Tutorial: multicamera and distributed video surveillance

Prof. Rita Cucchiara

Università di Modena e Reggio Emilia, Italy
This tutorial addresses **algorithms and techniques of computer vision and pattern recognition** for **multicamera** and **distributed video surveillance**.

When multiple (heterogeneous) cameras are connected in a forest of sensors, standard techniques used in single- fixed camera surveillance are not sufficient anymore.

Different approaches should be taken into account depending on the **camera layout** (e.g., with **overlapped or not overlapped field of view**), the **camera motion** (e.g., fixed or PTZ cameras), the network capability and the availability of computational resource in the smart camera for early processing.

The tutorial aims at presenting a short survey of the research activities in this area, mainly focusing on **people surveillance**; models and algorithms for object segmentation and tracking in multi-camera environments will be presented in details with several demos from **IMAGElab of Modena**.

Techniques for people detection in cluttered environment will be presented. Finally, recent advances in **trajectory analysis for people behaviour** classification in distributed cameras systems will be discussed.

**Benchmark videos with ground truth** and tutorial material will be available for the tutorial attendees.
Presentation of ImageLab

- Computer Vision for robotic automation
- Medical Imaging
- Video analysis for indoor/outdoor surveillance
- Off-line Video analysis for telemetry and forensics
- Digital Library content-based retrieval
- Multimedia: video annotation

**IMAGELAB-SOFTECH**
Lab of *Computer Vision*, *Pattern Recognition* and *MULTIMEDIA*

Dipartimento di Ingegneria
dell’Informazione
Università di Modena e Reggio Emilia
**Italy**

[Website](http://imagelab.ing.unimore.it)
# Imagelab: recent projects in surveillance

## Projects:

- **THIS** Transport hubs intelligent surveillance EU JLS/CHIPS Project 2009-2010
- **VIDI-Video:** STREP VI FP EU → (VISOR VideosSurveillance Online Repository) 2007-2009
- **BE SAFE** NATO Science for Peace project 2007-2008
- **Detection of infiltrated objects for security** 2006-2008 Australian Council
- **Bheave_Lib**: Regione Emilia Romagna Tecnopolo Softech 2010-2013
- **LAICA** Regione Emilia Romagna; 2005-2007
- **FREE_SURF** MIUR PRIN Project 2006-2008
- **Stopped Vehicles** with Digitek Srl 2007-2008
- **SmokeWave**: with Bridge-129 Italia 2007-2010
- **Sakbot for Traffic Analysis** with Traficon 2004-2006
- **Mobile surveillance** with Sistemi Integrati 2007
- **Domotica per disabili**: posture detection FCRM 2004-2005
Video surveillance

Video-surveillance concerns models, techniques and systems for

• **acquiring videos** about the 3D external world,
• **detecting targets** along the time and the space,
• **recognizing** interesting or dangerous situations,
• **generating real-time alarms**
• **recording** meaningful data about the controlled scene.
Video surveillance

**Research world**
- ‘60-70 Hardware
- ‘80 Military research
- ‘90 Traffic monitoring
- ‘00 People surveillance
- ‘10 ? Life surveillance

**Commercial world**
- ‘60-70 Analogue cameras
- ‘80 Digital CCTV systems
- ‘90 Digital surveillance systems
- ‘90 Network surveillance systems
- ‘10? Ubiquitous, WAN systems
Commercial System Generations: Hardware

Hardware:
Gh1: Time-Lapse Video Cassette Recorder System
Gh2: Digital Video Recorder System
Gh3: PC-based Digital Video Surveillance System
Gh4: (LAN) Digital Camera System
Gh5: Streaming LAN Smart Camera System
Commercial System Generations: Software

Gs1: JPEG, MJPEG, filtering
Gs2: MJPEG, MPEG2, PTZ, motion detection
Gs3: MPEG4, H264, ..., computer vision for video analytics

Research in surveillance: a brief history

‘80–90 Research on Military Environment
→ FOCUS ON SENSORS AND IMAGE PROCESSING FOR HUMAN-BASED ENHANCEMENT
→ FOCUS ON TARGET DETECTION

‘90–00 Research on Traffic Surveillance
→ MANUAL CALIBRATION OF THE ROAD
→ VEHICLE DETECTION WITH ADAPTIVE BACKGROUND \[ b_{\text{new}}(x) = \alpha b_{\text{old}}(x) + \beta \]
→ DOUBLE DIFFERENCE WITH GRADIENT ANALYSIS
→ REASONING FOR CROSS MONITORING
→ SURVEY


V. Kastrinaki, M. Zervakis, K. Kalaitzakis A survey of video processing techniques for traffic applications Image and Vision Computing 2003
Research in surveillance:

'00 Research on People surveillance:
→ SYSTEMS DEVELOPED IN UNIVERSITY CAMPUS (PFINDER, w4, VSAM, SAKBOT, KNiGTH…………………)

→ PRECISE ALGORITHMS FOR DETECTION AND TRACKING FROM SINGLE FIXED CAMERA

→ MOTION DETECTION WITH BACKGROUND SUPPRESSION

→ USE OF COLOR FOR PRECISE DETECTION AND SHADOW SUPPRESSION

→ SINGLE OBJECT TRACKING

→ ANALYSIS OF OCCLUSIONS FOR GROUP OF PEOPLE..

→ RESEARCH LABS IN INDUSTRIES (SIEMENS, HONEYWELL, IBM, SARNOFF, MERL..)

PFINDER

VSAM PROJECT

… many others..

W4

SAKBOT

KNiGHT
Background suppression: milestones et al.

- **Color Analysis**
- **Mixture of Gaussian**
- Background and Shadow detection
- Background and layered representation
- Background with **KERNEL DENSITY**
- .......... And other hundreds of papers...


A SHORT SURVEY


Tao Sawneir, Kumar. Object tracking with bayesian estimation of dynamic layer representation IEEE Trans on PAMI 24,1 2002

A.Elgammal, R.Duraiswami, D.Harwood, L.S. Davis
Techniques for people tracking: milestones

- Many SURVEYs

- Most used techniques:
  - MeanSHIFT
  - PARTICLE FILTERING
  - Appearance based tracking

Moeslund, Hilton, Kruger A survey of advanced in vision-based human motion capture and analysis Computer Vision and Image Understanding vol 104 2006

Yilmaz Javed Shah Object tracking a survey ACM COMPUTING Survey vol 33 n 4 2006

Dorin Comaniciu, Visvanathan Ramesh: Mean Shift and Optimal Prediction for Efficient Object Tracking. ICIP 2000


..........

Wu, H, Sankaranarayanan, A; Chellappa, R; Online Empirical Evaluation of Tracking Algorithms IEEE Trans. PAMI 2009 forthcoming

Tao Zhao Nevatia, R Tracking multiple humans in complex situations IEEE Trans. PAMI 2004

A.Senior Tracking people with appearance Models PETS 2002

R.Cucchiara, C.Grana, G.Tardini, R.Vezzani Probabilistic people tracking for occlusion handling, ICPR04, 2004
To parallel and distributed surveillance

- A review of commercial systems
- A review of hardware and software requirements
- A survey of multicamera and distributed surveillance

Valera, M. Velastin, S.A.


Enlarging the surveillance dimensions.. 3T space

The 3T space: T for Time analysis

- HistoricalPast (t0-nDt)..
- Recent Past (t0-Dt)..
- Present t0
- Near Future (t0+Dt)..
- Future (t0+nDt)..

**FORENSICS**
What happened ?
What is the result of the investigation?

**Surveillance**
What is happening?
Who is this one?
Which event can be detected now?

**Human driven Surveillance/forensics**
What is just happened?
Where he/she come from?
Which is the cause of the event ?

**Forecasting**
Statistical analysis
What will happen ?
Enlarging the surveillance dimensions.. 3T space

- **The 3T space: T for Topology** (in Space)

Partitioned FoV → 1 Moving FOV → 1 FOV → N overlapped FOV → Distributed FoV

- **Surveillance with Zoom Analysis**
  - Focus on faces
  - Surveillance PT moving camera
  - Hand-driven moving cameras
  - Cameras installed in moving systems

- **Surveillance Dear Single Static Fixed Camera**
  - Focus on event
  - Surveillance Multicamera systems

- **Wide Area Surveillance**
  - Focus on attention gathering
  - network of cameras
  - Distributed Surveillance
Enlarging the surveillance dimensions.. 3T space

- **The 3T space: T for Target**

  - **Body details**: 0.1
  - **Body parts**: 0.5
  - **Single individual**: 1
  - **Group of people**: 2-3, 10
  - **Crowd**: 100...

**Surveillance- Biometry**
- Face detection
- Fingerprint
- Expression analysis

**Surveillance**
- Detection, Identification
- Action and Interaction recognition
- Behavioral analysis

**Surveillance/forensics**
- People posture,
- Gait, activity detection

**Surveillance, Crowd analysis**
- People Flow modeling
Putting all together

Large Area (Network) Surveillance

One-Camera People surveillance

Forensics & Biometry

- Distributed FoVs
- Multiple Overlapped FoVs
- Single individual
- Group of people
- Crowd

Future (t0+nDt)

Body details: 0.1
Body parts: 0.5
Single individual: 1
Group of people: 2.3
Crowd: 100...
Focus on multicamera - distributed surveillance

1. Introduction

2. Multicamera surveillance
   - Detection, tracking
   - Trajectory analysis

3. Not overlapped cameras
   - Searching correspondences
   - Tracking

4. Moving cameras
   - PTZ synchronization
   - Cameras in motion
Architectural requirements

- **Multicamera (multiview) surveillance**
  - fully synchronized acquisition; 1 framegrabber with 1-20 fixed and PTZ cameras; 1 (multiprocessor) computer for many cameras
  - \(\rightarrow\) *shared memory* architecture
  - Challenges: More precision, 3D reconstruction, consistent labeling in multiview, occlusion handling, people identification, behavior analysis

- **Distributed (network camera) surveillance**
  - loosely coupled acquisition and processing; potentially thousands of nodes with *smart cameras* and traditional network cameras
  - \(\rightarrow\) *message passing* architecture
  - Challenges: Large coverage, communication, bandwidth, tradeoff-local computation and computation power; less precision, multiple hypothesis generation, search for similarity
  
  + in addition *freely moving cameras*
  on vehicles, moving infrastructure, hand-cameras
Multi camera surveillance with overlapped FoVs

- Using multiple sensors/cameras has many advantages:
  - Wider coverage of the scene
  - Multi-modal sensing
  - Redundant data (improved accuracy)
  - Fault tolerance

- Multiple cameras require consistent labeling

- Consistent labeling requires homography and techniques for solving tracking in overlapped FoVs

A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa,
Object Detection, Tracking and Recognition for Multiple Smart Camera

Saad M. Khan and Mubarak Shah;
Tracking Multiple Occluding People by Localizing on Multiple Scene Planes;
IEEE TRANS. ON PAMI, VOL. 31, NO. 3, MARCH 2009
Consistent labeling: overview

- **Appearance-based approaches** base the matching essentially on the color of the objects: **color histograms**, faces, ...

- **Geometry-based approaches** exploit geometrical relations and constraints between different views: **homography-based, epipolar geometry, calibration-less**, ...

- **Mixed approaches** combine information about the geometry with information provided by the visual appearance: **probabilistic information fusion, BBNs, ...**

References:


## Consistent labeling: overview

<table>
<thead>
<tr>
<th>Reference Paper</th>
<th>Color Information</th>
<th>Features</th>
<th>Geometric Information</th>
<th>Calib.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stauffer et al. (2003) [12]</td>
<td>—</td>
<td>—</td>
<td>Homography and TCM</td>
<td>No</td>
</tr>
<tr>
<td>Chang et al. (2001) [14]</td>
<td>Color appearance and HPCA</td>
<td>Object apparent height</td>
<td>Epipolar geometry</td>
<td>No</td>
</tr>
<tr>
<td>Yue et al. (2004) [15]</td>
<td>—</td>
<td>—</td>
<td>Homography</td>
<td>Yes</td>
</tr>
<tr>
<td>Khan and Shah (2003) [16]</td>
<td>—</td>
<td>—</td>
<td>EoFoV</td>
<td>No</td>
</tr>
<tr>
<td>Dockstader &amp; Tekalp (2001) [17]</td>
<td>—</td>
<td>Person’s body model</td>
<td>3D projection</td>
<td>Yes</td>
</tr>
<tr>
<td>Tsutsui et al. (1998) [18]</td>
<td>—</td>
<td>—</td>
<td>3D projection</td>
<td>Yes</td>
</tr>
<tr>
<td>Krumm et al. (2000) [19]</td>
<td>Color histogram</td>
<td>—</td>
<td>Stereo match</td>
<td>Yes</td>
</tr>
<tr>
<td>Zhou and Aggarwal (2005) [20]</td>
<td>—</td>
<td>—</td>
<td>3D trajectory projection and EKF</td>
<td>Yes</td>
</tr>
<tr>
<td>Black and Ellis (2005) [21]</td>
<td>—</td>
<td>—</td>
<td>3D location</td>
<td>Yes</td>
</tr>
<tr>
<td>Nummioaro et al. (2003) [22], Tan and Ranganath (2003) [23]</td>
<td>Color appearance</td>
<td>Face database</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>Mittal and Davis (2001) [24], Mittal and Davis (2003) [25]</td>
<td>Color Region</td>
<td>—</td>
<td>Epipolar geometry and 3D projection</td>
<td>Yes</td>
</tr>
<tr>
<td>HRCOT. (2007)</td>
<td>—</td>
<td>—</td>
<td>Epipolar geometry and homography</td>
<td>No</td>
</tr>
</tbody>
</table>

The solution at Imagelab

HECOL (Homography and Epipolar-based COnsistent Labeling)

- Ground-plane homography and epipolar geometry automatically computed from training videos
- Person’s main axis warped to the other view and Bayesian inference is used for validate hypotheses
Video surveillance at ImageLab

Sakbot
- Fixed camera
- Segmentation
- Tracking

Ad-hoc
- Fixed camera
- Segmentation
- Tracking

Hecol
- Geometry Recovery
- Homography & Epipoles
- Posture analysis
- Action analysis
- Trajectory Analysis
- HeadTracking
- Face selection
- High Resolution Detection
- Face Recognition & People Identification
- Face obscuration
- Video surveillance Ontology
- People Annotation
- Video surveillance Streaming
- Annotated video storage
- MPEG Streaming
- WEB
- Security control centers
- Mobile surveillance platforms

VISOR
- Consistent Labeling
- Multicamera tracking
- ROI and model-based Tracking
- Sensor data acquisition
- Mosaicing Segmentation & Tracking
- PTZ Control

Moses
- Video surveillance
- Ontology
Hecol on-line off line processes

1) off-line process for **HOMOGRAPHY AND EPIPOLE DETECTION** and construction of a Camera Transition Graph

2) detection and tracking in each single camera system and

3) consistent labeling at each new track detection
Homography

**One to One**

- One point in the 3D space is projected in a single point in the image plane.
- Thus, knowing the position of the point in the space and the camera parameters, I can know if the point can fall in the image plane.

**One to many**

- Instead, a point in the image plane can correspond to many points: to all the points of the line passing through the optical center and the source point.
But *if I have at least one constraint* (e.g. the point is at $z=0$) from a point $P$ in the image plane I can know its position in the real space
Rectification

- Hypothesis to have all the points in the ground plane
Homographic transformation matrix

- Homography is a linear non singular transformation of the projective plane in itself.

- Given a plane (e.g. ground plane) and given two bidimensional projections ( two image planes) the homographic transformation links the coordinates of the points in the two reference systems.

\[
\begin{bmatrix}
    x_1' \\
    x_2' \\
    1
\end{bmatrix} = \lambda \begin{bmatrix}
    x_1 \\
    x_2 \\
    1
\end{bmatrix}
\]

\[
H^{3 \times 3} = \begin{bmatrix}
    h_1 & h_2 & h_3 \\
    h_4 & h_5 & h_6 \\
    h_7 & h_8 & 1
\end{bmatrix}
\]

- \(H\) is defined at less of a scale factor and thus with 8 degrees of freedom.

- It can be defined with 4 points of corrispondence ( 4 points → 8 variables).
Homography computation

The **homography** is a planar projective transformation that relates point coordinates lying on the planes used to construct the warping matrix.

In projective coordinates a planar homography can be described by a 3x3 matrix and 8 independent parameters.

\[
\begin{bmatrix}
    x' \\
    y' \\
    z'
\end{bmatrix} =
\begin{bmatrix}
    h_{1,1} & h_{1,2} & h_{1,3} \\
    h_{2,1} & h_{2,2} & h_{2,3} \\
    h_{3,1} & h_{3,2} & h_{3,3}
\end{bmatrix}
\begin{bmatrix}
    x \\
    y \\
    z
\end{bmatrix}
\]

\[
\begin{bmatrix}
    H'
\end{bmatrix} =
\begin{bmatrix}
    h_{1,1}/h_{3,3} & h_{1,2}/h_{3,3} & h_{1,3}/h_{3,3} \\
    h_{2,1}/h_{3,3} & h_{2,2}/h_{3,3} & h_{2,3}/h_{3,3} \\
    h_{3,1}/h_{3,3} & h_{3,2}/h_{3,3} & 1
\end{bmatrix}
\]

- Selections of 4 far not aligned points:
  - **Manual selection** in initial calibration step
  - **Automatic selection** with human probe
Off-line stage

- Off-line process automatically computes ground-plane homography with $E^2oFoV$ and epipolar constraints
- Define the Entry Edge Field of Views


Automatic Homography computation 1/2

- Automatic learning phase to compute overlapping zones and ground-plane homography.
- Take many correspondences among ground plane support points (with a tracking algorithm for a single person).
- Define the Entry Edge Field of Views $E^2oFoV$ using Least Square Optimization.
- Define the overlapping zones and the extremes points.
- Compute the homography from points correspondences.
Examples

Engineering Campus of University Of Modena

From (http://www-sop.inria.fr/orion/ETISEO/)

G. Kayumbi and A. Cavallaro
Multiview Trajectory Mapping Using Homography with Lens Distortion Correction
EURASIP Journal on Image and Video Processing 2008
Epipolar geometry recovery

- Pure ground-plane homography-based matching is not reliable in case of segmentation errors and groups of people we need another 3D information.
- Using only homography we can detect only the presence in the planar world.
- For recovering the epipolar geometry we must obtain point correspondences not on the ground plane.

Thus we take the heads! (upper points of the blob)

Epipole computation is performed with RANSAC to improve numerical stability of the solution.
EPIPOLAR GEOMETRY

\[ l_{1,up} = \langle up_1, H \cdot up_2 \rangle = e_1 \times (H \cdot up_2) \]

- Where \( l \) is the epipolar line in the image plane of C1 passing through \( up \) projections
- \( H \) is the homography matrix from C2 to C1 ground planes
- \( \langle a, b \rangle \) Indicates the line passing through points a and b

**The intersections of two lines univocally identify the epipole**
- Point correspondences can be affected by errors if extracted automatically

- **RANSAC** is used to iteretively choose the two lines that give the best epipole locations

- **ALGORITHM:**
  1. during the previous $E^2oFoV$ training phase, correspondences between object’s head projections up in both cameras are sampled frame by frame.
  2. after choosing a camera, $C$, we randomly compute two epipolar lines taken from the sample set of $N$ frames using epipolar line equation and detect the epipole location.
  3. we evaluate the consensus of remaining samples counting how many samples are close enough to the epipolar lines computed using estimated epipole
  4. iterate from 2 until the consensus is above a fixed threshold.
On-line Stage: Consistency resolver

- HECOL defines a Bayesian-competitive approach with warping of vertical axis and a two-contributions check.
- Each hypothesis in two overlapped cameras is accounted both as single and as group.

Consistency resolver whenever a new object is detected (not only to the EoFoV)

an appearance-based tracking is needed
CTG: modeling the camera overlaps

- When a new track appears on one camera the search for possible matching tracks could be time consuming

- A Camera Transition Graph is built to speed up the search process:
  - The Graph incorporates the cameras topology learnt during off-line geometry learning stage
  - Each node consists of one camera
  - Arcs represent the overlapping constraint among cameras FoVs
  - Variables (tracks) at each node must satisfy the unary constraint of having different labels
  - When a new object appears (variable added at a node) the binary constraint that two instances of the same object on different nodes must have the same label is verified (Constraint Satisfaction Problem)
Example
Bayesian-competitive framework

- Given a new track $\tau_{\text{new}}^l$ on camera $l$ hypotheses’ space is build for each of the possible overlapped cameras in CTG.
- Each hypothesis in two overlapped cameras is accounted both as single and as group.
- Inter-camera MAP:

$$p(\gamma_k^{l,i} | \tau_{\text{new}}^l) = \frac{p(\tau_{\text{new}}^l | \gamma_k^{l,i}) p(\gamma_k^{l,i})}{\sum_{k=1}^n p(\tau_{\text{new}}^l | \gamma_k^{l,i}) p(\gamma_k^{l,i})}$$

prior
A hypothesis **consisting of a single object** will gain higher prior if the warped l.s.p. is far enough from the other objects’ support points.

A hypothesis **consisting of two or more objects** (i.e., a possible group) will gain higher prior if the objects composing it are close to each other after the warping, and, at the same time, the whole group is far from other objects.

\[
Id(\tau^l_a, \tau^l_b) = \begin{cases} 
\| (H^{1,1}l_p^a_a) \times (H^{1,1}l_p^b_b) \| & \text{if } \{\tau^l_a, \tau^l_b\} \in \gamma^l_{ik} \\
0 & \text{otherwise}
\end{cases}
\]

\[
Od(\tau^l_a, \tau^l_b) = \begin{cases} 
\| (H^{1,1}l_p^a_a) \times (H^{1,1}l_p^b_b) \| & \text{if } \{\tau^l_a, \tau^l_b\} \in \Sigma^{l,1}(\tau_{new}) \land \left(\tau^l_a \in \gamma^l_{ik} \land \tau^l_b \notin \gamma^l_{ik}\right) \\
0 & \text{otherwise}
\end{cases}
\]

\[
\sigma \left(\gamma^l_{ik}\right) = \min_{\tau^l_a, \tau^l_b \in \Sigma^{l,1}(\tau_{new}) \ a \neq b} \frac{Id(\tau^l_a, \tau^l_b)}{Od(\tau^l_a, \tau^l_b)} - \max_{\tau^l_c, \tau^l_d \in \Sigma^{l,1}(\tau_{new}) \ c \neq d} Id(\tau^l_c, \tau^l_d)
\]
Prior computation

- Priors must account for different probability in case of multiple hypotheses:

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$\sigma$ value [pxl]</th>
<th>Prior value</th>
<th>Hypothesis</th>
<th>$\sigma$ value [pxl]</th>
<th>Prior value</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ $\tau^2_{76}$ }</td>
<td>+52</td>
<td>0, 167</td>
<td>{ $\tau^2_{76}, \tau^2_{74}$ }</td>
<td>+61</td>
<td>0, 186</td>
</tr>
<tr>
<td>{ $\tau^2_{74}$ }</td>
<td>+52</td>
<td>0, 167</td>
<td>{ $\tau^2_{74}, \tau^2_{90}$ }</td>
<td>-84</td>
<td>0, 100</td>
</tr>
<tr>
<td>{ $\tau^2_{90}$ }</td>
<td>+113</td>
<td>0, 200</td>
<td>{ $\tau^2_{76}, \tau^2_{74}, \tau^2_{90}$ }</td>
<td>-61</td>
<td>0, 110</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>{ $\tau^2_{76}, \tau^2_{74}, \tau^2_{90}$ }</td>
<td>-136</td>
<td>0, 070</td>
</tr>
</tbody>
</table>
Likelihood is computed as the max of two contributions (forward - backward):

\[
p(\tau_{\text{new}}^i | \gamma_k^{i,j}) = \max(fP_{\text{backward}}(\tau_k^i), fP_{\text{forward}}(\tau_k^i))
\]

\[
fP_{\text{forward}}(\tau_{\text{new}}^i | \gamma_k^{i,j}) = \frac{\sum_{\tau_k^i \in I_{k}^i} \xi_{\gamma_k^{i,j} \rightarrow \tau_k^i}}{\text{card}(\sum_{i,j} (\tau_{\text{new}}^i)) (1 - S)}
\]

\[
fP_{\text{backward}}(\tau_{\text{new}}^i | \gamma_k^{i,j}) = \frac{\sum_{\tau_k^i \in I_{k}^i} \xi_{\gamma_k^{i,j} \rightarrow \tau_k^i}}{\text{card}(\sum_{i,j} (\tau_{\text{new}}^i)) (1 - S)}
\]

Fitness measure

\[
\rho(x_k^i, a_1^i, a_2^i, \tau^i) = \begin{cases} 
1 & \text{if } (x_k^i \in FG(\tau^i)) \land (x_k^i \in \langle a_1^i, a_2^i \rangle) \\
0 & \text{otherwise}
\end{cases}
\]

MAP label assignment

\[
\tau_{\text{new}}^l \rightarrow \tau_M^l \text{ with } M = \left\{ x_m^i(t) | \tau_m^i \in \gamma_{h}^{i,i} \right\} \text{ where } \\
h = \arg \max \limits_k \left( p(\gamma_{k}^{i,i} | \tau_{\text{new}}^i) \right) \text{ with } i = \arg \max \limits_n \left( p(C^n | \tau_{\text{new}}^i) \right)
\]
Experimental results (1)

- The system has been tested in a setup at our campus

<table>
<thead>
<tr>
<th>Name</th>
<th>Frame number</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Handoff Video (SHV)</td>
<td>12600</td>
<td>14 min</td>
</tr>
<tr>
<td>Miss segmentation Video (MSV)</td>
<td>16200</td>
<td>18 min</td>
</tr>
<tr>
<td>Group Video 1 (GV1)</td>
<td>9900</td>
<td>11 min</td>
</tr>
<tr>
<td>Group Video 2 (GV2)</td>
<td>14400</td>
<td>16 min</td>
</tr>
<tr>
<td>Mixed Video 1 (MV1)</td>
<td>22500</td>
<td>25 min</td>
</tr>
<tr>
<td>Mixed Video 2 (MV2)</td>
<td>9000</td>
<td>10 min</td>
</tr>
</tbody>
</table>
### Experimental results (2)

<table>
<thead>
<tr>
<th></th>
<th># Single HO</th>
<th># Segm.err.</th>
<th># Grp enter</th>
<th># Grp insd.</th>
<th># Simul. HO</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Correct</td>
<td>Total</td>
<td>Correct</td>
<td>Total</td>
<td>Correct</td>
</tr>
<tr>
<td>Only Homography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHV</td>
<td>137</td>
<td>128</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEV</td>
<td>28</td>
<td>28</td>
<td>76</td>
<td>30</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>GV1</td>
<td>30</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>GV2</td>
<td>78</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>MV1</td>
<td>39</td>
<td>34</td>
<td>21</td>
<td>10</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>MV2</td>
<td>52</td>
<td>50</td>
<td>38</td>
<td>16</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>Forward Contribution Only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHV</td>
<td>137</td>
<td>135</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEV</td>
<td>28</td>
<td>28</td>
<td>76</td>
<td>34</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>GV1</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>GV2</td>
<td>78</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>MV1</td>
<td>39</td>
<td>35</td>
<td>21</td>
<td>10</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>MV2</td>
<td>52</td>
<td>51</td>
<td>38</td>
<td>19</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>Backward Contribution Only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHV</td>
<td>137</td>
<td>131</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEV</td>
<td>28</td>
<td>28</td>
<td>76</td>
<td>65</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>GV1</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>GV2</td>
<td>78</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>MV1</td>
<td>39</td>
<td>36</td>
<td>21</td>
<td>16</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>MV2</td>
<td>52</td>
<td>52</td>
<td>38</td>
<td>35</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHV</td>
<td>137</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEV</td>
<td>28</td>
<td>28</td>
<td>76</td>
<td>75</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>GV1</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>GV2</td>
<td>78</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>MV1</td>
<td>39</td>
<td>39</td>
<td>21</td>
<td>19</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>MV2</td>
<td>52</td>
<td>52</td>
<td>38</td>
<td>36</td>
<td>30</td>
<td>15</td>
</tr>
</tbody>
</table>
Experimental results (3)

Examples of logging appearance

objects from camera 1

objects from camera 2
Park experimentation
Distributed surveillance

- **Network of (smart) cameras**: Not overlapped FoVs; loosely coupled.
- Problems of node communication
- If moving cameras: problems of calibration and tracking. The simultaneous localization and tracking (SLAT) problem, to estimate both the trajectory of the object and the poses of the cameras.

- Problem of tracking and **distributed consistent labeling**
  1) Tracking on single camera: multi hypothesis generation
  2) Camera world calibration
     - Geometry
     - Color
  3) Prediction into new cameras’ FoVs
  4) matching

---


Color calibration

- Methods:
  - Linear transformation
  - Independent channels
    \[
    \begin{bmatrix}
      A_R \\
      A_G \\
      A_B
    \end{bmatrix} = \begin{bmatrix}
      \alpha_1 B_R + \beta_1 \\
      \alpha_2 B_G + \beta_2 \\
      \alpha_3 B_B + \beta_3
    \end{bmatrix}
    \]
  - Full matrix (M computed with LSQ)
    \[
    \begin{bmatrix}
      A_R \\
      A_G \\
      A_B
    \end{bmatrix} = M \begin{bmatrix}
      B_R \\
      B_G \\
      B_B
    \end{bmatrix}
    \]
  - Look-up table
  - for non linear
  - transformation


## Example

<table>
<thead>
<tr>
<th>Camera A</th>
<th>Camera B</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Original" /></td>
<td><img src="image2.png" alt="Original" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Independent channels" /></td>
<td><img src="image4.png" alt="Independent channels" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Look-up table" /></td>
<td><img src="image6.png" alt="Full matrix" /></td>
</tr>
</tbody>
</table>
Feature to match

- Color (single / multiple)
- Shape (geometrical ratios / spline / elliptical models)
- Motion (speed, direction)
- Gait (Fourier transform)
- SIFT +, greylevel co-occurrence matrix, Zernike moments and some simple colour features
- Polar color histogram + Shape


Distributed Surveillance at ImageLab

- The problem:
- a people disappeared in the scene exiting from a camera FoV, where can be detected in the future?
- 1) tracking with a camera FoV
- 2) tracking in exit zones
- 3) matching in the entering zones

- Using **Particle Filtering + Pathnodes**
  - In computer graphic all the possible avatar positions are represented by nodes and the connecting arcs refers to allowed paths. The sequence of visited nodes is called **pathnodes**.
  - A weight can be associated to each arc in order to give some measures on it, such as the duration, the likelihood to be chosen with respect to other paths, and so on.
  - Weights can be defined or learned in a testing phase
To avoid all-to-all matches, the tracking system can exploit the knowledge about the scene

- Preferential paths -> Pathnodes
- Border line / exit zones
- Physical constraints & Forbidden zones NVR
- Temporal constraints
A possible path between Camera 1 and Camera 4
Pathnodes lead particle diffusion
Results with PF and pathnodes

Single camera tracking:
Recall=90.27%
Precision=88.64%

Multicamera tracking
Recall=84.16%
Precision=80.00%
Frame 431. a man #21 exits and his particles are propagated

Frame 452 a person # 22 exits too and also his particles are propagated

Frame 471 a people is detected in Camera #2 and the particles of both # 21 and #22 are used but the ones of #22 match and person à22 is recognized
And thus?

- Thus, providing a both multicamera and distributed consistent labeling we could
- identify people
- detecting their trajectories
- analyzing interaction
- defining abnormal behavior
- detecting suspicious situations.
- ..
- Reacting and defining alarm in realtime
- Forecasting crowd situation
An example: trajectory shape analysis

- Trajectory shape analysis for “abnormal behavior” recognition in open space
- Trajectory Shape similarity; invariant to space shifts
- Not only space-based or time-based similarity
Trajectory similarity and clustering

- **Trajectories similarity measures:**
  - Exact pairwise comparison
    - Euclidean Distance \([Fu \ et \ al. \ ICIP05]\)
    - Hausdorff distance \([Lou \ et \ al. \ ICPR02, \ Junejo \ et \ al. \ ICPR04]\)
  - Inexact alignment based comparison
    - LCSS (Longest Common Subsequence) \([Buzan \ et \ al. \ ICPR04]\)
    - DTW (Dynamic time warping) \([Keogh \ et \ al. \ ACM \ KDD00]\)
  - Statistical matching
    - HMM \([Porikli \ et \ al. \ CVPRW04], \ [Bashir \ et \ al. \ TIP \ 07]\)

- **Clustering techniques:**
  - Hard-clustering:
    - Vector Quantization and Spatial Saussians \([Mecocci \ et \ al. \ ICIP \ 05]\)
    - Graph Cuts \([Junejo \ et \ al. \ ICPR04]\)
    - Spectral clustering \([Hu \ et \ al. \ TIP07, \ Poikli \ et \ al. \ CVPRW04]\)
  - Soft Clustering:
    - EM \([Bennewitz \ et \ al. \ Intelligent \ Robot \ Systems \ 02]\)
    - Fuzzy K-means in spatial and temporal domains \([Hu \ et \ al. \ PAMI06]\)
1. **Trajectory description** with angle sequence
   \[ T_j = \{ \theta_{1,j}, \theta_{2,j}, \ldots, \theta_{n,j} \} \]

2. **Statistical representation** with a Mixture of Von Mises Distributions (MovM)

3. **Coding** with a sequence of selected vM pdf identifiers

4. **Code Alignment**

5. **Clustering** with k-medoids

---

A. Prati, S. Calderara, R. Cucchiara, "Using Circular Statistics for Trajectory Analysis" in *Proceedings of International Conference on CVPR 2008*
Training set and on-line classification

\[ T_j = \{\theta_{1,j}, \theta_{2,j}, \ldots, \theta_{n,j}\} \]

\[ \text{MovM}(T_j) \]

\[ <S = \{S_{ij}, S_{njj}\}, \text{MovM}(T_j)> \]

\[ M = \{M^1, \ldots, M^k\} \]

**Surveillance system**

**Trajectory repository**

**On-line EM for MoVM**

**Coding with MAP**

**Alignement**

**Clustering with Br distance**

**Classification with Br distance**

**Normal/abnormal**
Some tests

- 15 classes of synthetic trajectories
- 5 classes of real trajectories
- Experiments:
  - Periodicity
  - Multimodality
  - Sequence order
  - Noise
  - ..
Experimental results

- Results are evaluated in terms of:
  - classification accuracy
  - normal/abnormal accuracy

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
<th>Test 7</th>
<th>Test 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of test</td>
<td>Periodicity</td>
<td>Noise</td>
<td>Mono-modal</td>
<td>Multi-modal</td>
<td>Sequence</td>
<td>Learn Norm.</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Total n. traj.</td>
<td>Tr</td>
<td>250</td>
<td>400</td>
<td>450</td>
<td>680</td>
<td>650</td>
<td>150</td>
<td>2530</td>
</tr>
<tr>
<td></td>
<td>Te</td>
<td>150</td>
<td>250</td>
<td>400</td>
<td>430</td>
<td>400</td>
<td>200</td>
<td>1700</td>
</tr>
<tr>
<td>Average points/traject.</td>
<td>65</td>
<td>95</td>
<td>66</td>
<td>74</td>
<td>83</td>
<td>93</td>
<td>103</td>
<td>66</td>
</tr>
<tr>
<td>Training set</td>
<td>$C_3^<em>, C_4^</em>$</td>
<td>$C_1, C_{15}$</td>
<td>$C_1, C_2, C_3$</td>
<td>$C_1, C_2, C_3, C_5, C_8^*$</td>
<td>$C_1, C_8, C_{10}$</td>
<td>$C_1, C_{13}$</td>
<td>all training</td>
<td>$C_{R1}, C_{R2}, C_{R3}, C_{R4}^*$</td>
</tr>
<tr>
<td>Testing set</td>
<td>$C_3^<em>, C_4^</em>$</td>
<td>$C_{14}^*, C_{15}$</td>
<td>$C_1, C_2, C_3, C_4^*$</td>
<td>$C_1, C_2, C_3, C_4^*$</td>
<td>$C_8, C_{9}^*, C_{11}, C_{12}, C_{13}$</td>
<td></td>
<td>$C_{13}, C_{7}$</td>
<td>all testing</td>
</tr>
<tr>
<td>Class. accuracy</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>97.67%</td>
<td>94.59%</td>
<td>95%</td>
<td>99.60%</td>
<td>96.77%</td>
</tr>
<tr>
<td>Norm/Abn. accuracy</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>97.30%</td>
<td>100%</td>
<td>100%</td>
<td>96.77%</td>
</tr>
<tr>
<td>Class. accuracy HMM</td>
<td>75%</td>
<td>83%</td>
<td>94.1%</td>
<td>92.67%</td>
<td>89.1%</td>
<td>94.01%</td>
<td>86.40%</td>
<td>75.19%</td>
</tr>
<tr>
<td>Norm/Abn. accuracy HMM</td>
<td>84%</td>
<td>94.1%</td>
<td>93.02%</td>
<td>100%</td>
<td>85.1%</td>
<td>100%</td>
<td>82.4%</td>
<td>66%</td>
</tr>
</tbody>
</table>

- Compared with a HMM-based approach*

Example

- Example of clustering similar trajectories using Mixture of Von Mises and speed
Another Application

- Finding similar people at different time

S. Calderara, R. Cucchiara, A. Prati *Multimedia Surveillance: Content-based Retrieval with Multicamera People Tracking* Proc of VSSN 2006
Final example

- Example of 2 people detected in 3 different cameras
In addition...

- **Moving cameras:**
  - Problems of tracking with moving cameras
  - Problems of fast and low-bandwidth communication


**G. Gualdi, A. Prati, R. Cucchiara,** Covariance Descriptors on Moving Regions for Human Detection in Very Complex Outdoor Scenes in Proc of ICSDC 2009

- **Large view cameras:**
  - problem of people detection in large areas
Available data set

VISOR: Video Surveillance Online repository
http://Imagelab.ing.unimore.it/visor
ViSOR
Video Surveillance
Online Repository

http://www.openvisor.org
Aims of ViSOR

- Gather and make freely available a repository of surveillance videos
- Store metadata annotations, both manually provided as ground-truth and automatically generated by video surveillance tools and systems
- Execute Online performance evaluation and comparison
- Create an open forum to exchange, compare and discuss problems and results on video surveillance

http://www.openvisor.org
Video Surveillance Ontology

- People, fixed and mobile objects (e.g., trees, vehicles, buildings, and so on…)
- Actions by or interaction between people; events
- Metadata (e.g.: author of the video, camera information, calibration data, location, date and time, and so on…)

Concept List

- Concept
- Content
- Context
- Physical Object
- Action/Event
Different types of annotation

- **Structural Annotation**: video size, authors, keywords,…

- **Base Annotation**: ground-truth, with concepts referred to the whole video. **Annotation tool: online!**

- **GT Annotation**: ground-truth, with a frame level annotation; concepts can be referred to the whole video, to a frame interval or to a single frame. **Annotation tool: Viper-GT (offline)**

- **Automatic Annotation**: output of automatic systems shared by ViSOR users.
Actual Video Corpus set

• At the moment (August 2009) ViSOR contains about 200 videos grouped into 14 categories

• Some examples:
  • 5 videos for shadow detection
  • 14 videos for smoke detection
  • 40 Videos of different human actions
  • 6 videos of surveillance of entrance doors
  • 33 videos for consistent labeling on a set of partially overlapped cameras

• 1,160,698 frames – 13h
### Video corpus set: the 14 categories

<table>
<thead>
<tr>
<th>Video Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor Domotic Unimore D.I.I. setup</td>
</tr>
<tr>
<td>Long videos for human action recognition</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td>Outdoor - other</td>
</tr>
<tr>
<td>Outdoor Unimore D.I.I. setup - Multicamera</td>
</tr>
<tr>
<td>Shadows</td>
</tr>
<tr>
<td>Video for indoor people tracking with occlusion</td>
</tr>
<tr>
<td>Videos for Smoke detection</td>
</tr>
<tr>
<td>Videos for Stopped Vehicles Detection</td>
</tr>
<tr>
<td>Videos from the IseLab - Computer Vision Center...</td>
</tr>
<tr>
<td>Videos of different human actions</td>
</tr>
<tr>
<td>Videos used for the VSSN background competition</td>
</tr>
<tr>
<td>Videosurveillance of entrance doors</td>
</tr>
</tbody>
</table>
Outdoor multicamera

Synchronized views
# Surveillance of entrance door of a building

- **About 10h!**

## Videosurveillance of entrance doors

<table>
<thead>
<tr>
<th>Image</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>c3_196808</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>c4_196808</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Entrance_doors_01060_8</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /></td>
<td>entrance_doors_01060_8b</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td>Entrance_doors_31070_8a</td>
</tr>
</tbody>
</table>
## Videos for smoke detection with GT

<table>
<thead>
<tr>
<th>Videos for Smoke detection</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="01_ballistic.avi" alt="Video 1" /></td>
</tr>
<tr>
<td><img src="06_fumogeno3.avi" alt="Video 6" /></td>
</tr>
</tbody>
</table>

For problems or suggestions, contact Roberto Ver.
Videos for shadow detection

- Already used from many researcher working on shadow detection
- Some videos with GT

Videos for face detection with moving camera

Outdoor Video For Face detection

FACE_video1-original_.mpg  FACE_video2-original_.mpg  outdoor-cluttered.av_i
Action recognition

- 40 videos
- 10 different actions, such as “Taking off the jacket”, “walking”, “wearing glasses”, ...

We invite you to

- Contribute to the growing and population of both the video repository and the annotation set
- Join to the community using the forum
- Sending us comments and suggestions…

http://www.openvisor.org

Ing. Roberto Vezzani
Dipartimento d’Ingegneria dell’informazione, Modena
roberto.vezzani@unimore.it
Tel +39-059-2056270
For any other information

http://Imagelab.ing.unimo.it
Rita Cucchiara
Dipartimento di Ingegneria
dell’Informazione
+390592056136
rita.cucchiara@unimo.it

Thanks to Imagelab

ANDREA PRATI, ROBERTO VEZZANI, COSTANTINO
GRANA, SIMONE CALDERARA, GIOVANNI GUALDI,
PAOLO PICCININI, PAOLO SANTINELLI, DANIELE
BORCHIESANI, DAVIDE BALTIERI, SARA CHISSI,
RUDY MELLI, EMANUELE PERINI, GIULIANO
PISTONI