Introduction to Human Centered Computing

Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 3.0

Artificial Intelligence & Information Analysis Lab

Human Centered Computing

- **Semantic Video Content Analysis**
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization
- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Semantic Video Content Analysis & Description

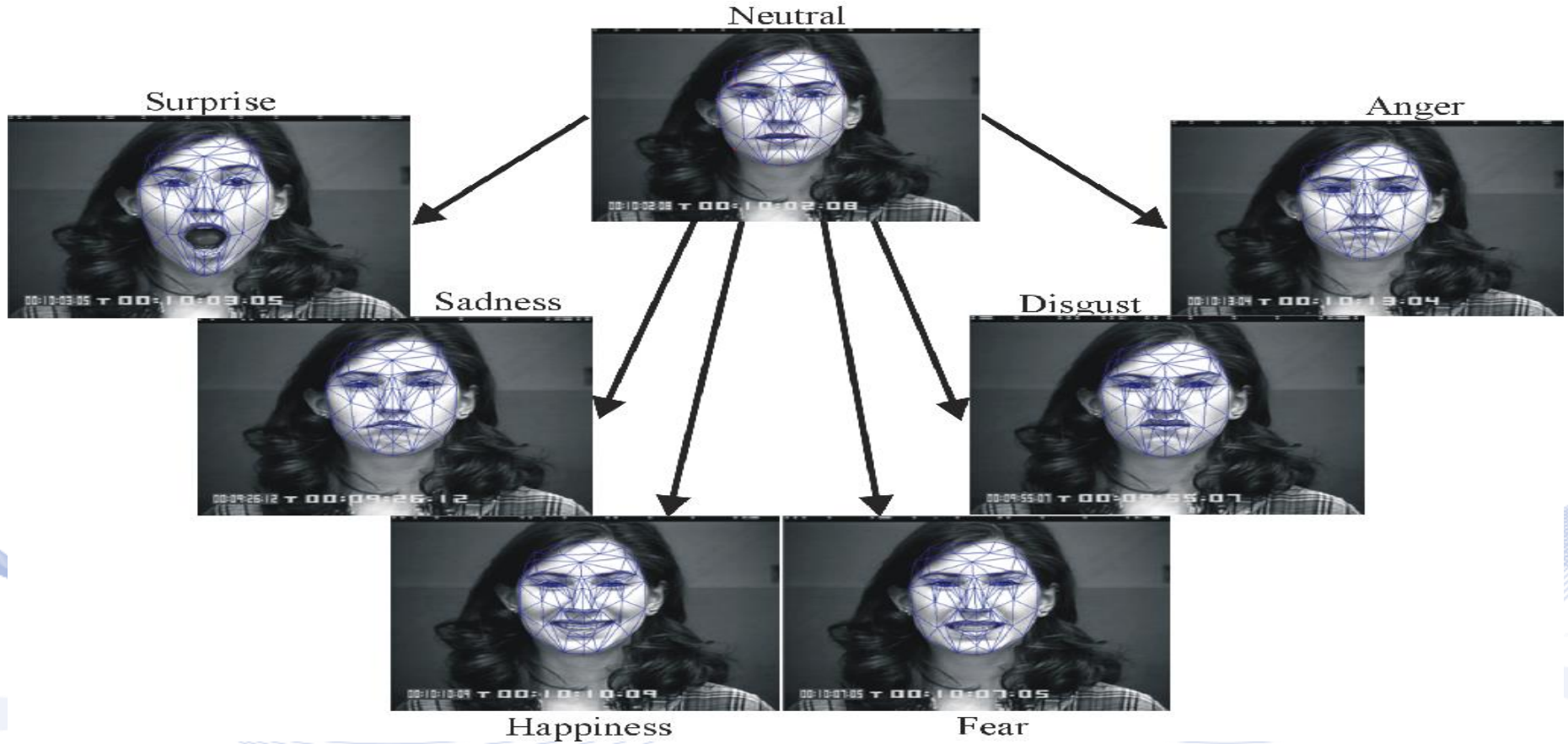
- Video content can be described through state and state transitions of individuals, interactions or communication between humans and physical characteristics of humans:
  - human presence (face, head body),
  - movement,
  - activity/gesture,
  - facial expressions,
  - face/body pose,
  - number of people in a video shot,
  - which people are in the shot (face recognition)
- Anthropocentric (human-centered) approach
  - humans are the most important video entity.
- Characteristics and behavior of other foreground entities such as objects can be also used for semantic description

Examples

1st frame | 6th frame | 11th frame | 16th frame

Face detection and tracking.

Examples

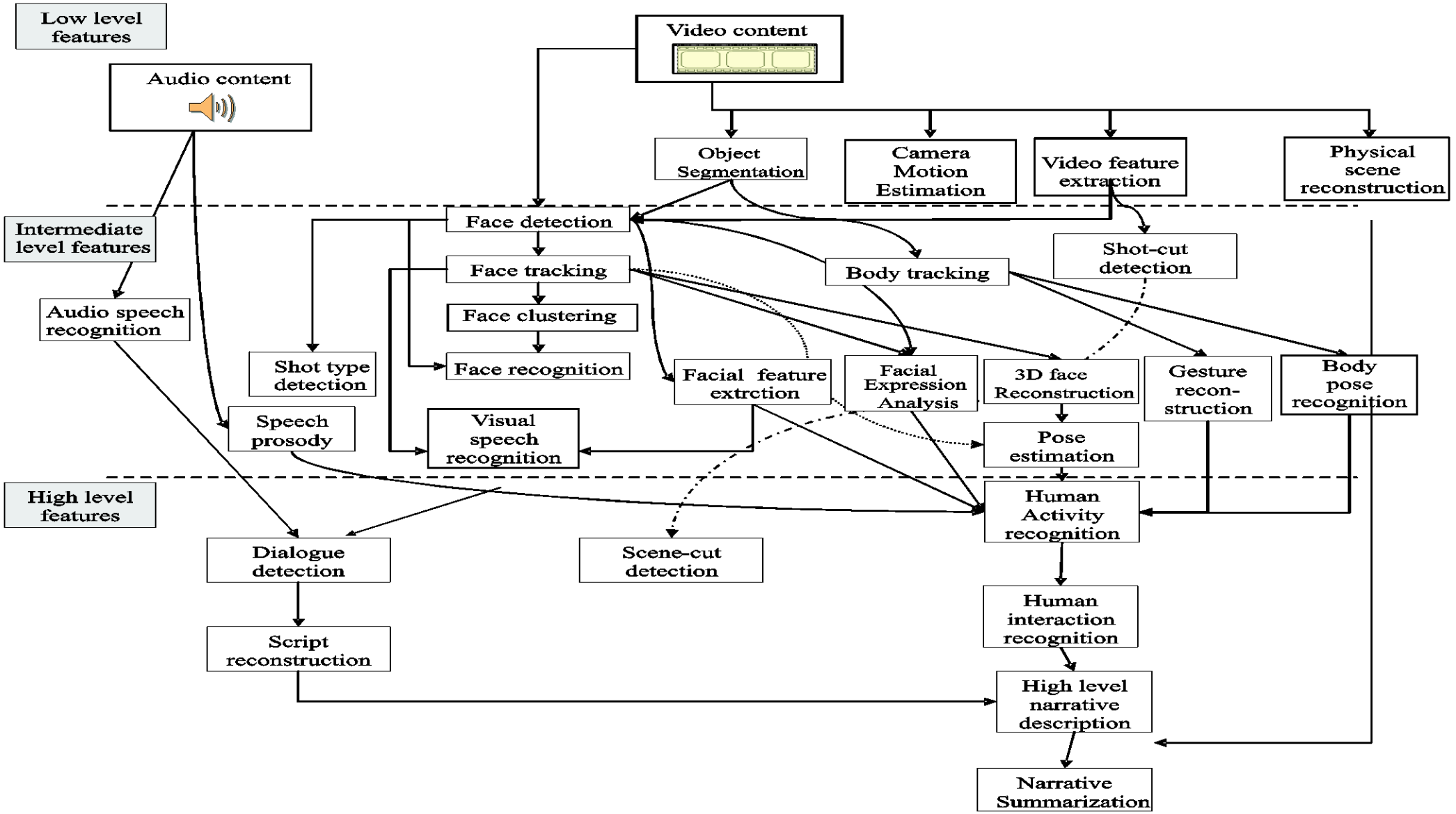


Facial expression recognition.

Relation between HCI operations

- All these HCI operations are related.
- The outputs of some of them are used as input to some other.

Semantic Video Content Analysis pipeline



Semantic Video Content Analysis & Description

Applications:

- Surveillance

  Biometrics, behavior analysis

- Video content search (e.g., in professional video archives or YouTube)

  Semantic annotation of video

  Video archival, indexing and retrieval

- Film and games postproduction:

  Matting, 3D reconstruction

Semantic Video Content Analysis Tasks

- Face/object detection and tracking
- Face clustering
- Face recognition
- 3D face reconstruction
- Facial expression recognition

Semantic Video Content Analysis Tasks

- Multiview human detection
- Eye detection
- Visual speech detection
- Activity/gesture recognition
- Field size (shot type) characterization
- Semantic video content description/annotation

Human Centered Computing

- Semantic Video Content Analysis
- **Face/object detection and tracking**
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Face Detection and Tracking

Problem statement:

- To detect the human faces that appear in each video frame and localize their ***Region-Of-Interest*** (***ROI***).
- To track the detected faces over the video frames.

Face Detection and Tracking

- Tracking associates each detected face in the current video frame with one in the next video frame.

- Therefore, we can describe the ***facial ROI trajectory*** in a shot in (x, y) coordinates.

- ***Actor instance definition*:** face region of interest (ROI) plus other info
- ***Actor appearance definition***: face trajectory plus other info

Face detection examples

Face Detection and Tracking

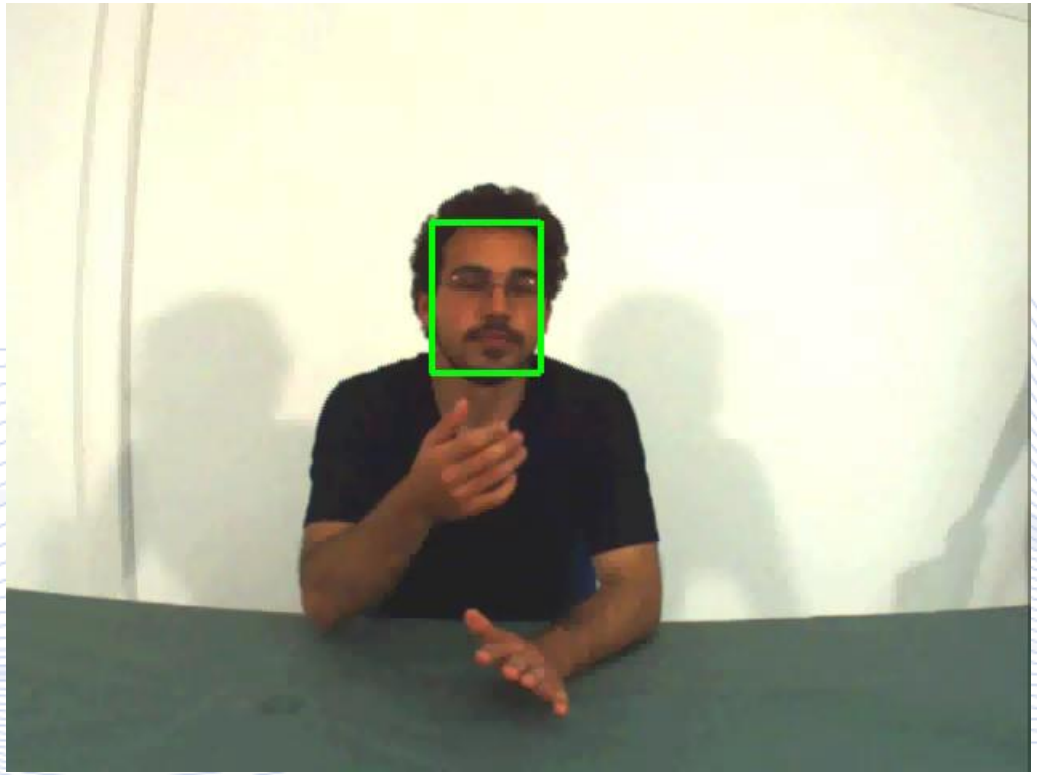
- Tracking failure may occur, i.e., when a face disappears.

- In such cases, face re-detection is employed.
- However, if any of the detected faces coincides with any of the faces already being tracked, the former ones are kept, while the latter ones are discarded from any further processing.

- Periodic face re-detection can be applied to account for new faces entering the camera's field-of-view (typically every 5 video frames).
- Forward and backward tracking, when the entire video is available.

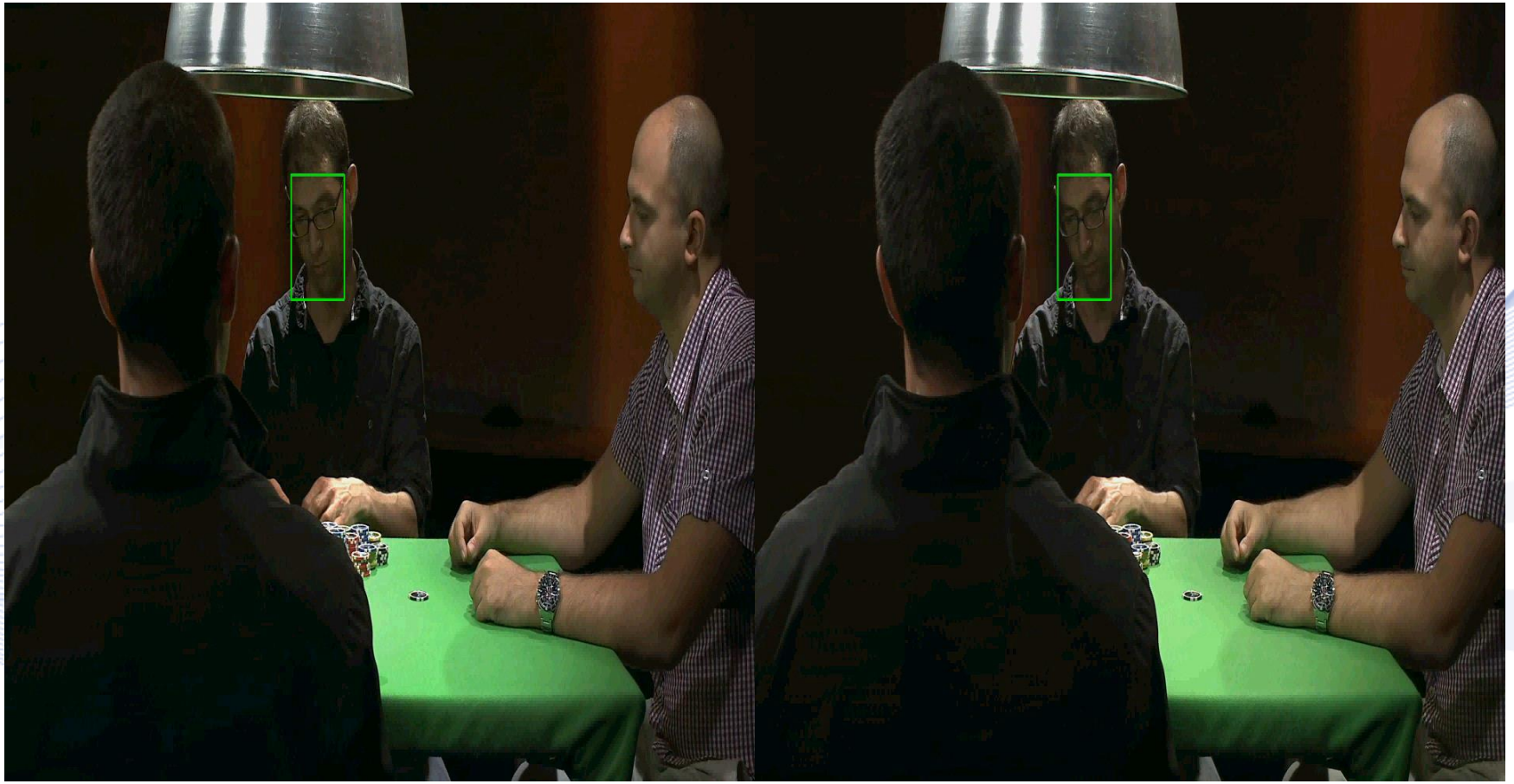
Face/Object Tracking

Experimental results



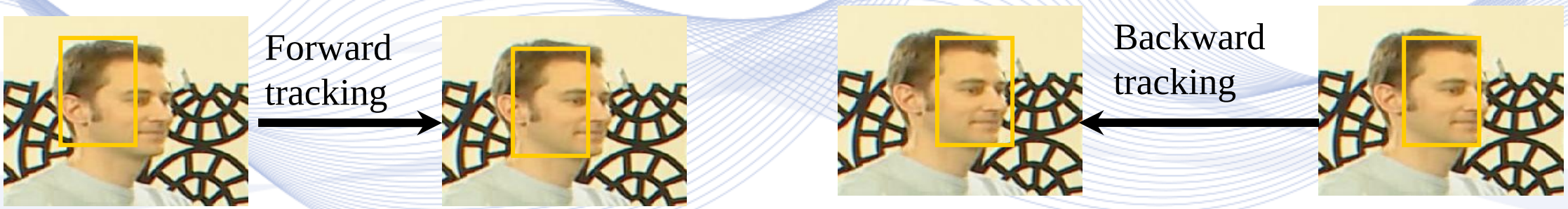
Face Tracking in Stereo Videos

Experimental results

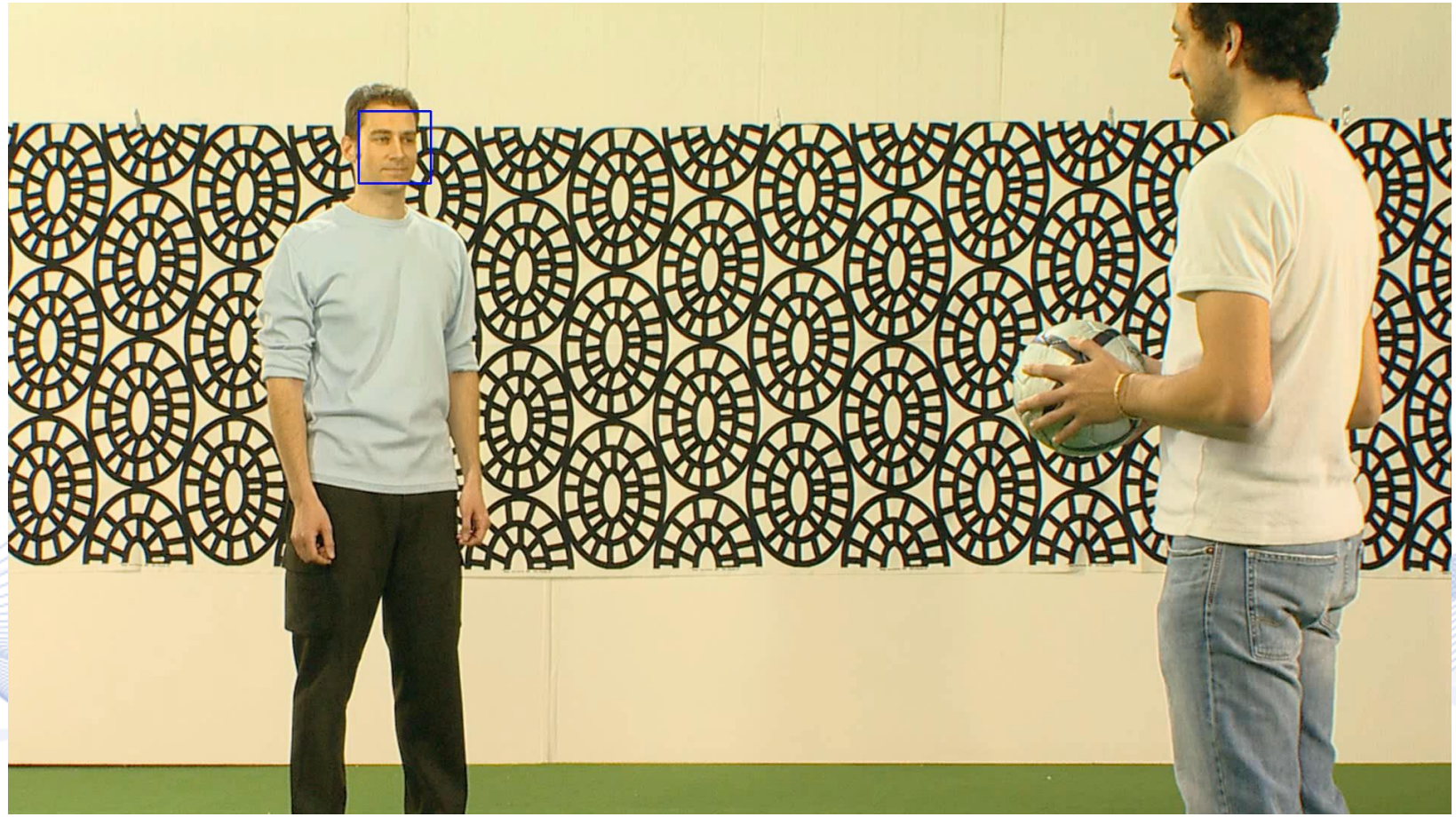


Forward-Backward Tracking Motivation

- Forward in time tracking: different results than backward in time tracking. One mode of tracking may succeed where the other fails
- The key point: post-process the results of a forward and a backward tracking jointly, in order to refine the tracking results
- Same reasoning for stereoscopic video: performing tracking on both channels of a video provides us with more information than using only one.



Forward-Backward Stereo Tracking Results



Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- **Multiview human detection**
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Multiview Human Detection

Problem statement:

- Use information from multiple cameras to detect bodies or body parts, e.g. head.

Camera 4

Camera 6

- Applications:

  Human detection/localization in postproduction.

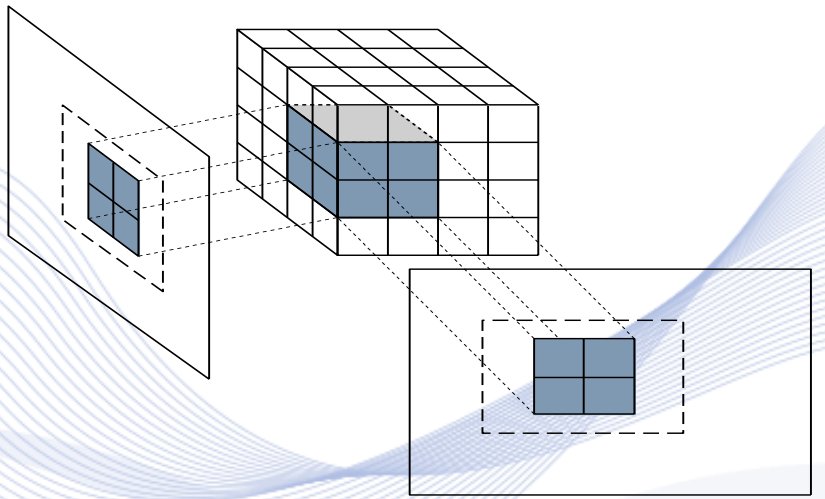
  Matting / segmentation initialization.

Multiview Human Detection

- Detected ROIs are projected back in the 3D space.

- A “probability volume” is created collecting “votes” from individual ROIs.

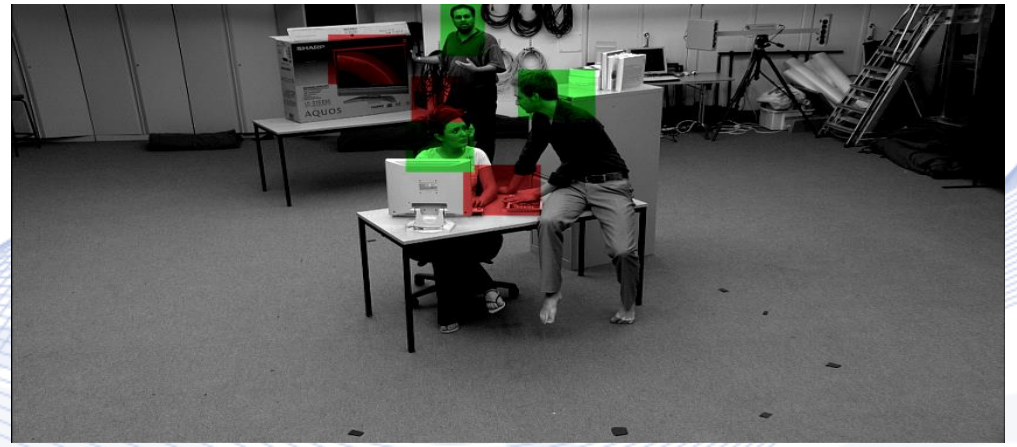
- High probability voxels correspond to the most probable head/body VOIs.



Multiview Human Detection

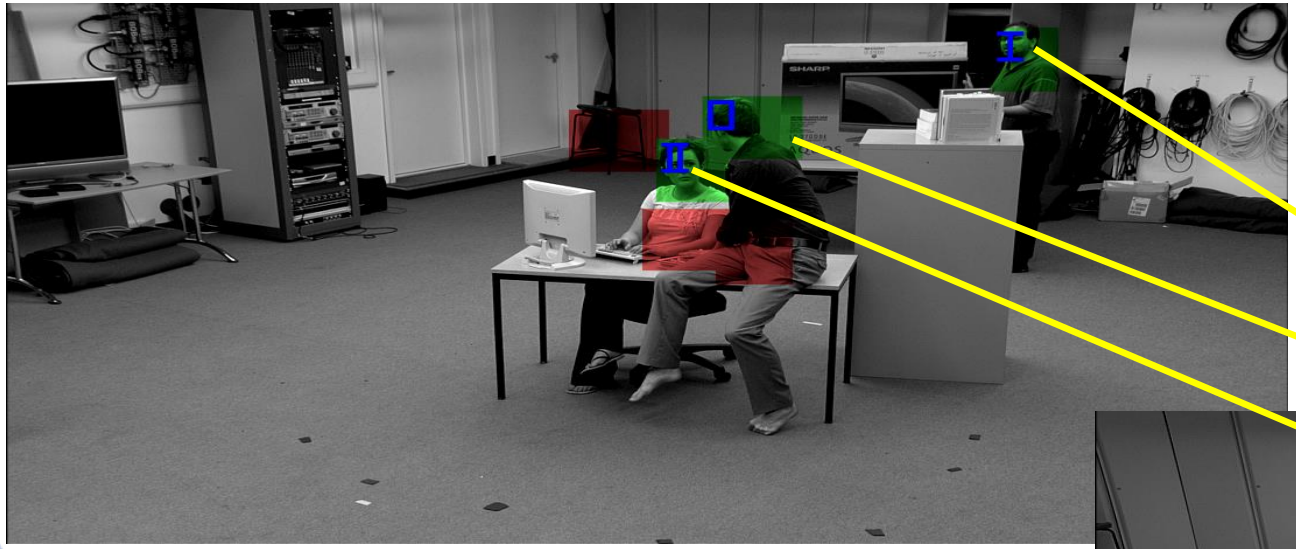
- The retained voxels are projected to all views.
- For every view we reject ROIs that have small overlap with the regions resulting from the projection.

Camera 2

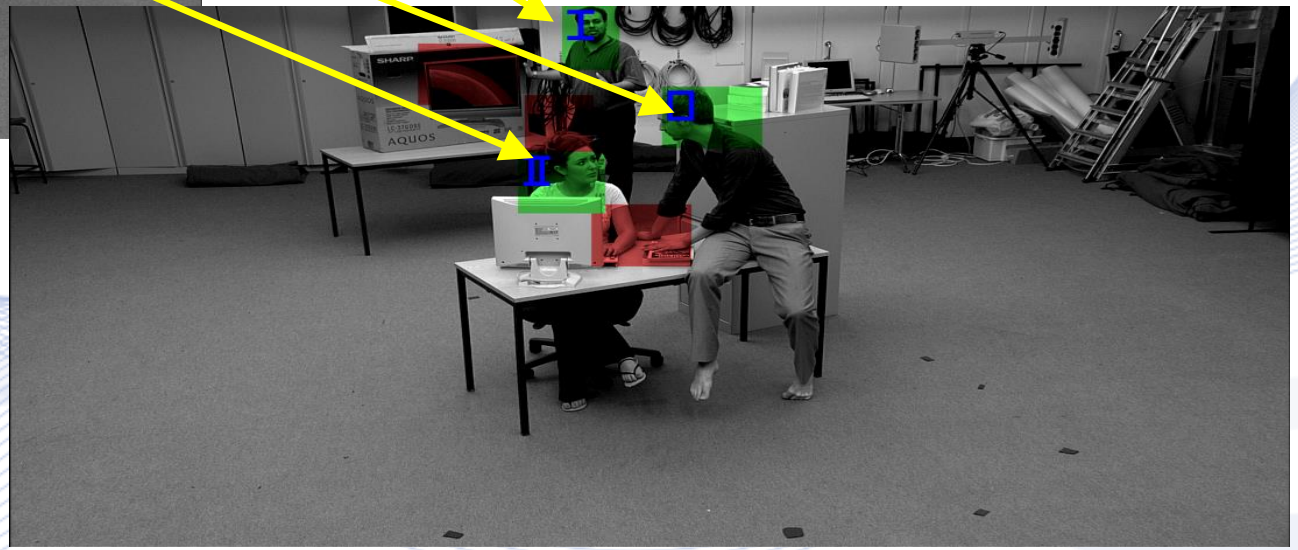


Camera 4

Multiview Human Detection



Camera 2



Camera 4

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- **Face detection obfuscation**
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Face detection obfuscation

A completely different engineering problem

- Goal: The detector does not output **correct** facial ROIs.
- Relatively new area
- Different cases:
  - The detector does not detect faces;
  - The detector detects faces, but classifies them as something else;
  - The detectors detects too many faces, where they do not exist.

- **It is a very useful privacy protection tool.**

Face detection obfuscation

The detector does not detect faces (object-vanishing attack)**:**

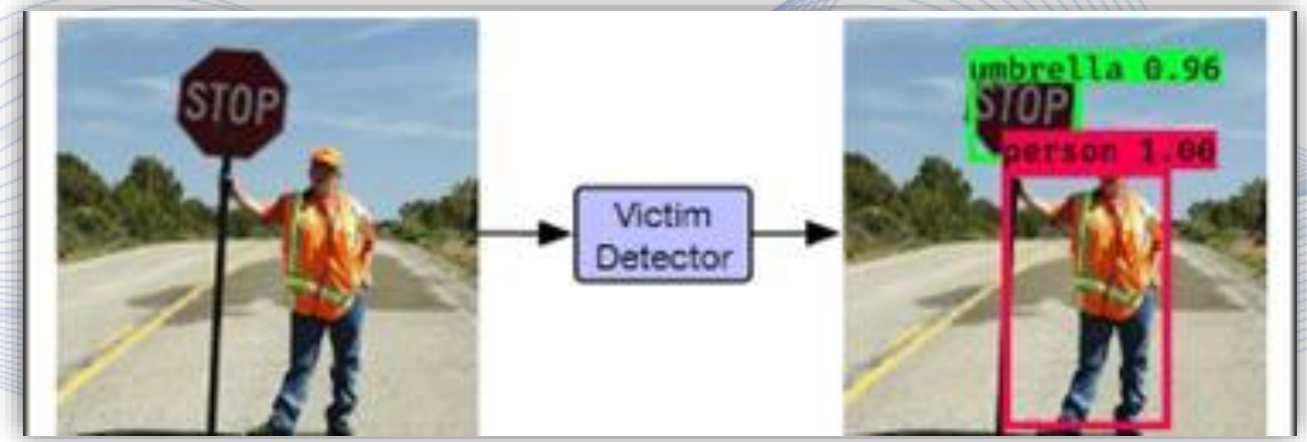
- all attacks against object detectors focus on objects with fixed visual patterns;
- do not take into account intra-class variety;
- adversarial patches can be used to fool person detectors;
- attack on targets with high level intra-class variety, like persons;
- detector does not detect any persons or objects
- “adversarial patch” used as a cloaking device to hide people from object detectors.

[24]. Object-vanishing attack. [CHO2020] K.-H. Chow, L. Liu, M. Loper, J. Bae, M.-E. Gursoy, S. Truex, W. Wei, Y. Wu, “Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems”, in *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 263-272, 2020.*

Face detection obfuscation

The detector detects faces, but classifies them as something else (object-mislabeling attack)**:**

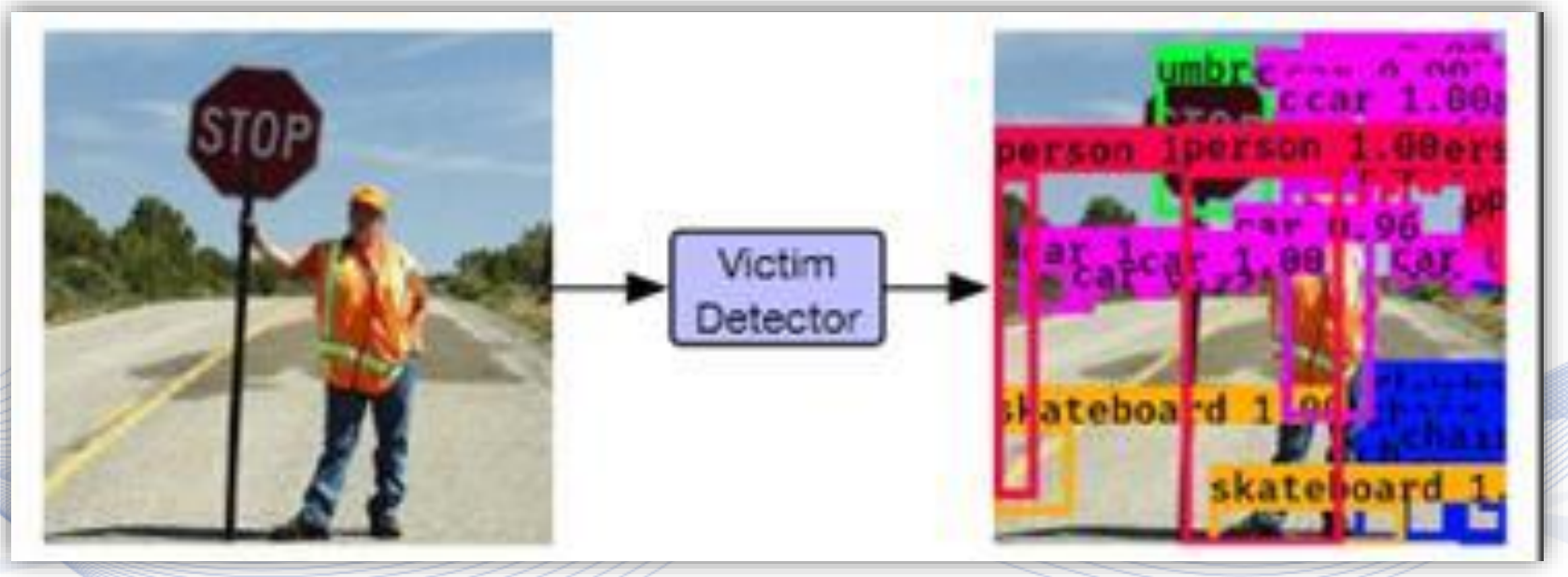
- generate adversarial examples without having access to any information about the network parameter values or their gradients.
- The only input their technique requires is the probabilistic labels predicted by the targeted model.
- adversarial attacks misleads the autoencoder to reconstruct a completely different image



[25]. Object-vanishing attack. https://khchow.com/media/TPS20_TOG.pdf

Face detection obfuscation

The detector detects faces, but classifies them as something else (object-fabrication attack)**:** the object-mislabeling attack fools the detector to mislabel detected objects (e.g., stop sign as an umbrella), which can result in disastrous consequences.



[26]. Object-vanishing attack. https://khchow.com/media/TPS20_TOG.pdf

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- **Face recognition**
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Face Recognition

Three general approaches:

- Deep Neural Networks (state of the art)
- Subspace methods
- Elastic graph matching methods.

Face Recognition

Subspace methods

- The original high-dimensional image space is projected onto a low-dimensional one.

- Face recognition according to a simple distance measure in the low dimensional space.

- Subspace methods: **Eigenfaces** (PCA), **Fisherfaces** (LDA), ICA, **NMF, Class Specific NMF (CSNMF)**.

- Main limitation of subspace methods: they require perfect face alignment (registration).

Face Recognition - NMF

- Original facial images are reconstructed using only additive combinations of the resulting basis images.
- Combination weights: coefficients in **H**.

- Consistent with the psychological intuition regarding the objects representation in the human brain (i.e. combining parts to form the whole).

Face Recognition

Elastic graph matching (EGM) methods

- Elastic graph matching is a simplified implementation of the Dynamic Link Architecture (DLA).

- DLA represents an object by a rectangular elastic grid.
- A Gabor wavelet bank response is measured at each grid node.

- Multiscale dilation-erosion at each grid node can be used, leading to Morphological EGM (MEGM).

Face Recognition

Output of normalized multi-scale dilation-erosion for nine scales.

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- **Face clustering**
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Face Clustering

Problem statement:

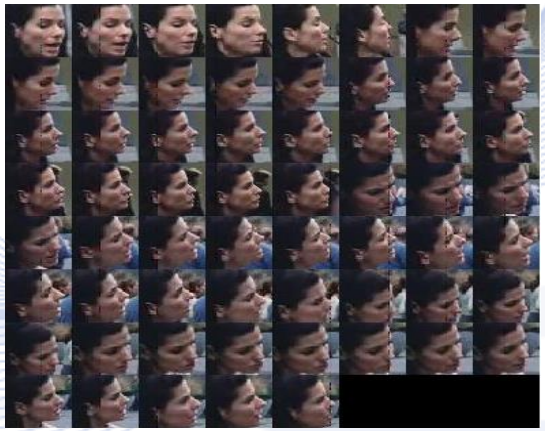
- To cluster a set of facial ROIs
- Input: a set of face image ROIs
- Output: several face clusters, each containing faces of only one person.
- Applications

Cluster actor images, even if they belong to different shots.

Cluster various views of the same actor.

Generate the cast of a movie.

Semi automatic face recognition



Face Clustering

A four step clustering algorithm is used:

- Face detection and/or tracking.
- Calculation of a similarity matrix in order to create a similarity graph. Similarity calculation between facial images using either mutual information or local steering kernels or any other similarity criterion.
- Application of face trajectory heuristics, to improve clustering performance.
- Similarity graph clustering.

Facial image clustering in videos

- In videos facial images can result from the application of face detection and tracking algorithms
- This leads to "**facial trajectories**": series of facial images of (usually) the same person over time

- Each such facial trajectory can be represented by any of the images included in it.
- Facial image clustering in videos: cluster facial trajectories by using their representative images

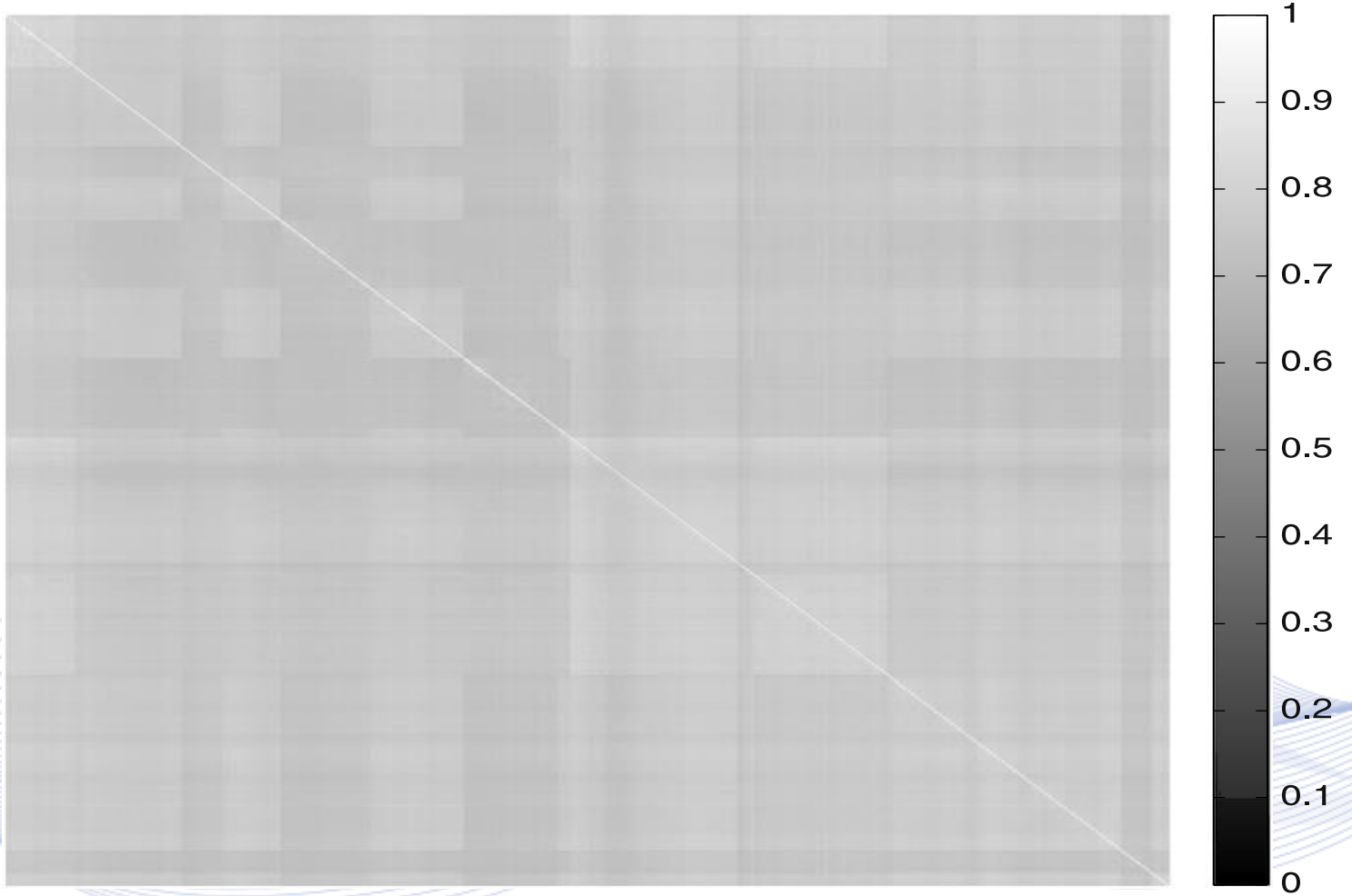
Facial image trajectory representatives

- Two approaches to choose trajectory representative:

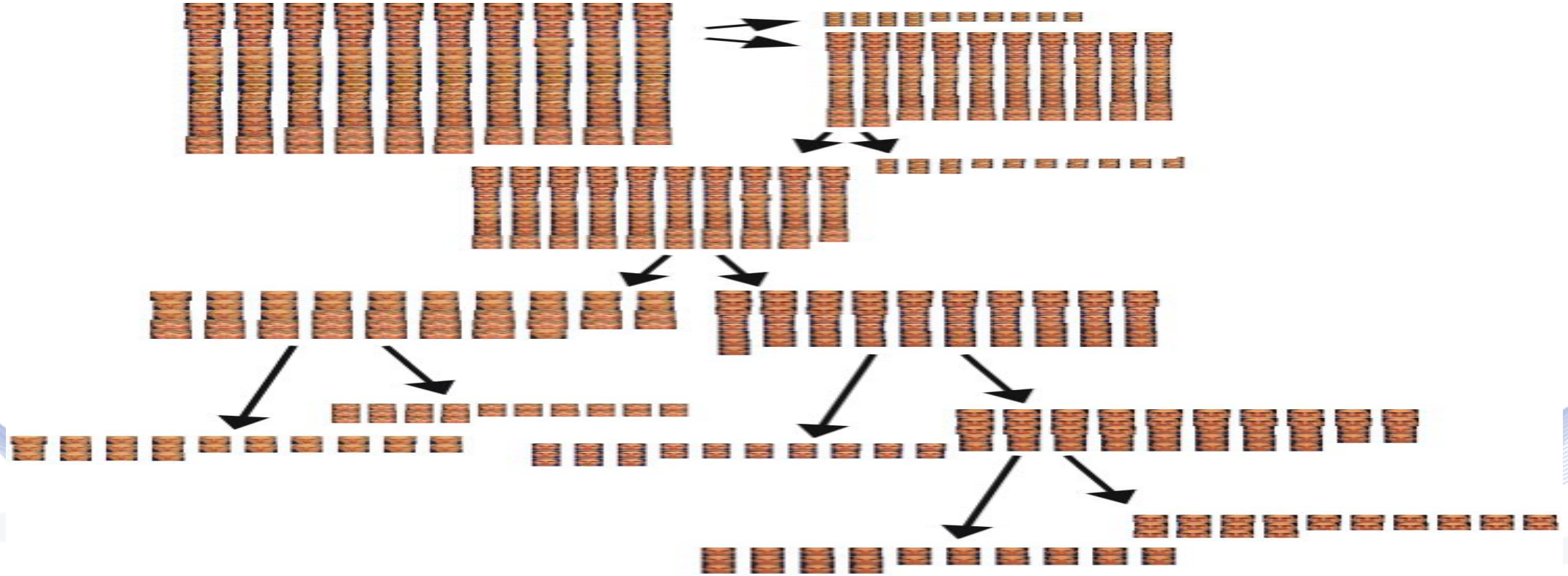
  A **single** detected facial image

  Multiple images from the trajectory

Face Similarity Matrix



N-Cut Graph Clustering (2-way Partitioning) :



Cluster Examples

Images of different scales

Images of different illumination and poses

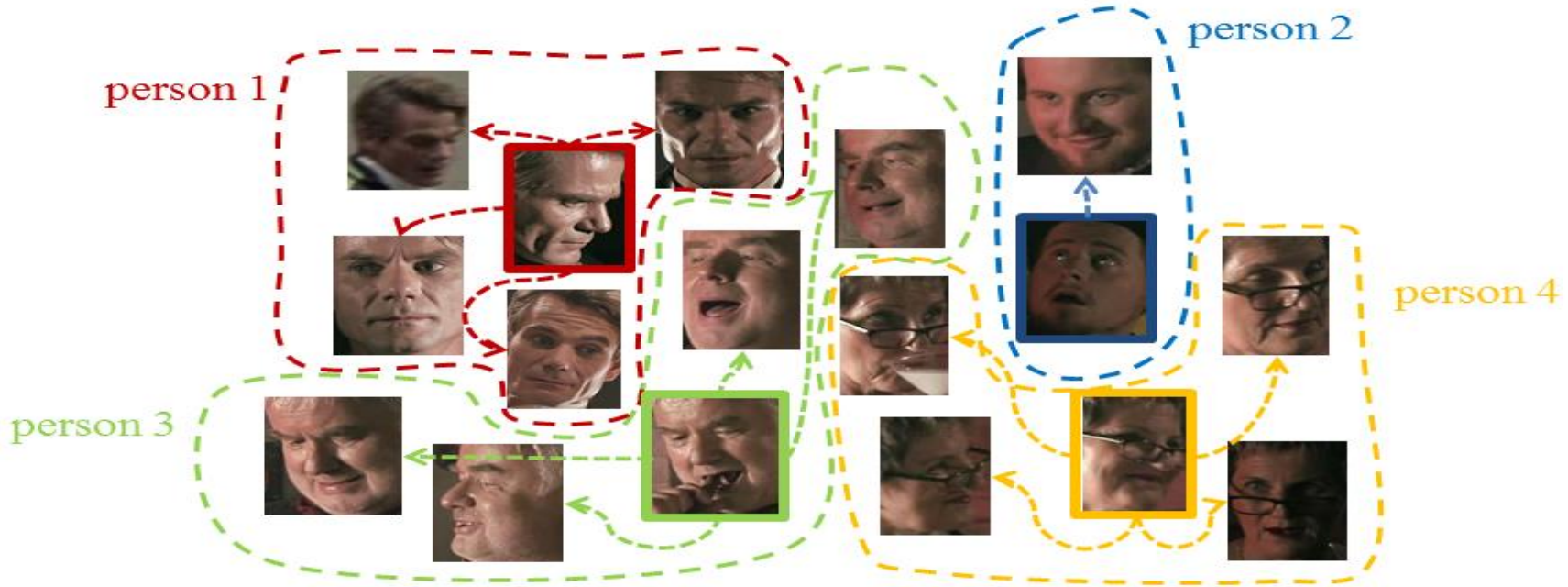
Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- **Label propagation on videos**
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Label propagation on videos

- Label propagation is a label diffusion process from a small set of labeled data $\mathrm{X_L} = \{\mathrm{x_i}\}_{i=1}^{n_l}$ to a larger set of unlabeled data $X_U = \{x_i\}_{i=1}^{n_u}$



Label propagation on videos

Label propagation algorithms satisfy the following conditions:

- They preserve the labels of the initially labeled data.
- They assign the same label to similar samples or to samples that lie in the same structure of the feature space.

Label propagation on videos

- The performance of label propagation algorithms depends highly on
  The data representation method (the data graph construction)
  The selection of the initially labeled data set

Human Centered Computing

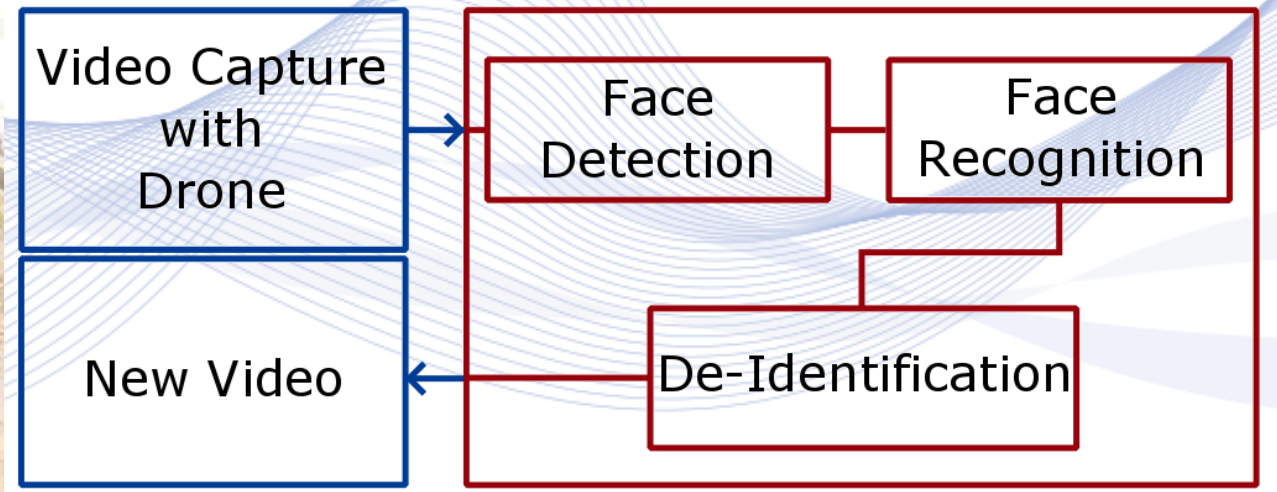
- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- **Face De-identification**
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

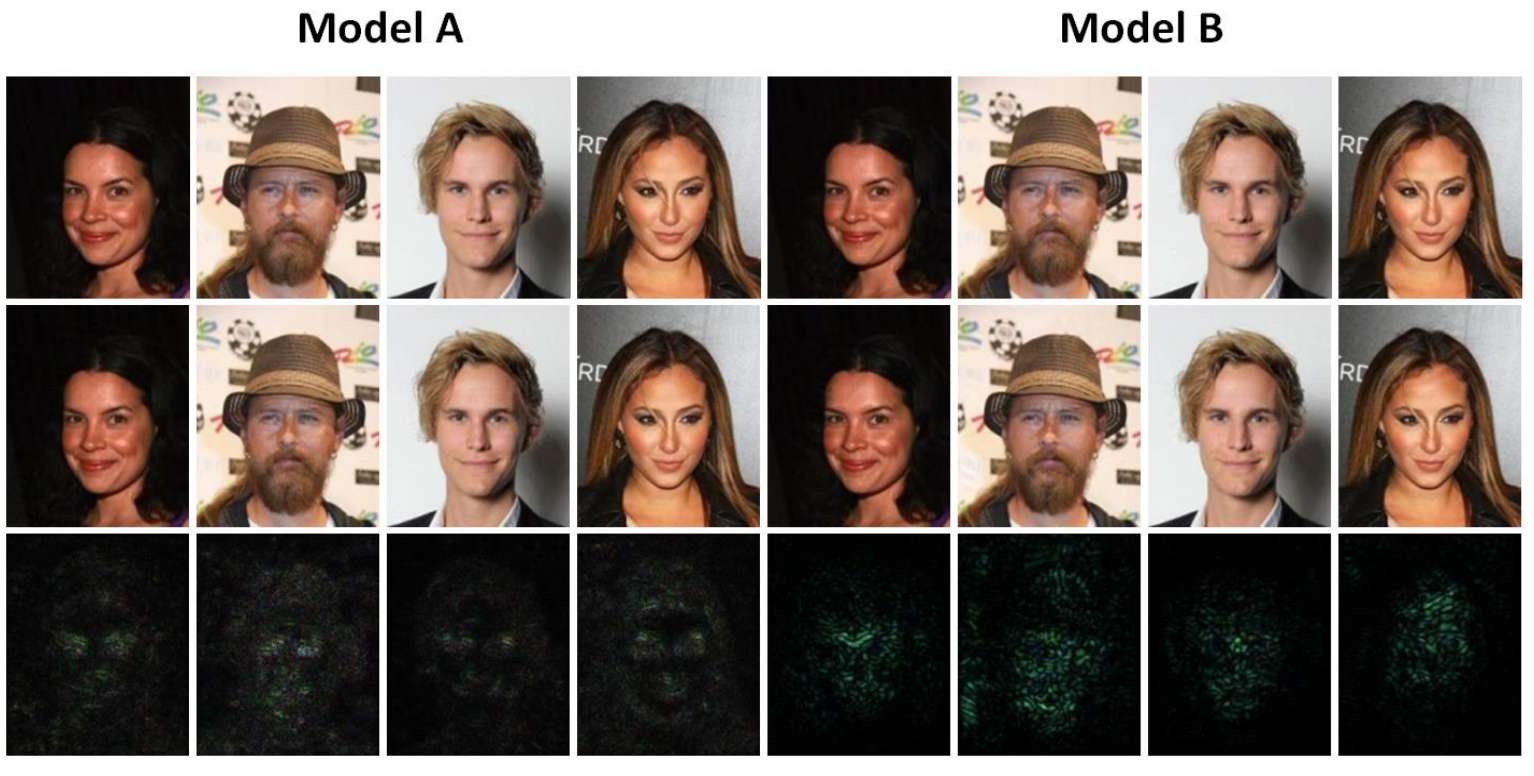
Face De-identification

- ***Face de-identification*** (Face recognizer obfuscation):
  - Corrupt the facial region so that deep NN face classifiers fail.
- Developed methodology:
  - Simple/Naive approaches (additive noise, impulsive noise)
  - Reconstruction-based (SVD, PCA, hypersphere projections, auto-encoder-based) approaches.
  - Autoencoder face de-identification.
  - GAN face de-identification.
  - Adversarial face de-identification.

Face De-identification on drone videos



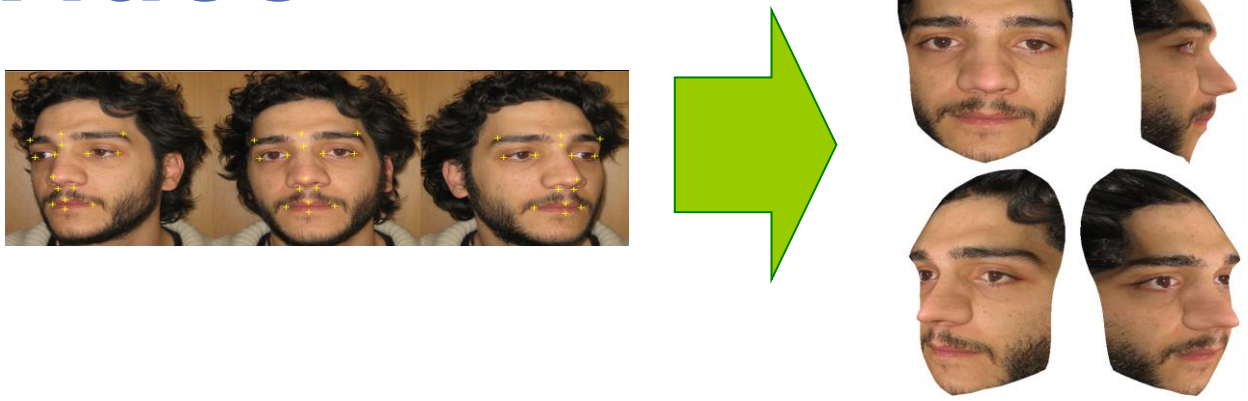
Face De-Identification



Adversarial Face De-Identification
First row: original image; Second row: de-identified image. Third row: adversarial perturbation absolute value (x10) [CHA2019].

3D Face Reconstruction from Uncalibrated Video

Problem statement:



- Input: facial images or facial video frames, taken from different view angles, provided that the face neither changes expression nor speaks.
- Output: 3D face model (saved as a VRML file) and its calibration in relation to each camera. Facial pose estimation.
- Applications:

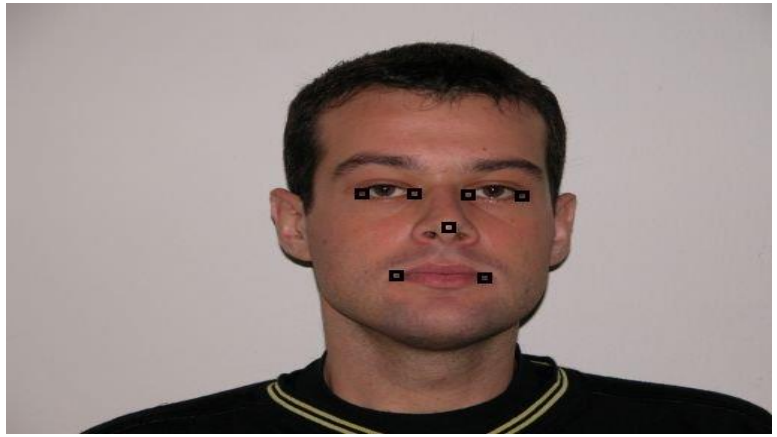
  3D face reconstruction, facial pose estimation, face recognition, face verification.

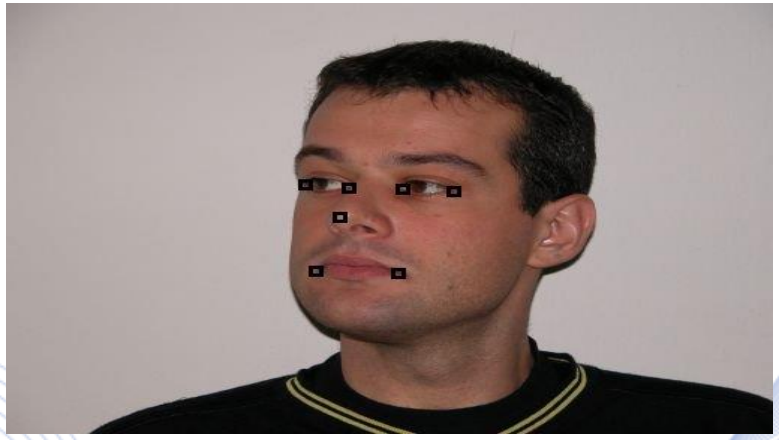
3D Face Reconstruction from Uncalibrated Video

Method overview

- Manual selection of characteristic feature points on the input facial images.
- Use of an uncalibrated 3-D reconstruction algorithm.
- Incorporation of the CANDIDE generic face model.
- Deformation of the generic face model based on the 3-D reconstructed feature points.
- Re-projection of the face model grid onto the images and manual refinement.

3D Face Reconstruction from Uncalibrated Video



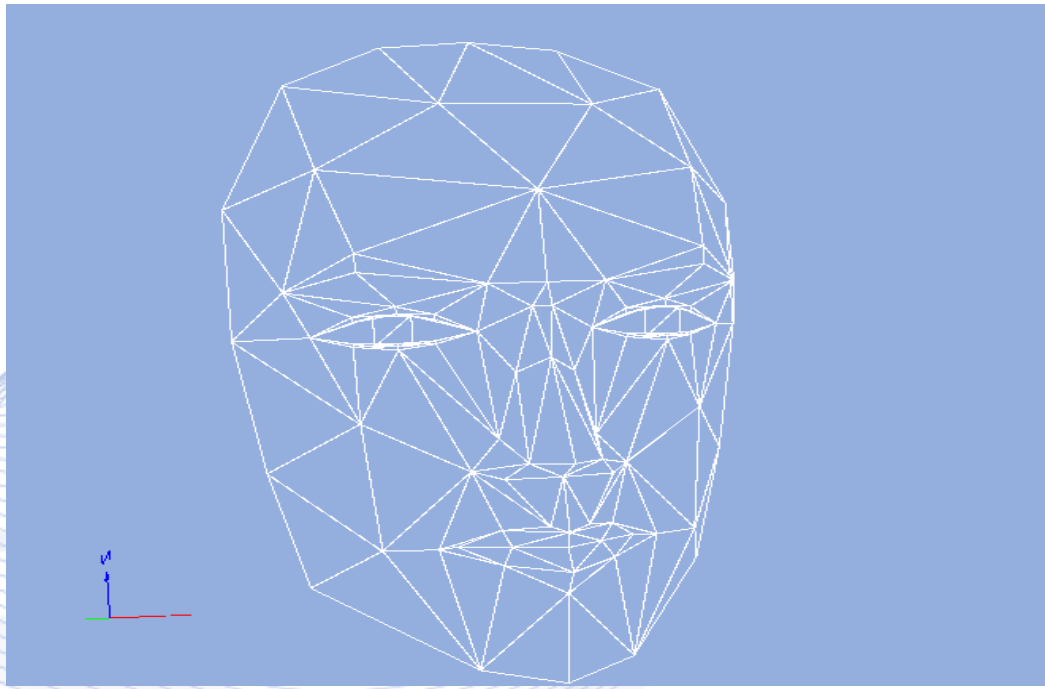


Input: three images with a number of matched characteristic feature points.

3D Face Reconstruction from Uncalibrated Video

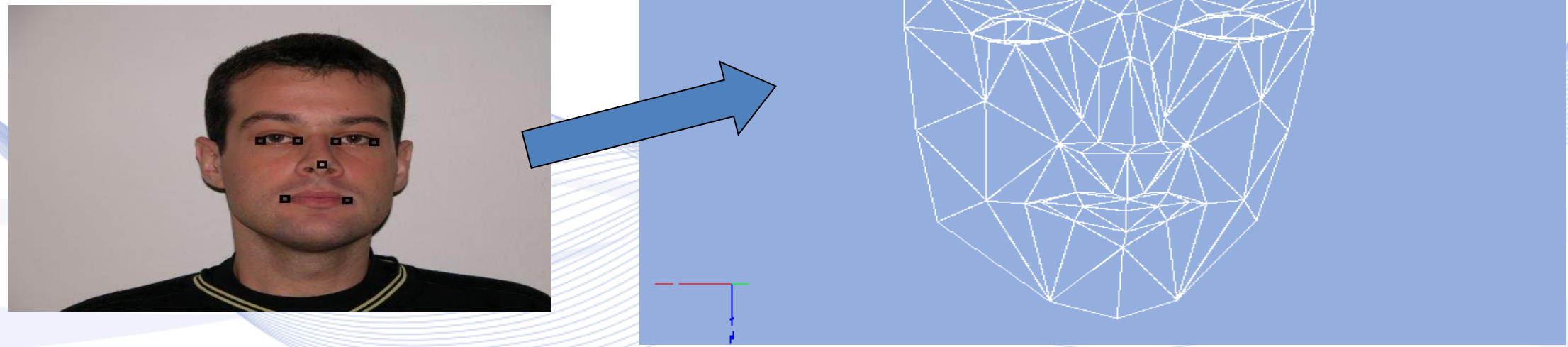
- The CANDIDE face model has 104 nodes and 184 triangles.

- Its nodes correspond to characteristic points of the human face, e.g. nose tip, outline of the eyes, outline of the mouth etc.

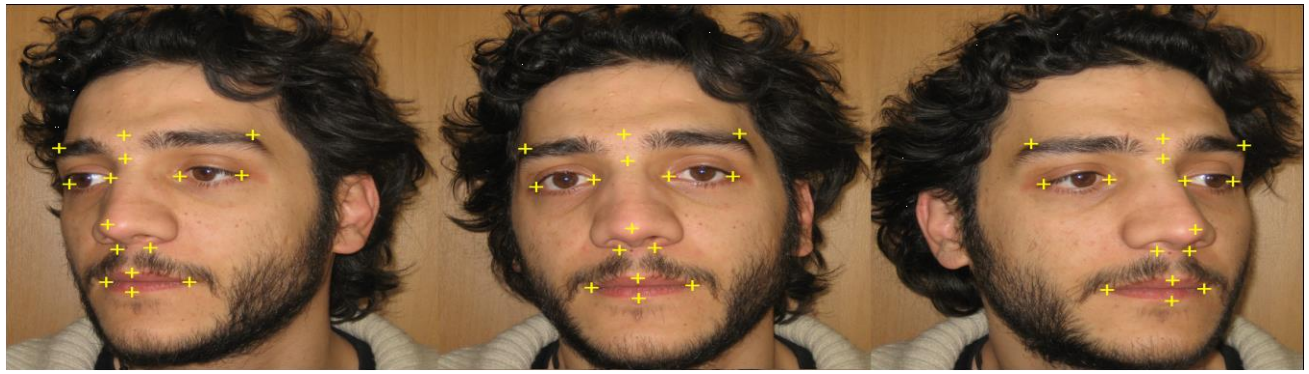


3D Face Reconstruction from Uncalibrated Video

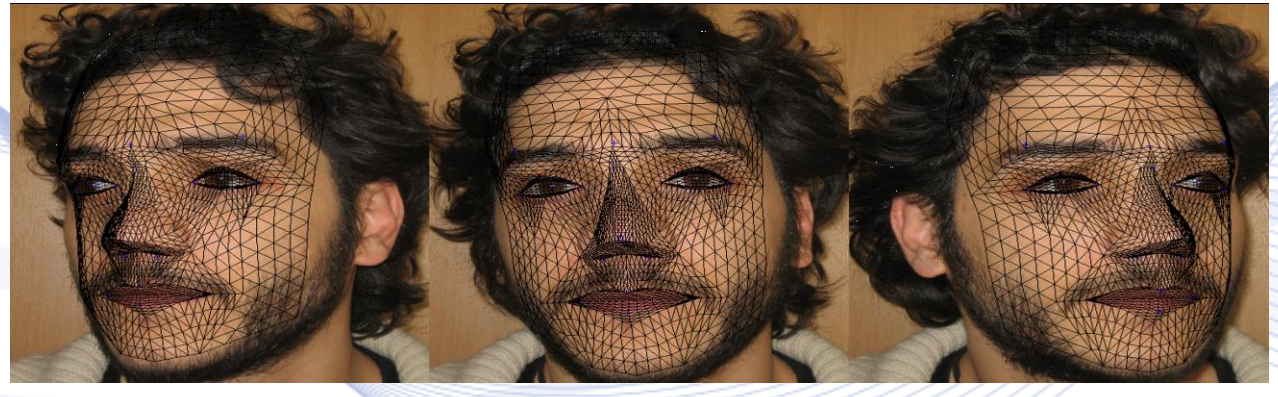
Example



3D Face Reconstruction from Uncalibrated Video



Selected features



CANDIDE grid reprojection

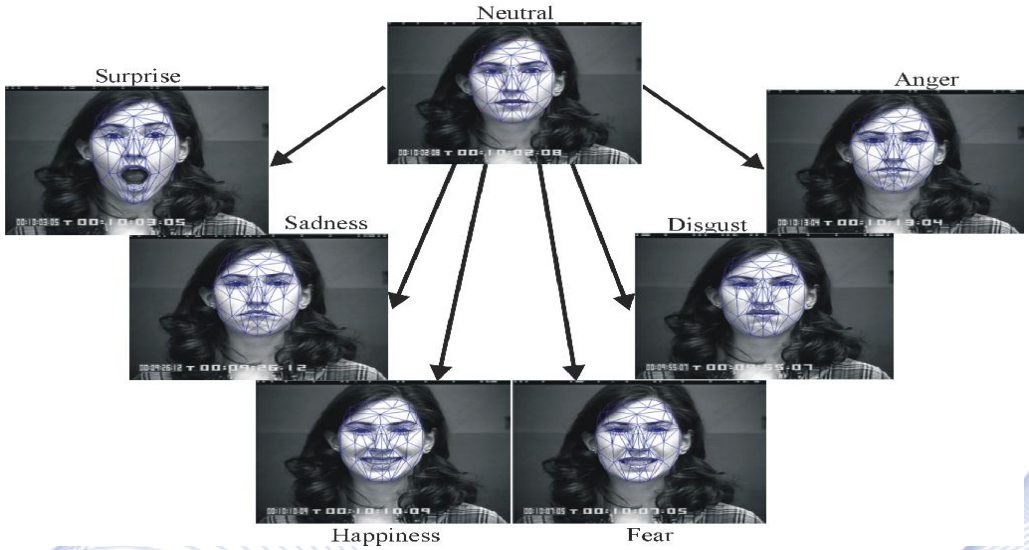
3D face reconstruction

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- **Facial expression recognition**
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Facial Expression Analysis



Problem statement:

- To identify a facial expression

in a facial image or 3D point cloud.

- Input: a face ROI or a point cloud
- Output: the facial expression label

(e.g. neutral, anger, disgust, sadness, happiness, surprise, fear).

Universal Facial Expressions

- Six universal facial expressions:
  - Anger
  - Fear
  - Disgust
  - Happiness
  - Sadness
  - Surprise

  - Neutral






Informative Content of Facial Expressions

- Human communication is mainly performed by nonverbal means (gestures and facial actions).
- Facial actions: important source for understanding human emotional state and intention.
- Key importance to various fields e.g. human behavior analysis, affective video content description, psychology, HCI, ambient intelligence, entertainment etc.

Facial Expression Analysis

- Latent facial image data dimensionality

  Facial image space dimensionality is much higher than required.

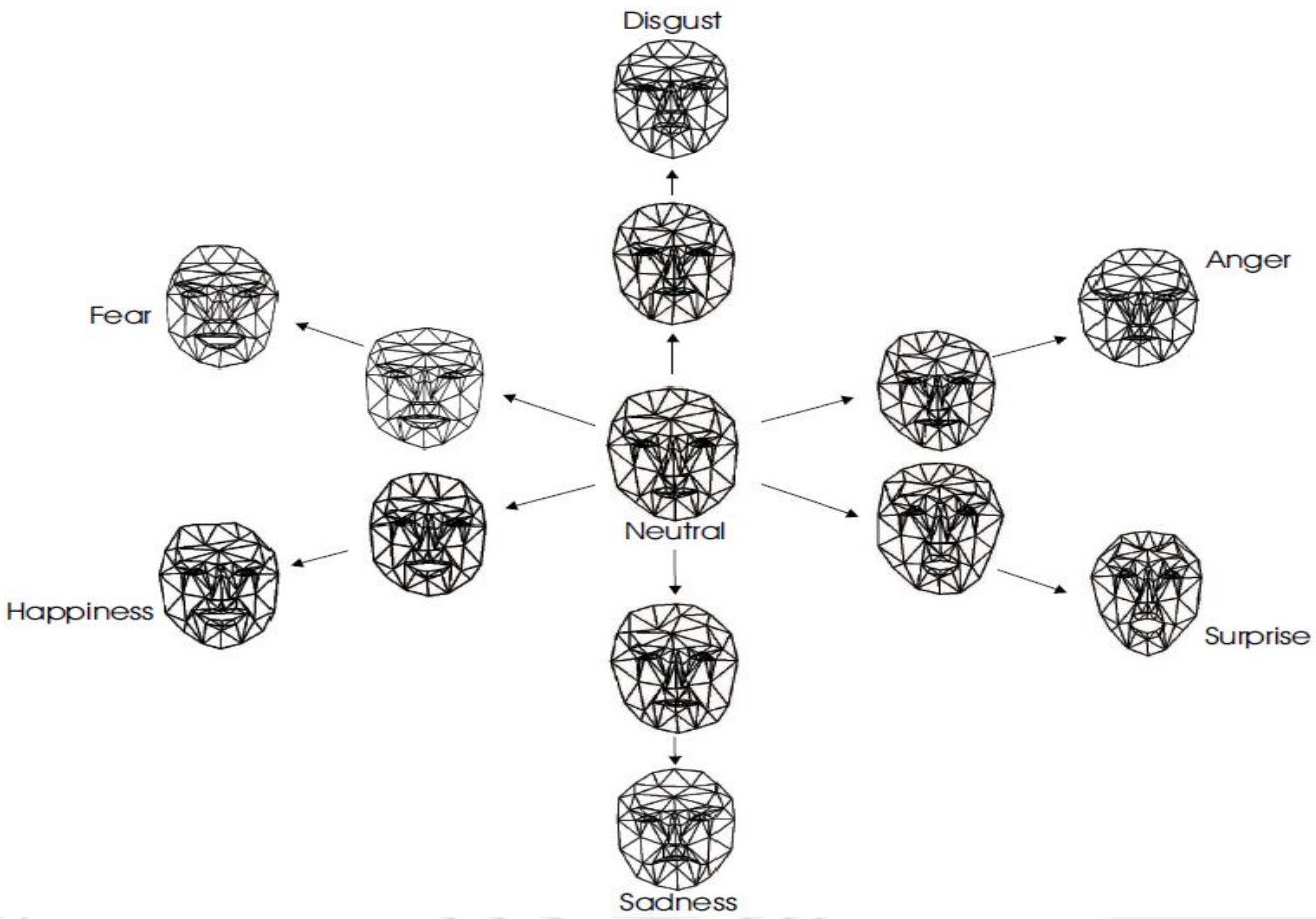
  Necessitates the use of efficient dimensionality reduction methods.

  Reduce complexity and boost performance of expression analysis algorithms.

- Three popular approaches to handle facial expression data high dimensionality:
  - Deep Neural Networks (state of the art)
  - Grid-Based Methods.
  - Subspace Learning Methods.

Grid-Based Methods

- A facial grid is a parameterized face mask specifically developed for model-based coding of human faces.
- A popular facial wireframe model is the Candide grid.
- Facial expression information extraction is performed by facial feature point tracking.

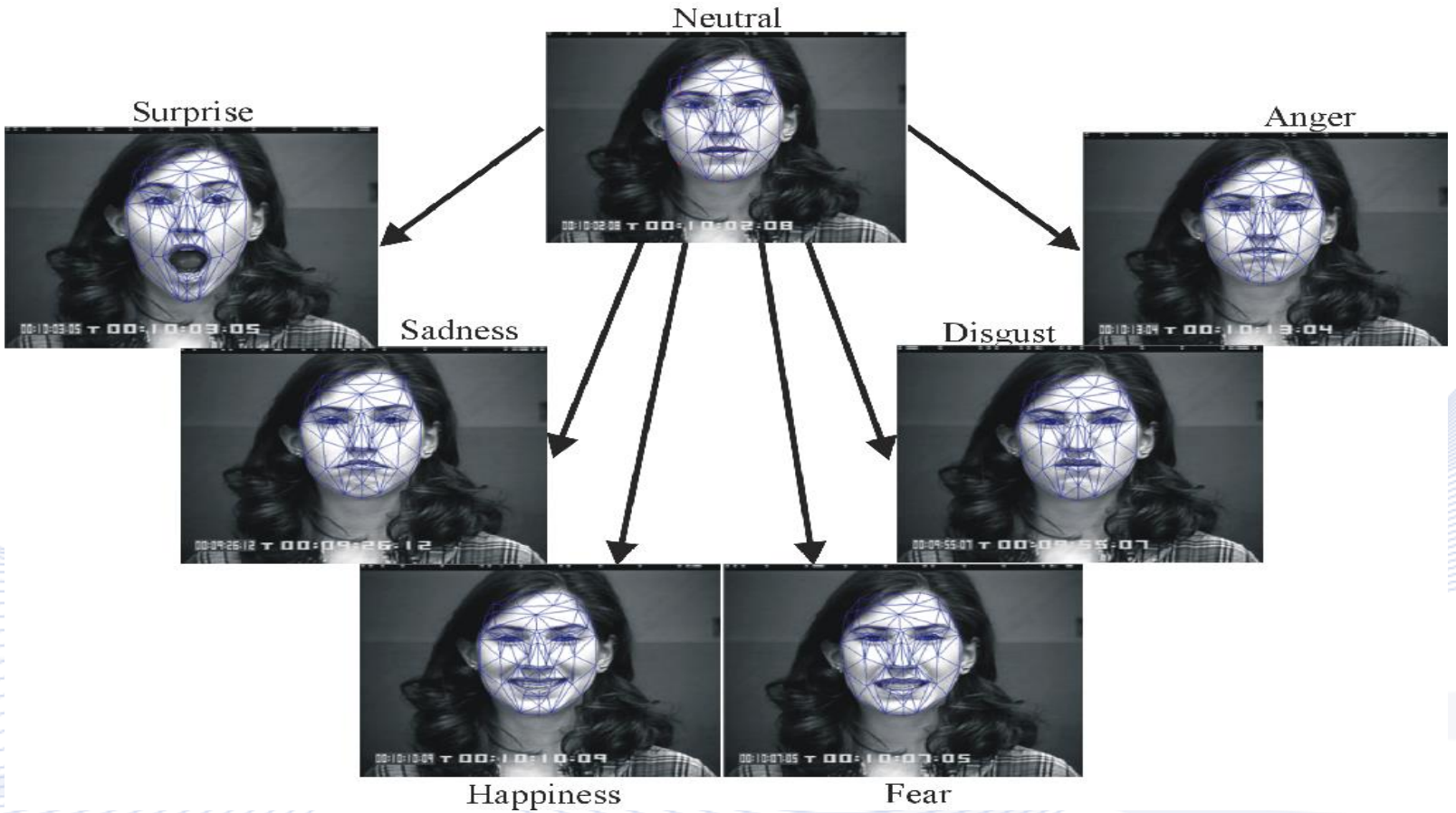


Grid-Based Methods

- Expression recognition:

  Track geometric positions of the grid nodes corresponding to fiducial points on a face;

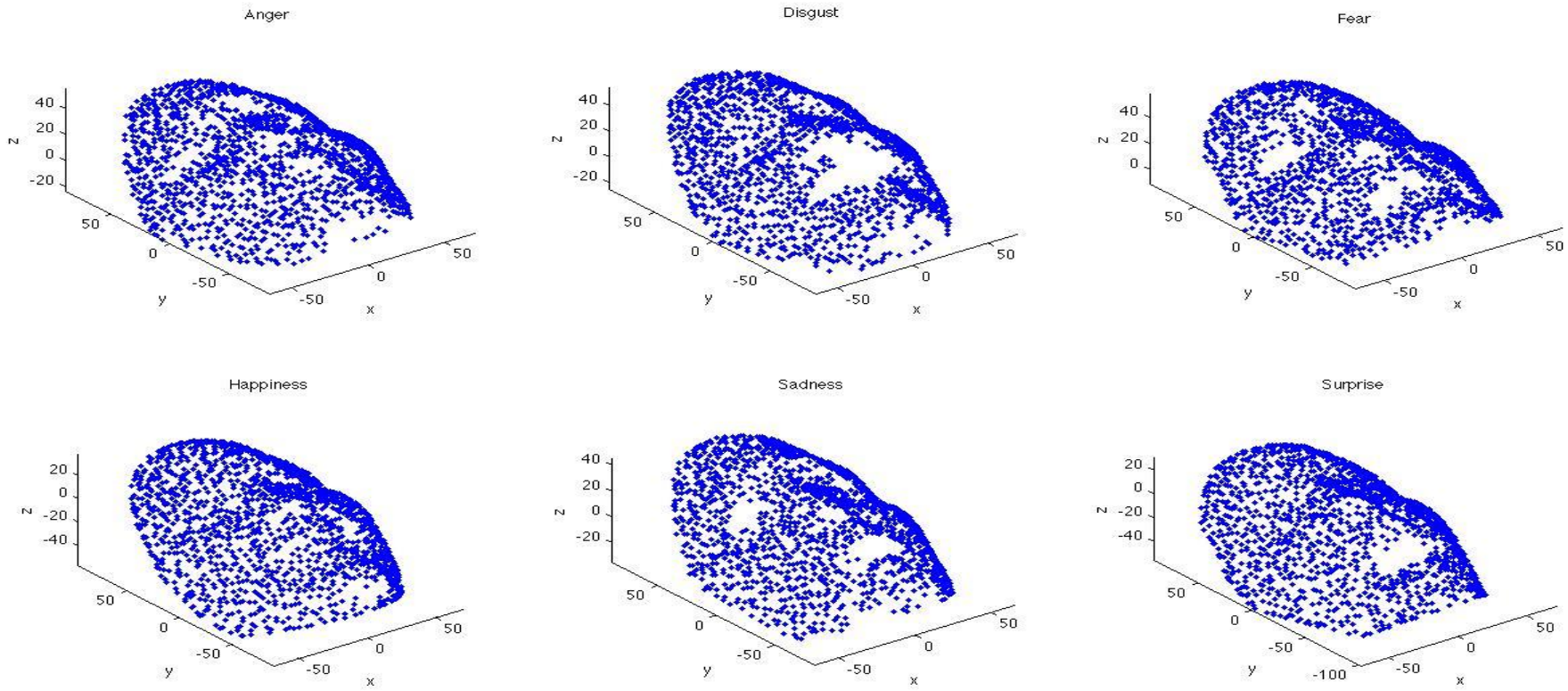
  Grid nodes displacements are used as classification features.



Experimental Results

3D Facial Expression Recognition

- Use of 3D facial point clouds.



3D Facial Expression Recognition

- Tackling global geometric transforms:

  Rotation, translation, scaling.
- A new coordinate system is created:

  Origin: point cloud center of mass.

  Axes: principal axes found by Principal Components Analysis.
- All points are projected in the new coordinate system.

3D Facial Expression Recognition

- A multi class SVM classifier is used for facial expressions recognition.

  Input: point clouds representation in the normalized PCA space.
- Correspondences between points in different clouds (faces) are required.
- The same method can be used for face recognition.

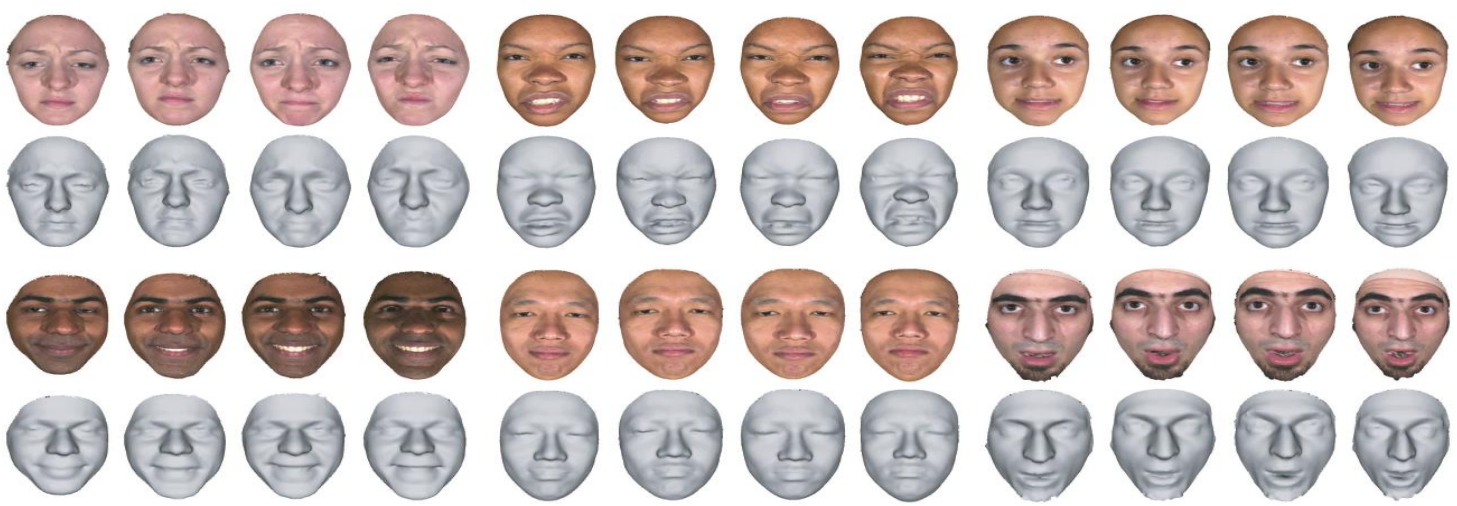
3D Facial Expression Recognition

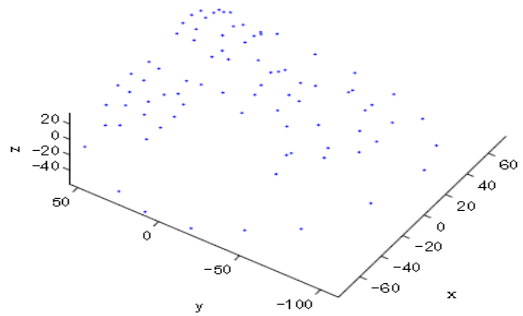
- Experiments on the BU-3DFE database

  100 subjects.

  6 facial expressions x 4 different intensities +neutral per subject.

  83 landmark points (used in our method).





Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- **Facial feature detection**
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Facial feature detection

Face images having noted facial landmarks

Facial feature detection

Eye Detection:

- To detect eye regions in an image.
- Input: A facial ROI produced by face detection or tracking
- Output: eye ROIs / eye center location.
- Applications:

  face analysis e.g. expression recognition, face recognition, etc

  gaze tracking.

Facial feature detection

Training data

a) The left eye training images

b) The right eye training images

c) The length maps for the left image eye

d) The length maps for the right image eye

e)The angle maps for the left eye image

f) The angle maps for the right eye image

Facial feature detection

Face and mouth within bounding boxes, and mouth of both persons extracted

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- **Visual speech detection**
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Visual Speech Detection

Problem statement:

- To detect video frames where a persons speaks using only visual information

- Input: A mouth ROI produced by mouth detection
- Output: yes/no visual speech indication.

- The mouth is detected and localized by a similar technique to eye detection.

Visual Speech Detection

- Applications

  Detection of important video frames (when someone speaks)

  Lip reading applications

  Speech recognition applications

  Speech intent detection and speaker determination

  - human--computer interaction applications
  - video telephony and video conferencing systems.

  Dialogue detection system for movies and TV programs

Visual Speech Detection

(a) Face detector outcomes

(b) Mouth detector outcomes

(a) Two active speakers

(b) One active speaker

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- **Shot type characterization**
- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Shot Type Characterization

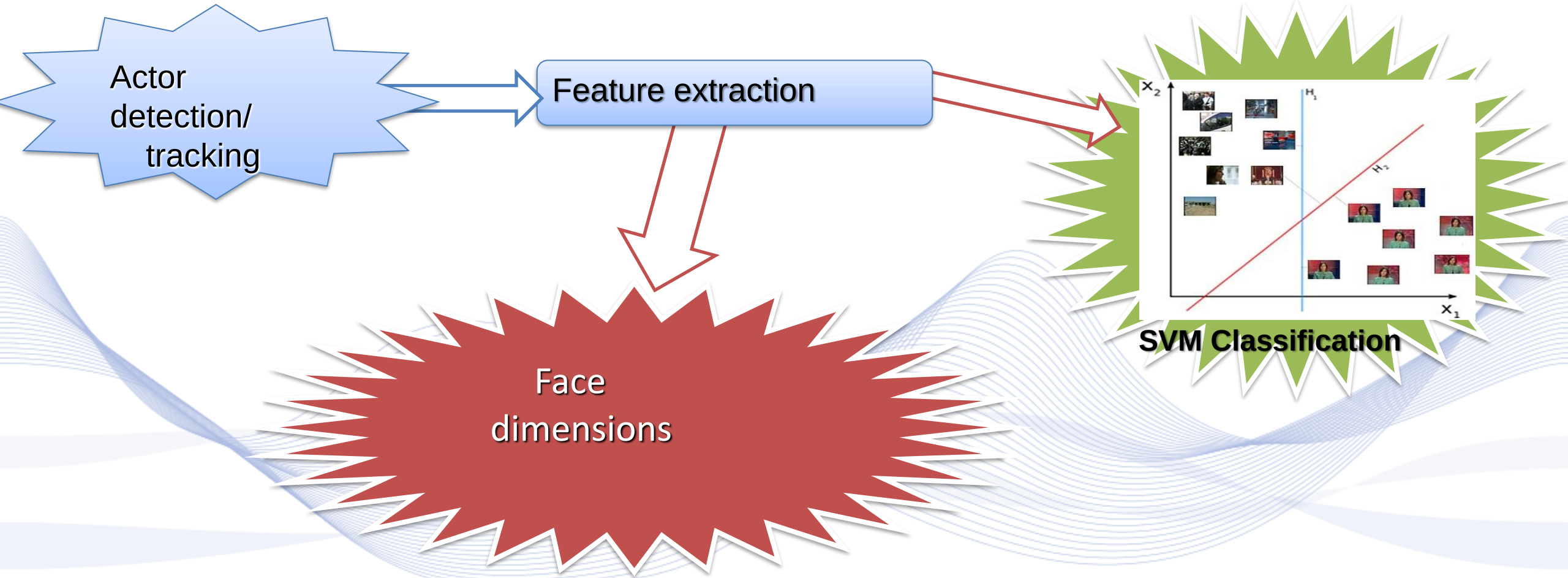
- **Problem statement**: Classify a video shot into one of the predefined classes :
  - eXtreme Close Up shot (XCU),
  - Medium Shot (MS),
  - Long Shot (LS),
  - eXtreme Long Shot (XLS), etc.
  - Over The Shoulder (OTS)

Basic shot types

Close up shot

Medium Shot

Shot Type Characterization

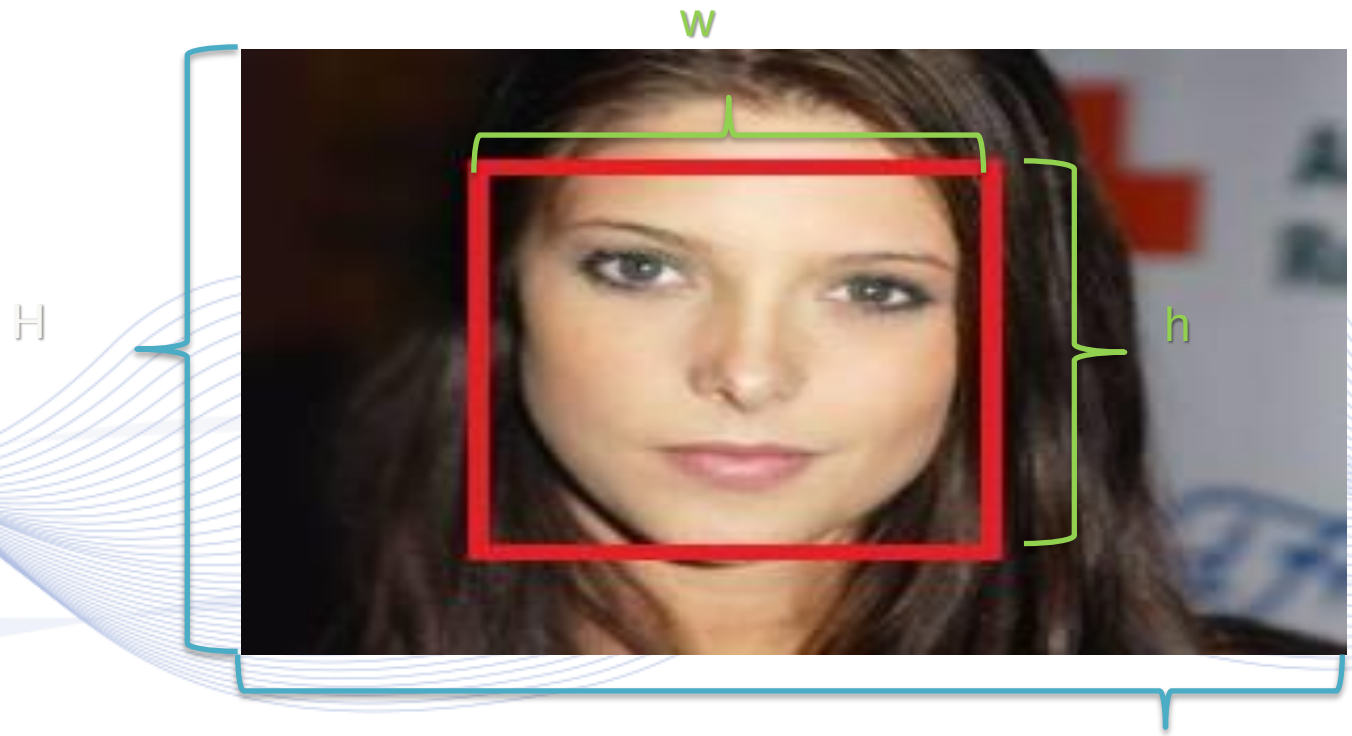


Classification steps.

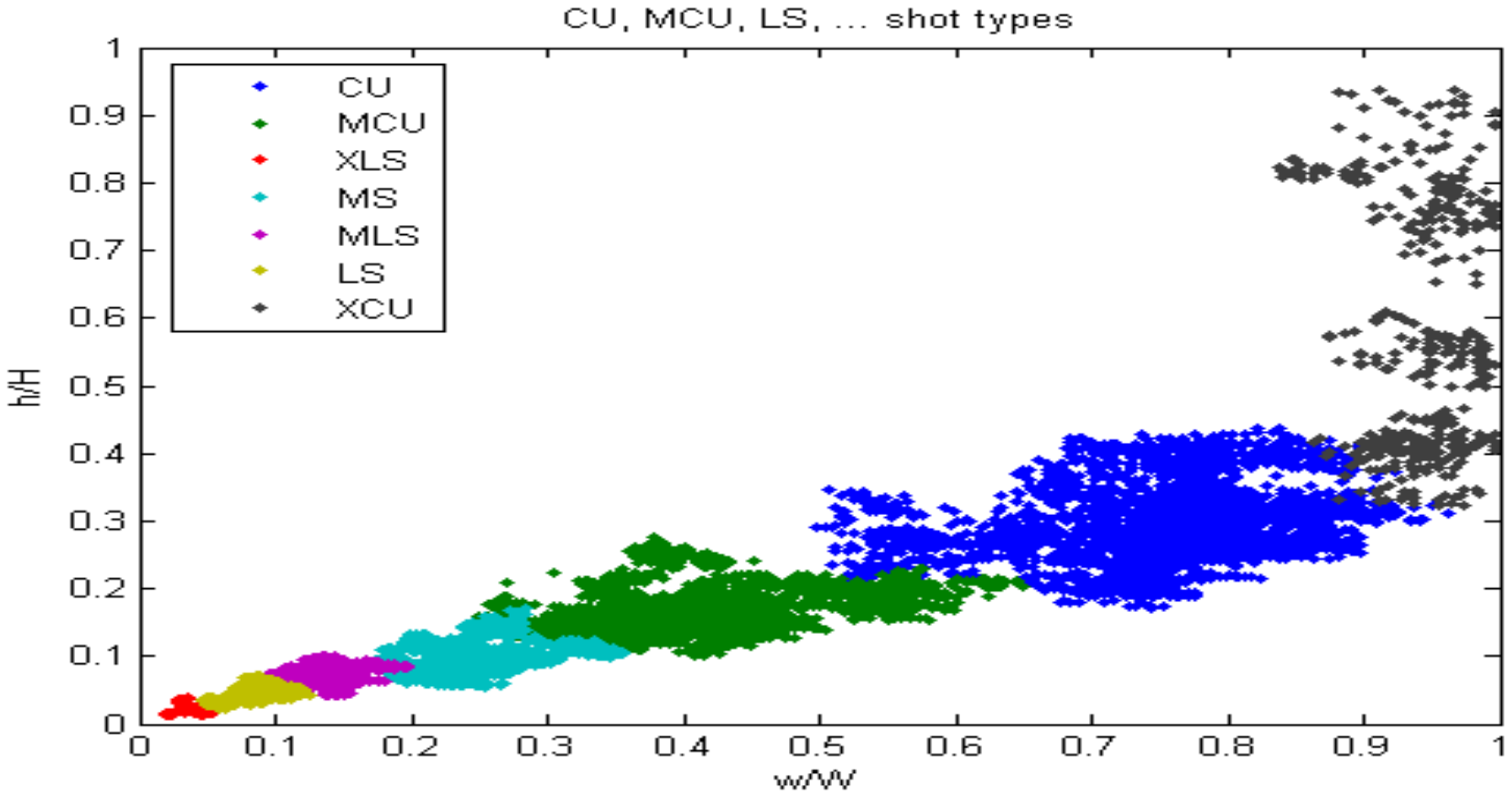
Shot Type Characterization

Facial image features:

- Face dimensions in relation to video frame dimensions.
- Feature Vector: $\mathbf{v} = \left[\frac{h}{H}, \frac{w}{W}\right]^T$.



Shot Type Characterization



Shot type characterization using two image features.

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

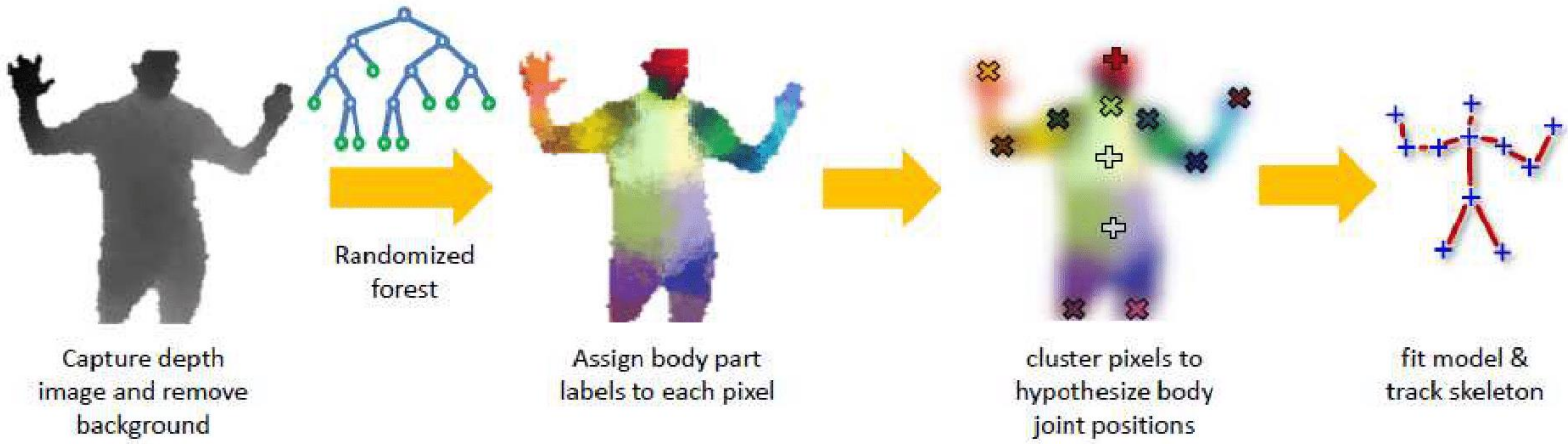
- **Human body posture and pose estimation**
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

2D Human Body Posture Estimation

- ***Posture*** means an intentionally or habitually assumed position of the joints.
- Human key posture points include rich 2D appearance, angular point, and multi-point autocorrelation.
- In order to achieve the 2D Posture Model, first we need to detect the human body.

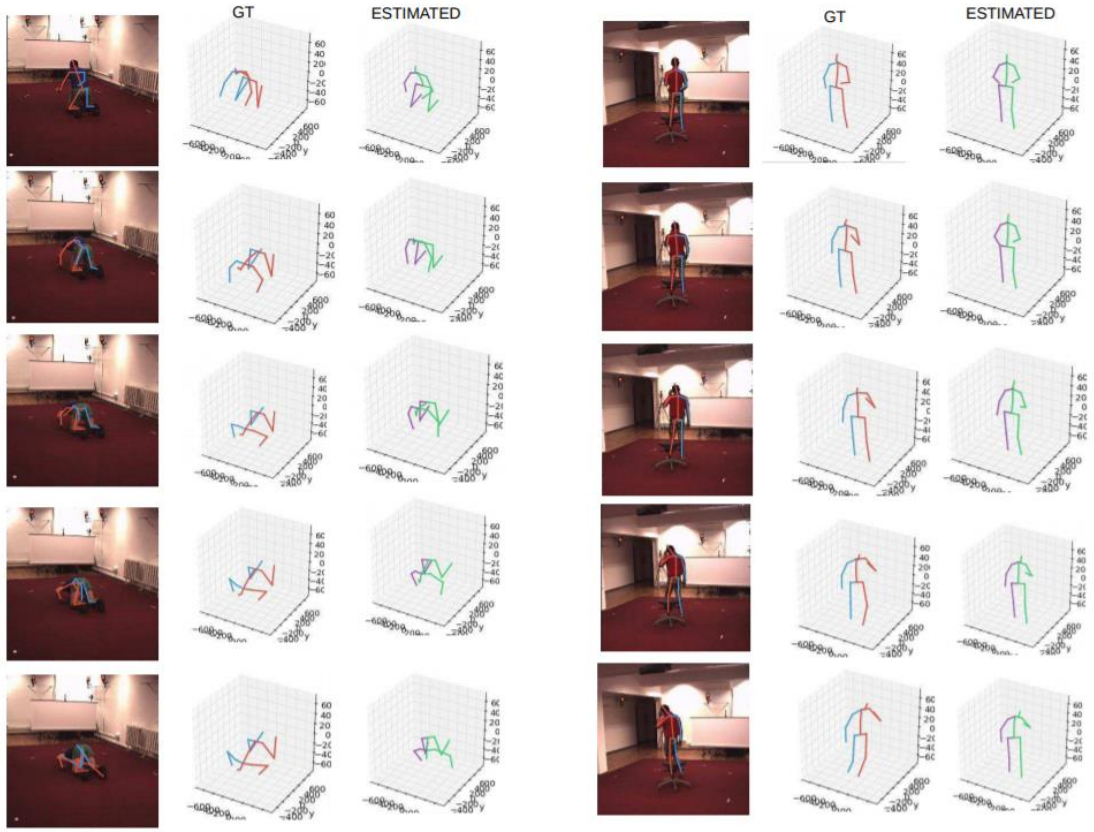
2D Human Body Posture Estimation

- Kinect sensors to get a skeleton sequence.



Kinect Skeleton Sequence. [SMIS2011]

3D Human Body Posture Estimation



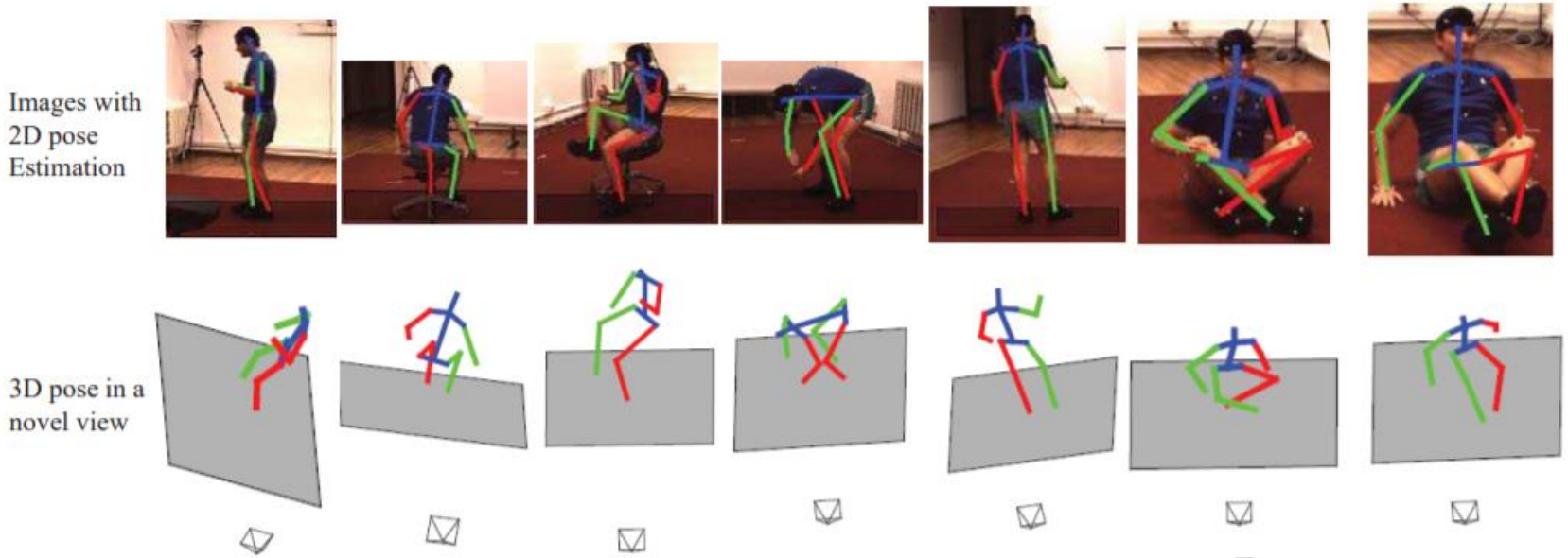
Video frames, ground truth and prediction. [RAY2018]

Human Body Pose Estimation

- ***Pose*** is the combination of *position* and *orientation* of an object vs the camera.
- It can be a considered as a regression problem
- Deep Learning Methods.

3D Human Body Pose Estimation

- ***Probabilities:***



Probabalistic Method Results.[CHEN2017]

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- **Activity/gesture recognition**
- Athlete Motion Analysis
- Soccer Video Analysis
- Semantic video content description/annotation

Action Recognition

Problem statement:

- To identify the action (elementary activity) of a person.
- Input: a single-view or multi-view video or a sequence of 3D human body models (or point clouds).
- Output: An action label belonging to a set of N_A action classes (walk, run, jump,…) for each frame or for the entire sequence.


run


walk


jump p.


jump f.


bend


sit

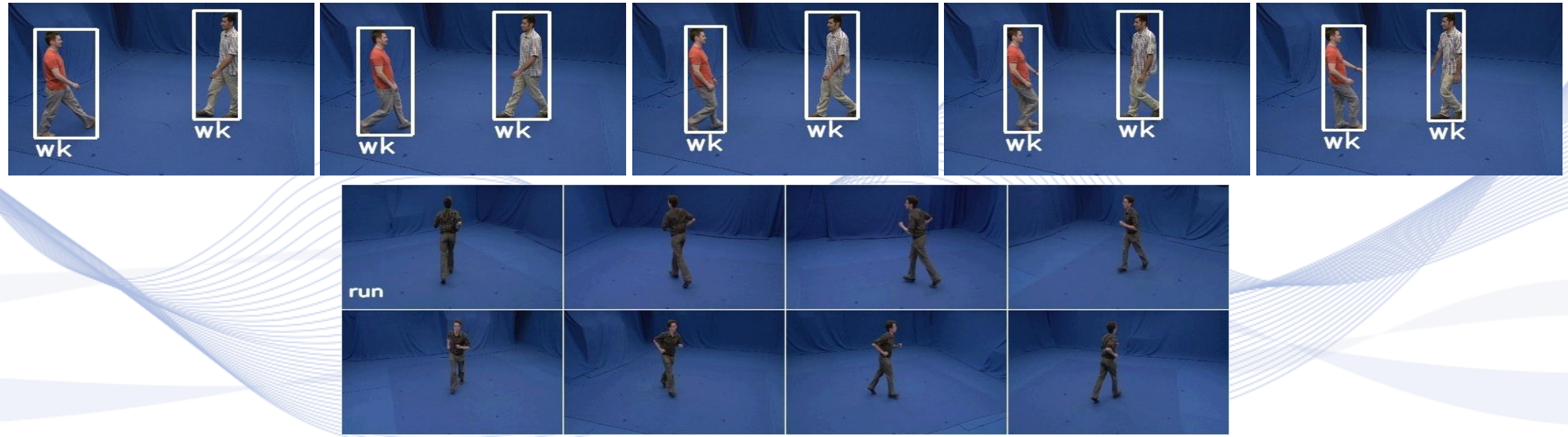

wave


fall

Action Recognition

Applications:

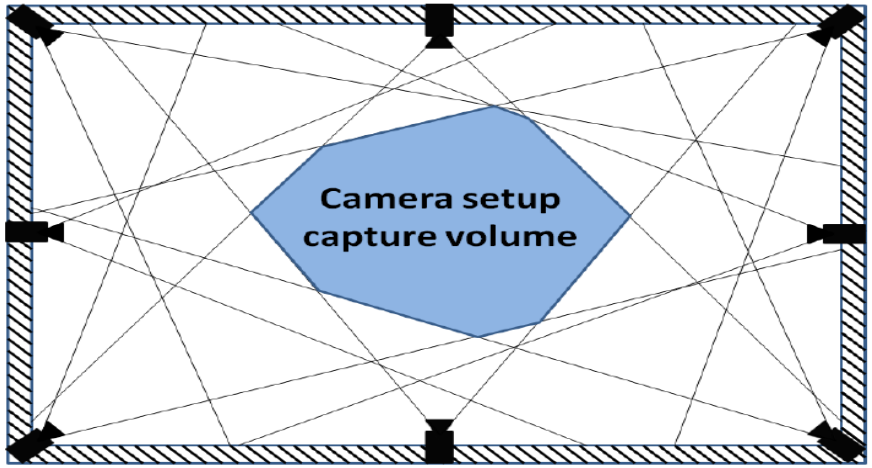
- Semantic video content description, indexing, retrieval.
- Video surveillance.
- Human – Computer Interaction (HCI).



Action Recognition Methods Categorization

- Single-view: methods utilizing one camera.
- Multi-view: methods utilizing multiple cameras forming a multi-camera setup.

An eight-view camera setup (N_C=8).



- Single-view methods are special cases of multi-view ones, i.e., for N_C=1.

Action Recognition on Video Data

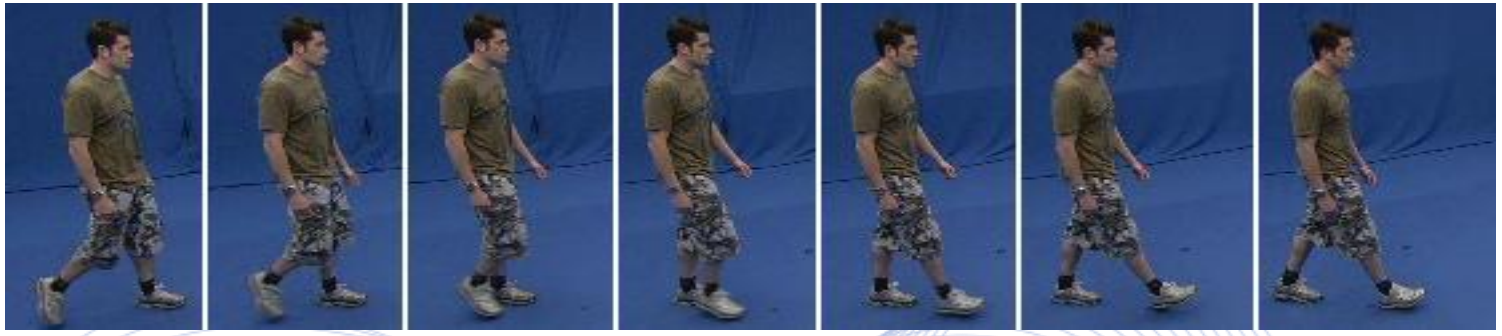
- Input feature vectors: binary human body images resulting from coarse body segmentation on each video frame.

run walk jump p. jump f. bend

- Segmentation techniques: background subtraction, chroma keying, motion detection.

Action Description

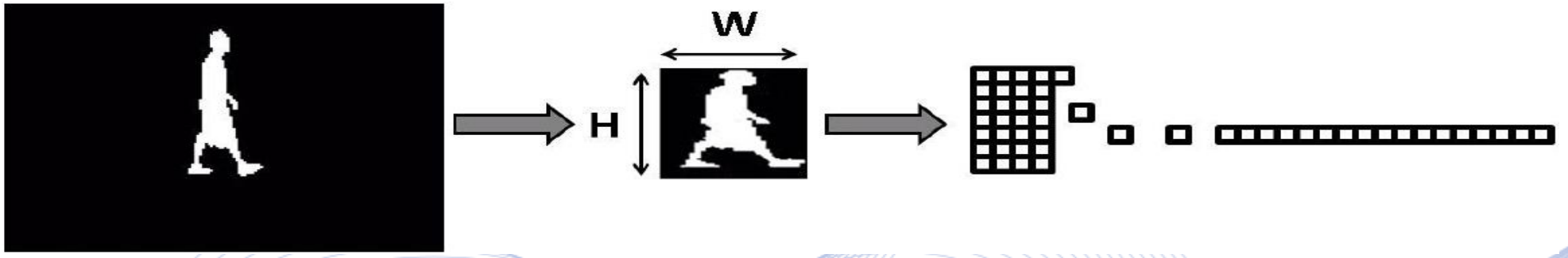
- A series of successive human body poses.



- Human body poses are represented by binary posture images.

Action Representation

- Posture vectors creation.



Binary human body image

Posture image of fixed size

Posture vector

$$\mathbf{p}_{ij} \in \mathbb{R}^{HW}$$

action video index

frame index

Action Representation

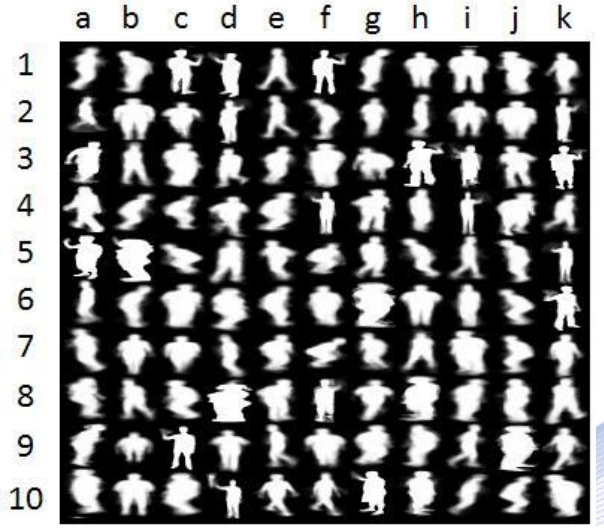
Dyneme-based representative human body poses:

- Dynemes calculation: Cluster all the training posture vectors $\mathbf{p}_{ij}$ in D clusters without exploiting the available action labels.
- Clustering techniques:
  - K-Means.
  - Self Organizing Map (SOM).
- Dynemes can be considered as representative human body poses.

Dynemes Calculation

- K-Means: a fast clustering technique minimizing the intra-cluster variance.

  D = 110 dynemes resulted by clustering the posture vectors of the i3DPost eight-view database

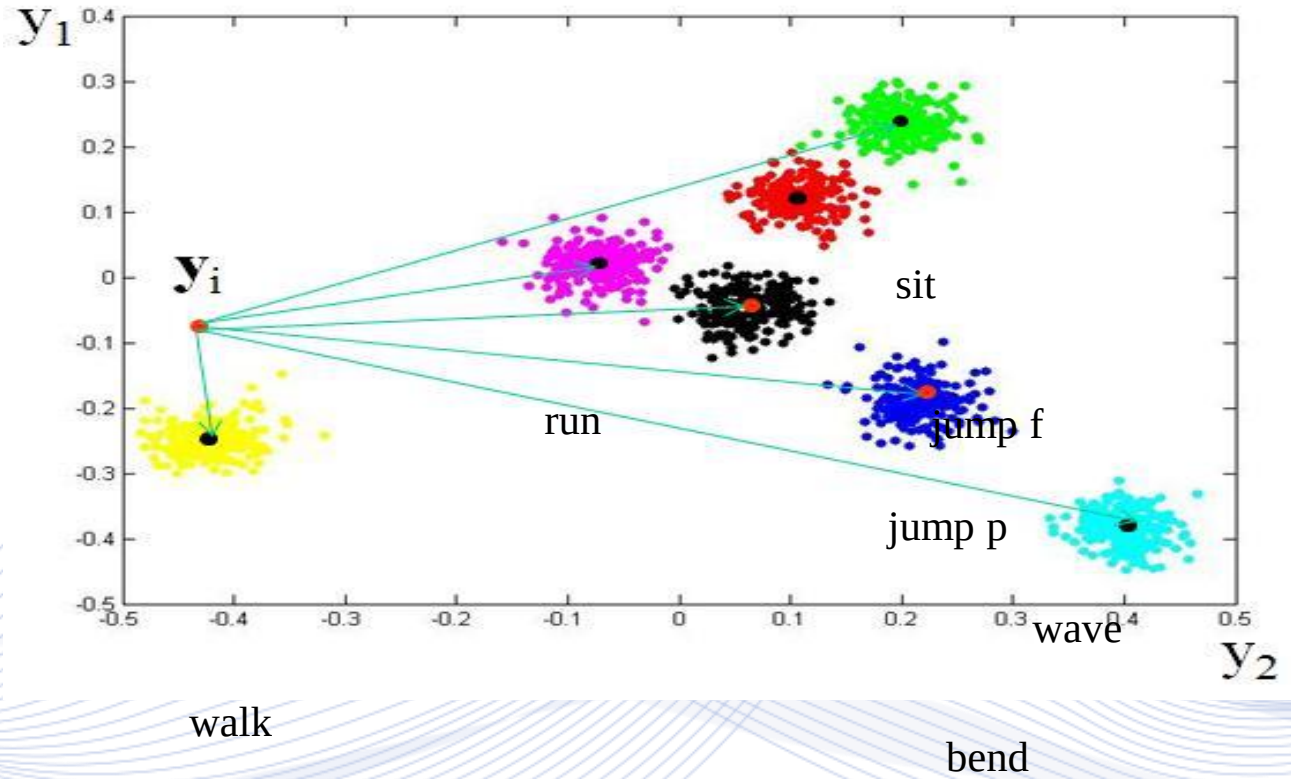


Dynemes are evaluated as the mean vectors of the resulting clusters (cluster centres) $v_k \in \mathbb{R}^{HW}$

Action Classification

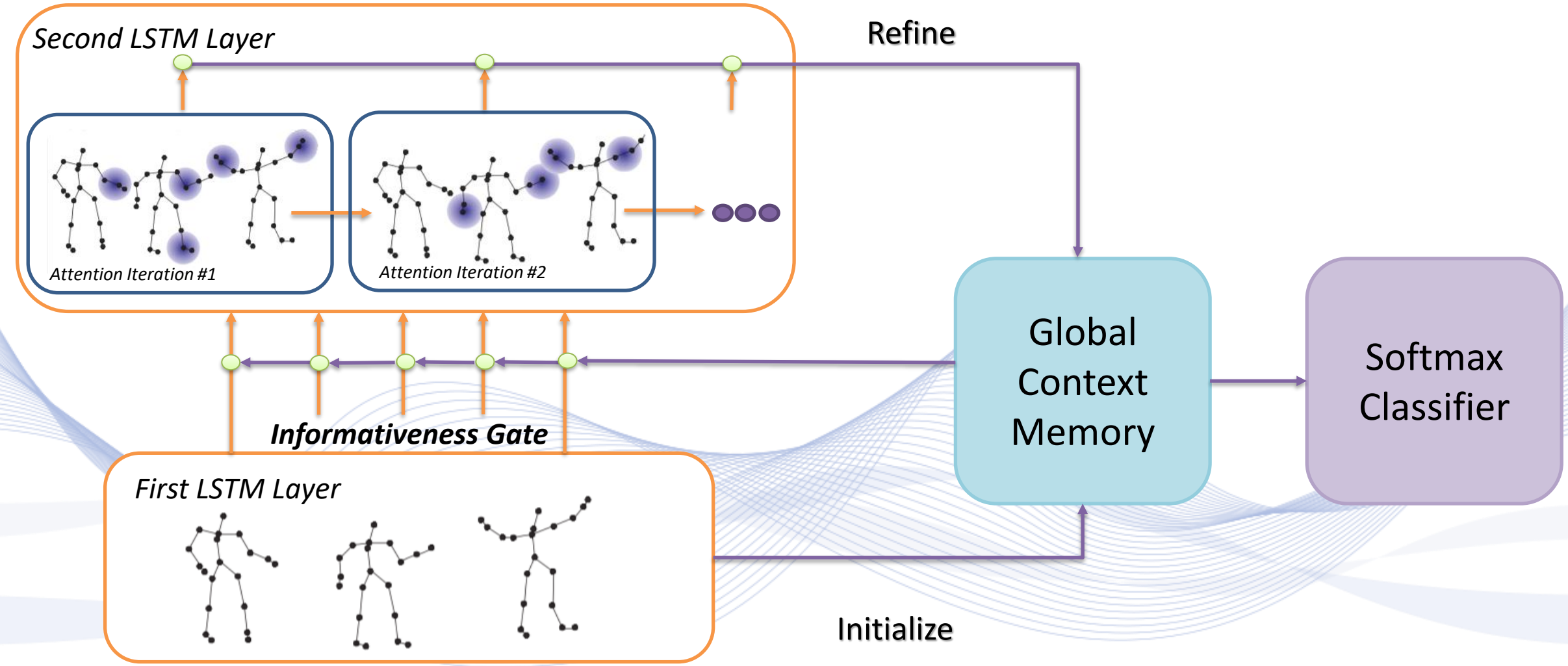
- Each action video is represented by the corresponding action vector.
- Various representation techniques can be used:
  - 2D or 3D CNN feature vectors
  - Dyneme-based feature vectors.
- Action classification is reduced to the corresponding action vector classification problem.
- Vector classification:
  - Deep Neural network techniques (2D/3D CNNs)
  - Graph Convolution Networks
  - Dimensionality reduction based classification.

Action Classification



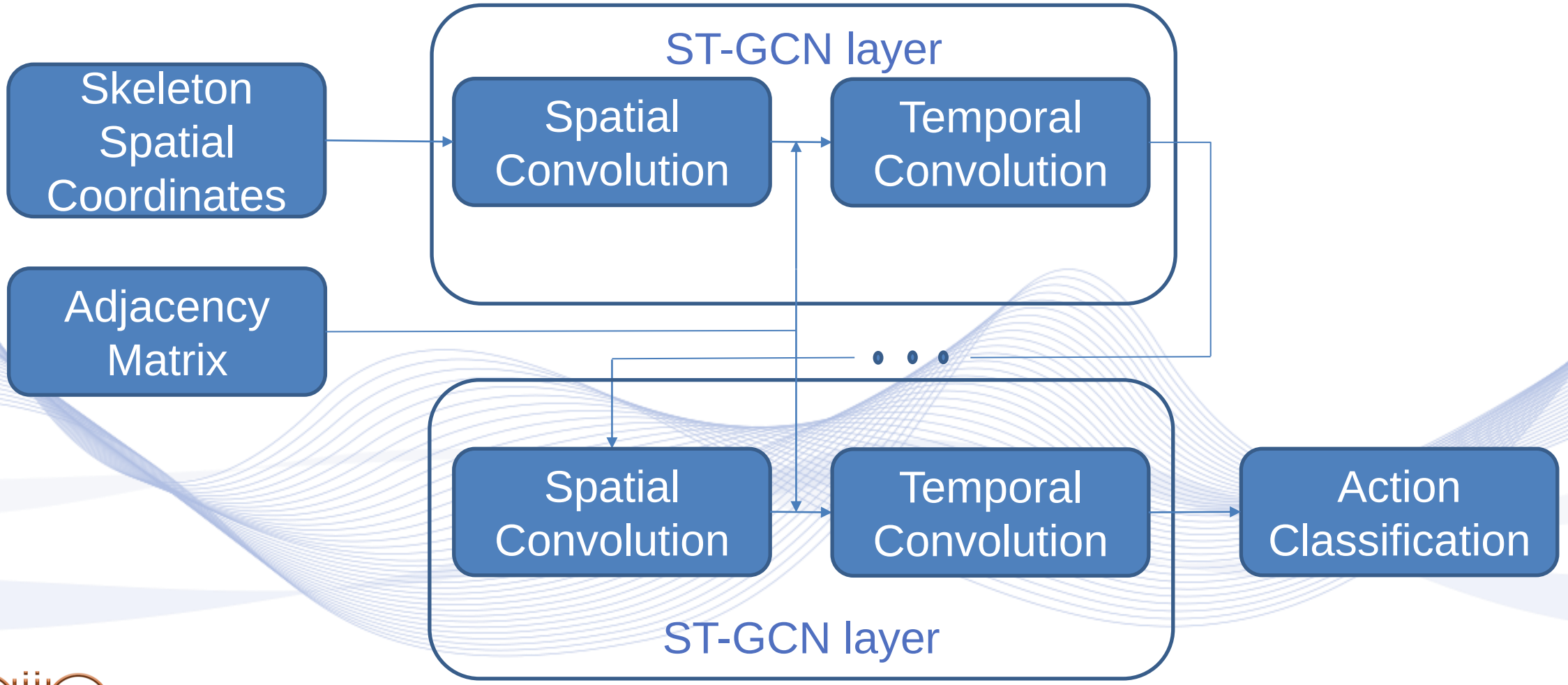
Classification based on the smallest Euclidean distance from all the mean action class vectors.

Neural Action Recognition



LSTM skeleton-based human action recognition.

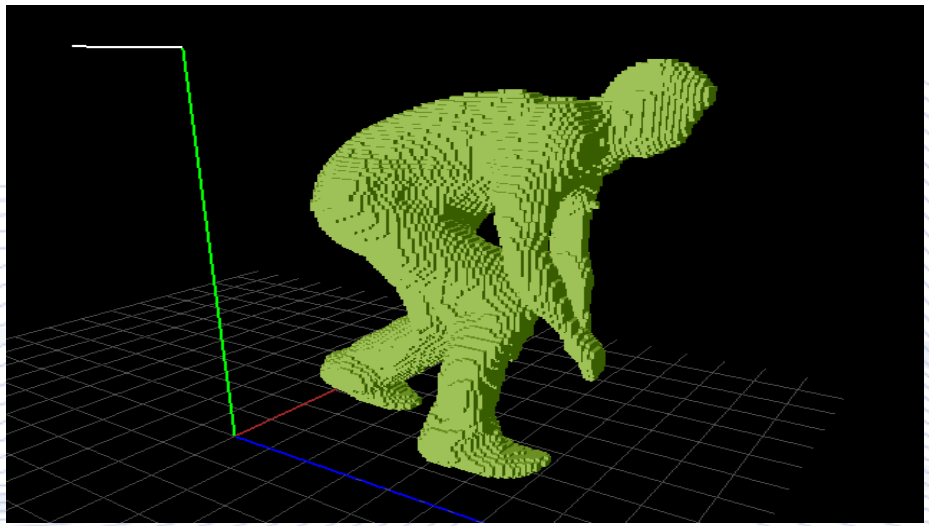
Spatio-Temporal GCN Action Recognition



3D Action Recognition

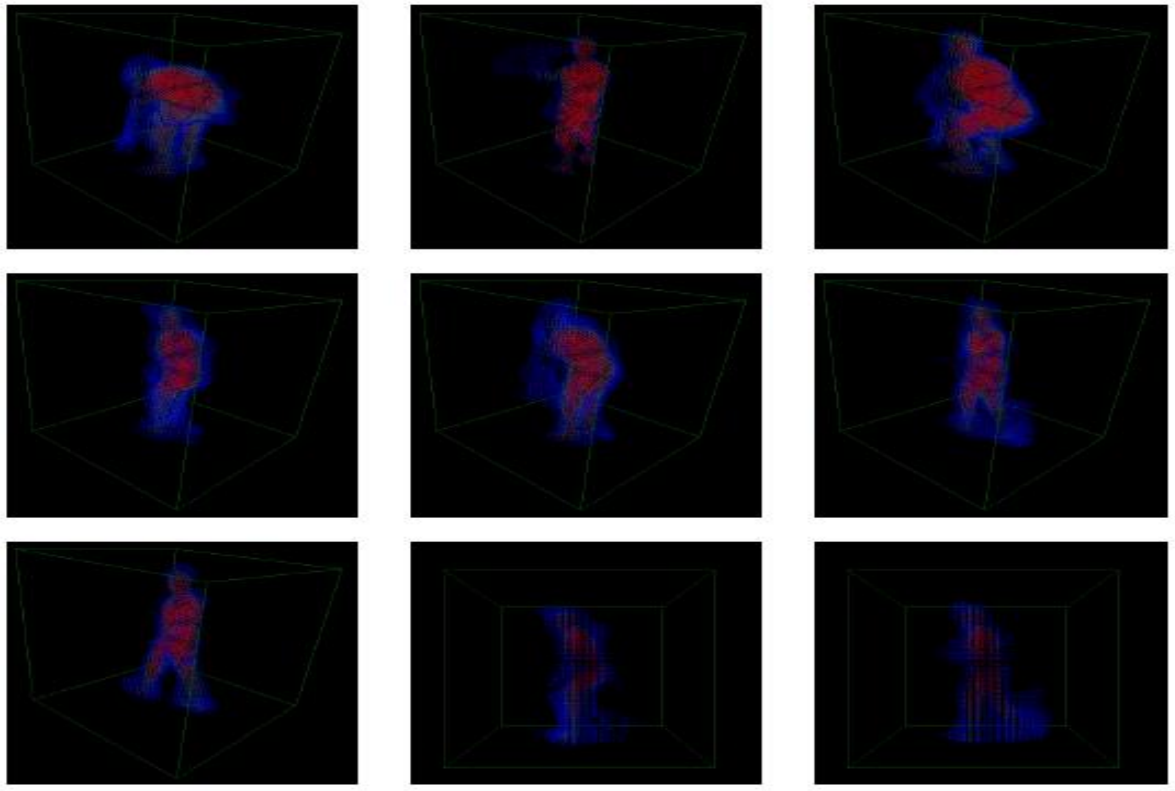
Action recognition on 3D data

- Extension of the video-based activity recognition algorithm
- Input to the algorithm: binary voxel-based representation of frames



3D Action Recognition

- 3D action characterization based on 3D “dynemes” (representative poses) derived through clustering along with LDA.



Action Recognition Examples

Action Recognition Examples

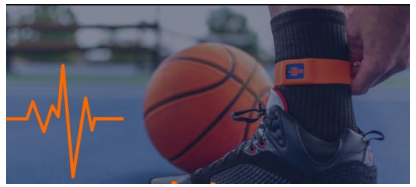
Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- **Athlete Motion Analysis**
- Soccer Video Analysis
- Semantic video content description/annotation

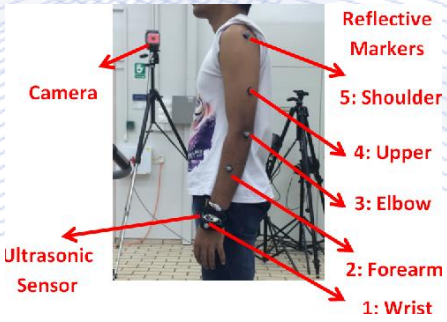
Athlete Motion Analysis

- Cameras
- Mechanical trackers

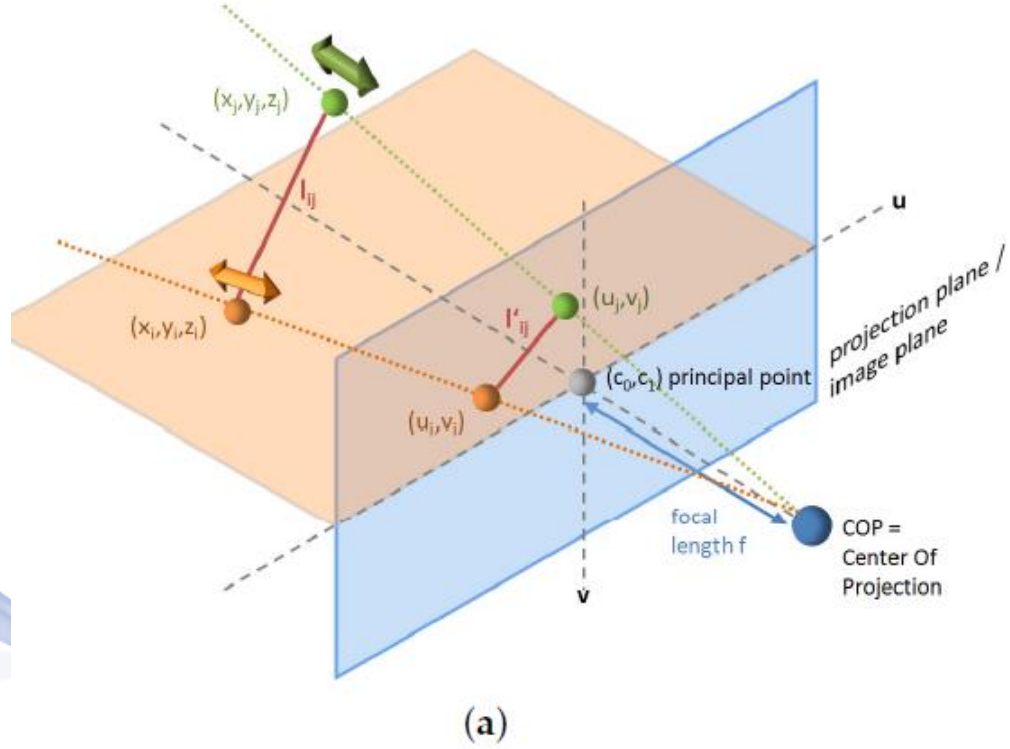


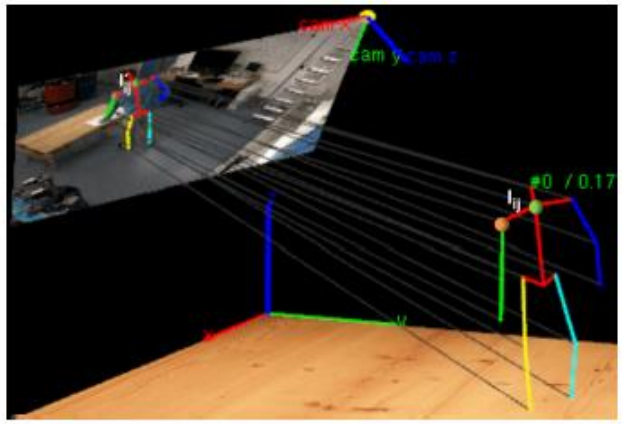
- Inertial trackers

- Electromagnetic devises
- Ultrasonic Trackers



Athlete Motion Analysis





3D Athlete Body Motion Estimation.

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- **Soccer Video Analysis**
- Semantic video content description/annotation

Soccer Video Analysis

- Background subtraction is required for objects detection and tracking

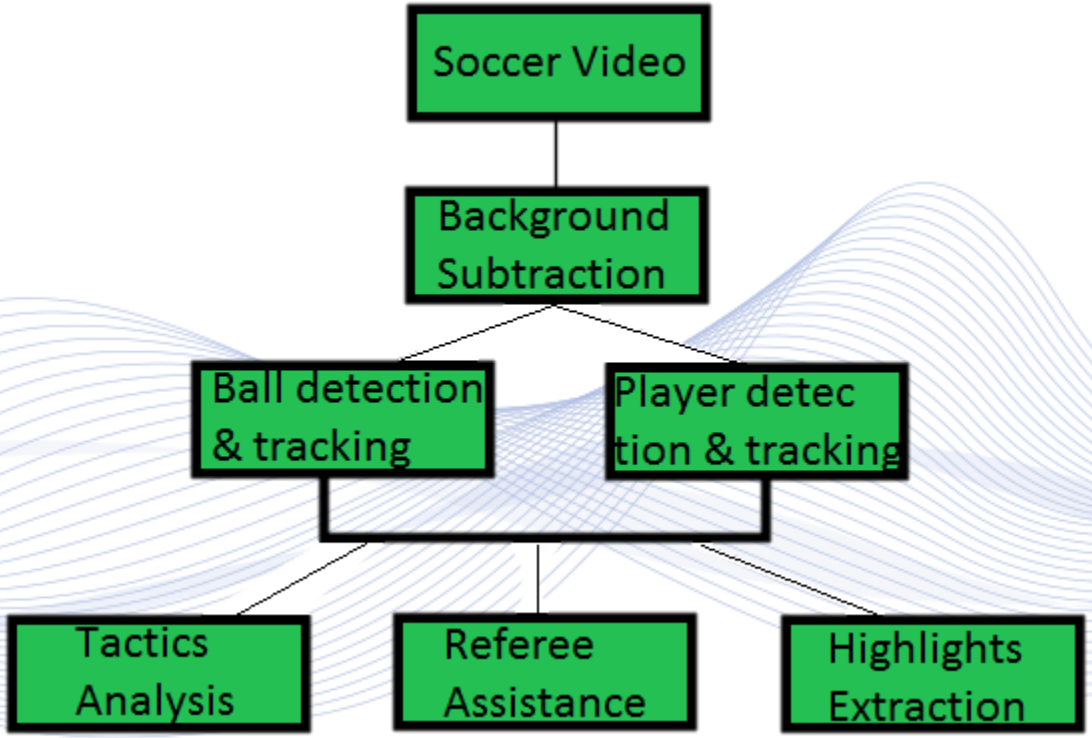
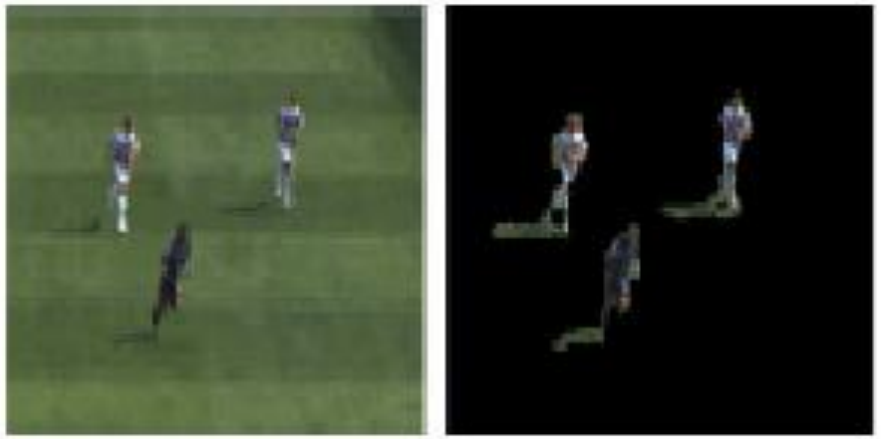


Figure 1: Video Analysis Process

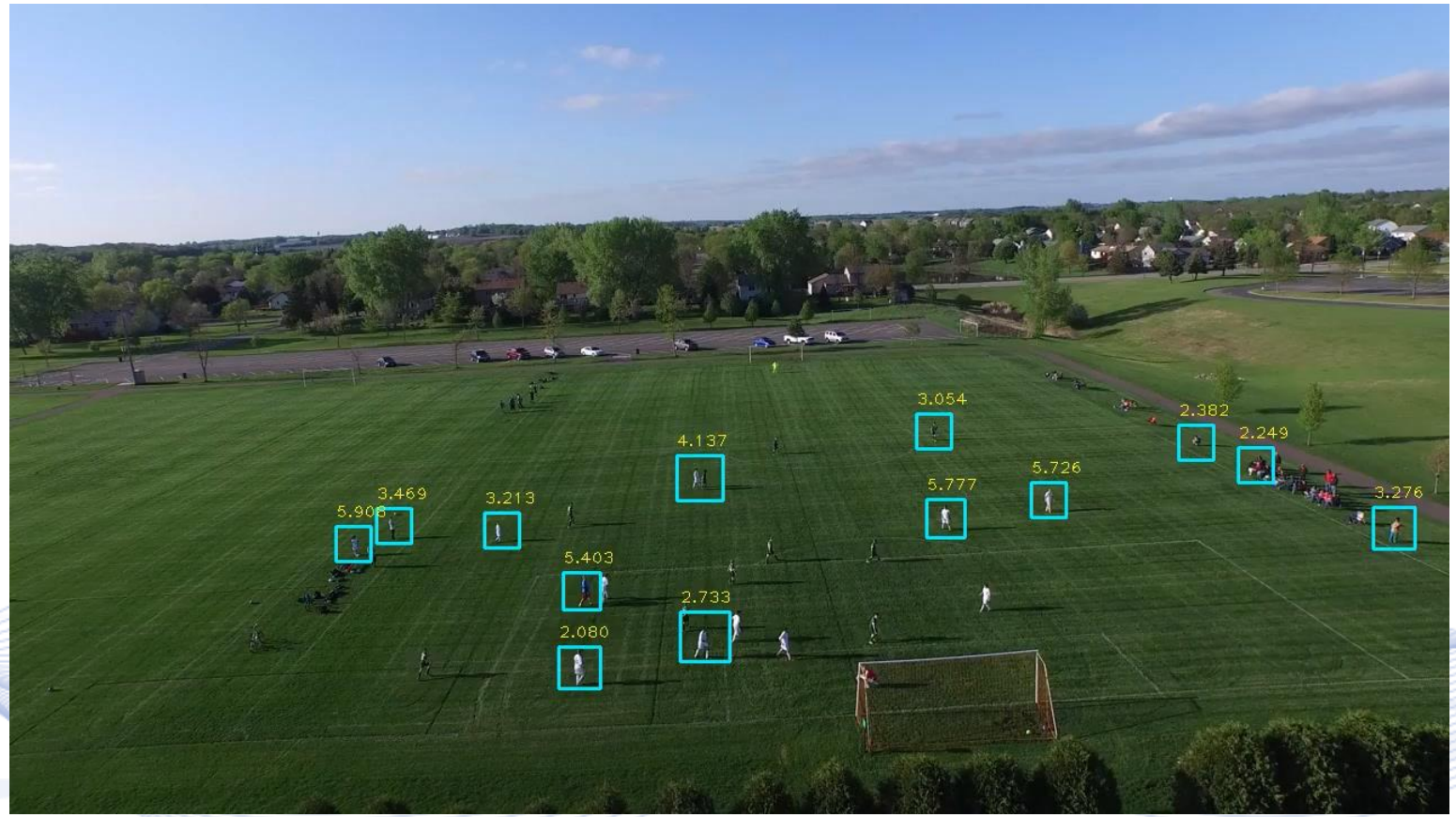
Soccer Video Analysis

- Segmenting each frame into background and foreground.
- Background corresponds to the playfield.
- Binarized image: 0 for playfield, 1 for moving objects.



Playfield Detection.

Soccer Video Analysis



Tracking soccer players.

Soccer Video Analysis

- ***Tactic Pattern Analysis***: Use of the extracted information (ball segments, dribbling segments, active play regions) to classify the attack events.

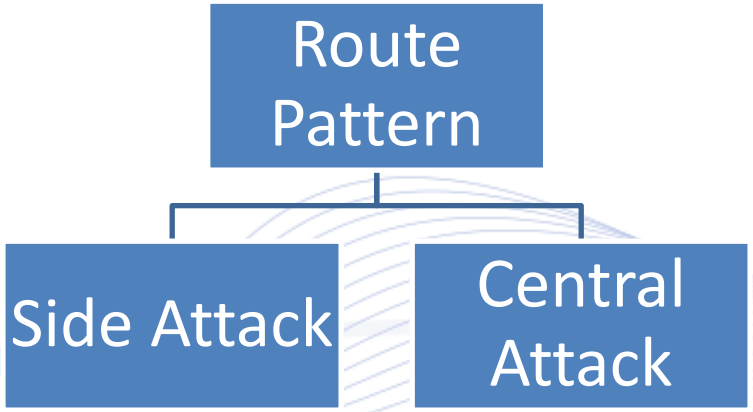


Figure 10: Route Pattern

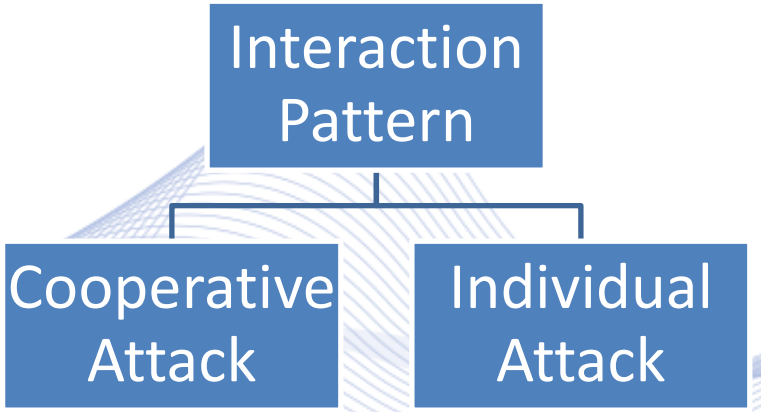


Figure 11: Interaction Pattern

- Route pattern and interaction pattern are complementary information.

Human Centered Computing

- Semantic Video Content Analysis
- Face/object detection and tracking
- Multiview human detection
- Face detection obfuscation
- Face recognition
- Face clustering
- Label propagation on videos
- Face De-identification
- 3D face reconstruction
- Facial expression recognition
- Facial feature detection
- Visual speech detection
- Shot type characterization

- Human body posture and pose estimation
- Activity/gesture recognition
- Athlete Motion Analysis
- Soccer Video Analysis
- **Semantic video content description/ annotation**

MPEG-7 standard

- Multimedia content semantic analysis results should be stored in a standardized, consistent & structured way

  Reasons: exchange of information, ingestion in 3D content databases/archives/MAM systems.
- MPEG-7 standard defines a description framework for handling multimedia content annotation and description
- A considerable amount of effort has been invested over the last years to improve MPEG-7 ability to deal with semantic content description, resulting in various MPEG-7 profiles

Audio-Visual Description Profile (AVDP)

- Audio-Visual Description Profile (AVDP) is an MPEG-7 profile introduced to describe results of multimedia analysis algorithms:

  Person identification, genre detection, keyframe extraction, speech recognition etc.

  Several results can be described in multiple timelines.
- Recently standardized from the MPEG Committee.
- Descriptions are in XML format, following the AVDP Schema Definition (XSD).

Audio-Visual Description Profile (AVDP)

- AVDP defines a subset of the MPEG-7 description tools needed for storing audiovisual content analysis results.
- AVDP also defines the semantics of some MPEG-7 description tools to suit audiovisual content description.
- AVDP was authored having in mind mainly single-view (“2D”) video & mono/stereo audio

Its semantics do not include ways for dealing with certain aspects of 3D video / multichannel audio content description.

Audio-Visual Description Profile (AVDP)

- A new framework for using the MPEG-7 AVDP profile for 3D content (e.g. 3DTV audiovisual content) description was proposed within European research project 3DTVS

  A subset of the description tools available in the AVDP has been selected

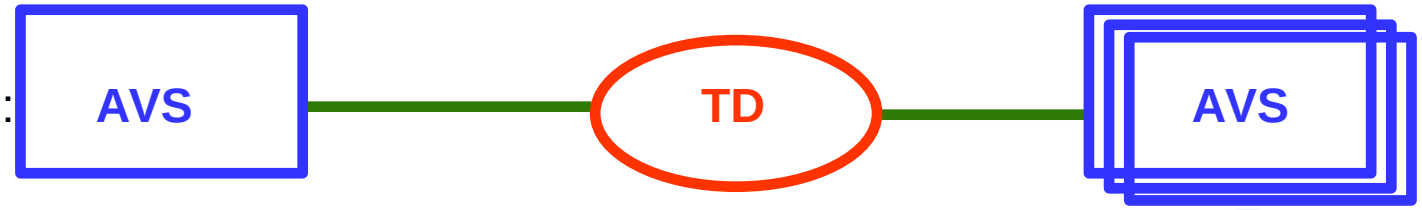
  A description procedure that can be used to store 3D content analysis results has been defined.

- The framework also details procedures for storing 2D content annotations, not foreseen in the AVDP initial guidelines

Useful AVDP/MPEG 7 Tools

- TemporalDecomposition type (TD): decomposes a VideoSegment type (VS), AudioVisualSegment type (AVS), or AudioSegment type into temporal segments.

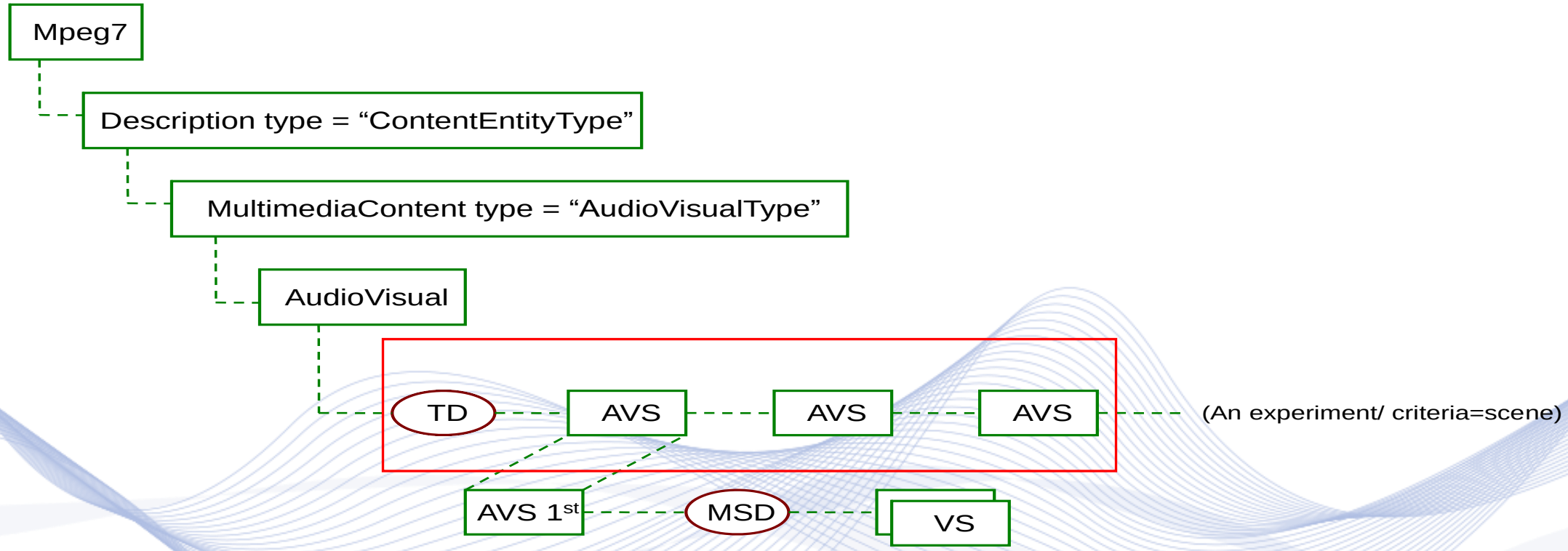
- Example:



movie

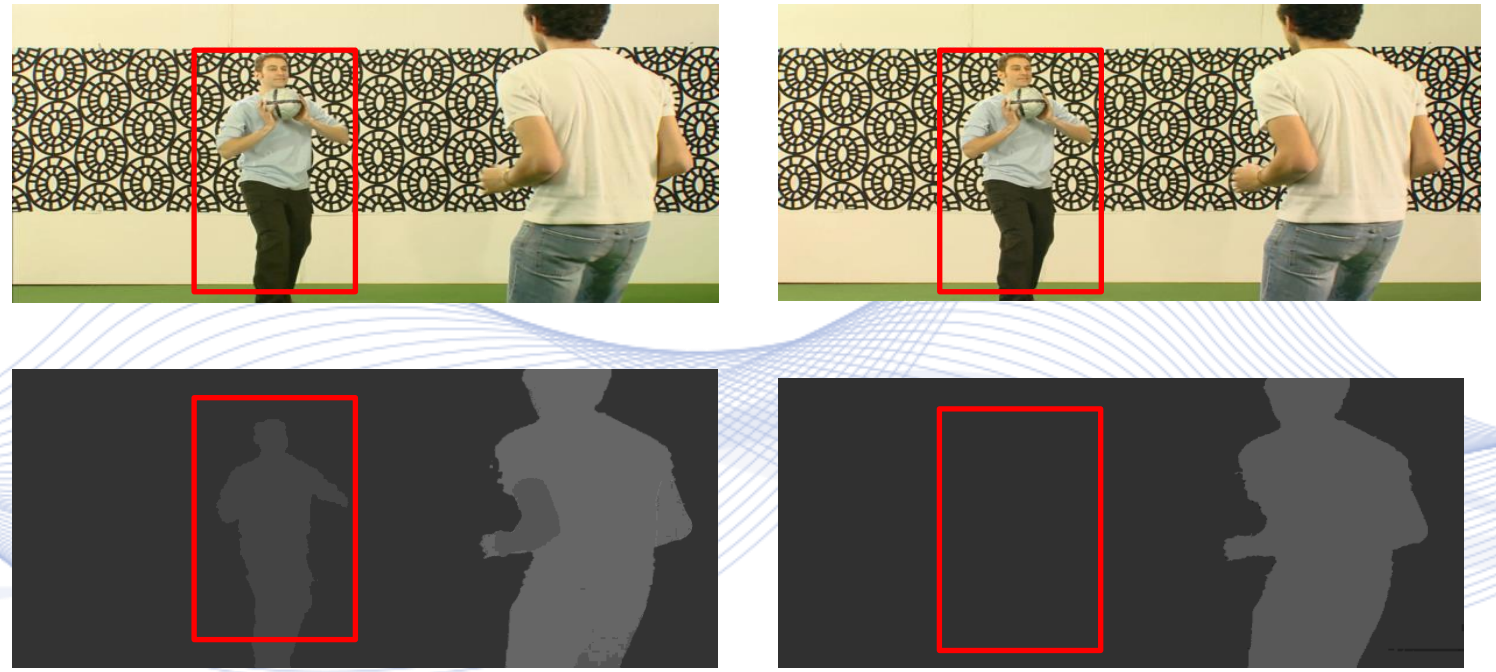
shots

Scene/Shot Boundaries Detection

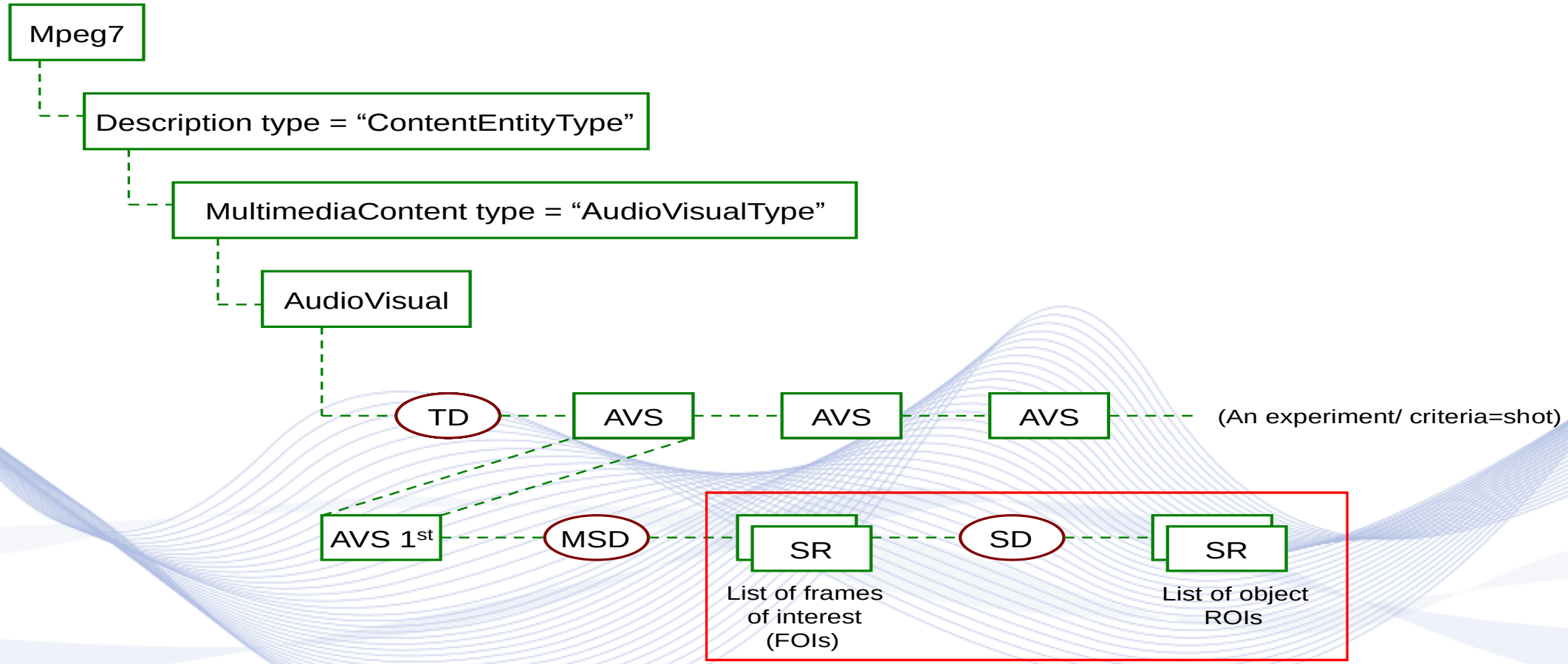


Human/Object Detection

- Human/object detection: localizing this entity in a frame.
- Generates a bounding box that includes the detected entity.
- Results are stored in StillRegion types (at least) in the channel(s) where detection takes place.

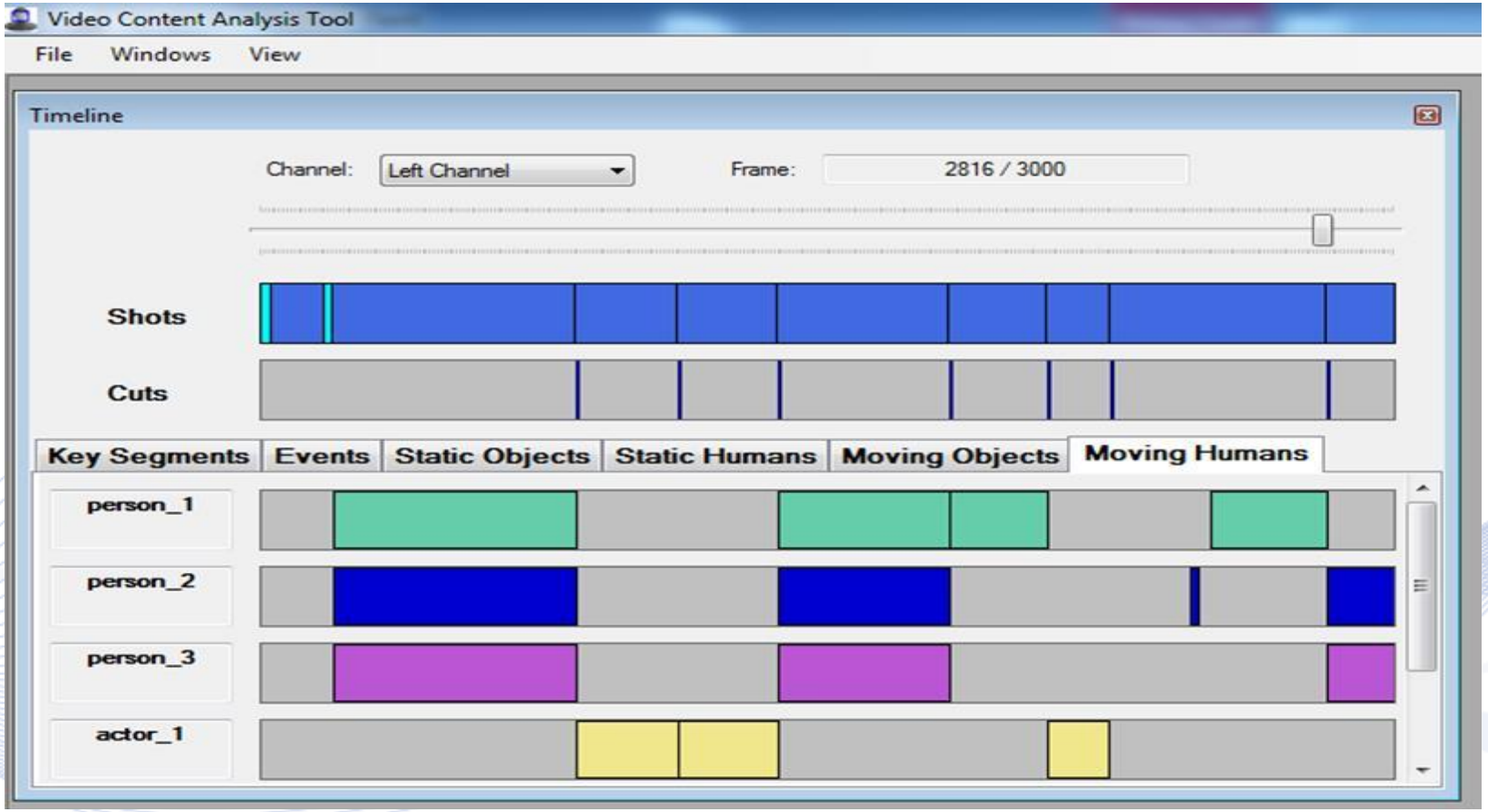


Human/Object Detection

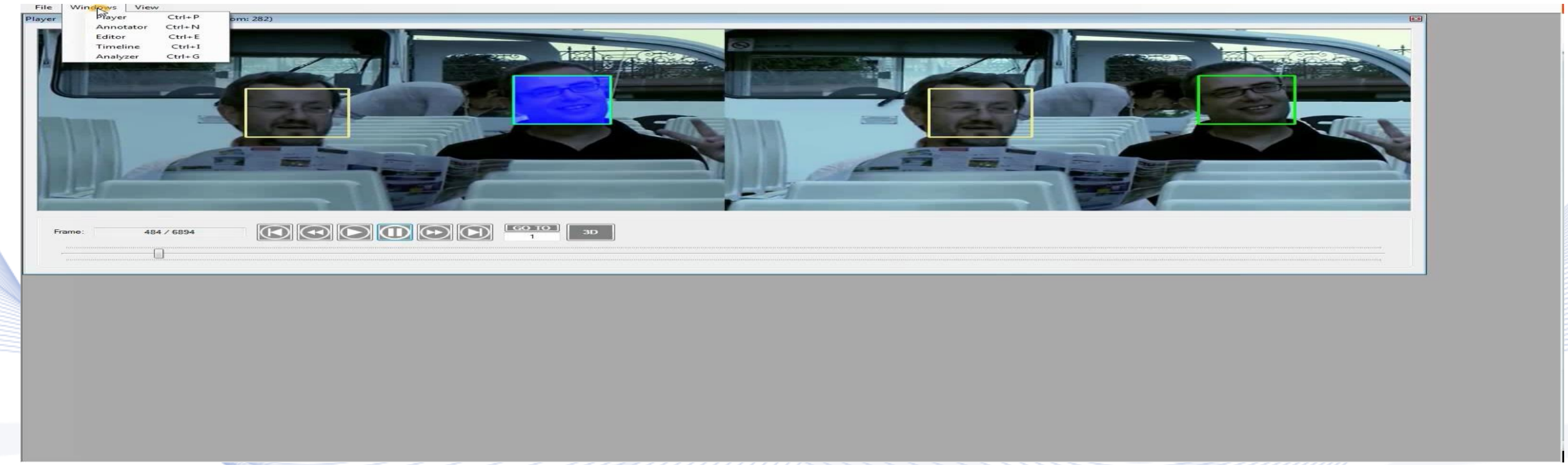


Timeline

- Shots & transitions
- Persons/objects appearances
- Events
- …



Timeline


File
Windows
View
Player
Player Ctrl+P
Annotator Ctrl+N
Editor Ctrl+E
Timeline Ctrl+I
Analyzer Ctrl+G
Frame:
484 / 6894
GO TO
1
3D

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr