Foreground detection and tracking in 2D/3D

José Luis Landabaso

Montse Pardàs

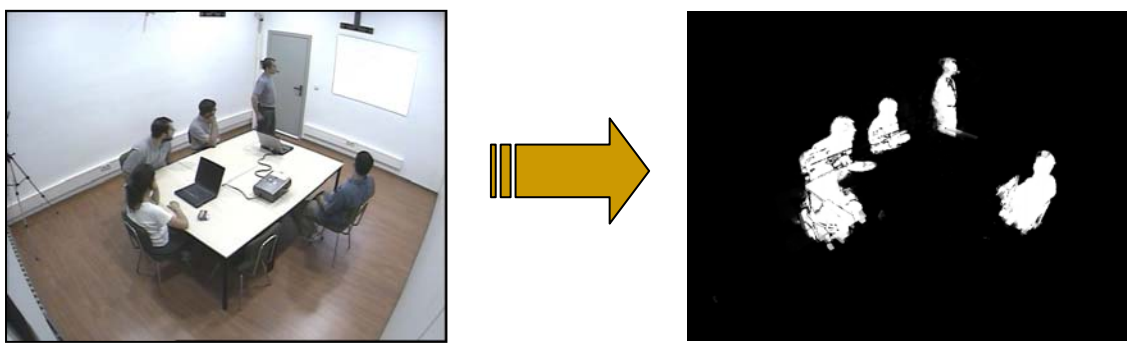
Outline

- 2D Foreground Detection
- 2D Object Tracking
- 3D Foreground Detection
- 3D Object Tracking

2D Entity Detection: Stauffer &Grimson

- We model each pixel by a mixture of K Gaussians in RGB color space (each Gaussian characterizes different color appearances)
- Means, variances and weights in Gaussians are adapted based on the classified incoming pixels

- Background pixels are characterized by Gaussians with high weight at low variances

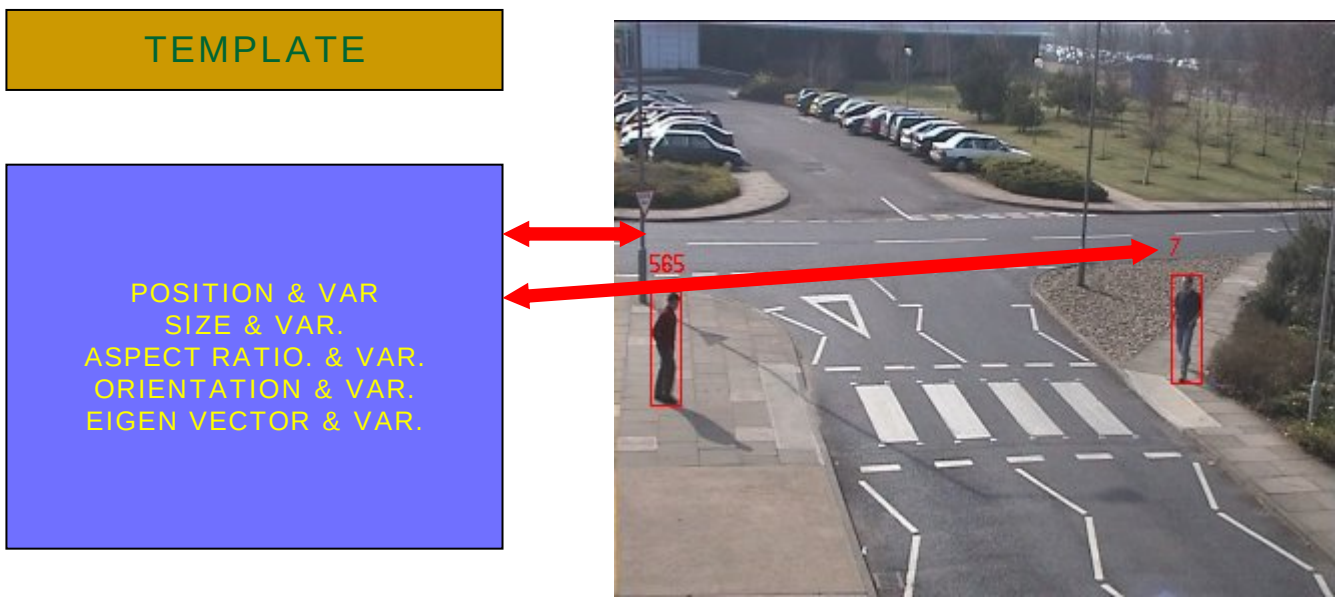


2D Model Extraction

- Each group of connected pixels (blob) is tagged as an object candidate
- Each object is represented by a template of features:
  - Vertical and horizontal position & velocity
  - Size of the blob
  - Aspect ratio
  - Orientation of the blob
  - Colour information
- Features are predicted with a Kalman filter prior to the matching

2D Entity Tracking I

- Distance is calculated between each blob (candidate) and all object templates
- It's a match if the minimum distance is lower than a certain threshold

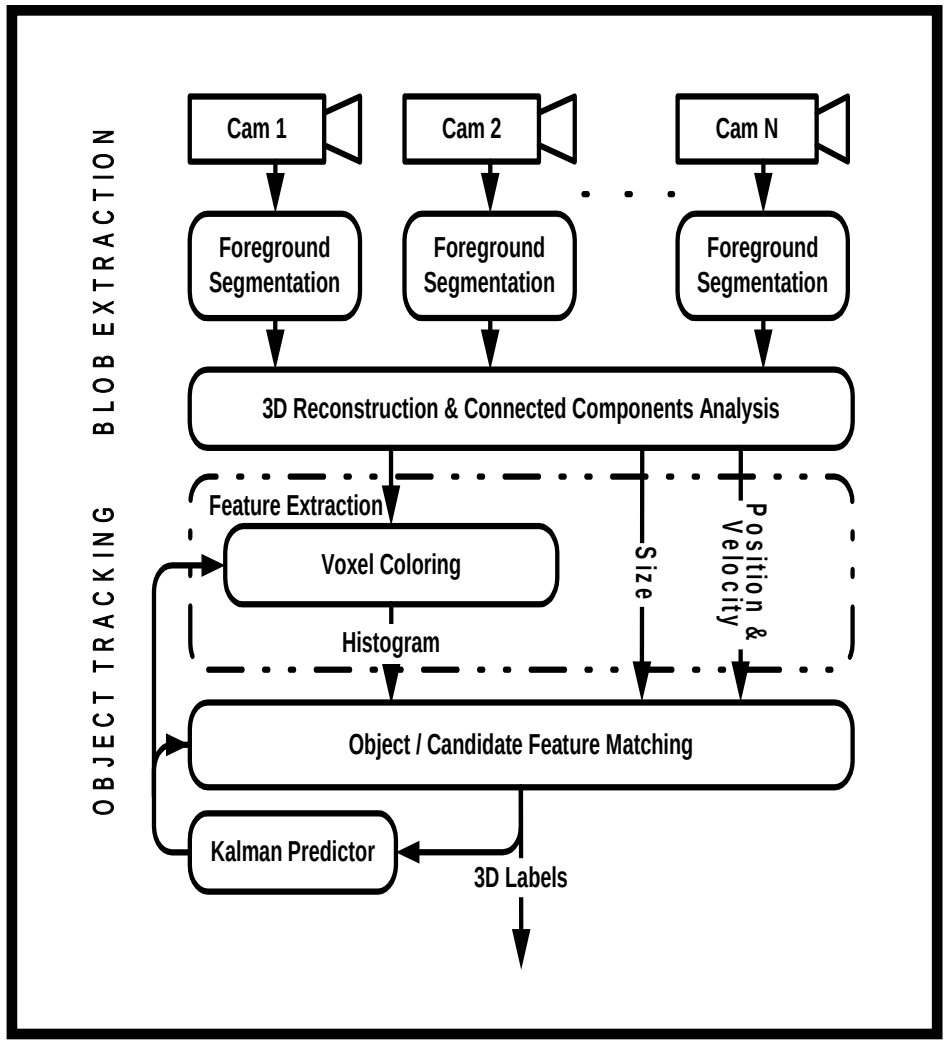


2D Entity Tracking II

- VIDEO

3D Extension

- The method uses a foreground separation process at each camera
- A 3D-foreground scene is modeled and discretized into voxels (VOlumetric piXELS) making use of all the segmented views
- Voxels are grouped into 3D blobs, whose colors are modeled for tracking purposes
- Color information together with other characteristic features of 3D object appearances are temporally tracked using a template-based technique, similarly as in the 2D case



Shape from Silhouette

- In multi-camera systems, Shape-from-Silhouette (SfS) is a common approach taken to reconstruct the Visual Hull, i.e. the 3D-Shape, of the bodies.
- Silhouettes are usually extracted using foreground segmentation techniques in each of the 2D views.
- The Visual Hull is formally defined as the intersection of the visual cones formed by the back-projection of several 2D binary silhouettes into the 3D space.

3D Entity Detection

(Shape from Silhouette)

Those parts of the volume which are in the intersection of ALL the Visual Cones are marked as ‘occupied’

- VIDEO

3D Entity Detection II

(Shape from Silhouette)

In principle, the more Visual Cones, the better object detection

- VIDEO

3D Entity Detection III

(Shape from Silhouette)

Resulting detection

- VIDEO

3D Entity Detection IV

(Shape from Silhouette)

8 Cameras

- VIDEO

3D Entity Detection V

(Shape from Silhouette)

8 Cameras

- VIDEO

3D Entity Detection VI

(Shape from Silhouette)

Can we just introduce more cameras to obtain more accurate 3D detections?

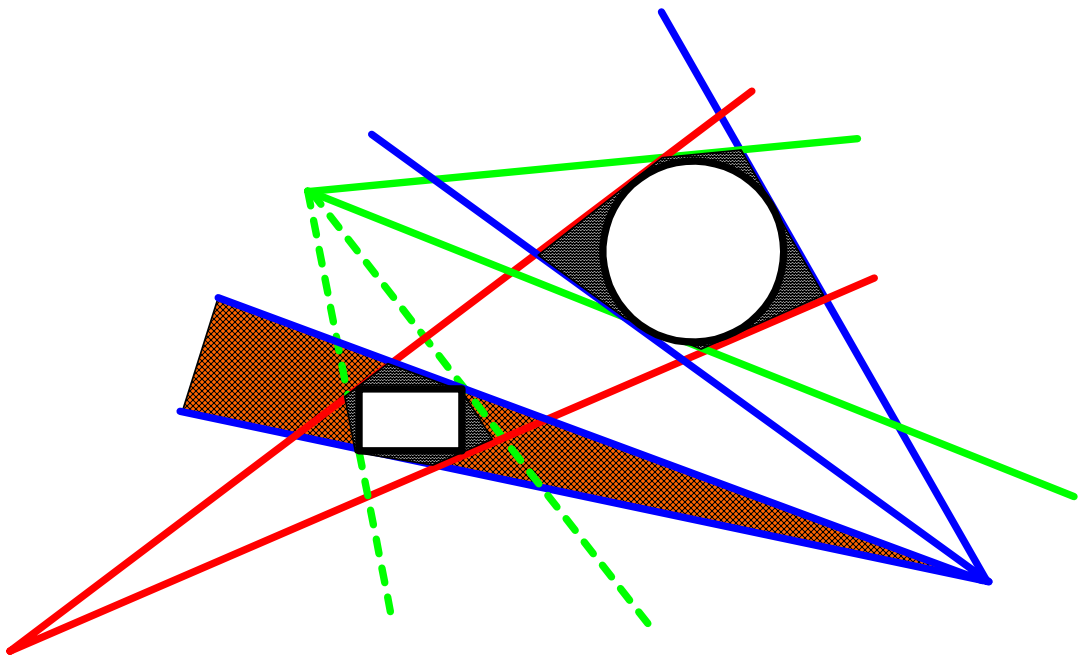
NO (that easy)

A single cone misdetection leads to an unreconstructed shape. As more cones are introduced, it is also more probable that one of the cameras will misdetect a foreground entity

Can we do something about this?

Shape from Inconsistent Silhouette I

- The geometric concept of Inconsistent Hull (IH) is introduced as the volume where there does not exist a reconstructed shape which could possibly justify the observed silhouettes.



In the figure above, the IH is shown in patterned orange after one camera (green) failed to detect the silhouette of the rectangle object

Shape from Inconsistent Silhouette II

We obtain the minimum number of foreground projections (T) so that we can guarantee that a 3D point in the Inconsistent Hull is better explained as foreground.

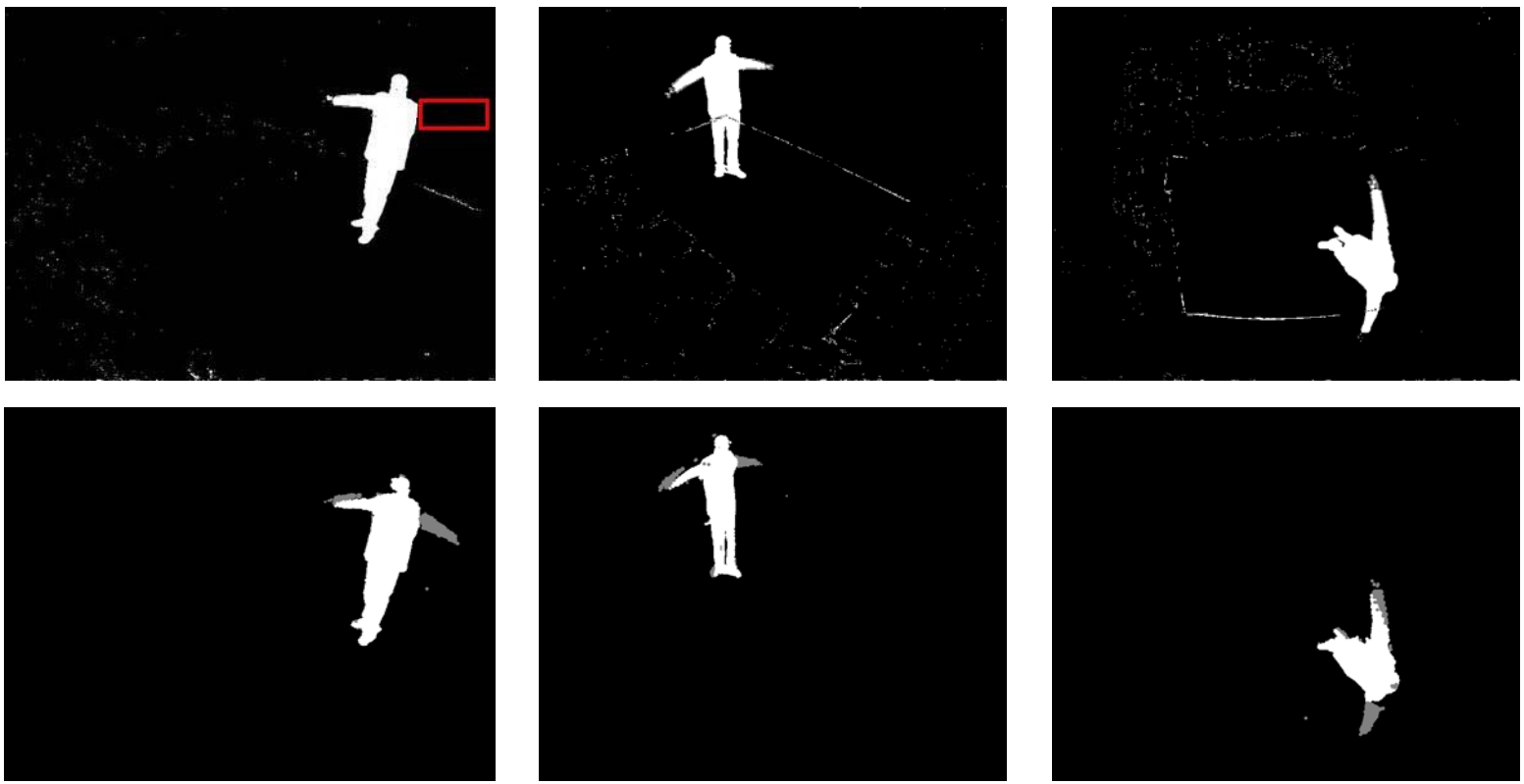
T Is obtained by minimization of the misclassification probability. $P(Err_{3D}) = P_B P(FA_{3D}) + P_S P(M_{3D})$

$$P(FA_{3D}) = \sum_{i=max(T,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(FA_{2D})^i (1 - P(FA_{2D}))^{C-i} \qquad P(M_{3D}) = \sum_{i=max(C-\mathcal{O}-T+1,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i}$$

The misclassification probability is convex under certain (reasonable) conditions, which allows obtaining T in real-time

Unbiased Hull

- Based on the Inconsistent Hull, we isolate conflictive areas in which we do further processing to obtain the Unbiased Hull:

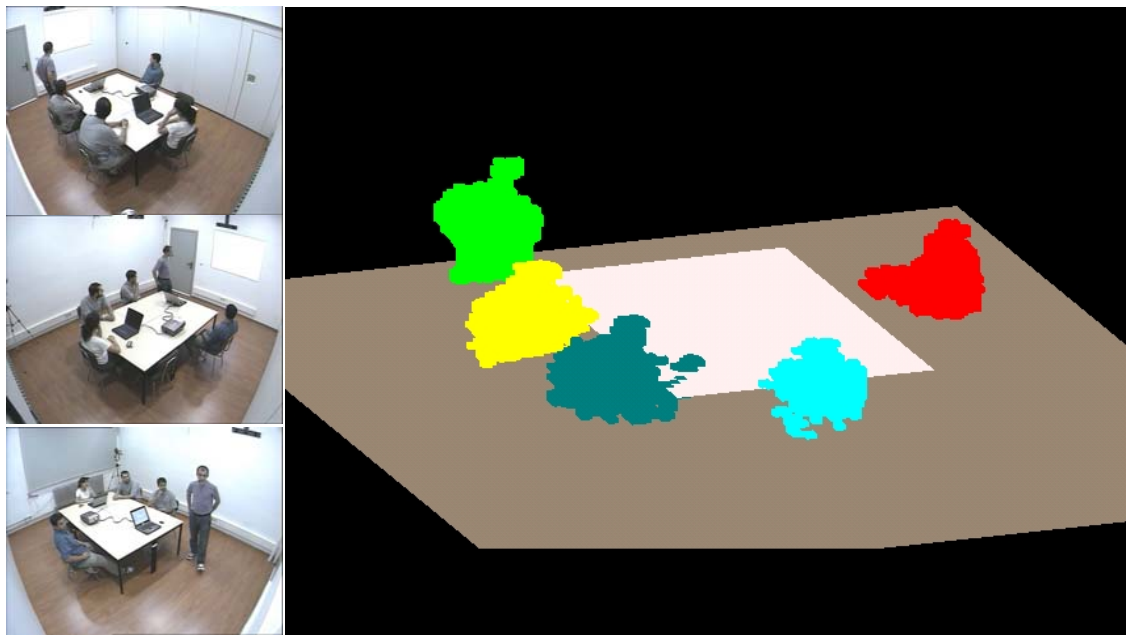


•Originals masks are shown on the top row. Note that some part of the mask in the image on the top-left has not been detected
•Projection of traditional SFS reconstruction is shown on the bottom row
•SfIS error correction is handy at the 2D background model update stage

3D Entity Tracking

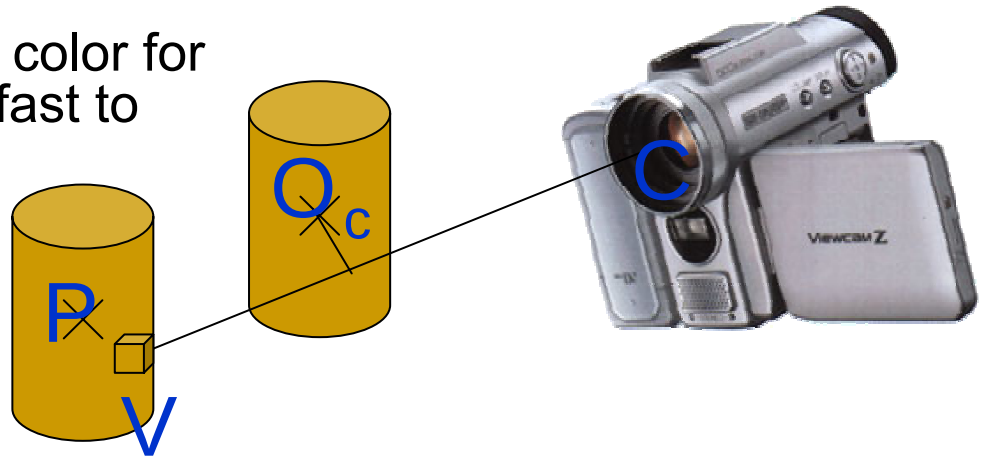
After marking the voxels a connectivity analysis is carried out:

- We choose to group the voxels with 26-connectivity (contact in vertices, edges, and surfaces)
- We consider only the blobs with a number of connected voxels greater than a certain threshold (B_SIZE), to avoid spurious detections



3D Blob Characterization

The blobs are characterized with their color for tracking purposes. This must be very fast to achieve real time operation



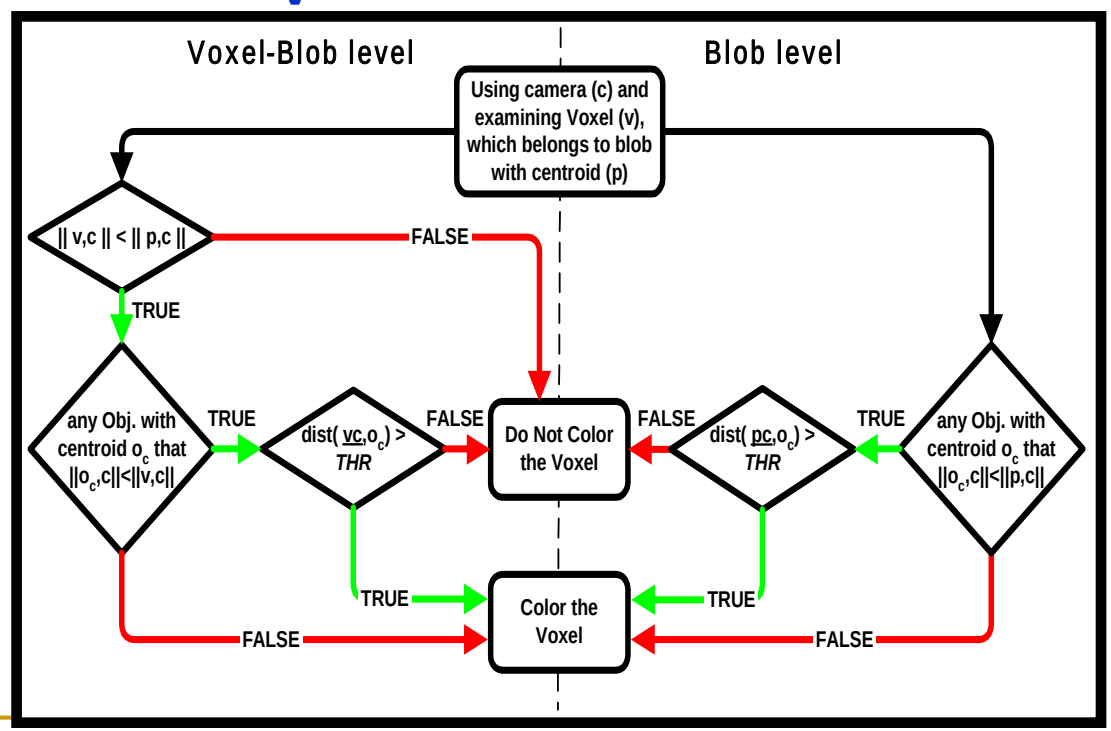
FAST VOXEL COLORING

VOXEL-BLOB LEVEL

- Intra-object occlusions are determined by verifying that the voxel is more distant to the camera than the centroid of its blob
- Inter-object occlusions in a voxel are determined by finding objects (represented by their centroid) in between the camera and the voxel

BLOB-LEVEL (faster)

- The voxels are approximated by the position of the centroid of the blob they belong to



3D Entity Tracking I

- Each object of interest in the scene is modeled by a temporal template of persistent features (velocity, volume, histogram)
- The template for each object has a set of associated Kalman filters that predict the expected value for each feature

3D Entity Tracking II

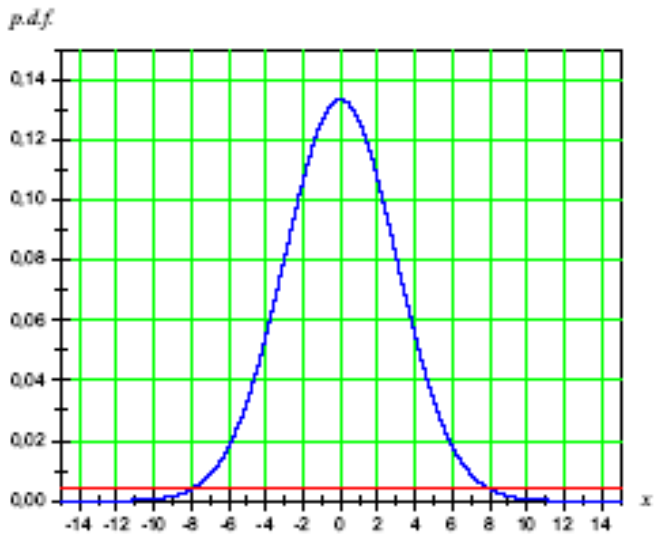
- VIDEO

Conclusions

- A system able to create a 3D-foreground scene, characterize objects with 3D-blobs and track them, preventing the difficulties of inter-object occlusions in 2D trackers
- 3D detections are obtained using Shape from Inconsistent Silhouette to allow using a large number of cameras with noisy 2D detections
- The system uses a fast voxel coloring scheme which allows fast object histogram retrieval used later with other features in a parallel matching technique during the tracking

2D Silhouette Extraction I

- Where do we get the silhouettes from?
- We define probabilistic models of the background and foreground stochastic processes in each camera and perform the classification using a simple maximum a posteriori (MAP) setting:
  - The background process of each pixel is characterized by a Gaussian pdf.
  - For the sake of simplicity we do not model the foreground pixels. Therefore, the stochastic foreground process can be simply characterized by a uniform pdf of value $1/256^3$ in the RGB colorspace



The mean and variance of each Gaussian is adapted based on the value of the pixels that are classified as background as by online expectation maximization

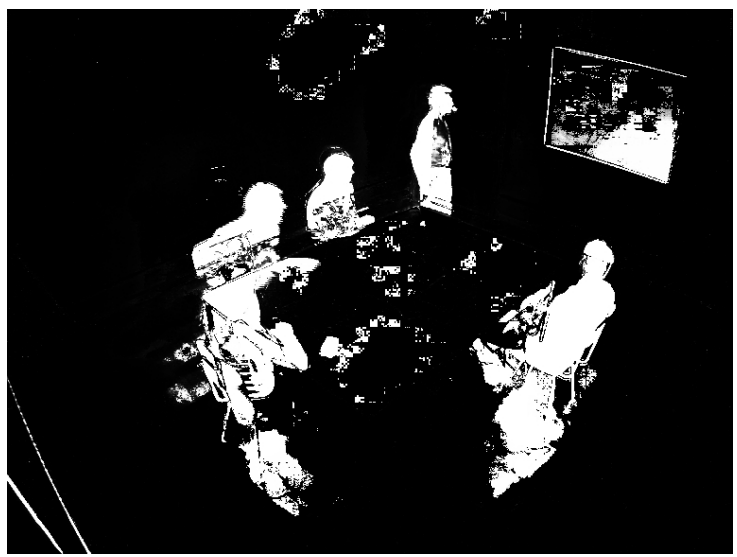
2D Silhouette Extraction II

The probability that a pixel x belongs to the foreground ϕ given an observation **I(x)**, can be expressed in terms of the likelihoods of the foreground ϕ and background β processes as follows

$$P(\phi|\mathbf{I}(\mathbf{x})) = \frac{P(\phi)p(\mathbf{I}(\mathbf{x})|\phi)}{p(\mathbf{I}(\mathbf{x}))},$$

In the MAP setting, a pixel is classified into the foreground class if

$$P(\phi)P(\mathbf{I}(\mathbf{x})|\phi) > P(\beta)P(\mathbf{I}(\mathbf{x})|\beta)$$



Cooperative Background Modeling I

Voxel-based Shape-from-Silhouette can also be thought as a classification problem:

Consider a pattern recognition problem where, in a certain view $\mathbf{I_i}$, a voxel in location **v** is assigned to one of the two classes ϕ (2D-foreground), or β (2D-background), given a measurement $\mathbf{I_i(x_i)}$, corresponding to the pixel value of the projected voxel: $\mathbf{v} \rightarrow \mathbf{x_i}$, in camera **i**.

Now, let us represent with super classes $(\Gamma_0, \cdots, \Gamma_K)$ all possible combinations of 2D-fore/background detections in all views ($\mathbf{i} = 1, \cdots, C$)

$$
\begin{array}{ll}
\Gamma_0 & = \{ \ \phi, \ \phi, \ \phi, \ \cdots, \ \phi \ \} \\
\Gamma_1 & = \{ \ \beta, \ \phi, \ \phi, \ \cdots, \ \phi \ \} \\
\Gamma_2 & = \{ \ \phi, \ \beta, \ \phi, \ \cdots, \ \phi \ \} \\
\vdots & \\
\Gamma_j & = \{ \ \Gamma_j[1], \Gamma_j[2], \Gamma_j[3], \ \cdots, \ \Gamma_j[C] \ \} \\
\vdots & \\
\Gamma_{C+1} & = \{ \ \beta, \ \beta, \ \phi, \ \cdots, \ \phi \ \} \\
\vdots & \\
\Gamma_K & = \{ \ \beta, \ \beta, \ \beta, \ \cdots, \ \beta \ \}
\end{array}
$$

Cooperative Background Modeling II

- A voxel of the 3D shape belongs to class Γ_0, while a voxel of the 3D background belongs to any of the other super classes

- According to Bayesian theory, given observations $\mathbf{I_i(x_i)}$, ($\mathbf{i}$ = 1,…,C), a super class Γ_j is assigned, provided the a posteriori probability of that interpretation is maximum

$$P(\Gamma_j|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = \max(P(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})))$$

Results I

The system has been evaluated using 5 synchronized video streams, captured and stored in JPEG format, in the smart room of our lab at the UPC.

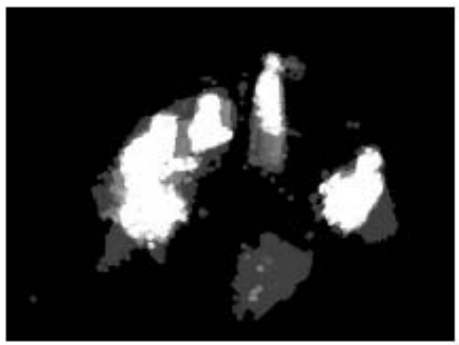
Apart from the compression artifacts, the imaging scenes also contain a range of difficult defects, including illumination changes due to a beamer and shadows.

In our experiments, the outlier model is used. We have used **e** = 0.5. The classification is performed setting a threshold to the probability of 3D-foreground by inspection of the projected probability, as discussed in the previous section.

Results II

(a)

(b)



(c)

(d)

The original image is show in (a). Picture (b), shows the foreground segmentation using conventional classification. In (c), the projected probabilities of the 3D-Shape are shown in gray scale. Finally, image (d) shows the foreground segmentation using the cooperative framework.

Conclusion and Future Work

- We have presented a vision-based system for accurate 2D and 3D foreground segmentation.
- The presented method is able to segment the foreground in a view using the evidence present in the rest of cameras.
- This leads to a better 2D and 3D segmentation, thus also leading to more accurate 2D background models.
- Some of the future works include adapting the presented Bayesian framework to other SfS approaches which are able to detect some 2D-classification errors based on the geometric constraints of Visual Hull (Shape From Inconsistent Silhouette).