UdeSantiago

Universidad Carlos III de Madrid

Some examples of the use of fusion in computer vision

Sergio A Velastin

sergio.velastin@ieee.org

IET

LACNEM 2015

Kingston University London
Digital Imaging Research Centre

Giaa

UdeSantiago

Universidad Carlos III de Madrid

Universidad de Santiago de Chile

- Original college founded 1842
- Oldest technical university in Chile
- #3 in Chile, #13 in Latin America
- Informatics: analysing large web data, medical imaging, affective computing, machine learning

giaa

Universidad Carlos III

- Within top 50 (37) in world's "under 50" group (25 years)
- Public university, Focuses on research
- Engineering, Social Sciences and Law
- Applied Artificial Intelligence group/Automatics: Data fusion, vision, agents, machine learning, automotive sensing (e.g. pedestrians)

Kings-ton (a place of Kings!)





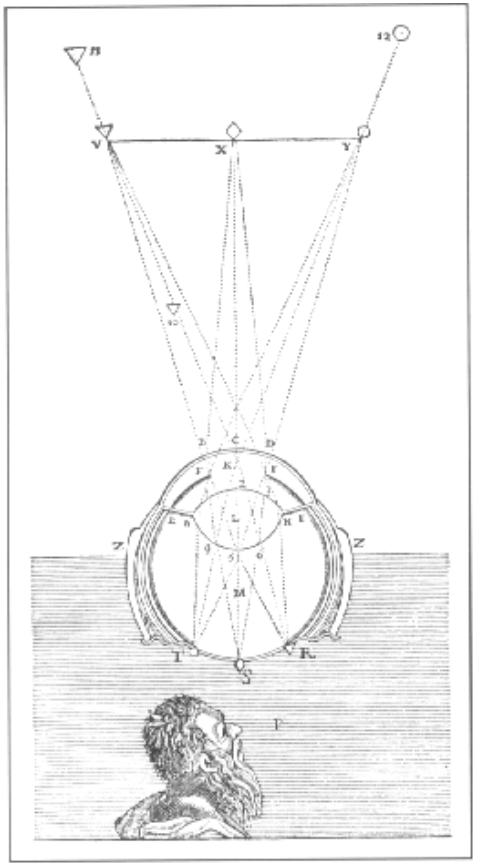
Royal Kingston

- Digital Imaging Research Centre (surveillance, medical imaging, human body motion, robotics)

Outline

- Introduction: Computer Vision and Fusion
- Registration
- Multi camera tracking
- High dimensionality
- Action recognition
- Context
- Final Remarks

Introduction



- "It is by looking and seeing that we come to know what is where in the world"

David Marr

(computational neuroscience)

Computer Vision

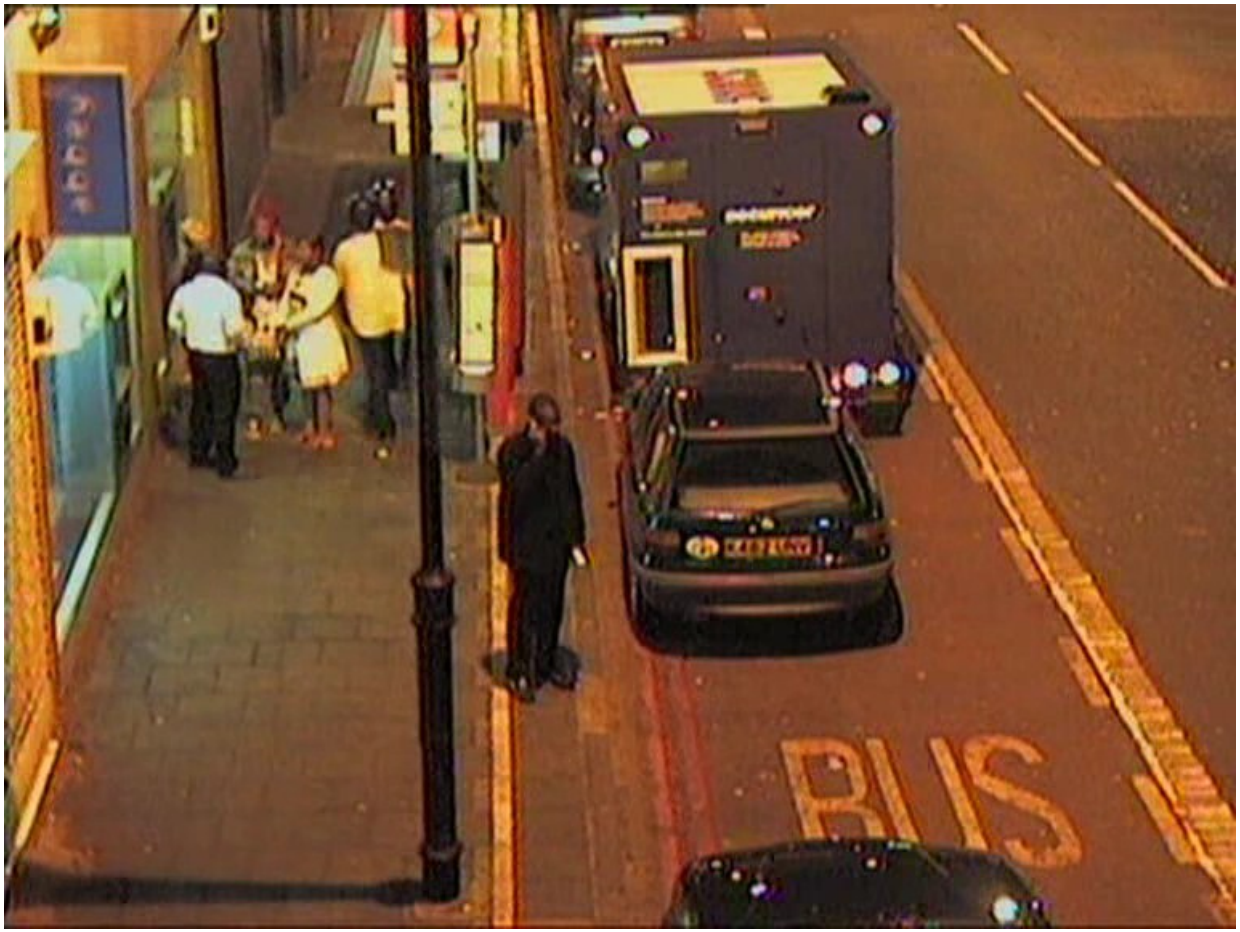
Turing's test …., generate textual **narratives**

So ....

- In many cases, computer vision is mostly about converting visual data to temporal/spatial **narratives** …

- But not always, e.g. medical imaging enhancement, vision-guided navigation (even here, **interpretation** is best represented textually e.g. "benign tumour")

Is one picture worth 1000 words?



Fusion

- Combine data/information to obtain something better than what can be obtained separately
- In many computer vision problems:
  - Different "descriptors" (features) can be extracted from the same object image
  - Different instances of the same type of objects (e.g. people,cars) "look different"
  - Different instances of the same class of objects "behave" (temporally) differently
  - The same object seen from different views (e.g. different cameras) and it looks different
  - Any combination of the above (including all)
- **Holy grail**: capture (eg many sensors), fuse, understand, fuse, context, ….→ better narratives

This talk

- Will show some examples of where fusion is useful in computer vision, taken from various groups, **at the expense of detail**
- Typical application areas for computer vision:
  - Object detection
  - Action recognition
  - Tracking (people, vehicles, …)
  - Biometrics (face recognition, gait, fingerprints, iris, …)
  - Robotics
  - Ambient Assisted Living
  - Media "analytics" (e.g. automated commentary generation)
  - Medical imaging
  - Bio mechanics (sports training, rehabilitation)
  - …

Aspects to consider

- **Early** fusion (data level, feature level)?
  - Registration (common reference frame e.g. via homography through feature correspondence …)
  - Multiple descriptors
  - Generates large feature vectors: curse of dimensionality: Dimensionality reduction (implicit fusion?)
- **Late** fusion (label level, decision level, …)?
  - With Multiple sensors (e.g. multiple cameras) or multiple modes (e.g. MRI/CT in medical imaging)
  - Probability models
  - Combine classifiers (labellers)
  - Trackers (e.g. Kalman)
- **Hybrid** (not easy to get a good taxonomy to organise work in CV!)

UdeSantiago

Universidad Carlos III de Madrid

An iterative integrated framework for thermal-visible **registration**, sensor fusion and people tracking for video surveillance applications (CVIU 2011)

Atousa Torabi, Guillaume Massé,

Guillame-Alexandre Bilodeau

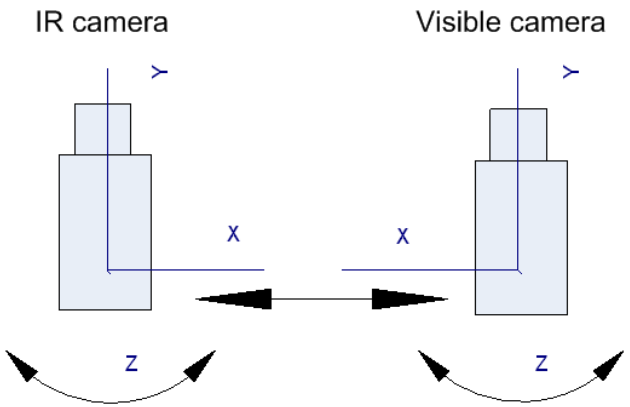
École Polytechnique de Montréal, Canada

POLYTECHNIQUE MONTRÉAL

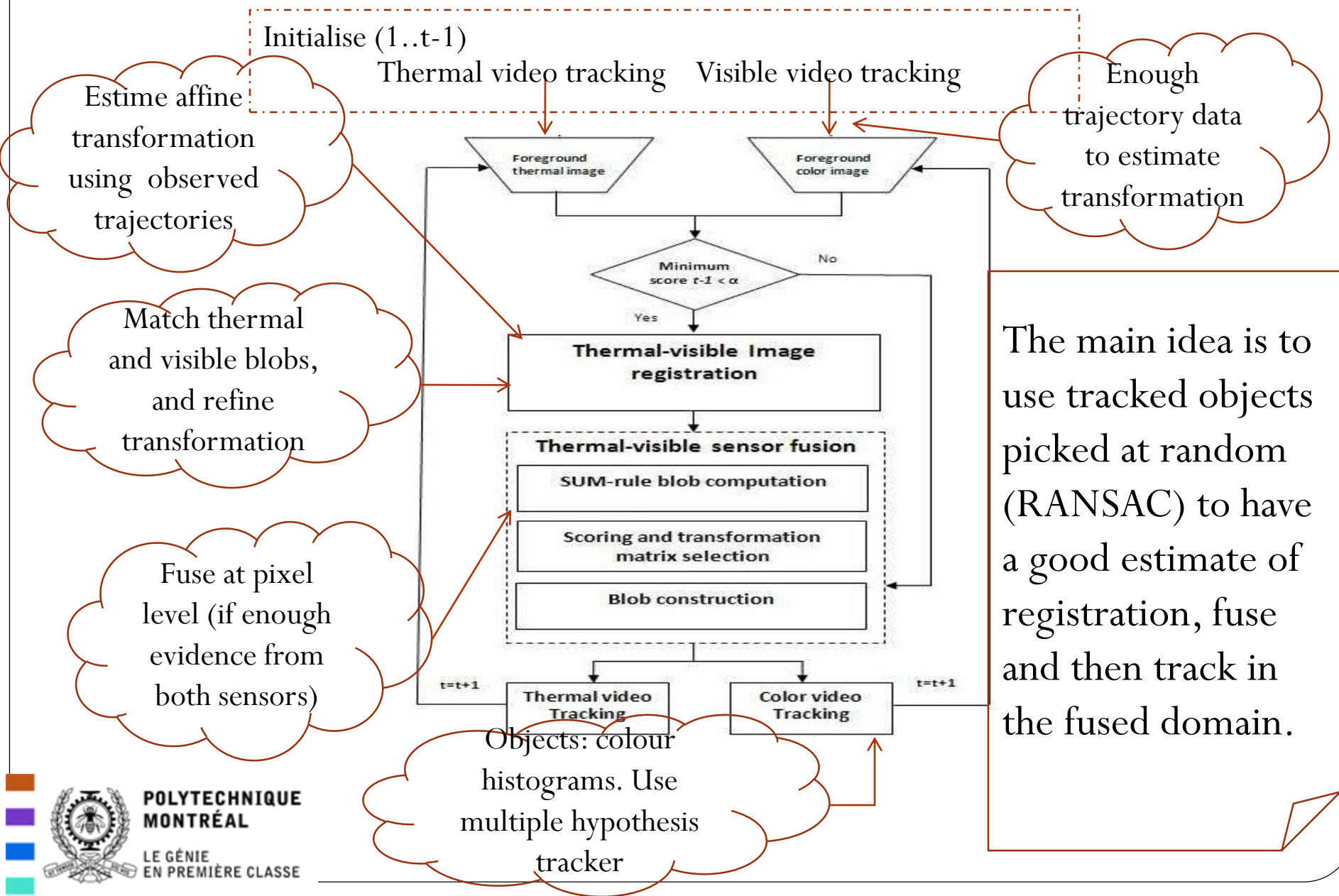
LE GÉNIE EN PREMIÈRE CLASSE

Camera setup

- Collocated thermal and visible cameras
- No explicit calibration
- Intersection of field of views (FOVs) of two cameras
- Cameras can be with different zooms
- Thermal and visible camera are synchronized



System Flowchart



Multi Camera Tracking

Problem: A wide area is monitored by multiple cameras (sometimes not overlapped). How to get a global view from the individual observations of each camera?

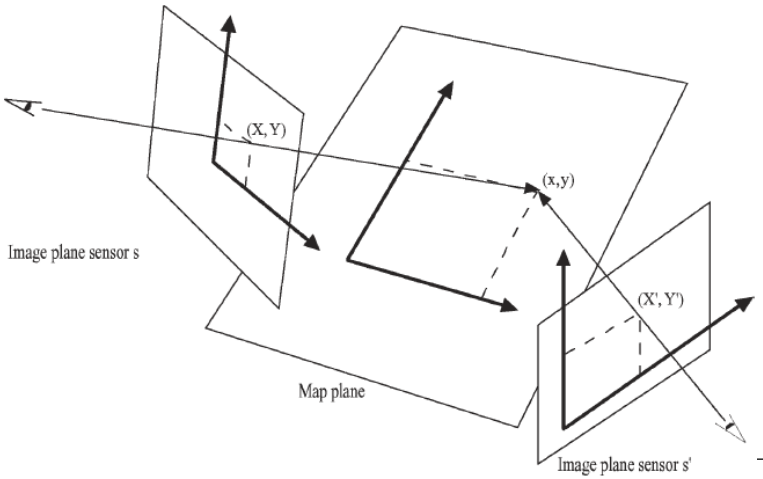
Possible approach: Registration on common ground plane (using observations), object tracking on that plane (fusing observations)

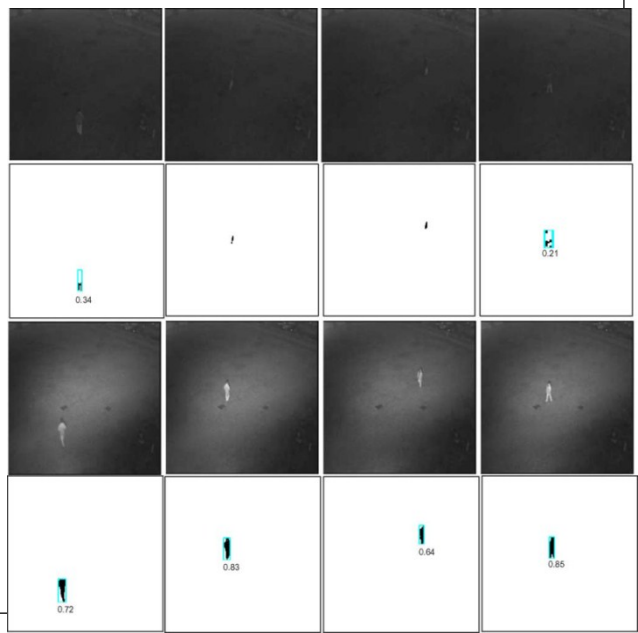
UNIVERSITÀ DEGLI STUDI DI UDINE

Quality-Based Fusion of Multiple Video Sensors

- Multisensor system for video surveillance
- Fusion of target's location from different sensors
- Map plane used as common reference system after **homographic** projection
- Metric estimates "quality" of the blob and adjusts the measurement covariance matrix accordingly

L. Snidaro, R. Niu, G. L. Foresti, and P. K. Varshney, "Quality based fusion of multiple video sensors for video surveillance", *IEEE Trans. System, Man, and Cybernetics Part B,* vol. 37, n°4, pp.1044-1051, August 2007.

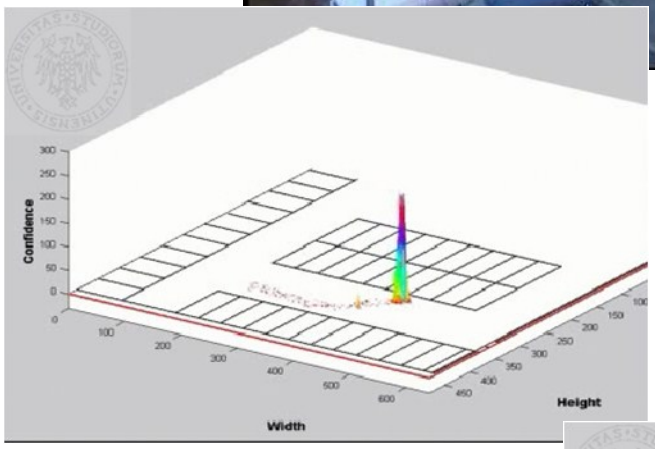


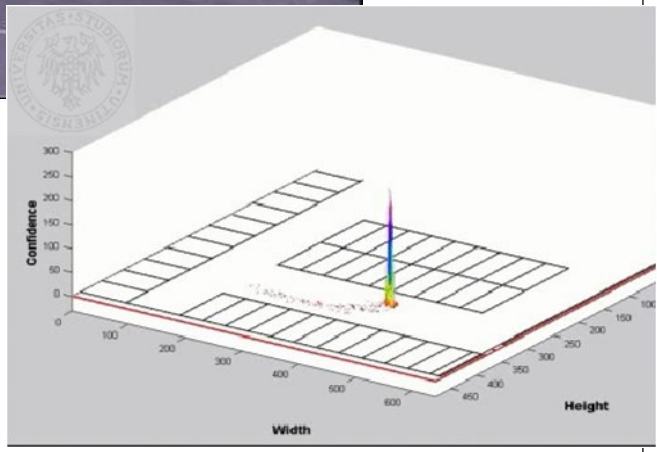


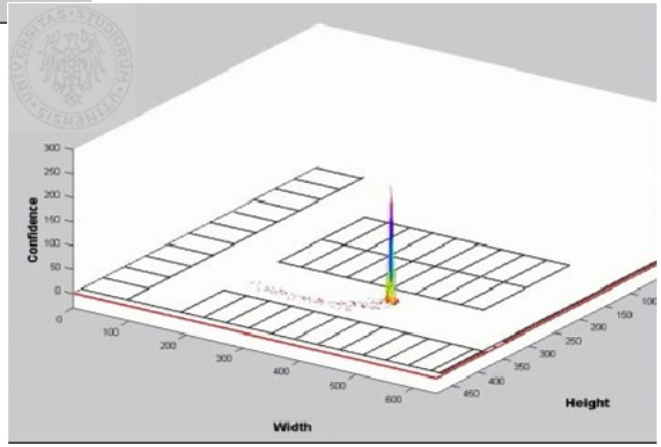
UNIVERSITÀ DEGLI STUDI DI UDINE

(1)

(2)

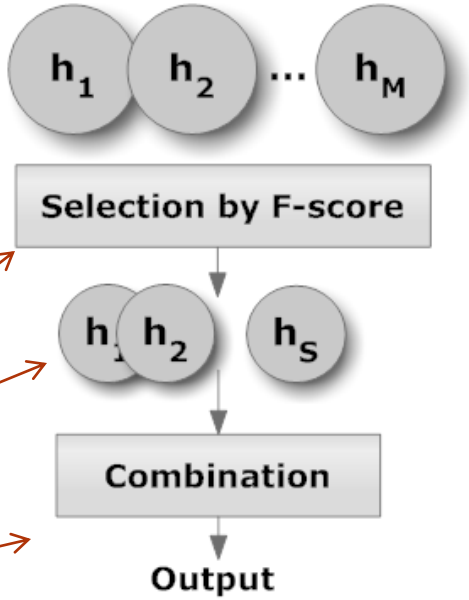




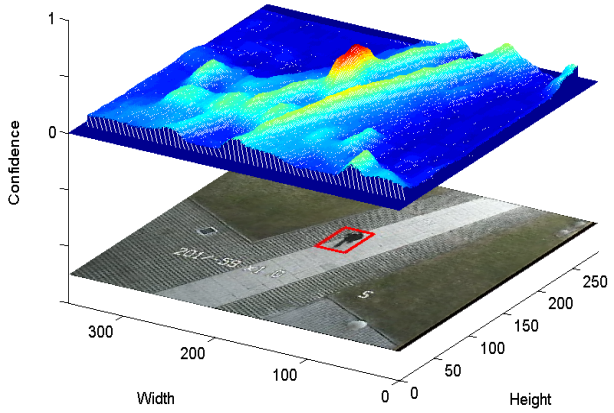


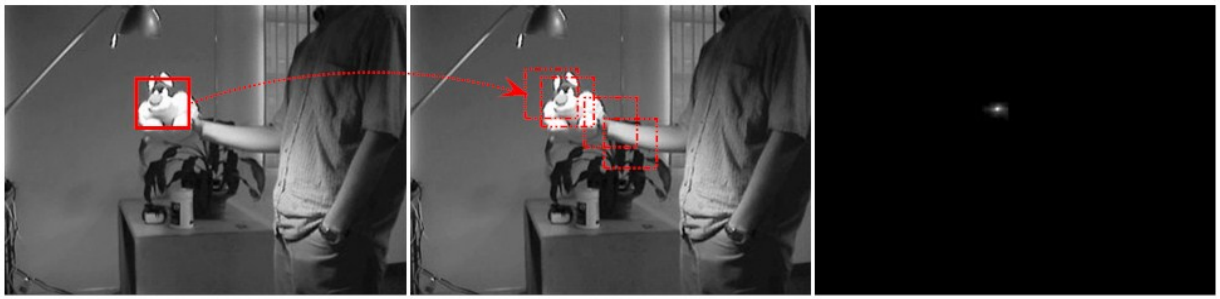
F-score based classifiers selection

- Fast, fuses classifiers (object vs. background)
- Exploits a performance evaluation metric (F-score) for classifier selection
- Uses selected classifiers to build an ensemble
- Tracks an object frame-by-frame using the ensemble



I. Visentini, L. Snidaro, G. L. Foresti, "Selecting classifiers by F-score for real-time video tracking", *Proceedings of the Thirteenth International Conference on Information Fusion*, Edinburgh, U.K, July 26-29, 2010 (**Fusion 2010 Best Student Paper Award runner up**).





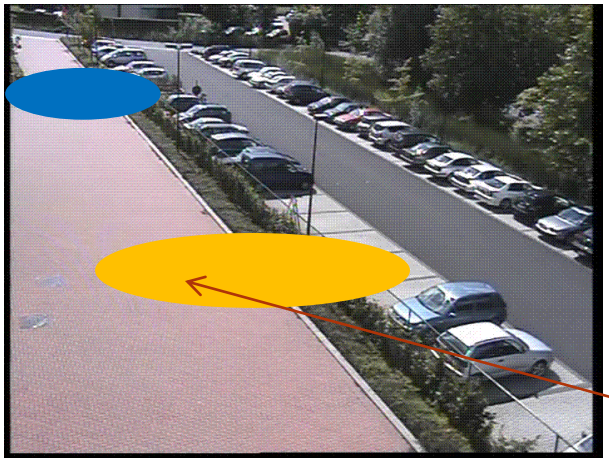
UdeSantiago

Universidad Carlos III de Madrid

Object correspondence and tracking across cameras with overlapping views

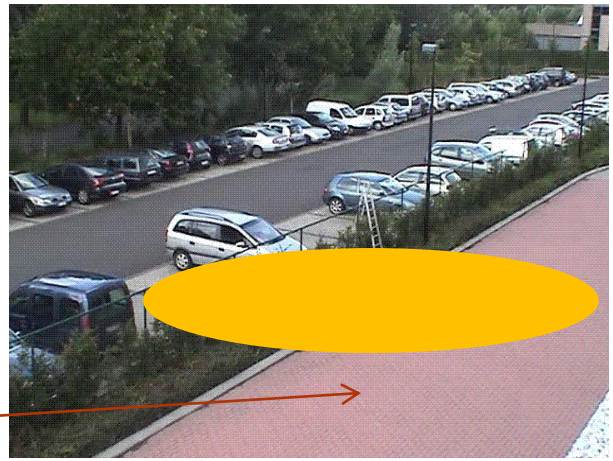
Fei Yin, D. Makris, S. A. Velastin, T. Ellis

MuCCD Dataset

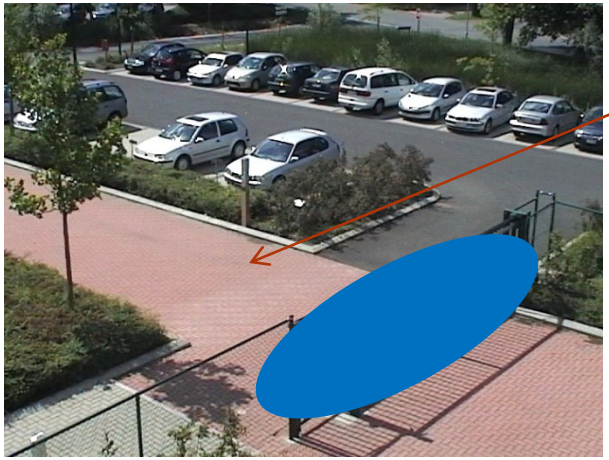


West

Common Ground Plane

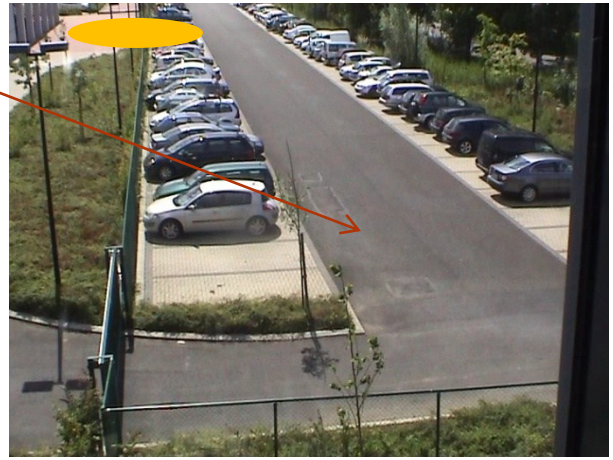


East



Gate 1

http://dipersec.kingston.ac.uk/MuCCD



Gate 2

Scene Calibration

- **Homography:** Image1→Image2 (i.e. Registration)
- Manually draw two lines
- Use tracking output to estimate homography, taking random samples until reasonable accuracy (reproject)

Generate Ground Plane

2: Gate1 view

H_{21}

4: Gate2 view

H_{43}

H_{31}

3: East view

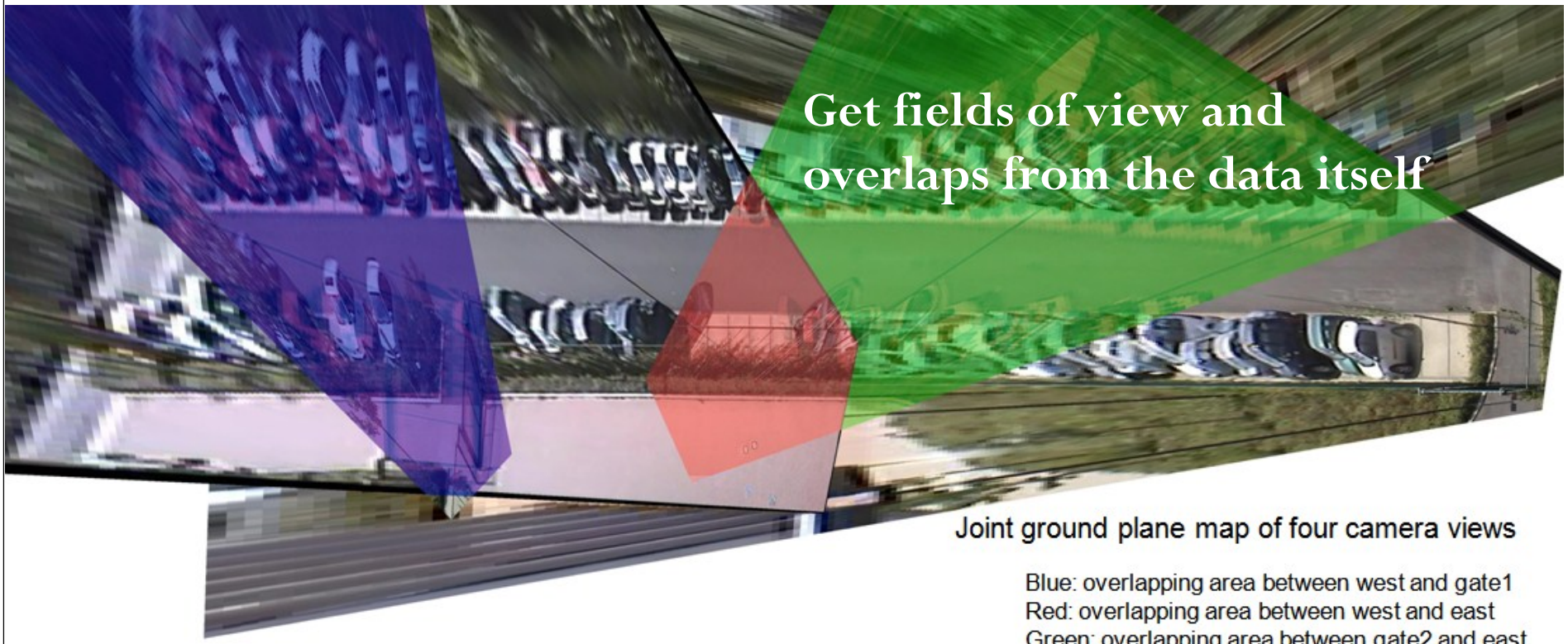
1: West view

H_{15}

Use one view to project points to the ground plane

5: Ground plan

Ground plane Map



Joint ground plane map of four camera views

Blue: overlapping area between west and gate1
Red: overlapping area between west and east
Green: overlapping area between gate2 and east

Camera View: west

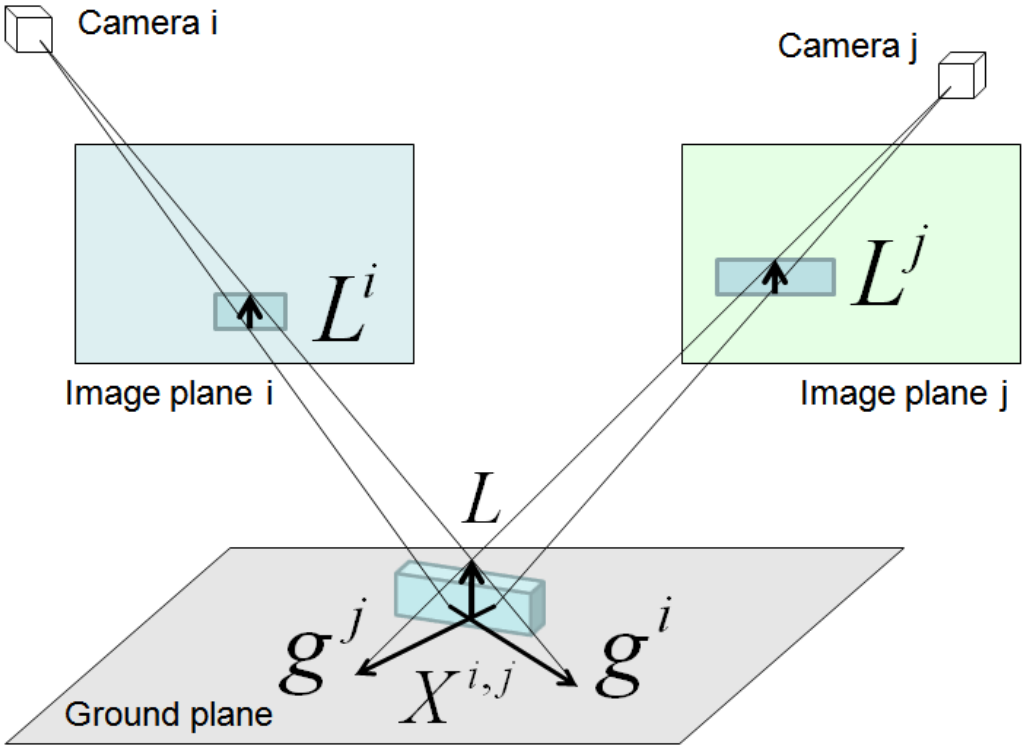
Camera View: gate1

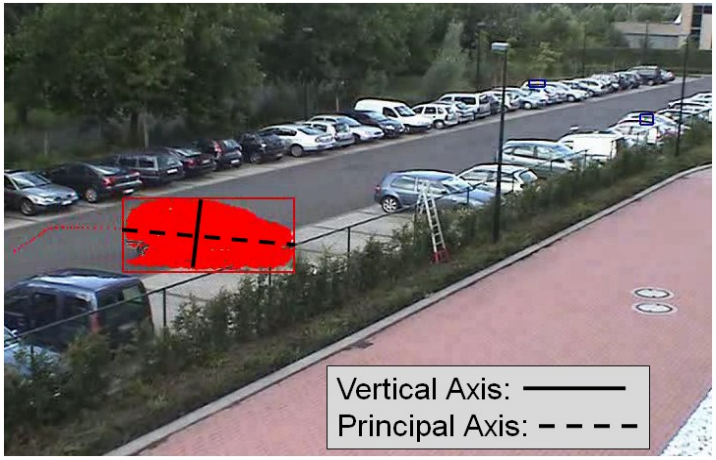
Camera View: east

Camera View: gate2

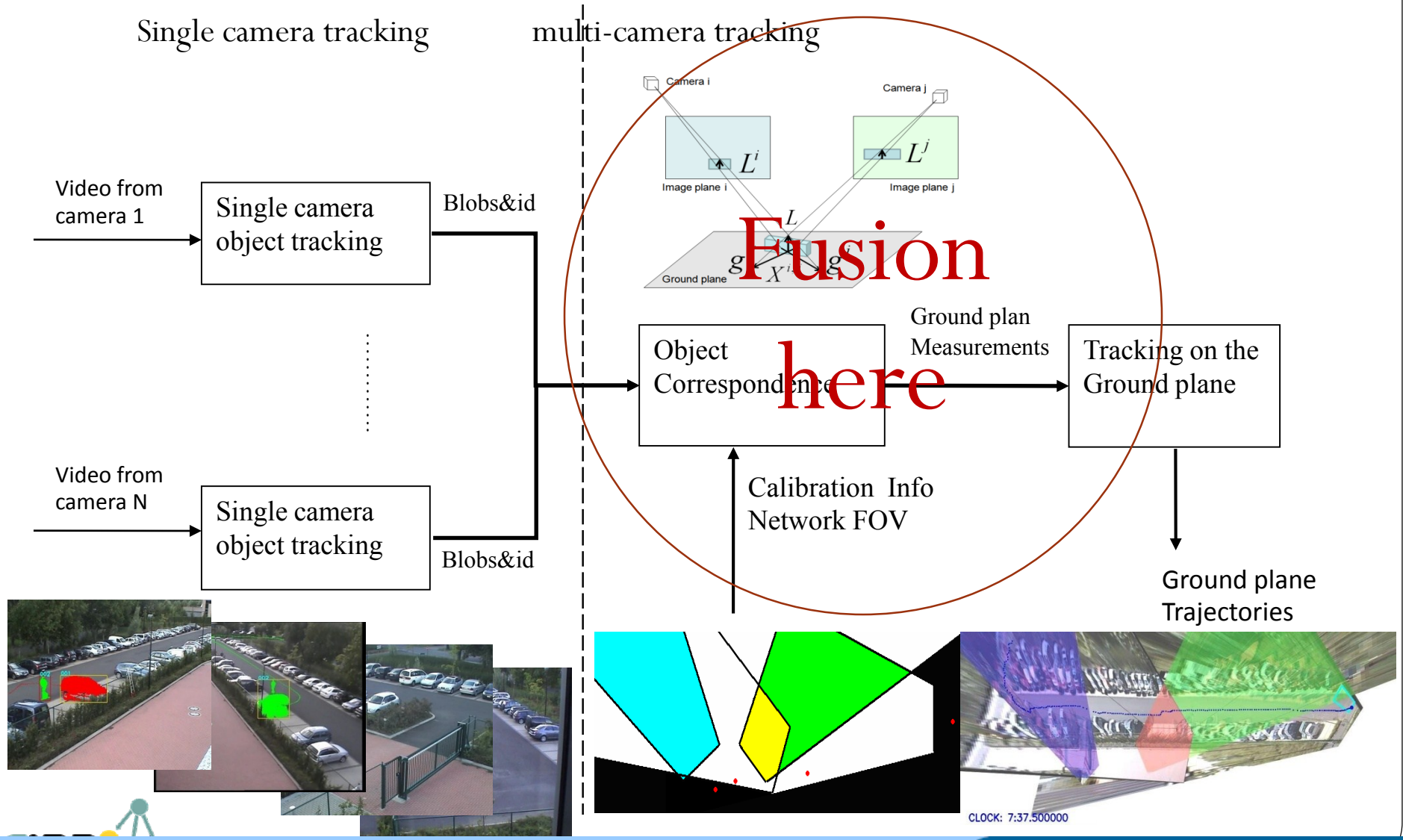
Vertical Axis and tracking

- Data seen by different cameras are fused and tracked on the ground plane using a **Kalman** filter
- Data seen by more than one camera has stronger evidence (i.e. to remove outliers)





Overview


Single camera tracking
multi-camera tracking
Video from camera 1
Single camera object tracking
Blobs&id
Video from camera N
Single camera object tracking
Blobs&id
Camera i
Camera j
Image plane i
Image plane j
Ground plane
Fusion
here
Object Correspondence
Ground plan Measurements
Tracking on the Ground plane
Calibration Info Network FOV
Ground plane Trajectories
CLOCK: 7:37.500000

But what happens when the world is not flat .....?

UdeSantiago

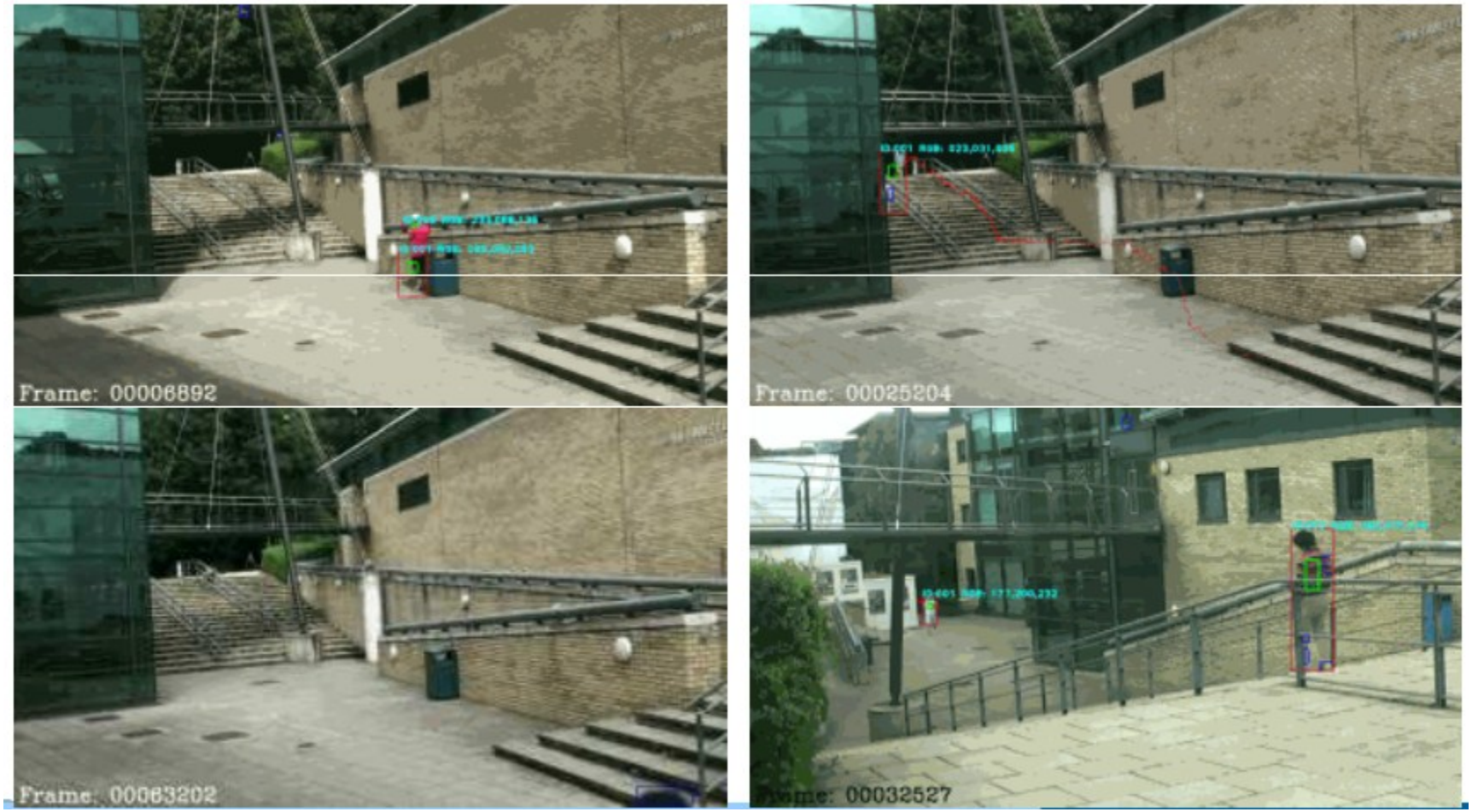
Universidad Carlos III de Madrid

Learning Multi-Planar Scene Models in Multi-Camera Videos

F. Yin, D. Makris, T. Ellis
S. A. Velastin

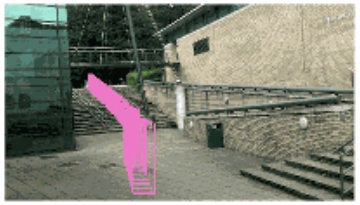
Kingston Hill dataset

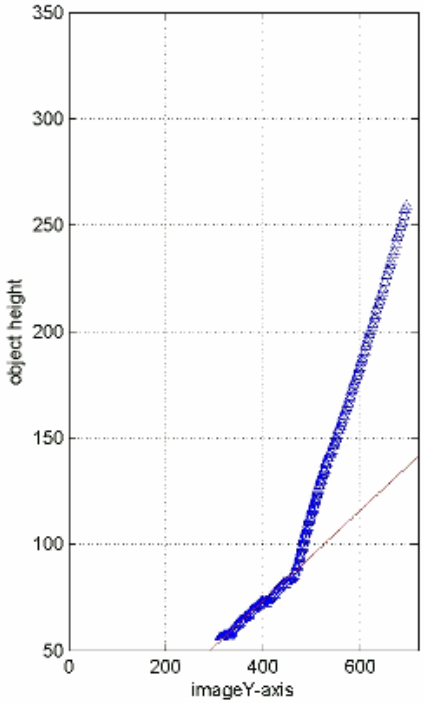
http://dipersec.kingston.ac.uk/MCGMdata



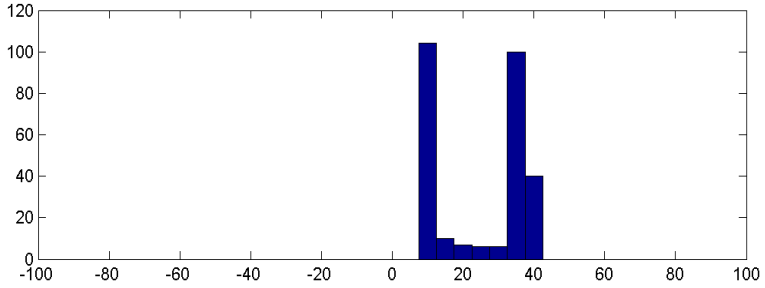
Walkable regions from tracking

Planes from height variations

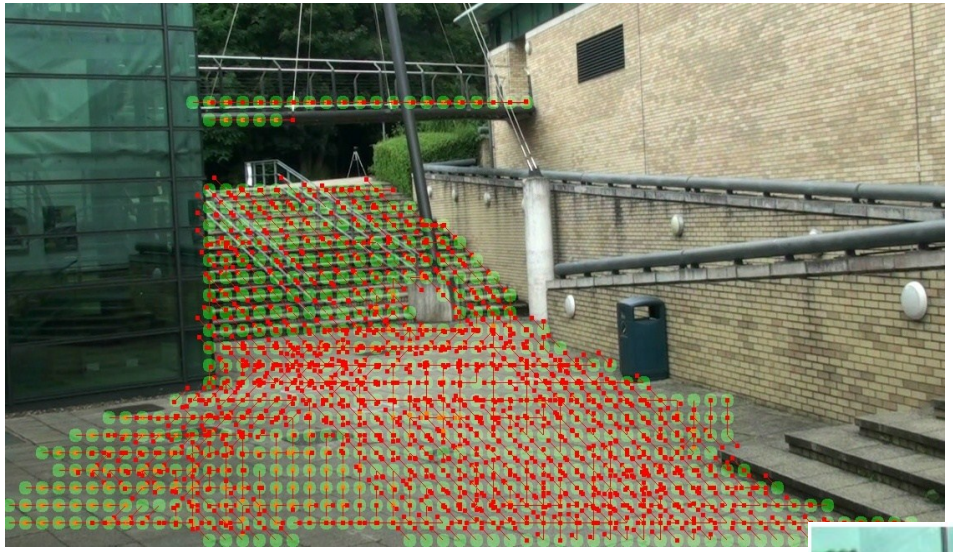


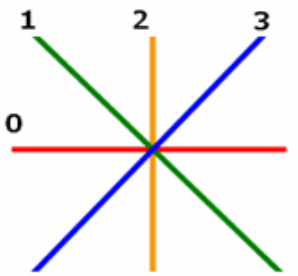


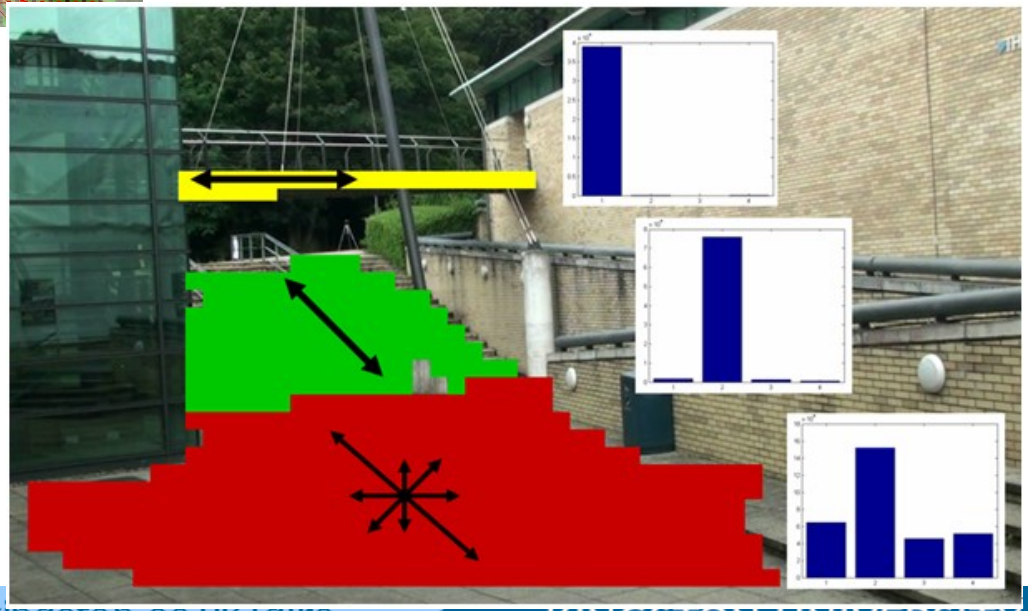
Histogram of line angles θ_i



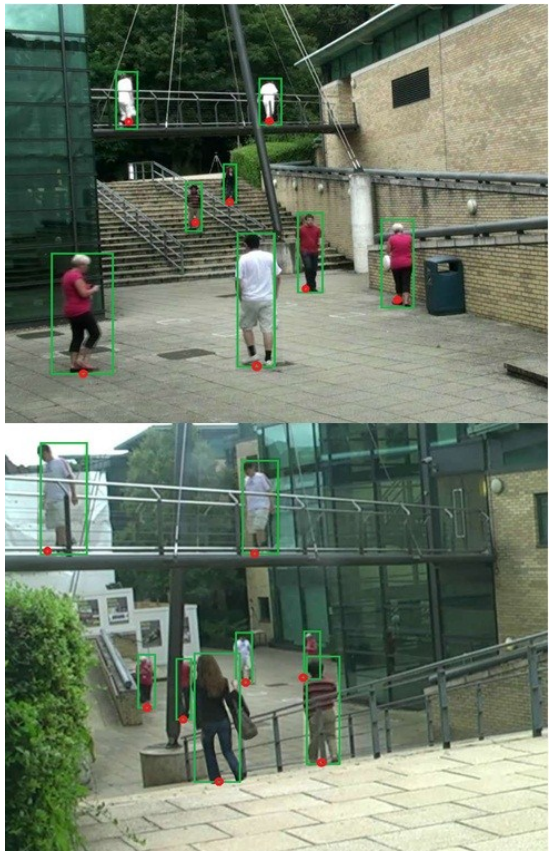
Motion variety/plane segmentation: Reference Plane

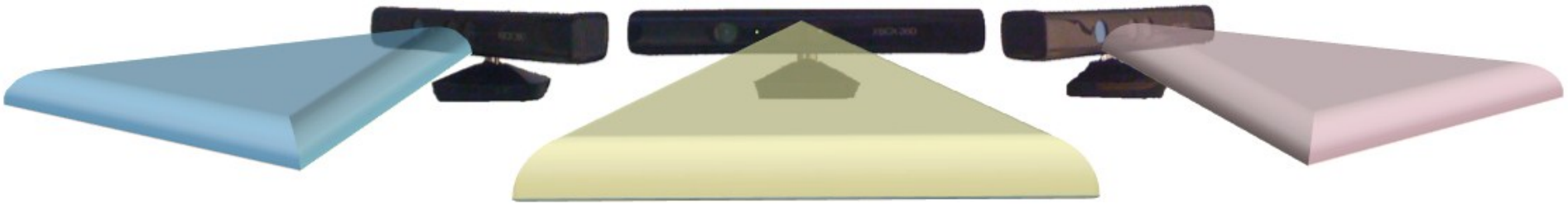






Multi camera tracking





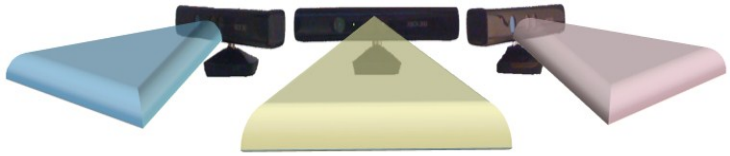
Tracking People in Range Data across Multiple *Kinect* Sensors

Emilio J. Almazán

Graeme A. Jones

Main aim and Issues

- To create a wide-view depth-based sensor which can *track* multiple individuals within a large indoor space.



- Issues

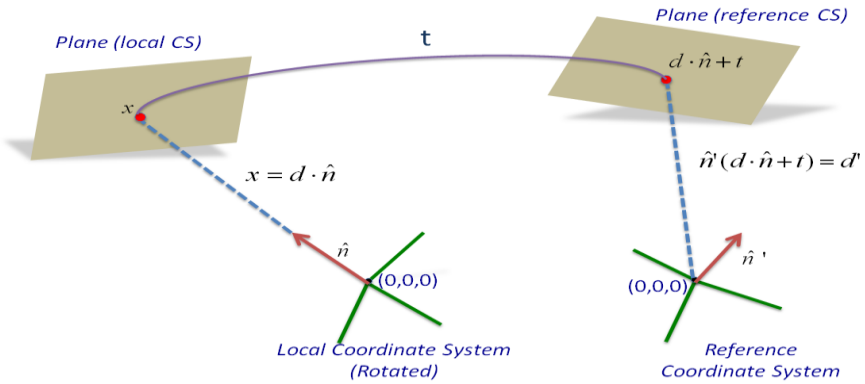
- Integration of non-overlapping viewpoints
- Crowded scenes
- Static and dynamic occlusions
- Different illumination conditions
- Noise and limited resolution of RGB-D sensors

System Geometry and Calibration

Combine 3 Kinects to create a 180° sensor

- Non-overlapping view volumes
- Maximize monitoring area
- Minimize interferences

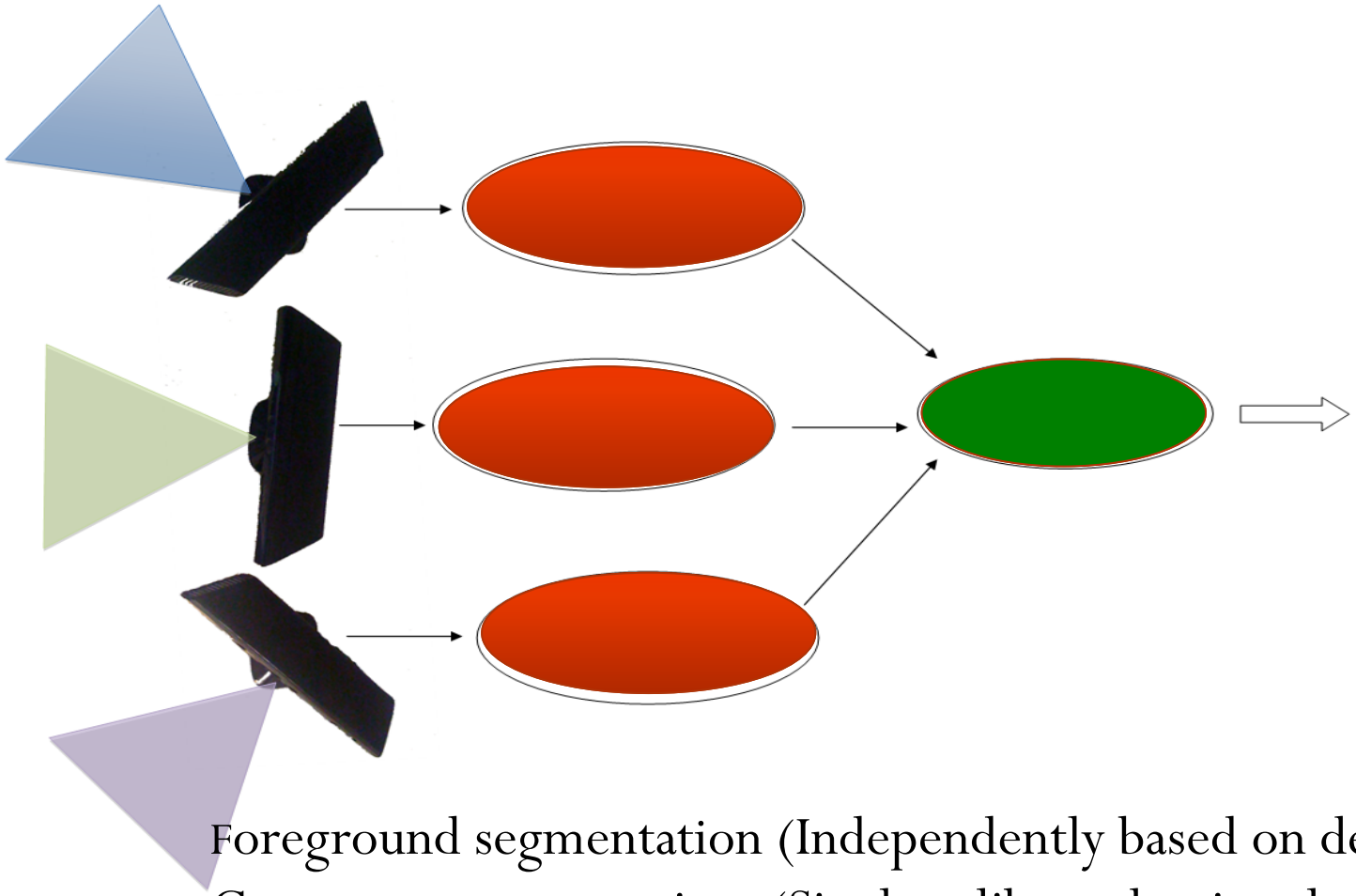
System Geometry and Calibration



Calibration using planes

- A pair of sensors at a time
- Search for common planes
- Plane fitting – easy task with depth data
- Calibration parameters (rotation and translation)

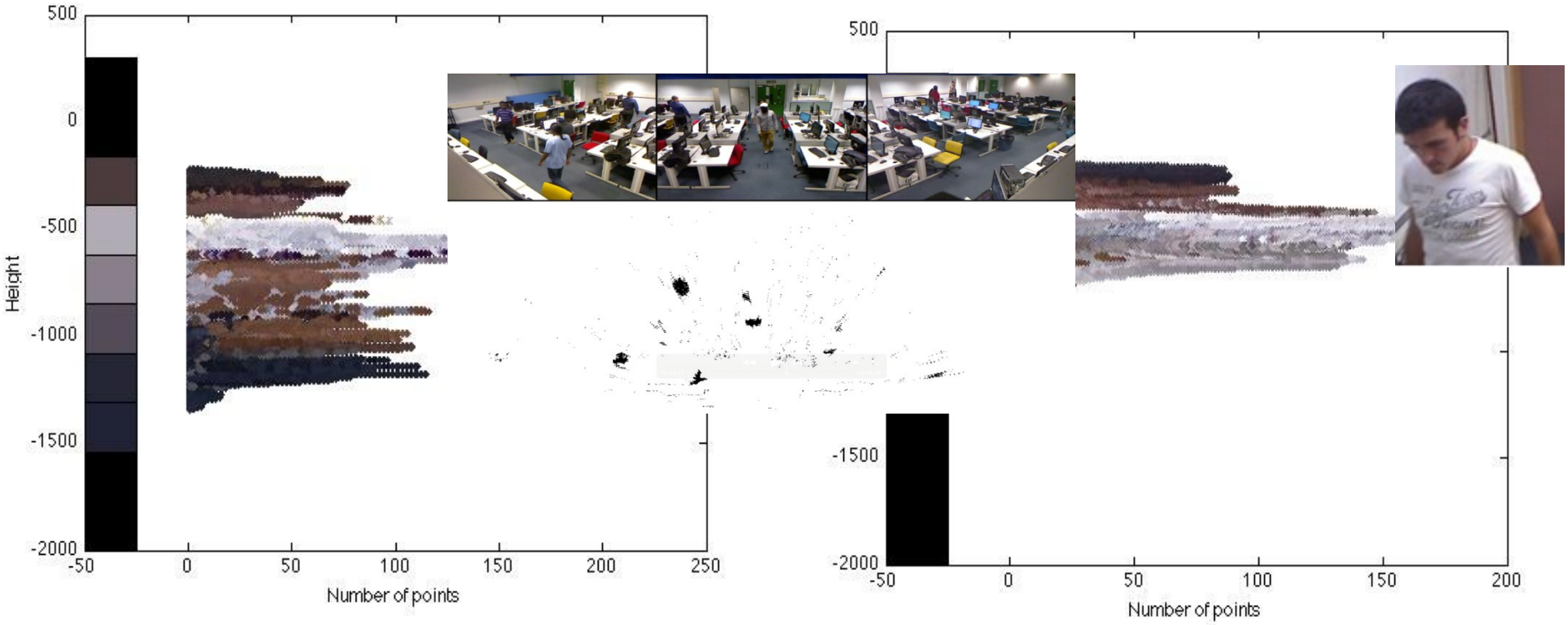
People Detection



Foreground segmentation (Independently based on depth)
Common representation (Single calibrated point cloud)
Blob Detection

Tracking

- Appearance model: Augmented histogram (height + colour)
- And Linear Kalman Filter (on common ground plane)



UdeSantiago

Universidad Carlos III de Madrid

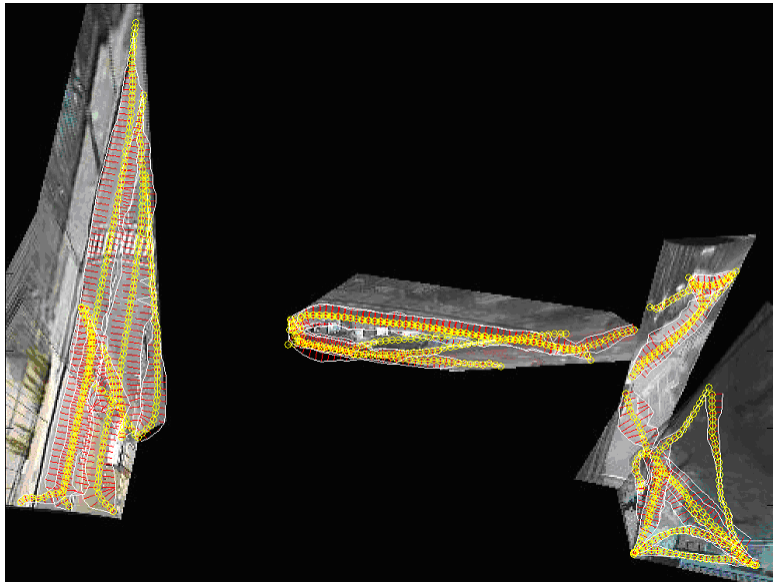
Bridging the Gaps between Cameras (for non overlapping cameras)

Dimitrios Makris, James Black, Tim Ellis

Multi-Camera Integration

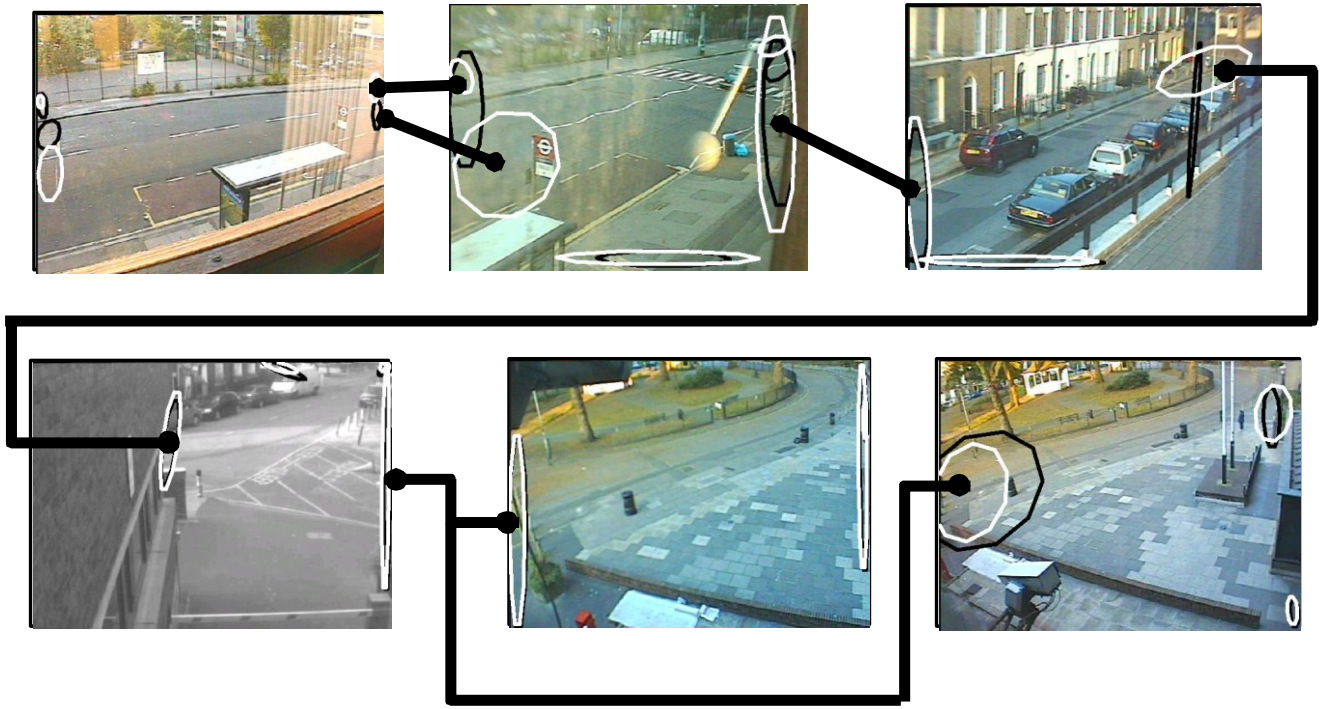
- How to integrate information from multiple cameras?

Common Ground Plane Map



- Ground plane map requires manual calibration of all the cameras
- Valid only when all activity is coplanar
- No model for the "blind" areas of the scene

Tempo-probabilistic links



Multiple Camera Activity Network

- Camera views are connected using tempo-probabilistic links between entry/exit zones
- The network is learnt automatically by correlated events in different camera views

- Learn Entry/Exit zones for each camera view using the tracked objects

- Zones->Nodes of a Probabilistic network
- Learn probabilities from the data

Multi camera Tracking

UdeSantiago

Universidad Carlos III de Madrid

Autonomous active-camera control architecture based on multi-agent systems for surveillance scenarios

http://giaa.inf.uc3m.es

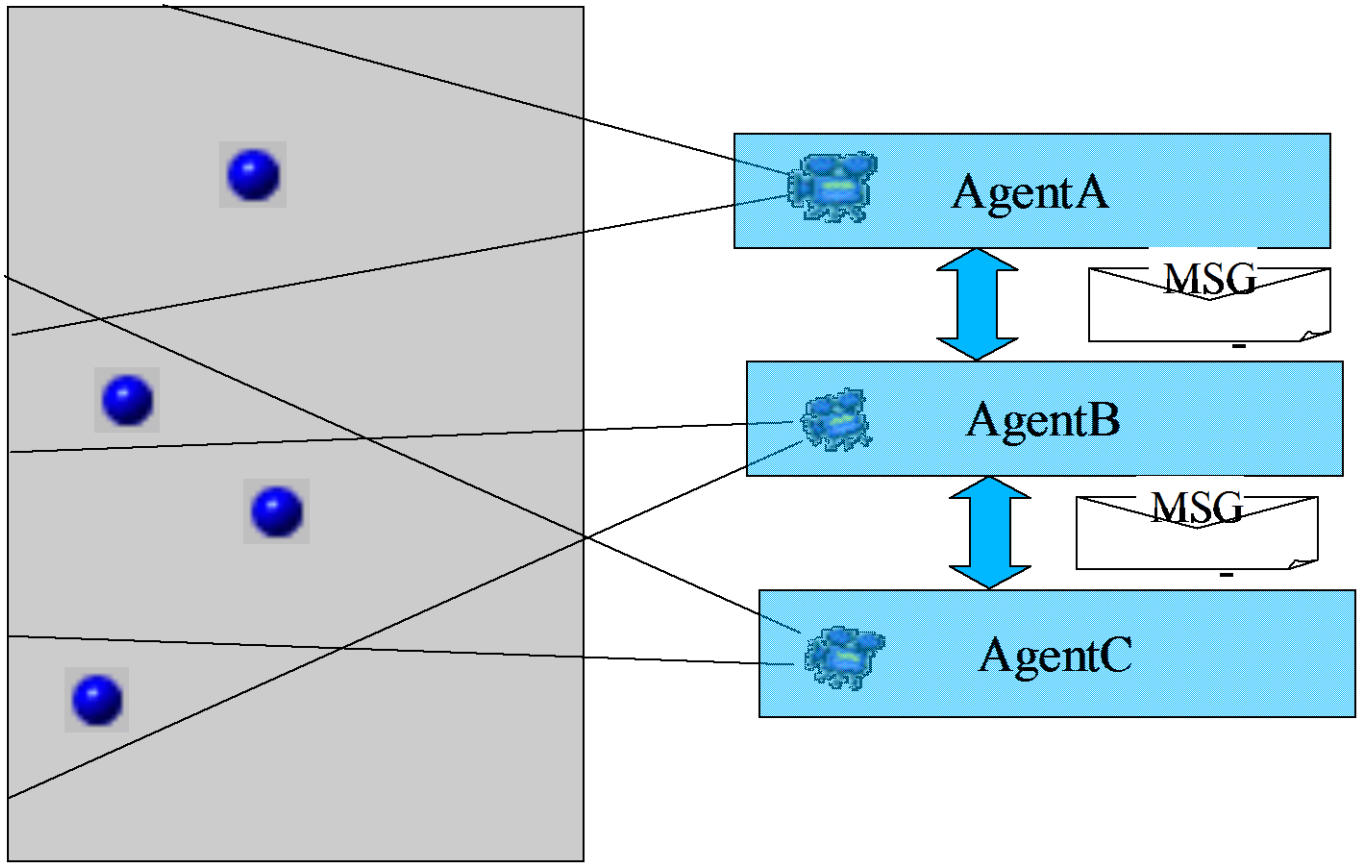
José Manuel Molina López

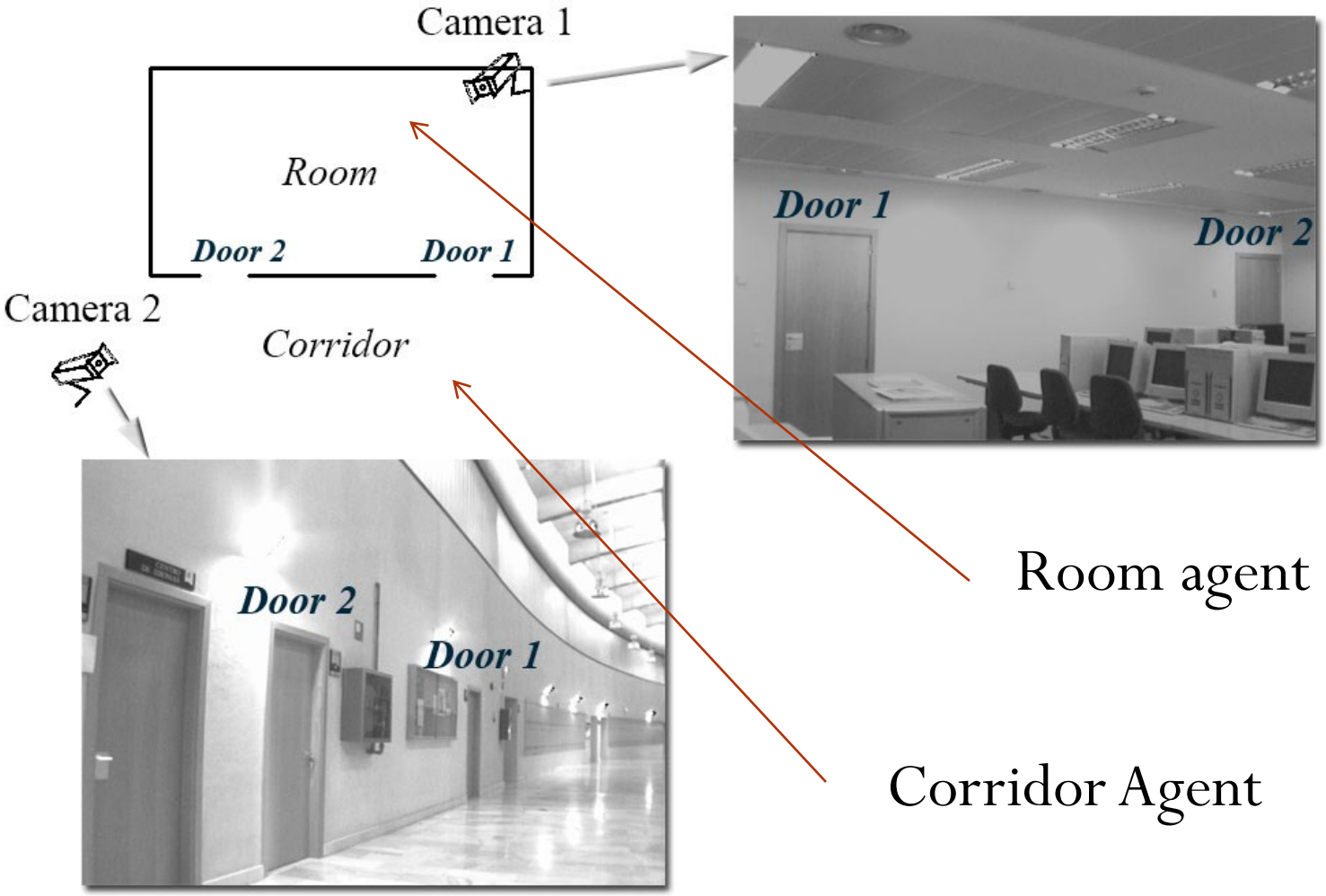
Alvaro Luis Bustamante

Visual Sensor Networks using a Multiagent Platform

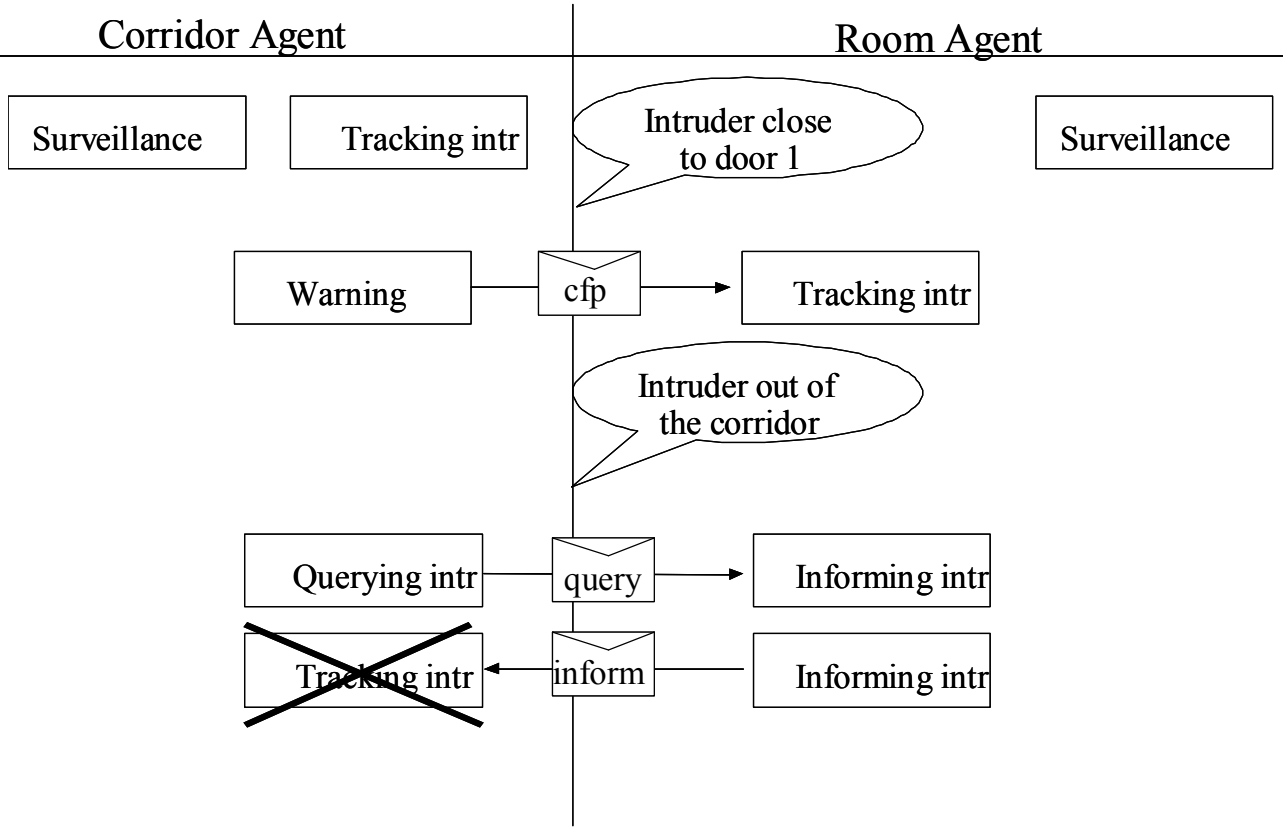
- Distributed Knowledge-> Robust
- Scalability
- Fault Tolerance
- Explicit Knowledge. Coopearting Agents
- Platform standarization
- Easy integration of different technologies

Geographical distribution




Camera 1
Room
Door 2
Door 1
Camera 2
Corridor
Door 1
Door 2
Door 2
Door 1
Room agent
Corridor Agent

Timeline


Corridor Agent
Room Agent
Surveillance
Tracking intr
Intruder close to door 1
Surveillance
Warning
cfp
Tracking intr
Intruder out of the corridor
Querying intr
query
Informing intr
Tracking intr
inform
Informing intr

UdeSantiago

Universidad Carlos III de Madrid

High Dimensionality

D. Makris*, V. Bloom*, M. Lewandowsky,
S.A. Velastin

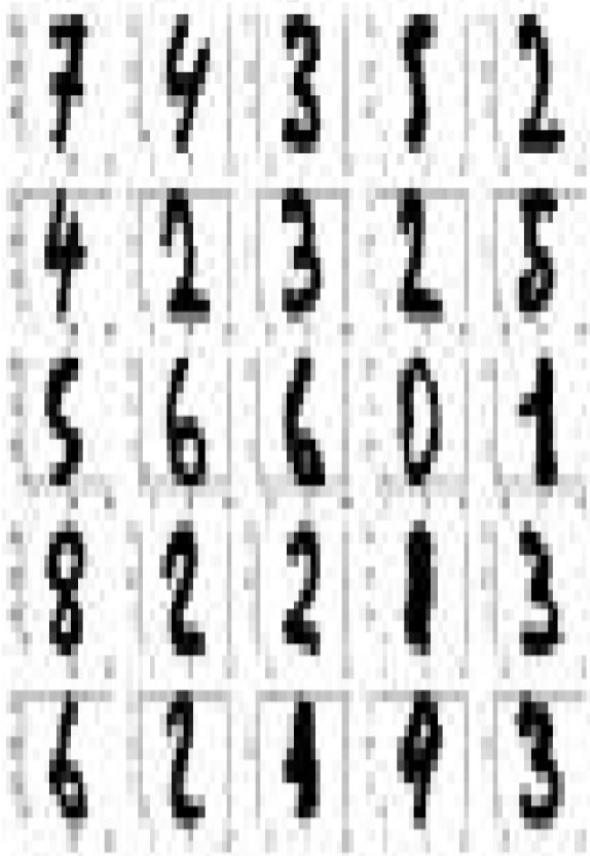
*Kingston University London

giaa

High-dimensional data

Human Motion

Face images



Handwritten digits

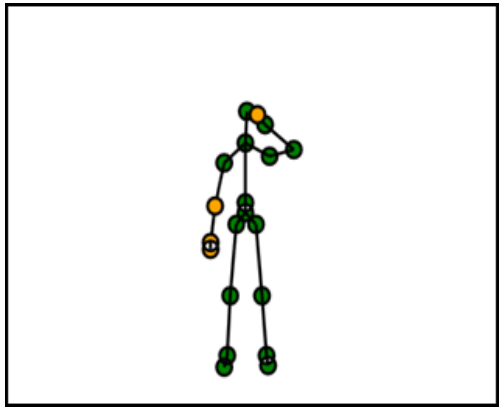
Why dimensionality reduction?

- "**Curse of Dimensionality**":
  - Data spreads out, machine learning becomes more difficult
  - Computational load increases
- The **intrinsic** dimension may be small. For example, walking could be simply described by 2 or 3 dimensions
- Easier to **visualise**
- Better **data compression**

Example: Action Recognition

- G3D: Gaming action dataset, http://dipersec.king.ac.uk/G3D/
- MuHaVi: Multi Camera Human Action Dataset, http://dipersec.king.ac.uk/MuHAVi-MAS/





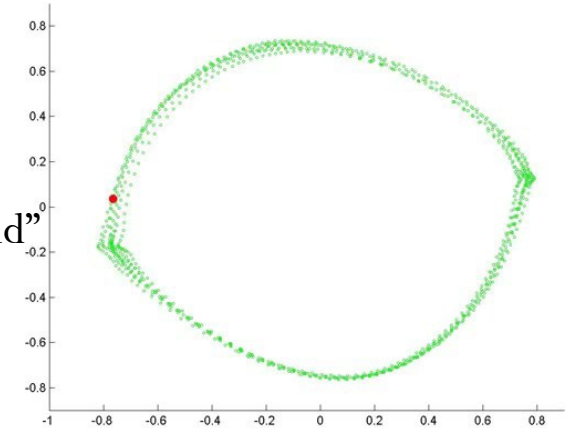
RGB, depth, skeleton (joint angles): Lots of data!

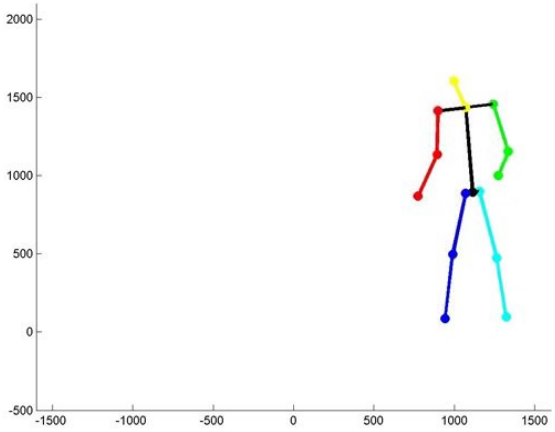
e.g. 640x480 RGB, 640x480 (11 bit) depth, 13 joints (**52** parameters)

Using low level image features (blobs, histograms, …): ~3K

But intrinsic dimensionality around 2 or 3 dimensions, so **worth doing! How?**

"Manifold"

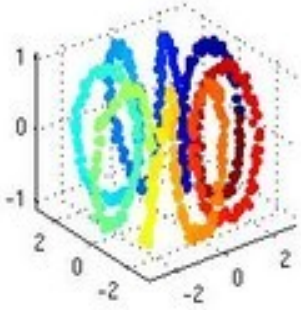




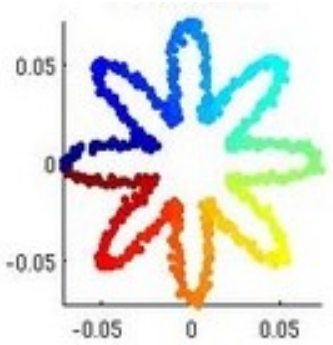
So ....

- Feature Extraction
  - Find a new set of *d* dimensions that are combinations of the original *D* dimensions, **with minimum loss of information** (*d* << *D*).
- Methods
  - Linear
    - Principal Components Analysis (PCA)
    - Linear Discriminant Analysis (LDA)
  - Nonlinear
    - Isometric feature mapping (Isomap)
    - Laplacian Eigenmaps (LE)

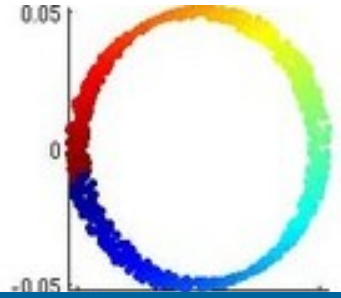
Toroidal Helix



PCA

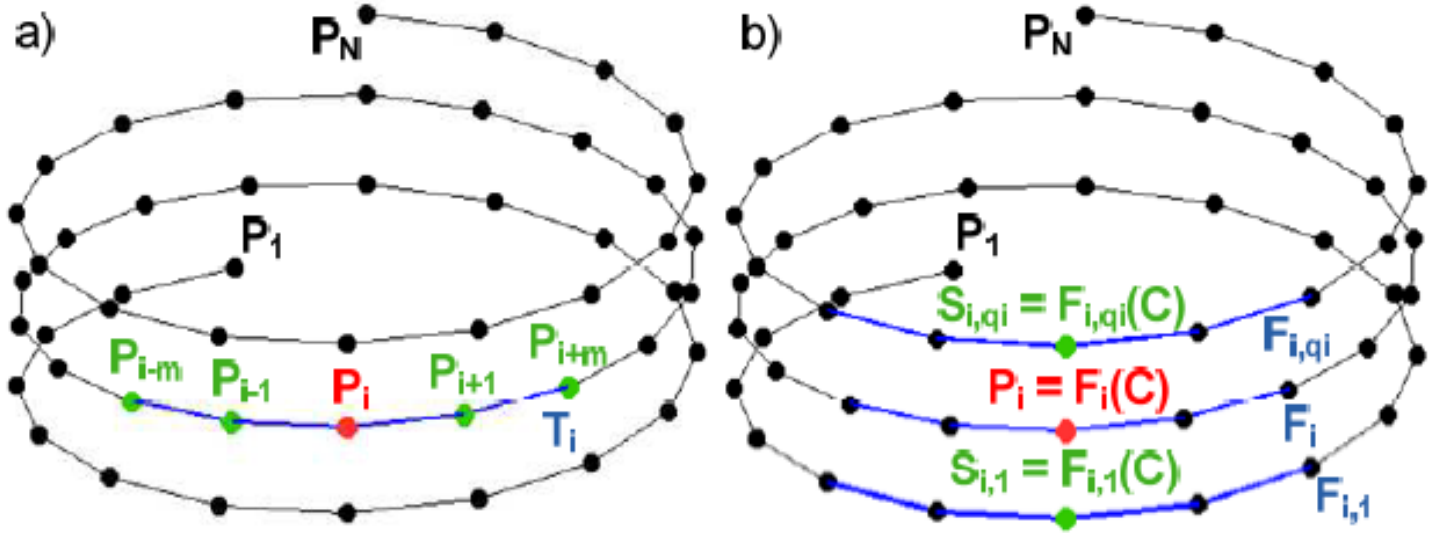


LE

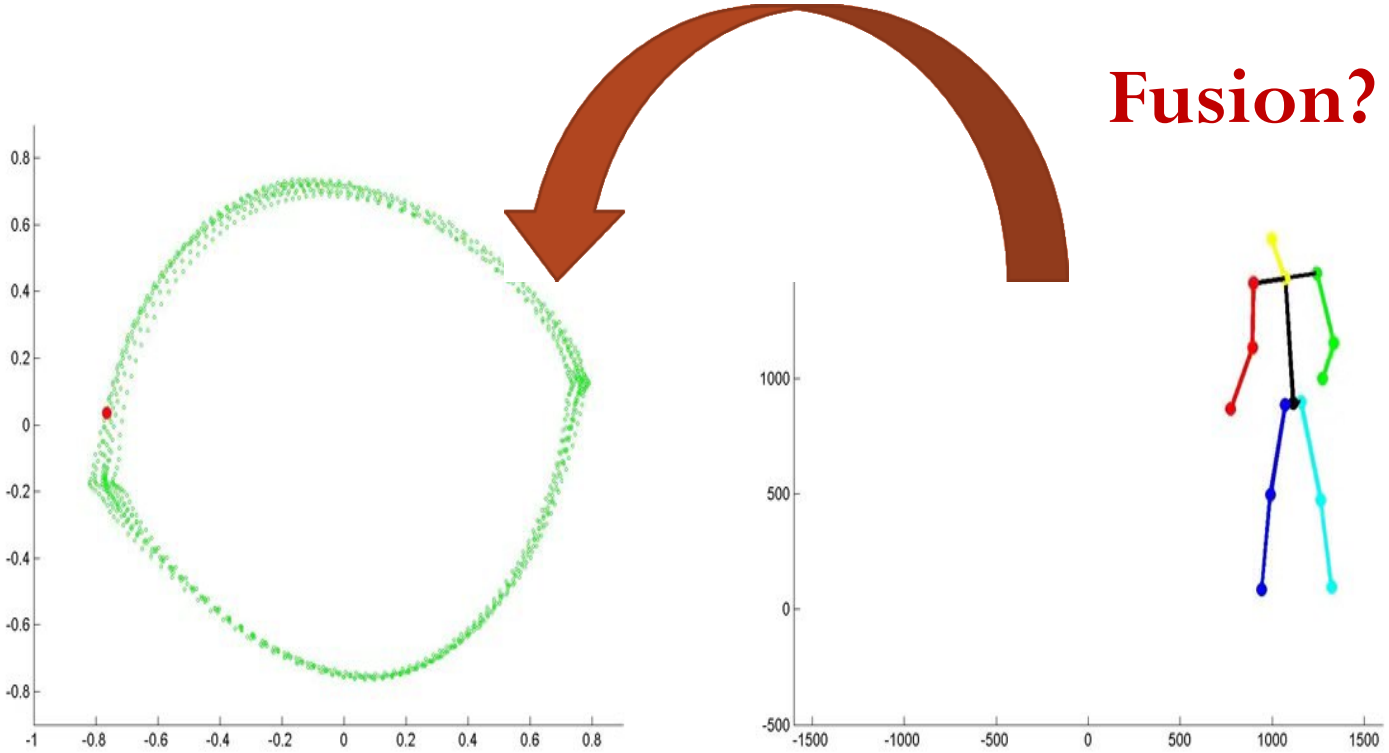


- Temporal Laplacian Eigenmaps (TLE)
  - Preserve the temporal structure of time series data in the low dimensional space by constructing 2 graphs.

Temporal neighbours **Spatio-temporal repetition neighbours**



And we get this …



To see how: M Lewandowski, D. Makris, S.A. Velastin, J.C. Nebel, "Structural Laplacian Eigenmaps for modelling sets of multivariate sequences" in 'IEEE Transactions on Systems, Man and Cybernetics, Part B', DOI: 10.1109/TCYB.2013.2277664, (2013).

UdeSantiago

Universidad Carlos III de Madrid

An alternative approach

Carlos Orrite, Mario Rodriguez, Elías Herrero, Gregory Rogez
(University of Zaragoza, Spain)
Sergio A. Velastin

Carlos Orrite, Mario Rodríguez, Elías Herrero, Gregory Rogez, Sergio Velastin, "Automatic Segmentation and Recognition of Human Actions in Monocular Sequences" in *22nd International Conference on Pattern Recognition (ICPR)*, 24-28 Aug, Stockholm, Sweden (2014)

giaa

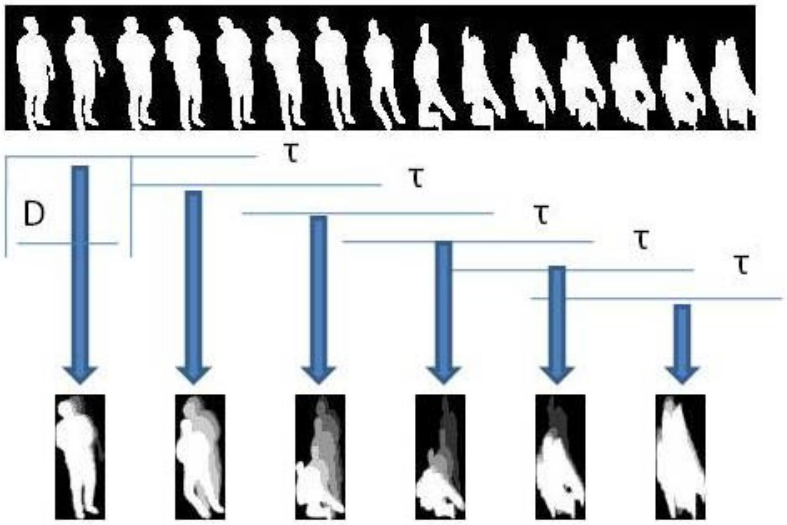
Universidad Zaragoza

Motion History Image

Capture motion information in images
Encode how recently motion occurred at a pixel

$$MHI_t(x,y) = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ \max(0, MHI_{t-1}(x,y) - 1) & \textit{otherwise} \end{cases}$$

More recently moving pixels are brighter

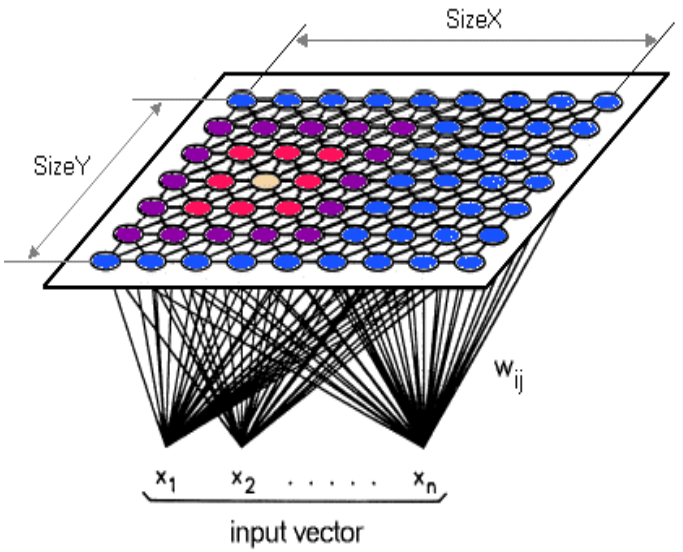


Self-Organizing Map (SOM)

Main problem of temporal representation: **High dimensional non-linear space**
Capture temporality relations (important for actions: here through MHI)

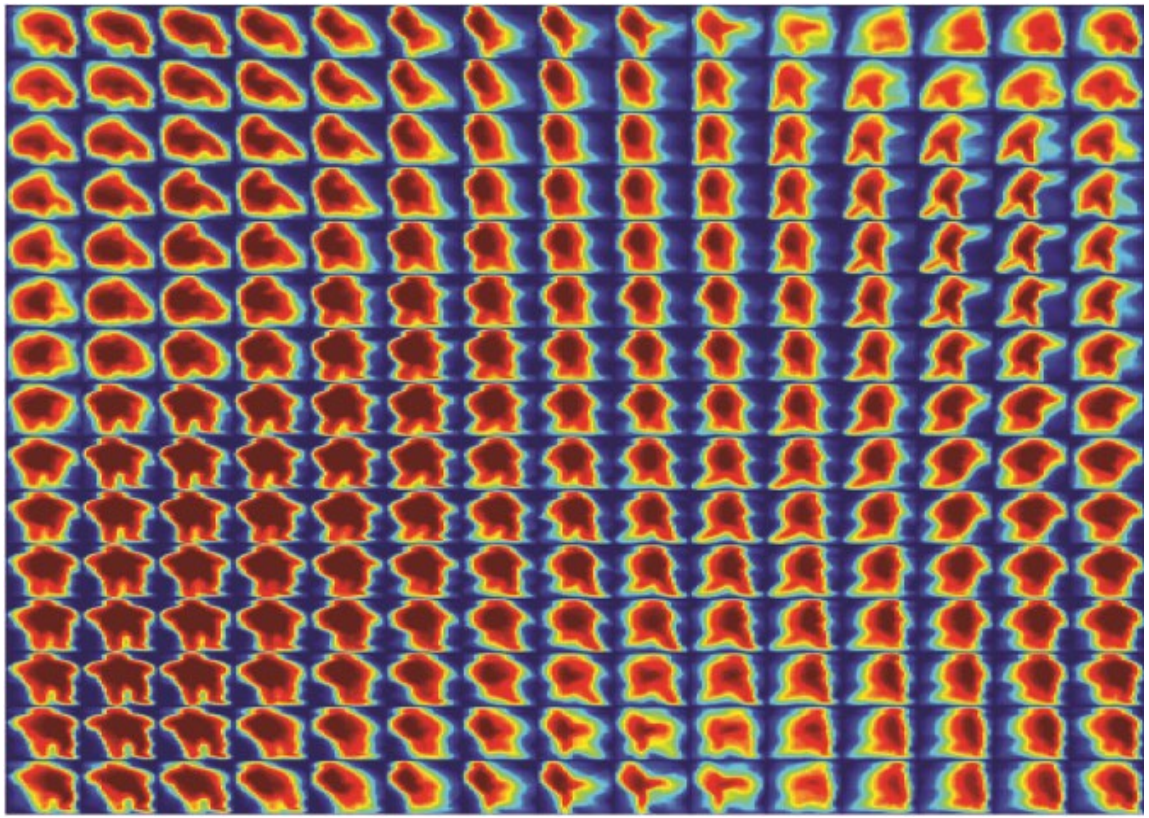
SOM is an unsupervised neural network

Maps a set of n-dimensional vectors to a two-dimensional topographic map (so, dimensionality reduction: a common approach)



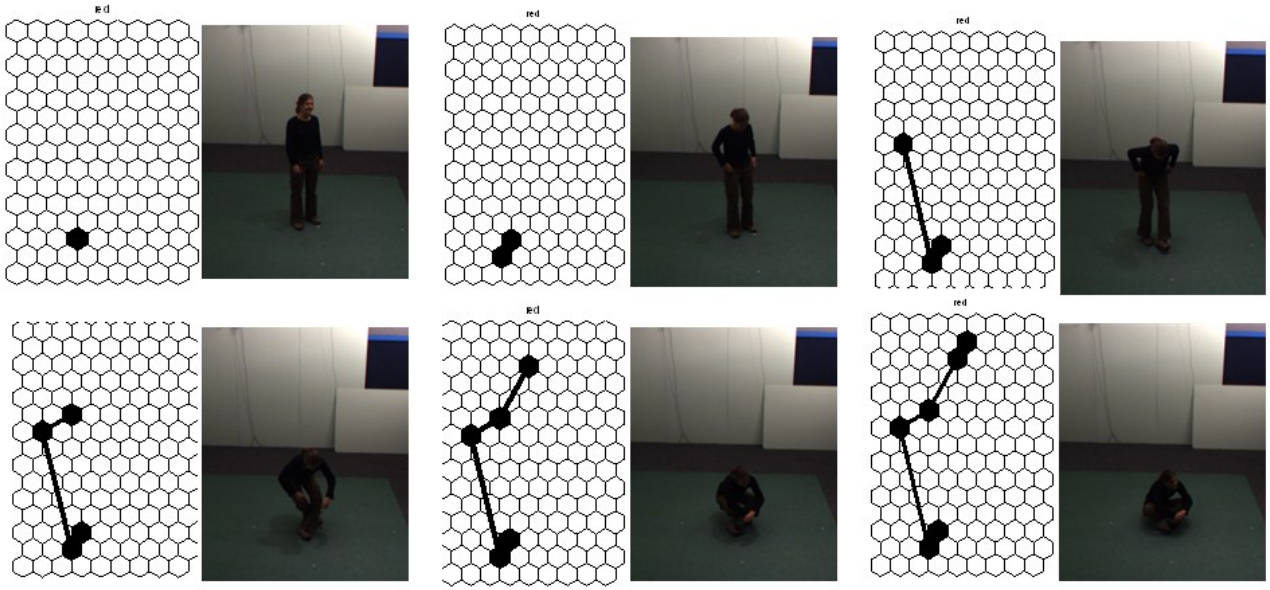
Similar data items are located close to each other on the map.

Self-Organized Maps (SOM)

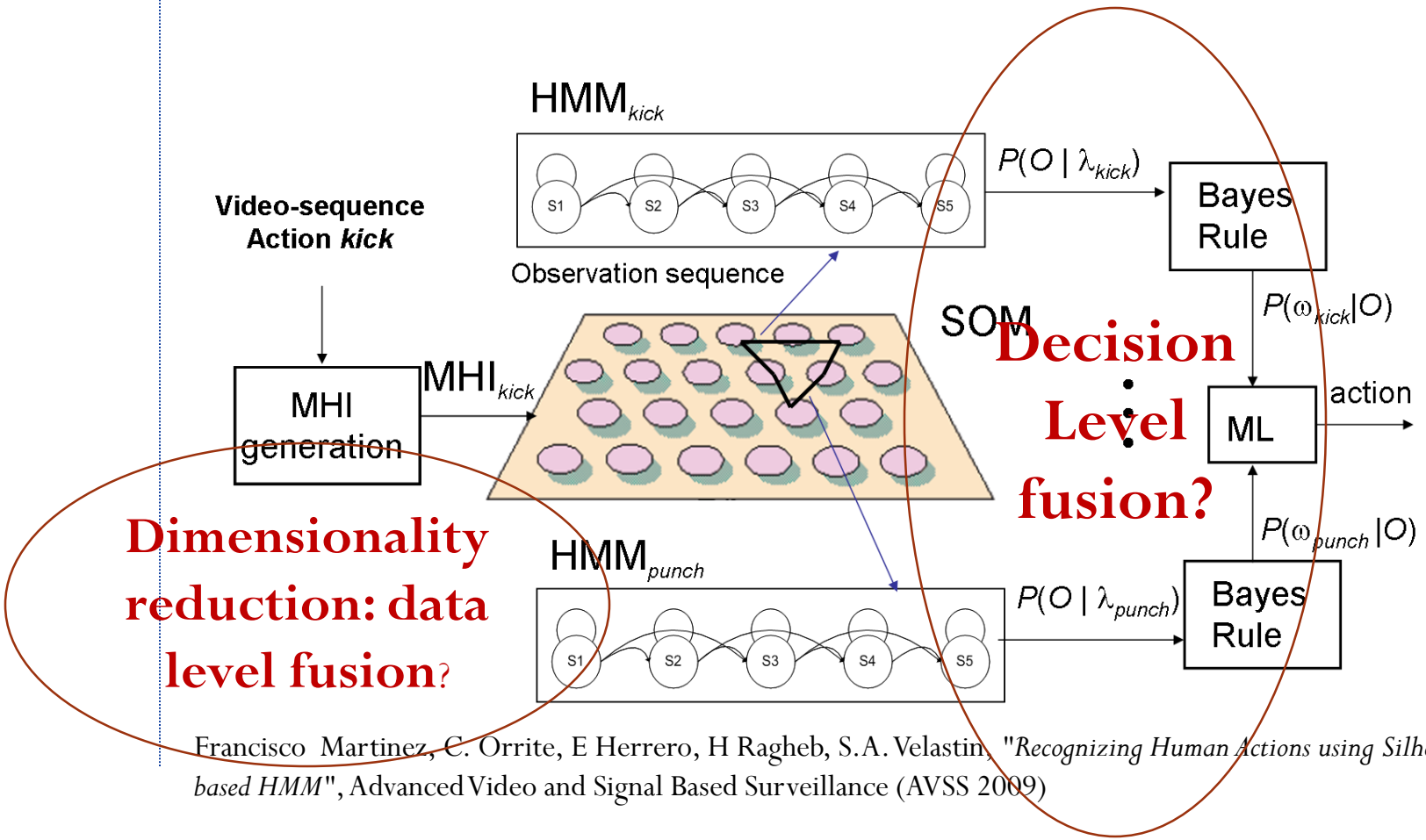


SOM + HMMs

A single SOM is trained with all the different actions.
The outputs of the SOM are inputs of Action Specific HMMs
Action specific HMM is a statistical model which gives the probability that the input sequence belongs to this action.



System Overview



Francisco Martinez, C. Orrite, E Herrero, H Ragheb, S.A. Velastin, "*Recognizing Human Actions using Silhouette-based HMM*", Advanced Video and Signal Based Surveillance (AVSS 2009)

Results

MuHAVi Database:

The overall recognition rate is 98.44%.

Action	1	2	3	4	5	6	7	8
Rate (%)	100	100	100	100	93.75	93.75	100	100

Multicamera Human Action Video - Manually Annotated Silhouette

- 25 fps
- 8 actions
- 2 actors
- 2 camera views
- 136 video sequences

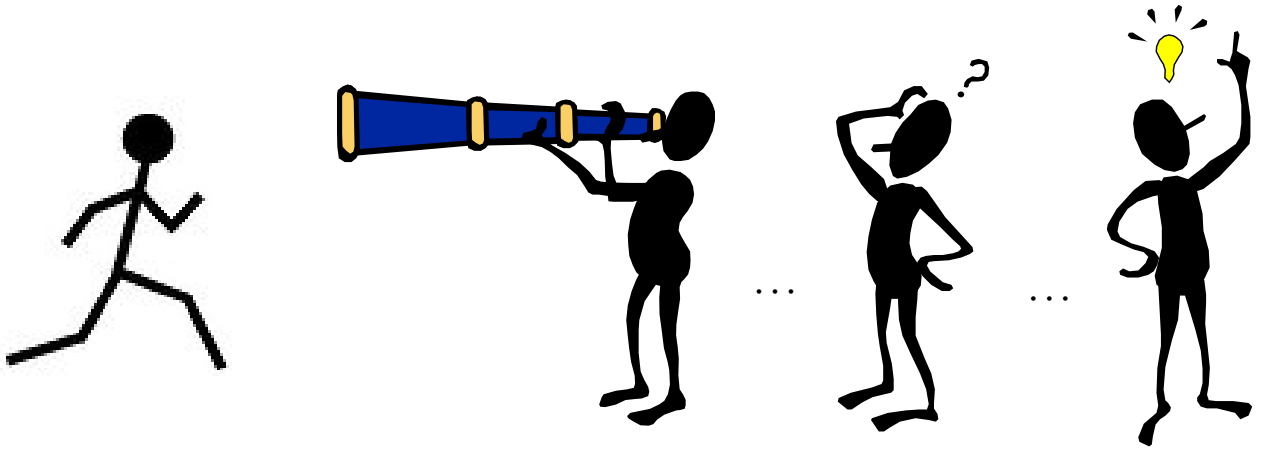
http://dipersec.king.ac.uk/MuHAVi-MAS

UdeSantiago

Universidad Carlos III de Madrid

Action Recognition

GIAA



An Efficient Approach for Multi-view Human Action Recognition Based on Bag-of-Key-Poses

ɔros Andre Chaaraoui, Pau Climent-

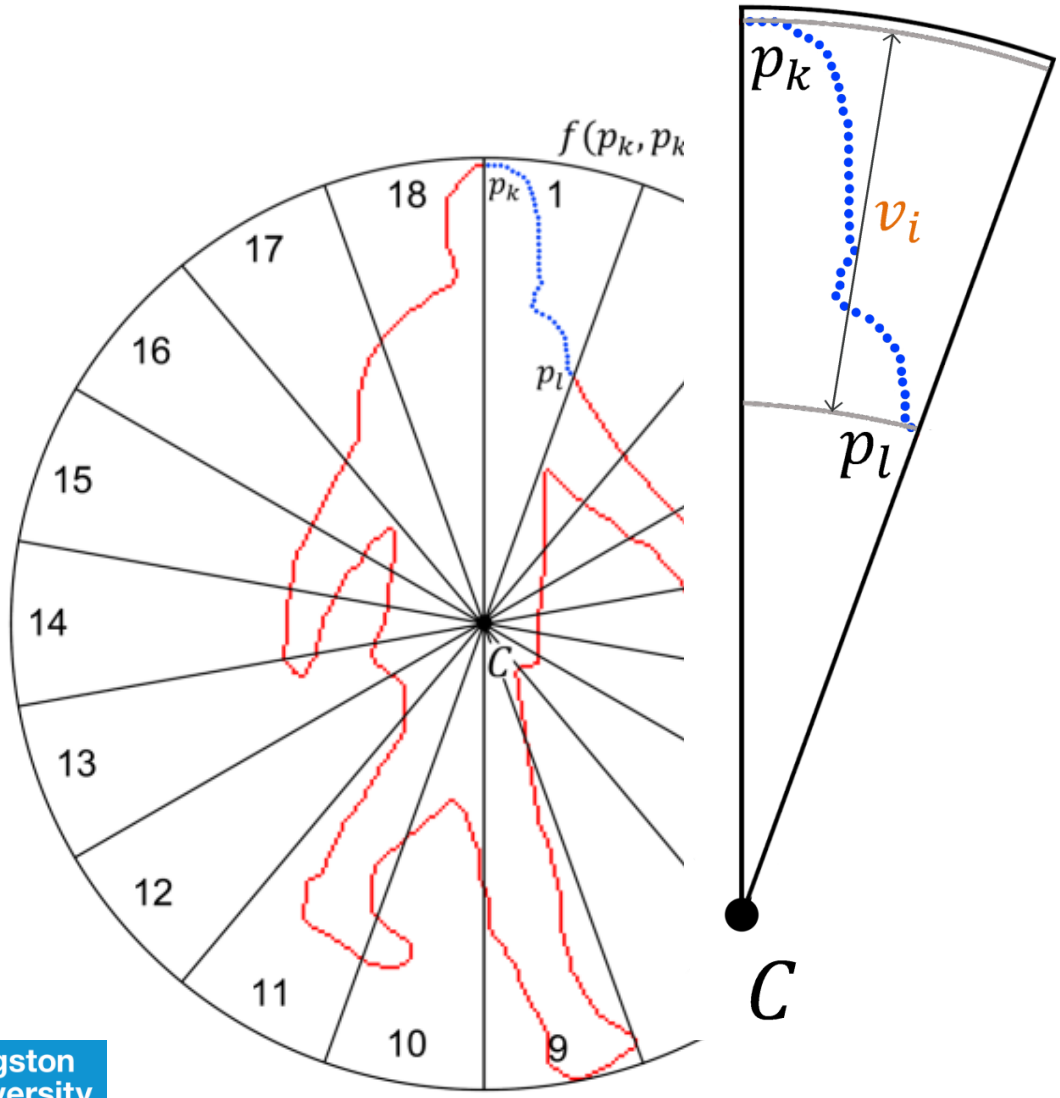
nd Francisco Flórez-Revuelta

Kingston University London

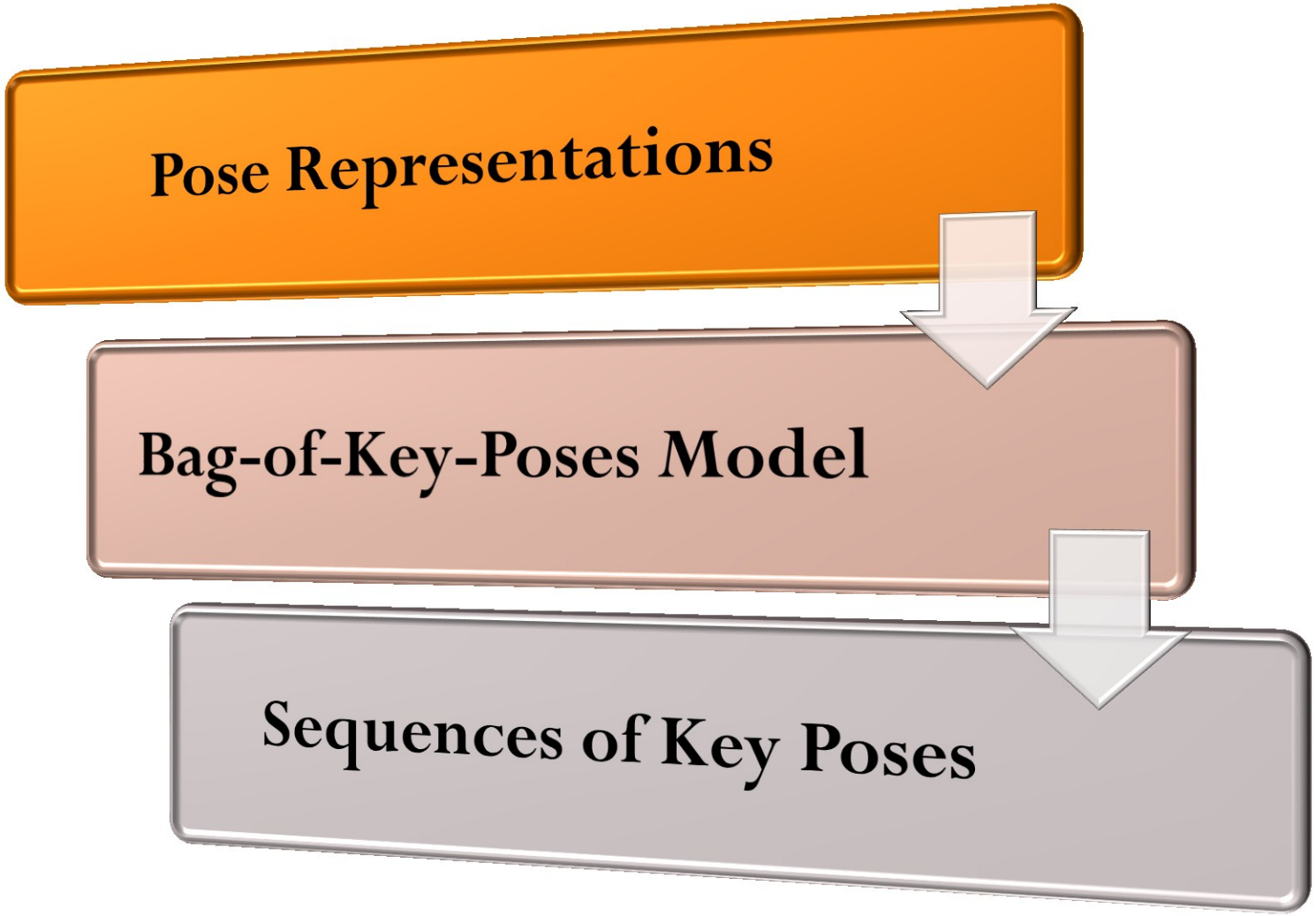
Universitat d'Alacant
Universidad de Alicante

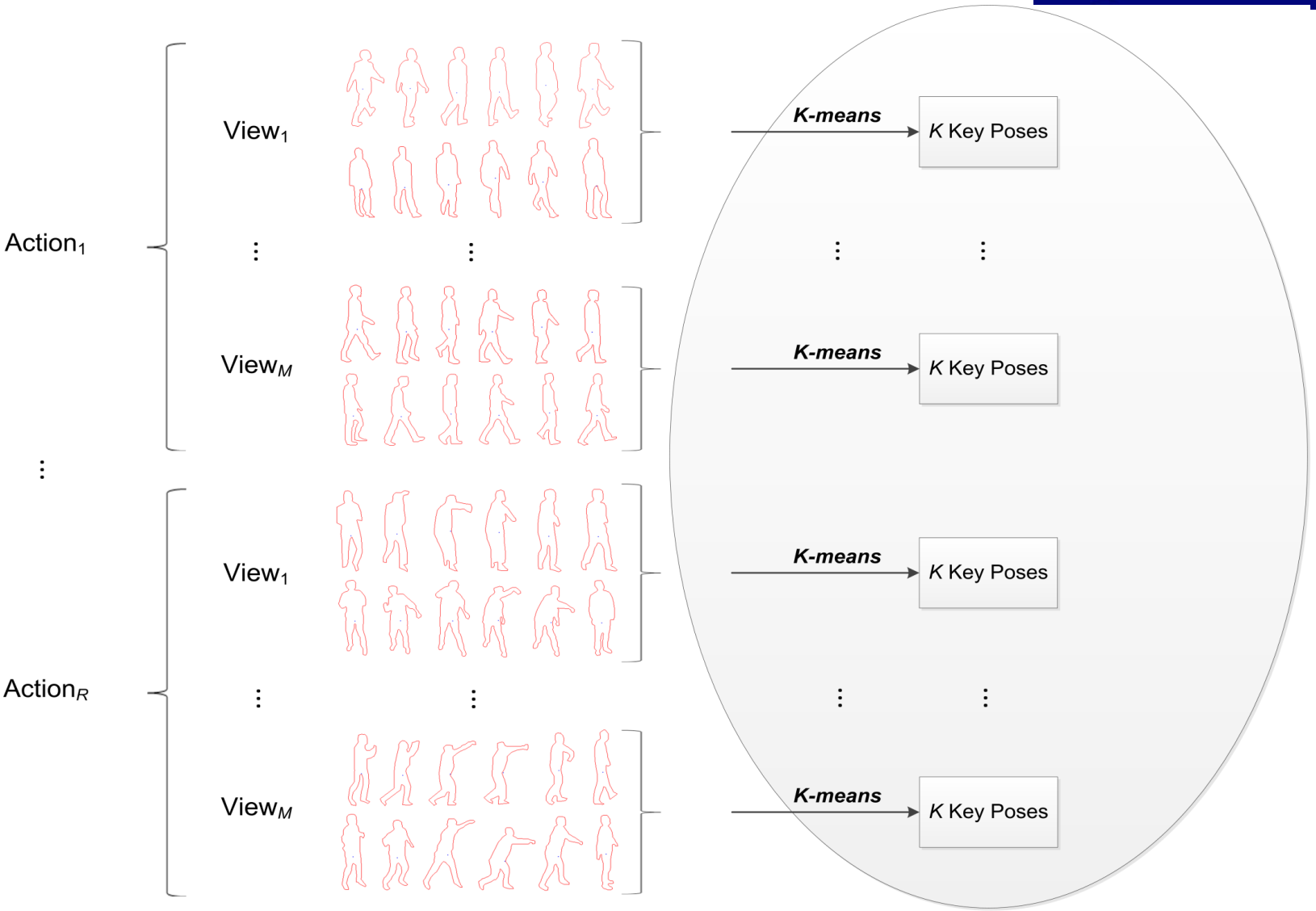
DAI
DOMÓTICA AMBIENTES INTELIGENTES

Radial Silhouette Feature



Human Action Recognition Method





Bag-of-Key-Poses

Sequence Recognition

- Sequences of key poses
- Nearest-neighbour key poses
- Sequence matching (dynamic time warping)

Walk

Turn right

Dealing with multiple views

- "Feature" fusion (aggregation): concatenate feature vectors from different views into a single feature vector:
  - Easy to implement without changing recognition scheme
  - But need to have all cameras to recognise an action
- "Model" fusion: Train separately for each view and recognise by seeing which action/view is more likely
  - Can recognise an action from a single view
  - Takes more time to determine which action/view fits best
- "Weigthed" feature fusion: for any given action give more weight to cameras with a "better" view

Results

- Tested on the MuHAVi-MAS Dataset (Singh et al.)
  - Two versions with 14 and 8 actions
  - Manually Annotated Silhouettes
  - We also tested *Feature Fusion* and *Without Fusion*

***Leave-one-sequence-out* cross validation**

Approach	MuHAVi-14		MuHAVi-8	
Singh et al. [21] (baseline)	82.4%		97.8%	
Cheema et al. [22]	86.0%		95.6%	
Martínez-Contreras et al. [23]	-		98.4%	
Eweiwi et al. [24]	91.9%		98.5%	
Without Fusion	$(L = 600, K = 120)$	85.3%	$(L = 350, K = 60)$	95.6%
Feature Fusion	$(L = 200, K = 140)$	92.6%	$(L = 300, K = 100)$	97.1%
Model Fusion	$(L = 450, K = 60)$	94.1%	$(L = 250, K = 75)$	98.5%

Results

***Novel Actor* Test**

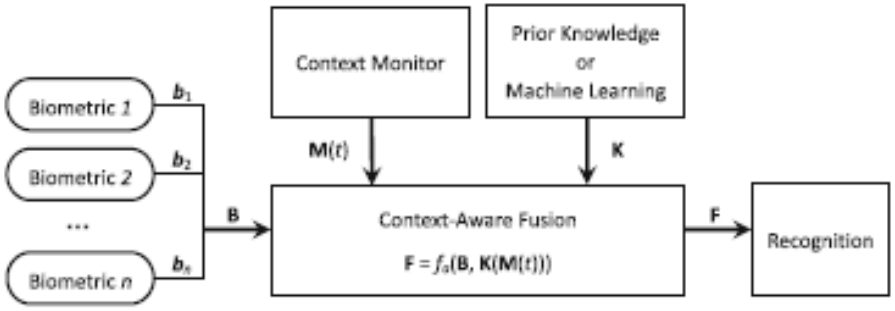
Approach	MuHAVi-14		MuHAVi-8	
Singh et al. [21] (baseline)		61.8%		76.4%
Cheema et al. [22]		73.5%		83.1%
Eweiwi et al. [24]		77.9%		85.3%
Without Fusion	$(L = 200, K = 80)$	81.6%	$(L = 300, K = 60)$	92.6%
Feature Fusion	$(L = 200, K = 100)$	80.9%	$(L = 200, K = 100)$	91.2%
Model Fusion	$(L = 450, K = 60)$	86.8%	$(L = 250, K = 75)$	95.6%

- Real-time suitability, 51 - 66 FPS (39 – 50 FPS at training).

Context Awareness

"any information that can be used to characterize the situation of entities" (Dey 2001)

Sense external (situational, environmental) factors that might affect decisions and adapt accordingly



Geng et al (2010), Context/aware fusion: a case study on fusion of gatin and face for human identification in video

What does a picture MEAN?

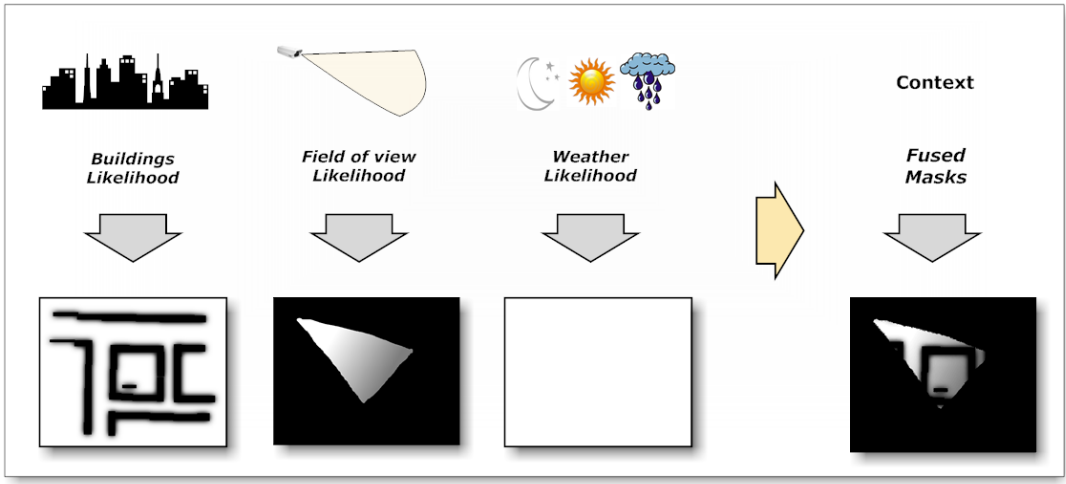
- Meaning implies **context** and experience (incl. non-visual).
- We are still not sure how to best represent and manipulate these.

An approach to context fusion

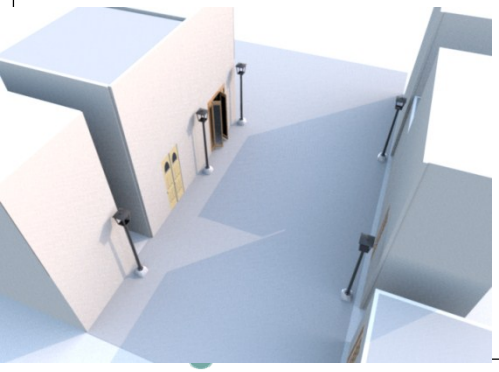
University of Udine, Italy

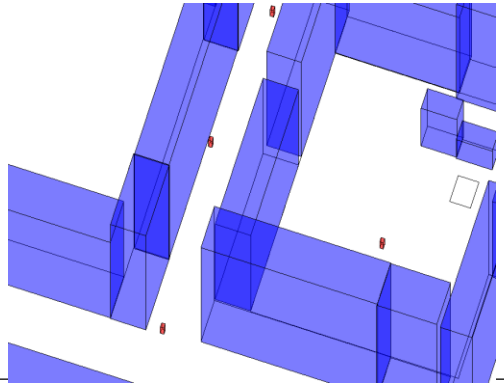
Lauro Snidaro
Ingrid Visentini
Gian Luca Foresti

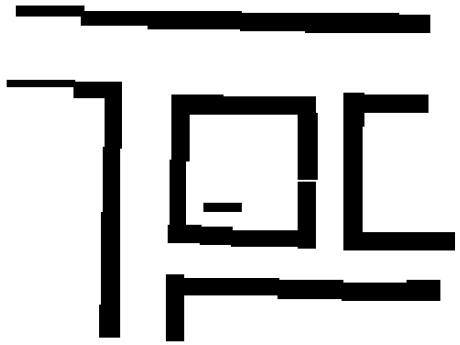
Context-enhanced Fusion



- **Likelihood** masks can be integrated with **observation likelihoods** for target tracking (e.g. context representing buildings, sensors fields of view and performance, etc.)
- Each mask can model the detection capabilities of a sensor regarding a particular contextual aspect
- Bayesian combination







Final Remarks

- Only able to illustrate some of the work in this field
- Fusion for Computer Vision is an active research area
- Many application areas: multi mode, multi sensor, action recognition
- How to use context and reasoning is a very interesting area
- CV community would benefit from more interaction with fusion experts

- Enough to do for many lifetimes!

UdeSantiago

Universidad Carlos III de Madrid

Thank you very much!

sergio.velastin@ieee.org

giaa