
  - Low-level (e.g., motion or color)
    - Quantitative
    - Qualitative (e.g., slow&fast)
  - High-level semantic (e.g., events or activities)
    - Context-independent (e.g., removal)
    - Context-dependent (e.g., intention)

Processing: rigular, data-independent ←→ irrigular, data-dependent

What information is meaningful?

Motivation: What info?

➜ (Information is in "the eye of the beholder")

➜ Video is `rich` and `complete meaning` depends on context

➜ Content `selection` depends on application

Motivation: What info?

➜ (Information is in "the eye of the beholder")

➜ Video is `rich` and `complete meaning` depends on context

➜ Content `selection` depends on application

⇒ Many on-line application: high-level object content

⇒ Wide applicability: context-independent content

Processing levels to meaningful content

Video shot ◯ ◯ ◯ ◯ ◯ ◯

Content extraction

Global features

Moving objects ← Video analysis

Meaning
(e.g., events) ← Video interpretation
- context independent -

Object behavior ← Video understanding
- context dependent -
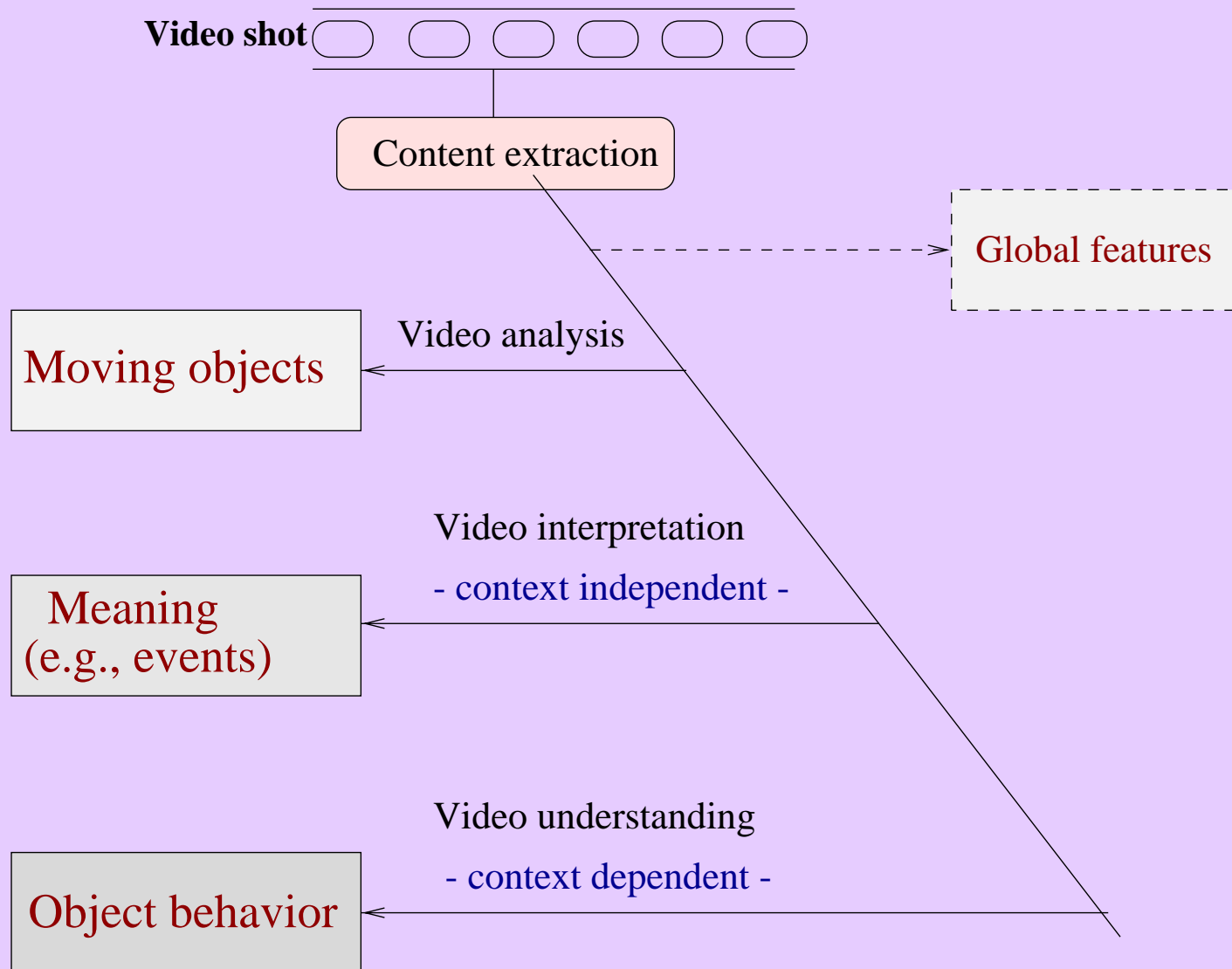
Processing levels to meaningful content

➜ Much work in enhancing low-level methods:

- Lower-level features are more difficult to extract
- Despite extensive research: no accurate object segmentation

➜ Little work on context-independent methods

Our framework: an overview

➜ End-to-end automated real-time stable systems
that extract
moving objects and their low-level & high-level features
independently of the context of the input video

Our framework: an overview

➜ End-to-end automated real-time stable systems that extract moving objects and their low-level & high-level features independently of the context of the input video

➜ Motivation:

⇒ Wide applicability: fast & context-independent

⇒ On-line applications: stable extraction foregoes precision

⇒ People focus on and memorize

   ○ "who" = moving objects

   ○ "what" = their activities (e.g., events)

   ○ ("where" = location & "when" = time)

Related work

$\Rightarrow$ Little work on context-independent or end-to-end systems

➜ AVI: Courtney, 1997 (Texas instruments)

- Basic events: *stop*, *deposit*, *removal*..
- Simplistic: indoor, no occlusion, *stop* = same position for two images
- Noise sensitive: motion detection, tracking

➜ Stringa, Regazzoni, 1998 (Uni. Genova, Italy)

- Classification: *abandoned object*, *person*
- Simple environments: indoor, no occlusion

➜ Haering et al., 1999 (Uni. Central Florida)

- Off-line processing
- Domain-specific events and classification

➜ W^4: Haritaoglu, Harwood, Davis, 2000 (Uni. Maryland, MD)

- Limited events: person carrying an object
- Restrictions on movements: upright, little occlusion

Outline

① Introduction √

---

② Extraction of video objects: a motivation √

③ Our framework: an overview √

④ Related work √

⑤ A modular framework
  ➜ Video enhancement: enhanced images
  ➜ Video analysis: moving objects
  ➜ Video interpretation: high-level content

⑥ Applications

⑦ Conclusion and outlook

---

⑧ Video demonstration

Our Framework

➜ Assumptions:

- We can measure & extract only part of the real world data
- Meaningful content is related to moving objects
- White noise signal
- Fixed (or moving) camera

Our Framework

→ Assumptions:

- We can measure & extract only part of the real world data
- Meaningful content is related to moving objects
- White noise signal
- Fixed (or moving) camera

→ Multi-layered modular framework:
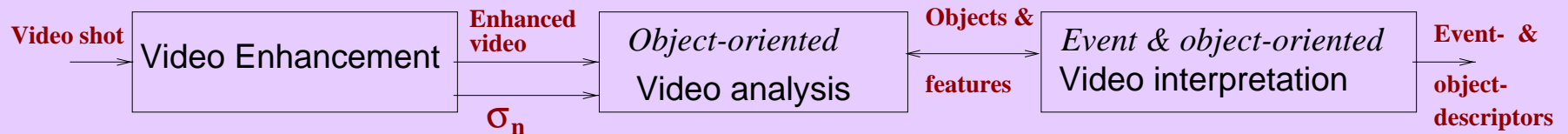
→ Layer interaction aims at balancing

- missing information can prevent complete information
- additional information may mask relevant information

Video Enhancement

Noisy image → **Noise estimation** σ_n → **Noise reduction** → Enhanced image
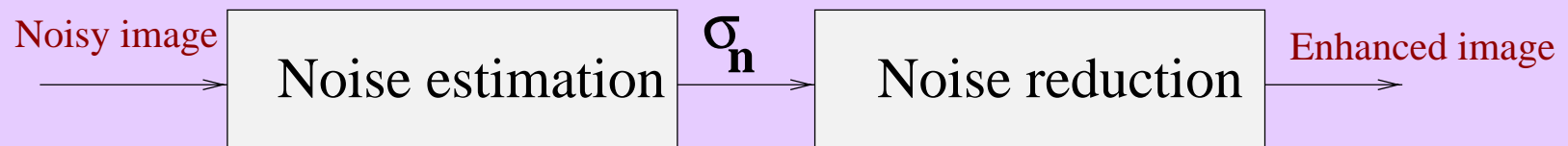
➜ Noise model: white Gaussian

observed image = ideal image + noise

➜ Noise sources:

- Camera (analog & digital)
- Transmission channel (analog)
- Storage device (analog)

Video Enhancement

➜ Noise estimation:

- Find Intensity-homogeneous blocks
  $\Rightarrow$ rejects blocks with line structure using new our masks to detect lines

➜ Noie reduction:

- Preserve line structure & noise adaptive

- Challenge: reliable line & noise estimation

- Solution: multiple masks

Our Framework
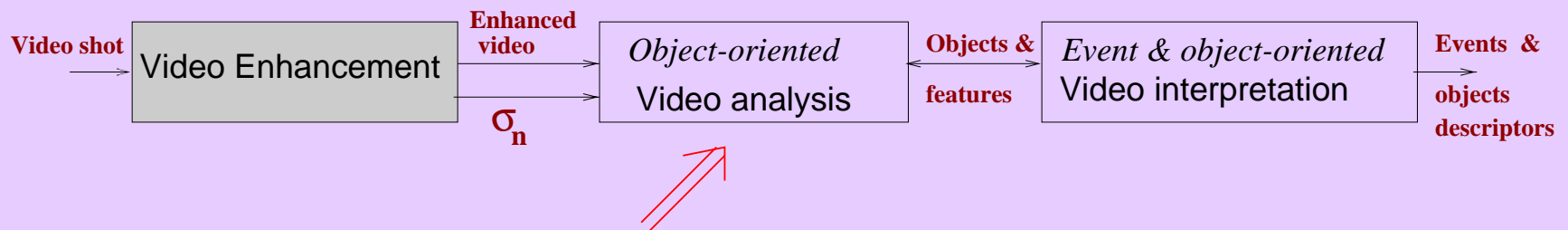
Video shot → **Video Enhancement** → Enhanced video → σ_n → *Object-oriented* Video analysis → Objects & features → *Event & object-oriented* Video interpretation → Events & objects descriptors

Our Video Analysis

➜ **Goal:** extract moving objects and low- & mid-level features

➜ **Trade-off:** generality - stability - real time

➜ **Related work:** complex to be practical or controlled situations

➜ **Focus:** stability that foregoes precision (& complex operations)

Our Video Analysis

➜ **Goal:** extract moving objects and low- & mid-level features

➜ **Trade-off:** generality - stability - real time

➜ **Related work:** complex to be practical or controlled situations

➜ **Focus:** stability that foregoes precision (& complex operations)

R(n)   I(n-1)   I(n)

Moving object segmentation

⟶ Pixels to objects

O(n-1)   O(n)

Feature extraction & Motion estimation

⟶ Objects to features

Low-level object features

Multi-feature object tracking

⟶ Objects to video objects

Mid-level object features

Goal: separate objects & spatial features

Flowchart:

input images → binarization (motion detection) [with z^{-1} feedback] → Binary images → morphological edge detection → Edges → contour analysis [with z^{-1} feedback] → Contours → object labeling (contour filling) → Objects
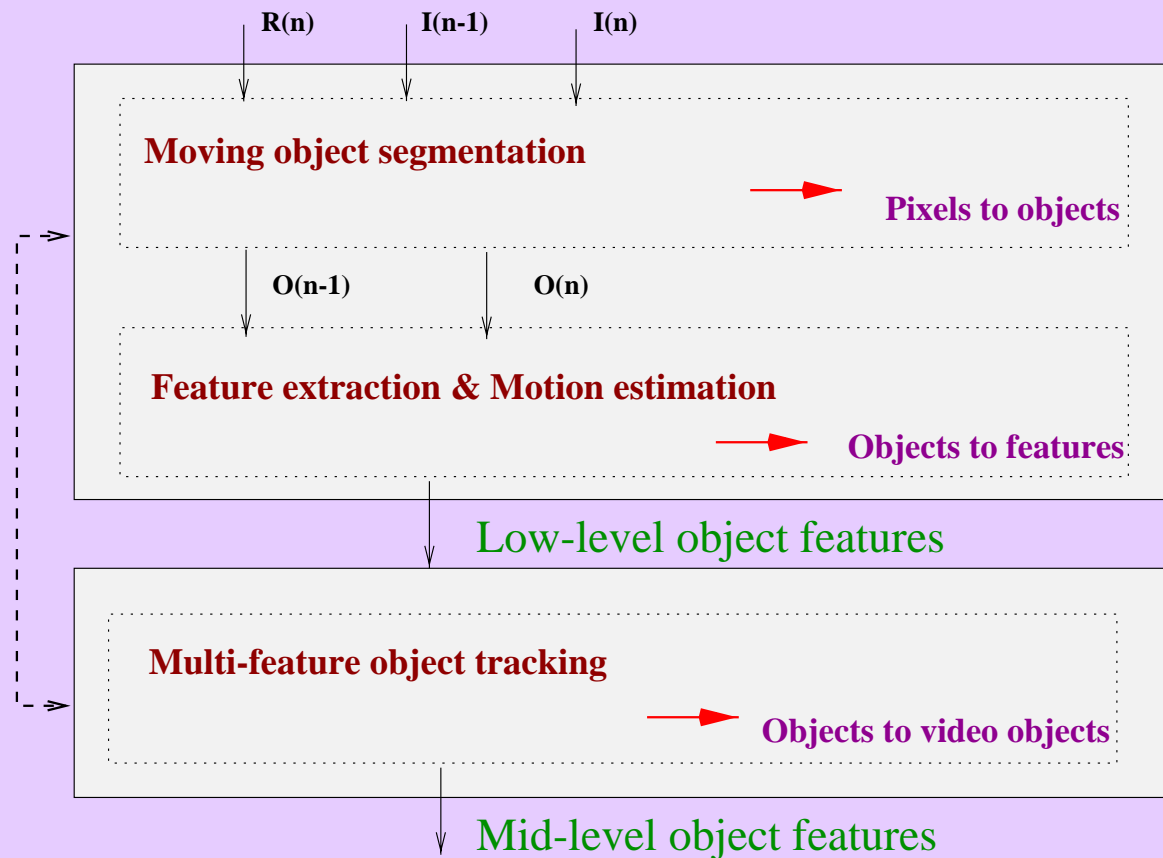
Video Analysis: Object Segmentation

Goal: separate objects & spatial features
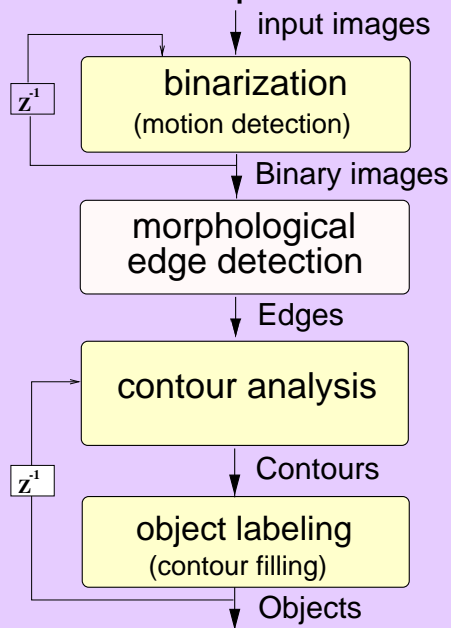
Video Analysis: Object Segmentation

Motion detection:

| Backgr. | Org. | Abs. | Ave. +Max | Thresh. |

Block diagram:

$R(n)$

$I(n)$ $\xrightarrow{+}$ $\oplus$ $\xrightarrow{-}$ **absolute value** → **spatial averaging filter** → **spatial MAX filter** $\xrightarrow{D(n)}$ **Spatio-temporal adaptation** $\xrightarrow{B(n)}$ $\xrightarrow{T(n)}$

Noise estimation σ_n

$T(n-T_z)$

z^{-1}
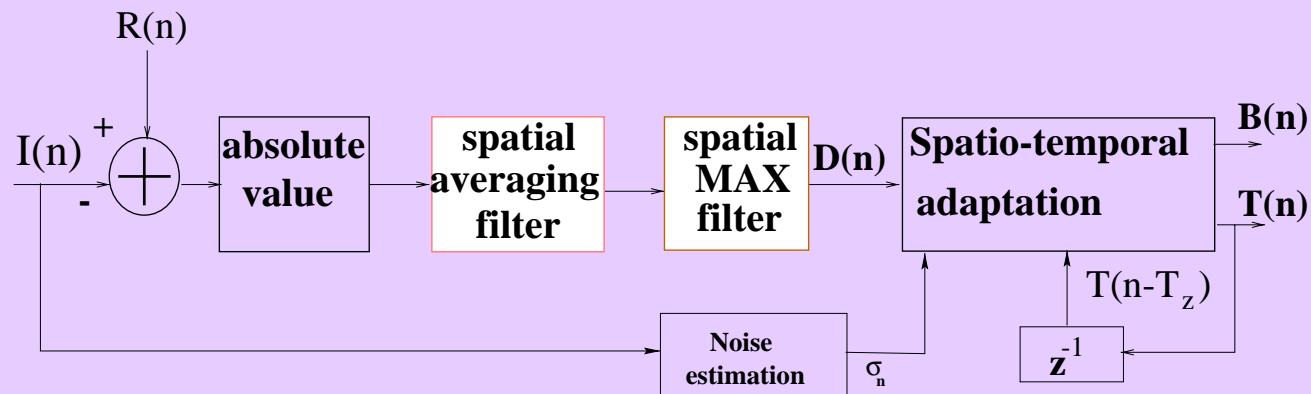
➜ Averaging: smoothes & reduces noise

➜ Maximum filter: stabilizes boundaries & reduces grain noise

Motion detection: spatio-temporal adaptation

1. Adaptation to noise (low sensitivity to small σ_n):
$$T_n = T_g + c \cdot \sigma_n^2, \quad c < 1$$

2. Quantization: compensates for illumination changes & causes spatio-temporal stability

$$T_q = \begin{cases} T_{\text{min}} & : \quad T_n \leq T_{\text{min}} \\ T_{\text{mid}} & : \quad T_{\text{min}} < T_n \leq T_{\text{mid}} \\ T_{\text{max}} & : \quad \text{otherwise} \end{cases}$$

3. Temporal integration: causes temporal stability

$$T(n) = \begin{cases} T_{\text{min}} & : \quad T_q \leq T_{\text{min}} \\ T(n-1) & : \quad T_q < T(n-1) \\ T_q & : \quad \text{otherwise} \end{cases}$$

$\Rightarrow$ Stable with respect to noise and illumination variation

Video Analysis: Object Segmentation

Motion detection: a comparison

➜ State-of-the-art: Aach et. al, 1993 & Ziliani et. al, 1999:
A threshold by a statistical test of hypothesis using a noise model

Org.   Prop.   Ref.

⇒ higher stability with illumination change and noise

⇒ lower computational cost

Subjective comparison

➜ Simulations with **same automatically adjusted parameter**

➜ Stability foregoes accurate boundaries

$I(37)$

| Our | MPEG-2 | Noisy | COST-AM |

Subjective comparison

➜ Simulations with **same automatically adjusted parameter**

➜ Stability foregoes accurate boundaries

$I(37)$

Our   MPEG-2   Noisy   COST-AM

$I(197)$
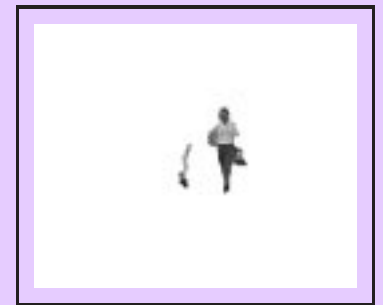
⇒ Stable in corrupted images and throughout a video

Video Analysis: Object Segmentation

Objective comparison: reference to estimated object masks
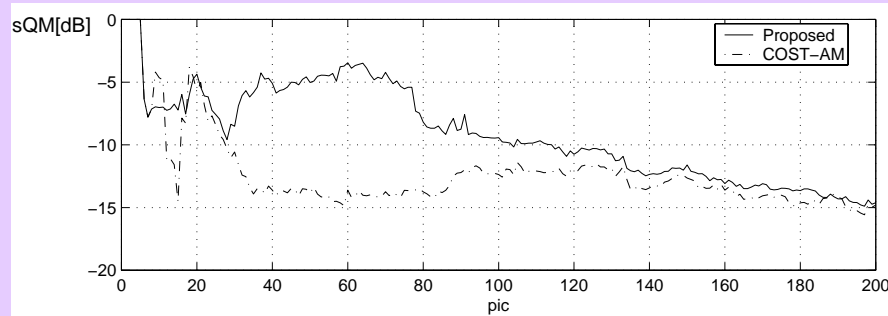
Spatial accuracy
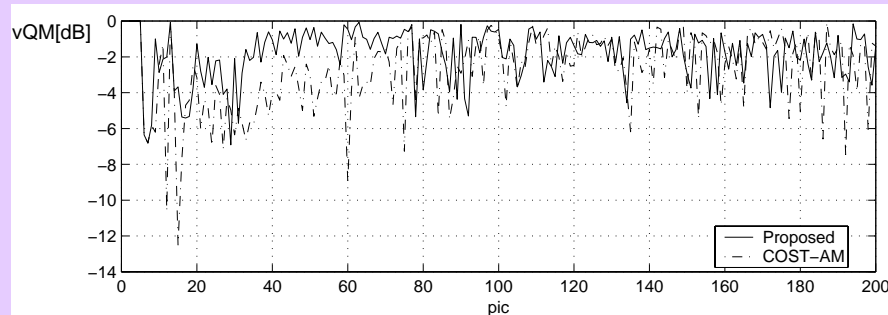
Temporal stability
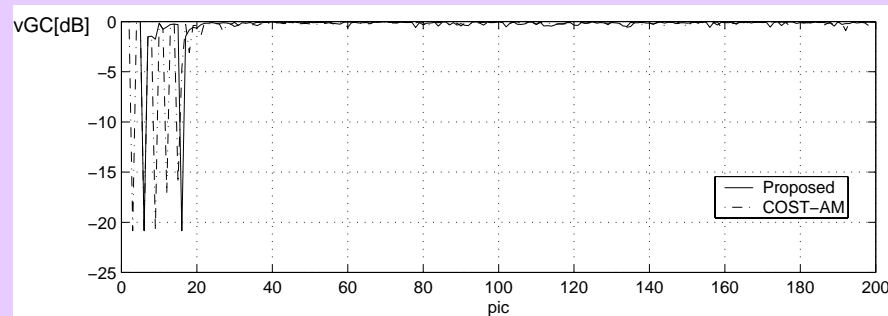
Temporal coherency

$\Rightarrow$ Our method better with respect to all three criteria

Goal: temporal object features throughout the video
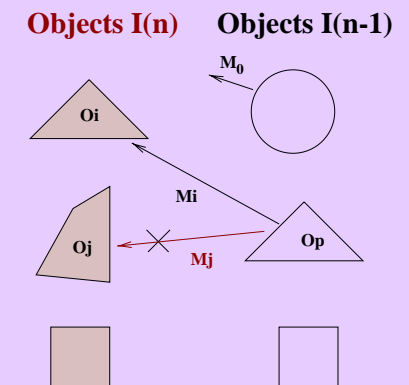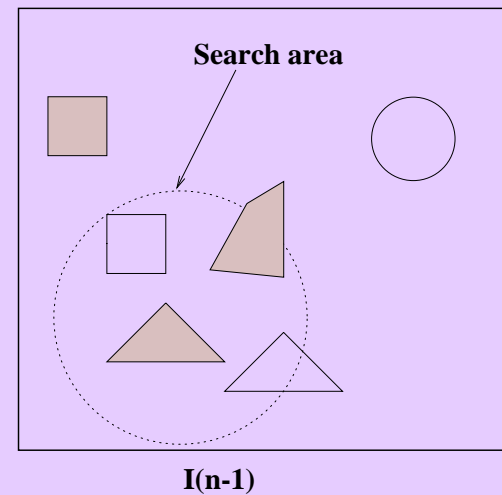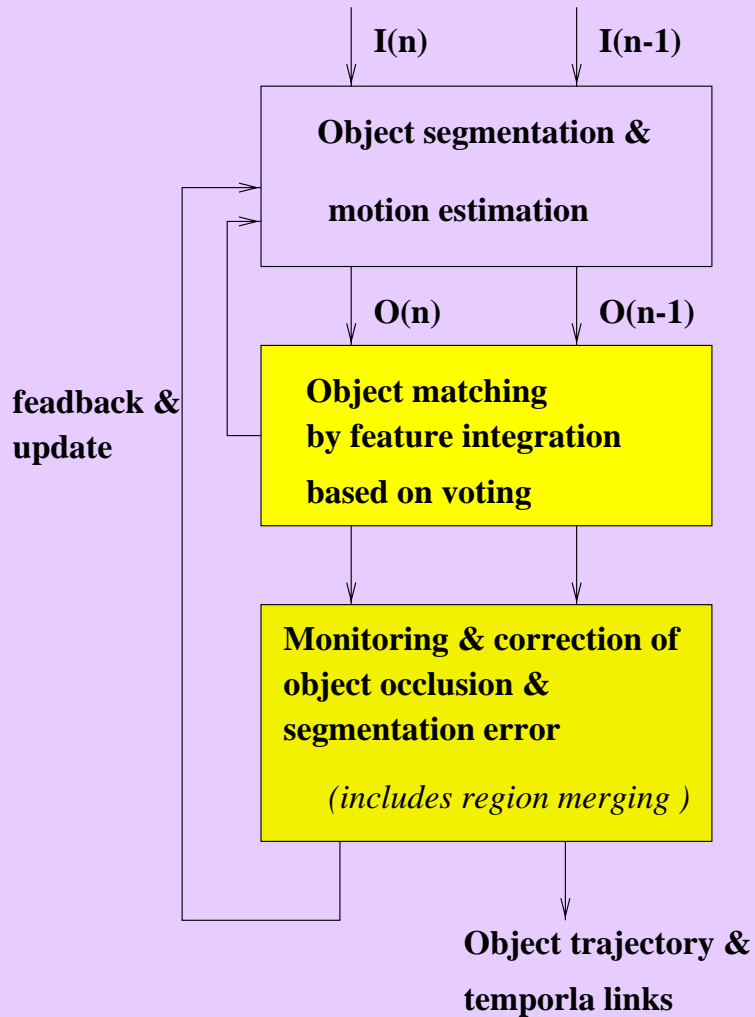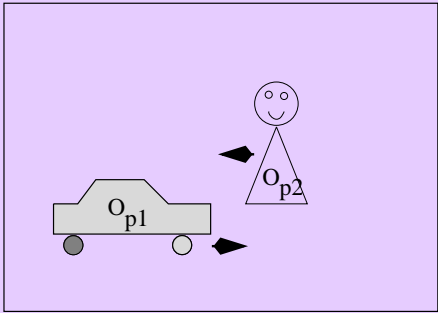
Video Analysis: Object tracking

➜ Two step feature voting : object voting & match voting

- Features: distance, size, shape, motion (direction & magnitude)
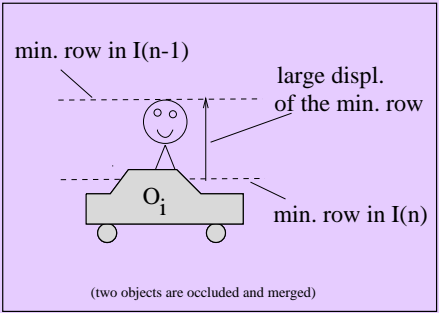
Video Analysis: Object tracking

➜ **Two step feature voting** : object voting & match voting

- Features: distance, size, shape, motion (direction & magnitude)

➜ Example of **object voting** :

- Shape features for $O_p \in I(n-1)$ and $O_i \in I(n)$: compactness c_p (c_i), irregularity r_p (r_i), extent ratio: e_p (e_i)
- $d_{e_i} = |e_p - e_i|$, $d_{c_i} = |c_p - c_i|$, and $d_{r_i} = |r_p - r_i|$

$$s_{++} \quad : \quad d_{e_i} \leq t_s \quad \vee \quad d_{c_i} \leq t_s \quad \vee \; d_{r_i} \leq t_s$$

$$d_{++} \quad : \quad d_{e_i} > t_s \quad \vee \quad d_{c_i} > t_s \quad \vee \; d_{r_i} > t_s$$

no vote : otherwise

➜ Confidence measure: $\zeta = \frac{s}{d}$ degree of confidence of a correspondence M_i

➜ If an **object has two matches** , a **match voting** is performed

➜ Object occlusion:

I(n-1)

min. row in I(n-1)

large displ.
of the min. row

O_i

min. row in I(n)

(two objects are occluded and merged)

I(n)

→ Object occlusion:

I(n-1)

min. row in I(n-1)

large displ.
of the min. row

O_i

min. row in I(n)

(two objects are occluded and merged)

I(n)

→ Object splitting:

I(n)

O_{i1} O_{i2}

(object is split into 2 regions)

O_{p1}

I(n-1)

01[age45]: Moves   60
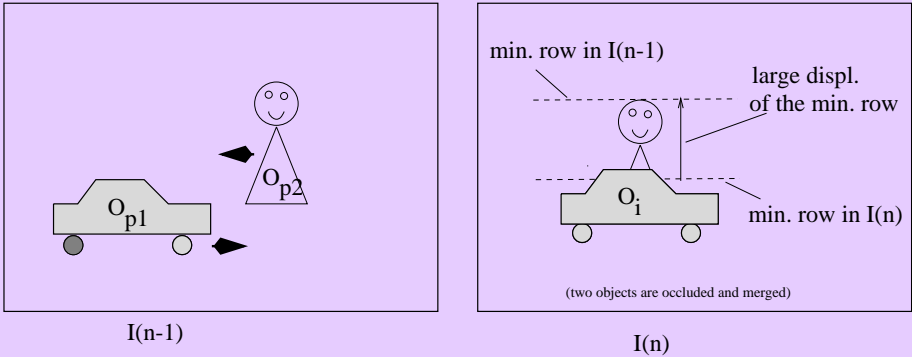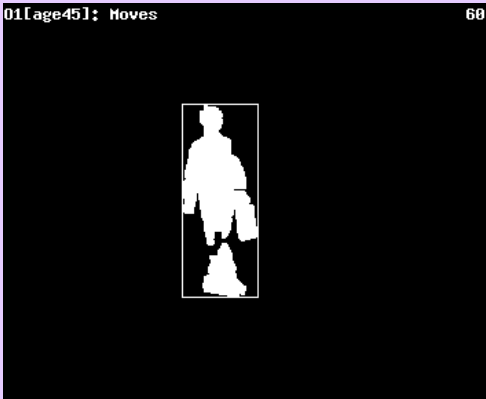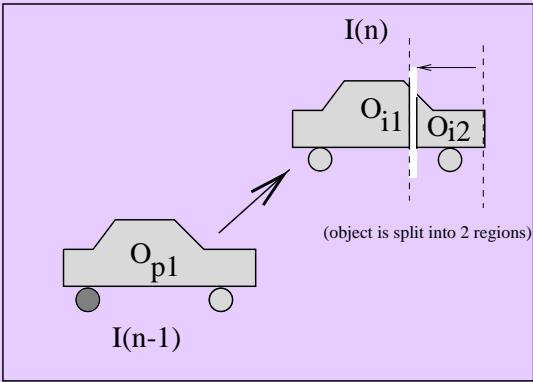
Video Analysis: Parameter adaptation

➜ Parameters handle variations

- due to feature estimation errors or
- due to characteristics (e.g., object size, frame rate, or frame size) of the input video.

➜ Parameters handle variations

- due to feature estimation errors or

- due to characteristics (e.g., object size, frame rate, or frame size) of the input video.
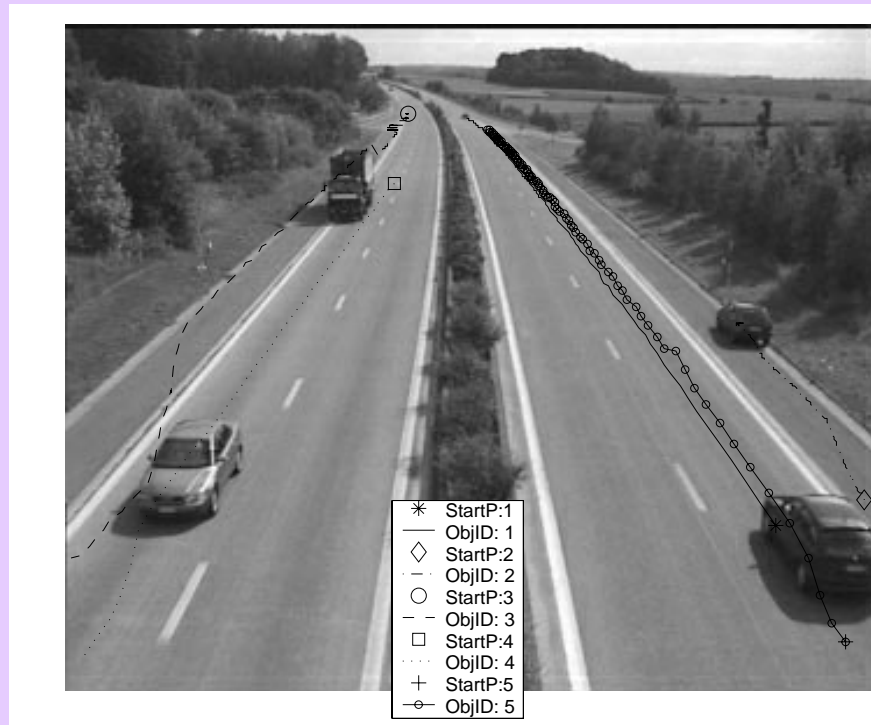
➜ Plausibility rules:

- With small objects (or frames), an error is significant

$$
t = \begin{cases}
t_{\min} & : \quad A \leq A_{\min} \\
t_{\min} + \frac{(t_{\max} - t_{\min})}{A_{\max}} A + \frac{t_{\min}}{F_{\max}} F : & A_{\min} < A \leq A_{\max} \\
t_{\max} & : \quad A > A_{\max}
\end{cases}
$$

- Distances and displacements thresholds:
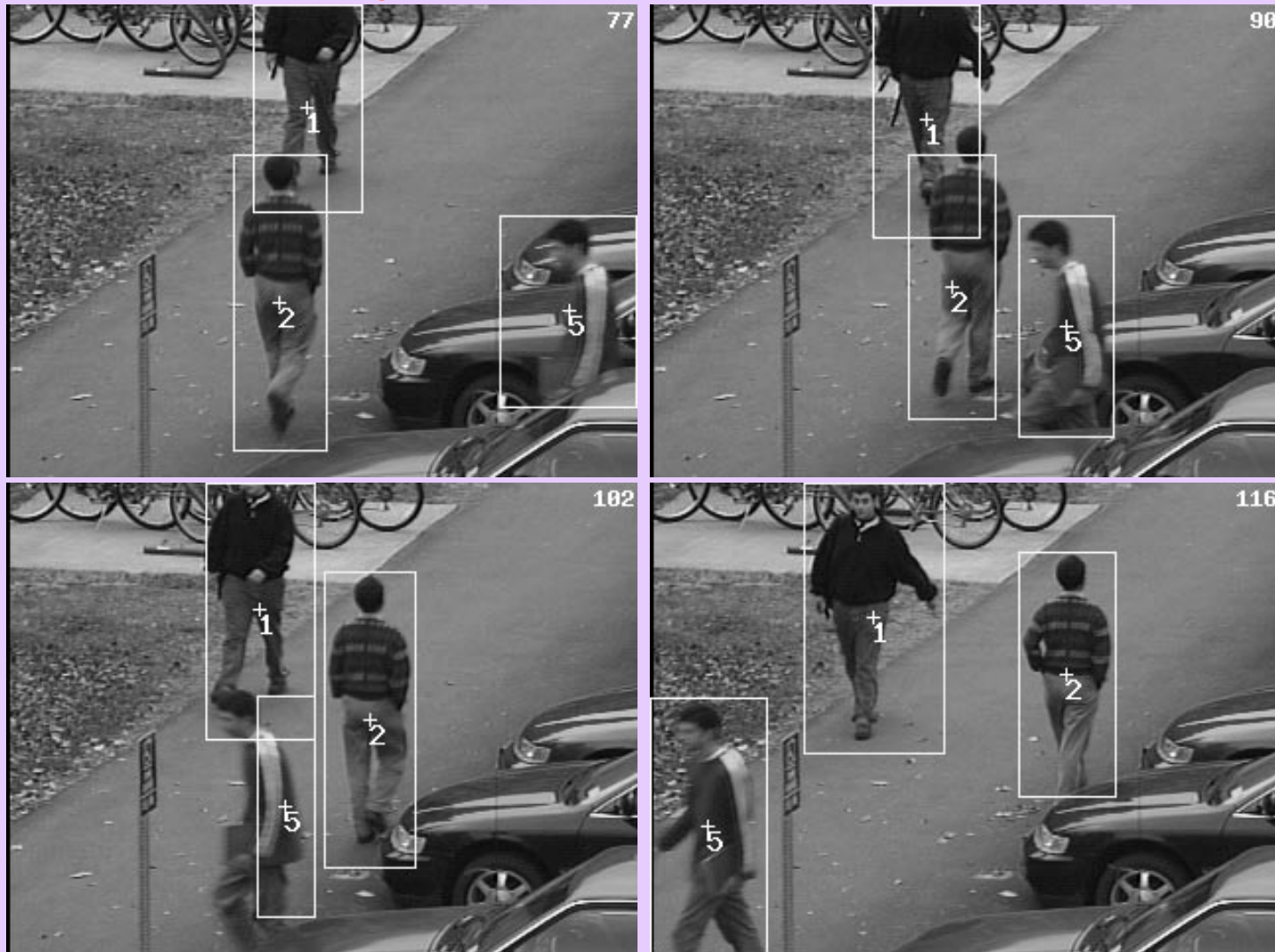 adaptive to the frame rate (lower thresholds for higher rates)

Video Analysis: Object tracking

Results: object trajectories (or paths)

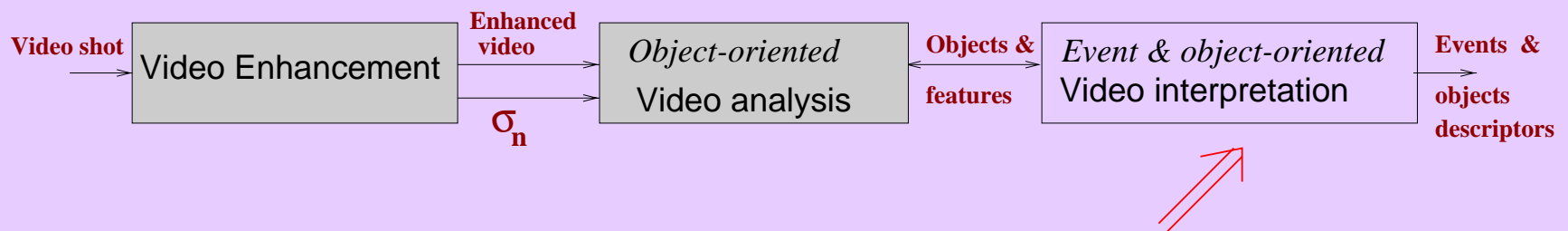Results: multi-object occlusion

Our Framework

Video shot → **Video Enhancement**

Enhanced video → σ_n → *Object-oriented* **Video analysis**

Objects & features → *Event & object-oriented* **Video interpretation** → Events & objects descriptors

Event-oriented Video Interpretation

➜ **Goal:** to specify meaning related to object movement

➜ What meaning? ⇒ context-independent & fix

- *Event*: a particular behavior of a finite set of objects

- *Deposit*: A fixed meaning = an object is added

- Variable meaning: context-dependent (e.g., where, intention)

Event-oriented Video Interpretation

➜ **Goal:** to specify meaning related to object movement

➜ What meaning? ⇒ context-independent & fix

- *Event*: a particular behavior of a finite set of objects
- *Deposit*: A fixed meaning = an object is added
- Variable meaning: context-dependent (e.g., where, intention)

➜ Event detection: analysis of object motion & interrelations

➜ Detected events:

- Enter, Appear, Exit (leave), Disappear, Move, Stop
- Occlude/occluded, Remove/removed, Deposit/deposited
- Abnormal movements: *stays for long, moves too fast*
- Dominant object: event, largest size (speed or age)

Video Interpretation - event detection

Approximate but efficient world models:

➜ Input for a O_i:

- *Age* - g_i the time interval when the object is tracked
- *Size* - (initial, average, and current)
- *Shape* - (initial, average, and current)
- *Location* - (initial and current)
- *Motion* - (initial, average, and current)
- *Corresponding object* - M_i a temporal link

Approximate but efficient world models:

➜ Input for a O_i:

- *Age* - g_i the time interval when the object is tracked
- *Size* - (initial, average, and current)
- *Shape* - (initial, average, and current)
- *Location* - (initial and current)
- *Motion* - (initial, average, and current)
- *Corresponding object* - M_i a temporal link

➜ O_i *moves* at time instant n if

- $O_i \in I(n)$,
- $M_i : O_p \to O_i$ where $O_p \in I(n-1)$, and
- Median of the motion of O_i in the previous k images is $> t_m$

➜ Deposit/deposited:
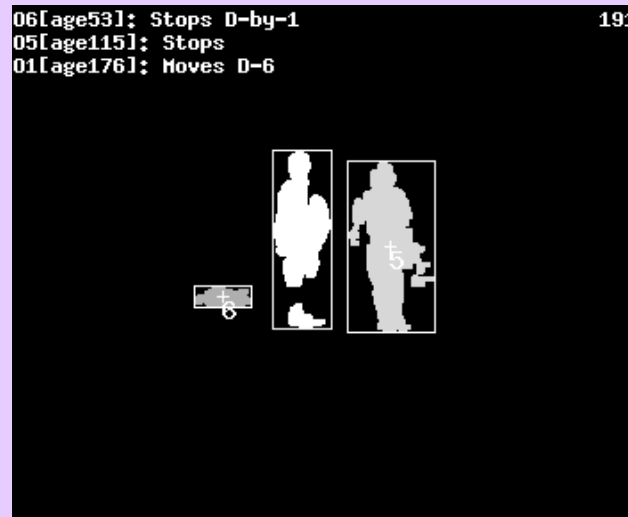O_i *deposits* O_j if:

- $O_p \in I(n-1)$, $O_i, O_j \in I(n)$, and $M_i : O_p \to O_i$.

- $g_i > t_g$,

- $O_j \notin I(n-1)$, i.e., zero match $M_0 :\dashv O_j$,

- $\frac{A_j}{A_i} < t_a$, $\quad t_a < 1$,

- $A_i + A_j \simeq A_p \quad \wedge \quad [(H_i + H_j \simeq H_p) \vee (W_i + W_j \simeq W_p)]$

- O_j is close to a side, s, of the MBB of O_i

- H_i or W_i changes between $I(n-1)$ and $I(n)$ at s.

Video Interpretation - event detection

Examples of reducing false alarms

➜ Differentiate between deposit and segmentation error

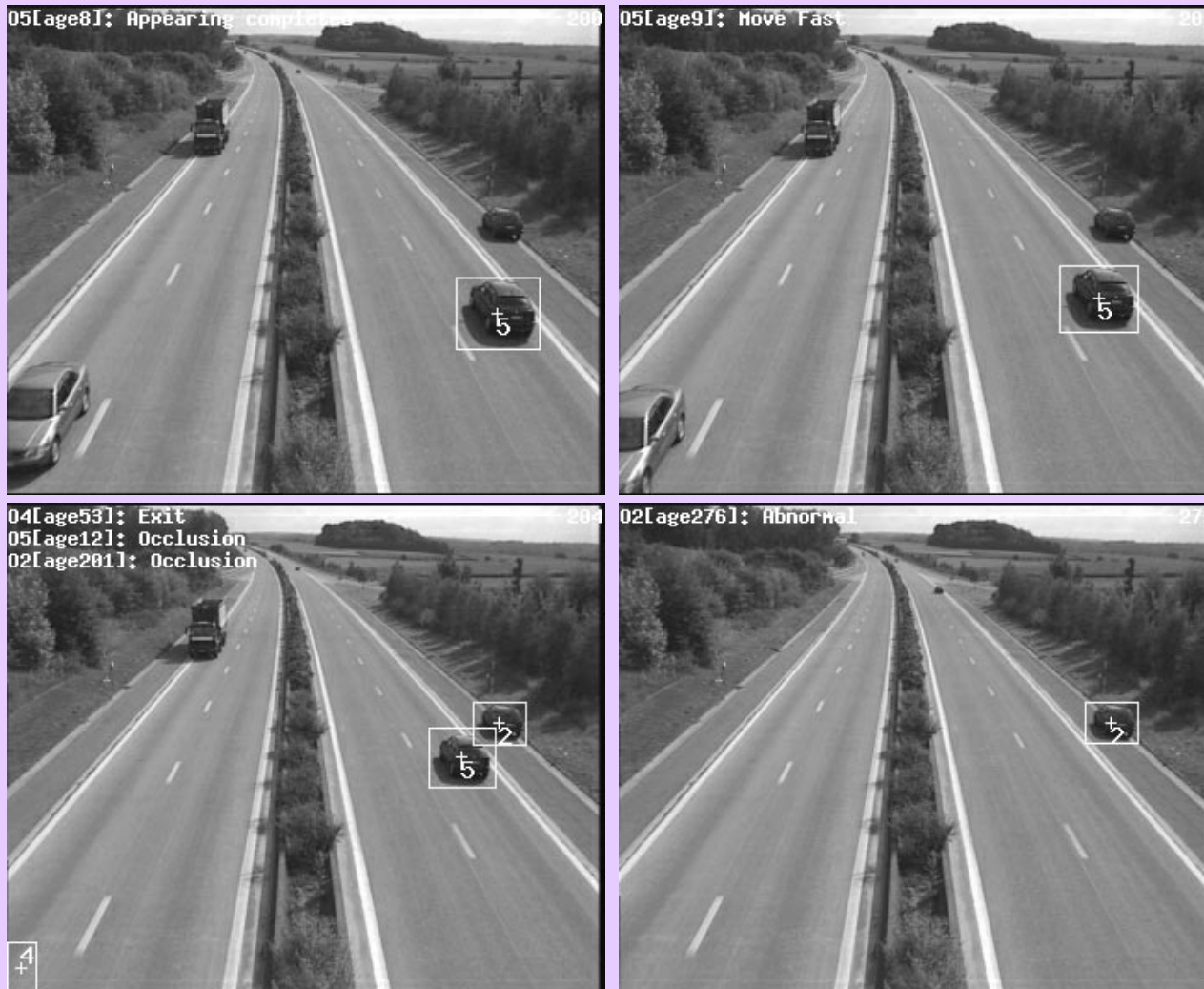
(e.g., object split)

➜ *Deposit* is declared if:

- the deposited object remains for some time in the scene

- the distance from the depositor increases

➜ Differentiates between stopping and deposited objects

➜ A moving object is declared if its age larger than a threshold

Event-based summary of the 'Highway' sequence (300 images)

Our Framework
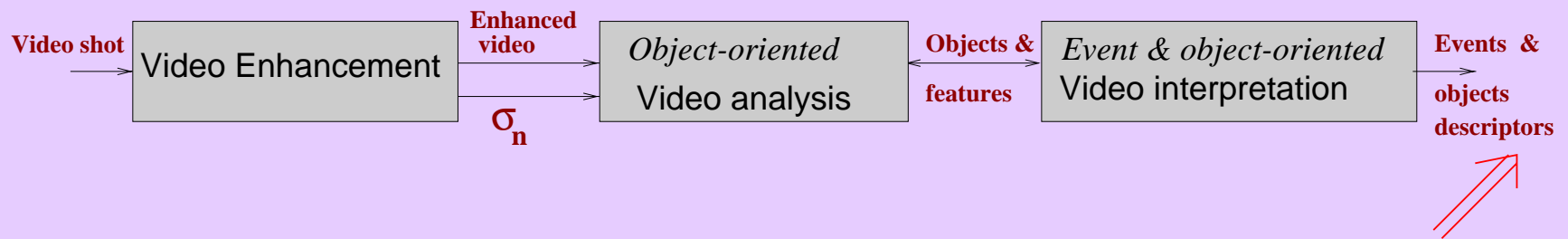
Video shot → **Video Enhancement** →(Enhanced video)→ *Object-oriented* Video analysis →(Objects & features)→ *Event & object-oriented* Video interpretation → Events & objects descriptors

σ_n

Applications

➜ Object motion analysis:

- Detection of natural or biological events (e.g., hunts in Wildlife)
- 'Smart' environments for human interactions
- Athletic activities in sports
- Dance performance

Applications

➜ Object motion analysis:

- Detection of natural or biological events (e.g., hunts in Wildlife)
- 'Smart' environments for human interactions
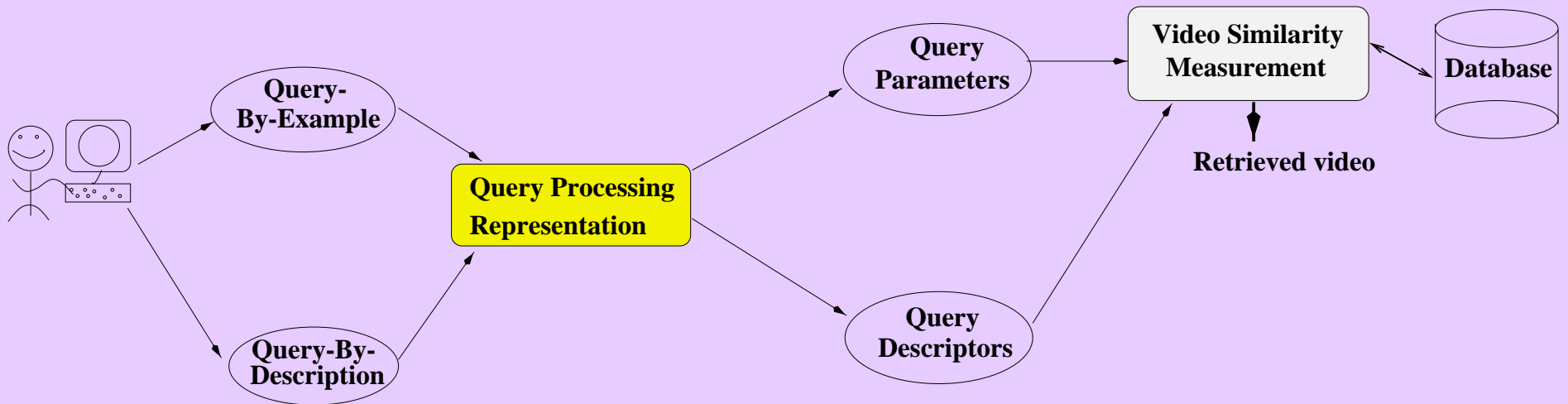- Athletic activities in sports
- Dance performance

➜ Entertainment & telecommunications:

- Dynamic video summarization
- Video editing & reproduction
- Browsing of video on Internet
- 'Smart' video devices (e.g., 'smart' cameras)

- Video surveillance
- Video compositing
- Visual fine arts

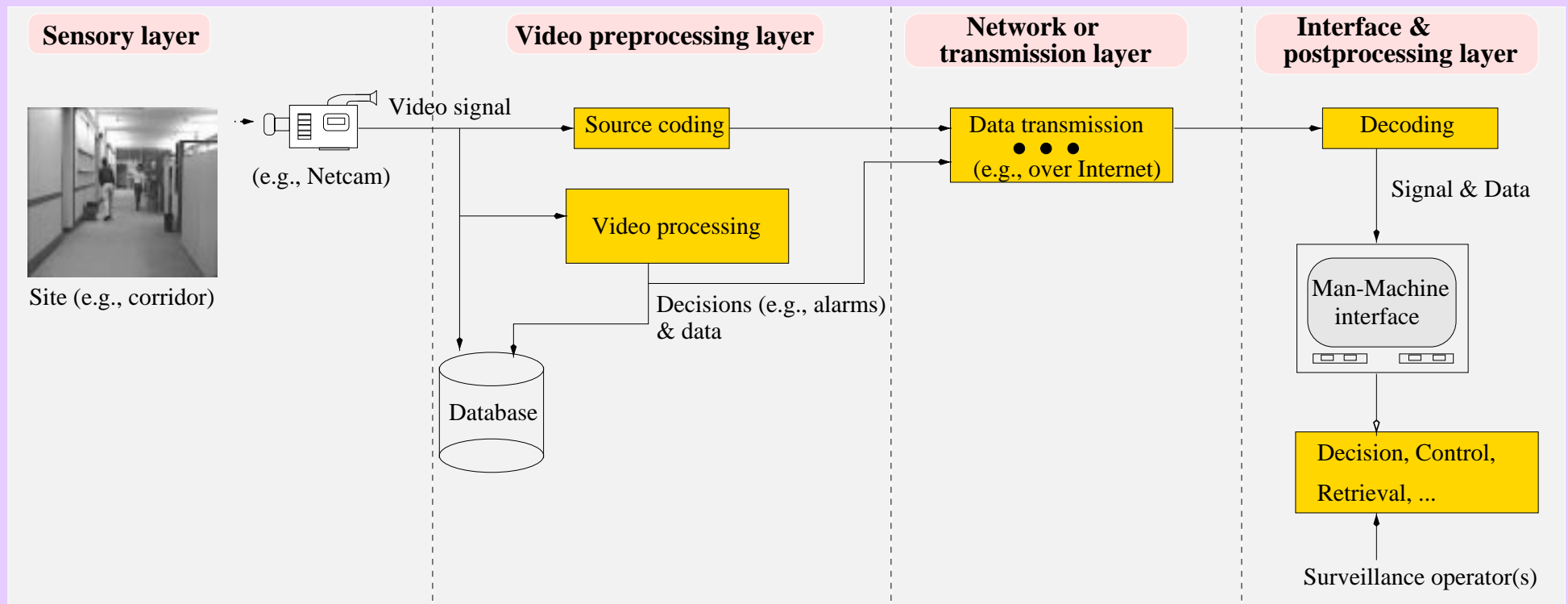# Applications: Video databases

➜ On-line retrieval based on events & objects.

Applications: Video surveillance

➜ Automated monitoring of object activities

➜ Goal: reduce the amount of information presented to inspector(s)

Applications: Visual fine arts

➜ Art-led technology research effort

➜ Explore new forms of visual-based expression

➜ Expand the range of expressive tools to creative professionals

➜ Develop new techniques for media processing and analysis
   e.g., to integrate video objects and recognized dynamic speech

➜ Goal:
   reduce or facilitate the amount of information presented to viewers

   The following image is courtesy of the NextText project

Figure 1.
A sketch showing Intralocuter. The speaker on the right is talking, and her words are 'soaking' into the listener on the left and settling down to the bottom of the silhouette. In the actual work, the typography will undergo additional visual manipulation depending on the affect of the speaker's speech.

Conclusion

➪ A modular framework to extract meaningful objects

➪ A practical solution for different applications

➪ Based on efficient and stable methods

➪ Extensive experiments on real video: occlusion, artifacts..

Conclusion

⇨ A modular framework to extract meaningful objects

⇨ A practical solution for different applications

⇨ Based on efficient and stable methods

⇨ Extensive experiments on real video: occlusion, artifacts..

⇨ Stability is due to

- Adaptation to artifacts and noise
- Combination of spatial and temporal analysis
- Error compensation where higher level information is given

⇒ Context-independency possible even with low-cost methods

⇨ Integration into real applications:
Video surveillance & Visual fine arts

Conclusion

Video shot

Video enhancement
Enhanced video

Noise estimation & reduction

Enhanced video

Image stabilization
Stable video

Global feature extraction

Background update

Global-motion compensation

$R(n)$ **Stable video** σ_n

Video analysis
Pixels to video objects

Motion-based object segmentation
=> Pixels to objects

Object-based motion estimation

Voting-based object tracking
=> Objects to video objects

Spatio-temporal object descriptors **Global shot descriptors**

Video interpretation
Video objects to events

Analysis & interpretation of low-level descriptors

Event detection & classification

Results (Events & Objects) **Requests**

Object & Event-based application
e.g., event-based decision-making

Outlook - Anticipated extensions

⇨ Integration into real applications

⇨ Global motion compensation

⇨ Global illumination changes

⇨ Objects in cluttered or crowded zones

⇨ Detailed analysis in case of **very high** noise levels

Outlook - Anticipated extensions

➪ Integration into real applications

➪ Global motion compensation

➪ Global illumination changes

➪ Objects in cluttered or crowded zones

➪ Detailed analysis in case of **very high** noise levels

➪ Classification: motion with and without purpose (e.g., trees)

➪ Extension to higher-level events, e.g., agitated behavior

➪ (Integrate coding data for better object segmentation)

➪ (Objective evaluation measures for segmentation)

Outlook - Anticipated extensions

⇨ How successful can context-independency be? Limits?

- Context-independency for context-dependent problems?

Outlook - Anticipated extensions

➪ How successful can context-independency be? Limits?

- Context-independency for context-dependent problems?

<div style="text-align:center">———————</div>

➪ Towards a unified set of algorithmic tools for many domains:

⇒ Which high-level features appropriate for many domains?

⇒ Is there a continuum of methods from real-time to off-line?

○ Must we use completely different algorithms?

○ Redesign/improve rather than new algorithmic tools?

Thanks For Your Attention