A Unified Framework for Consistent 2D/3D Foreground Object Detection

Jose-Luis Landabaso

Telefónica R&D, Vía Augusta, 177, 08021 Barcelona, Spain

Montse Pardàs

Technical University of Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain

Abstract

This paper addresses two-dimensional (2D) and three-dimensional (3D) active entity detection in video scenes. Active entities are the foreground parts in a stationary background scene and they typically correspond to the regions of interest in many applications such as automated video surveillance, object and person tracking and suspicious object detection, among others. We present a novel framework that permits obtaining 2D and 3D active entities as an inter-dependent probabilistic procedure. In the process of creating this framework, a study has been conducted to explore ways to generalize existing activity detection techniques to a Bayesian form. With regard to volumetric activity detection, very little work has been done in the field of Bayesian classification. Thus, in order to support the framework previously outlined, a new Bayesian 3D activity detection technique has been developed. A probabilistic analysis only accounts for half of the problem. The Bayesian framework gives a unified manner to interact between the planar and the volumetric detection tasks and helps to prevent the propagation of noisy pixel observations to the 3D space. However, when large systematic errors occur in the 2D detection level, a different approach has to be taken to correct them. In this respect, 2D/3D geometric relations can be exploited. Errors in the planar detection task often produce a set of incompatible foreground planar regions in the sense that they cannot be globally explained as the projection of the detected 3D volume. This is a key issue with significant implications that is not considered in most of current approaches. We use a new 3D foreground detection scheme that is able to correct errors in 2D planar detections by checking the consistency between 3D foreground detections and the set of corresponding 2D foreground regions [1].

[1]This work has been partially supported by the Spanish project ACERCA TEC-2004 and the Spanish Administration organism CDTI.

I. INTRODUCTION

Detecting active entitis in video scenes has been one of the most studied topics in computer vision. Active entities are the foreground regions in a stationary background scene. They typically correspond to persons and objects that move over static elements of the environment. Thus, areas of activity concur with the regions of interest in many applications, such as automated video surveillance, object tracking, human behavior modeling, immersive video-conferencing, suspicious object detection, etc. This great number of applications has motivated a lot of research and led to many significant achievements.

In recent years, as computing hardware became more powerful, the growth of 3D detection applications has been particularly noticeable. Several of the volumetric foreground detection techniques are built on top of planar foreground detectors. In practice, most volumetric detectors simply take planar detections as an input source, without considering which was the process that yielded the planar foreground regions. These systems have been used in recent years with great success. However, the dependency between the planar and volumetric approaches can be exploited and improved further in order to bridge the gap between both techniques.

This paper addresses the problem of precise 3D active entity detection by analyzing the 2D-3D interaction process. We present a novel framework that permits obtaining 2D and 3D active entities as an inter-dependent procedure. In order to create this framework, we have taken a Bayesian approach that unifies our new findings with many of the most successful current approaches.

State of the art in 2D foreground segmentation

Over the years, many works have been published on the two dimensional foreground segmentation task, describing different methods that treat to extract that part of the scene containing active entities. In most of the cases, the stochastic background process is modeled first, and then the foreground pixels are classified as an exception to the model [31], [13], [23], [11], [7]. In other setups, the foreground process is also modeled, and the scene is classified using maximum a posteriori (MAP) [33], [25], [15], [21]. Then, in order to guarantee accurate results along the time, the background models are continuously updated making use of all the pixel values that are classified as background. In this particular type of framework, all background observations contribute equally to update the background models. However, it seems reasonable to assume

that observations with higher background probabilities should weight more than those with lower probabilities.

Proposed Bayesian classification in 2D

In this paper we develop a theoretical framework in which pixel classification and update stages are fully explained as a Bayesian procedure, where one stage is probabilistically related to the other. This framework permits to obtain an estimate of the system's error rate and the probability of each classification, which is necessary for the 3D foreground detection proposed in this paper. We also provide a simple solution for not having to model the foreground appearances and still be able to use the probabilistic setting. The classification step of our scheme provides the probabilities that are then used to update the models. Moreover, this new scheme opens the doors to the possibility of incorporating other sources of information that provide solid information about pixel probabilities. One of these external sources of information, consisting in projecting more-informed 3D probabilistic maps, will be described along the paper.

State of the art in 3D foreground segmentation

Providing probabilistic justification to classical approaches significantly improves the performance of planar foreground detection. With regard to volumetric activity detection, the literature reveals that very few work has been done in the field of Bayesian classification. Several multi-camera 3D foreground detection systems are based on 2D foreground segmentation techniques that make use of 2D background modeling. Shape from Silhouette (SfS) for 3D model extraction is the approach taken in most of these systems. It consists on building binary volumetric models (the Visual Hull) from the set of binarized foreground masks. SfS was firstly introduced by Baugmart [2] in 1974, though it was not until 1991 when Laurentini [20] defined the geometric concept of Visual Hull (VH) as the maximal object silhouette-equivalent to the real object S, i.e., which can be substituted for S without affecting any silhouette.

Many algorithms have been developed for constructing volumetric models from a set of silhouette images. Once the silhouettes are extracted, the main step of all the algorithms is the intersection test. Some methods back-project the silhouettes, creating an explicit set of cones that are then intersected in 3D [22], [28], [10]. Others divide the volume into voxels [26], [16], [30], [4], [17]. Then each voxel is projected into all the images to test (using a projection test) whether they are contained in every silhouette. See [29], [6] for two surveys on volumetric-based methods.

Accurate silhouette extraction is crucial for good performance of SfS, independently of the algorithm used. Errors in the silhouettes affect the reconstructed Shape. Defects observable in the silhouettes can be categorized into two types: false alarms and misses. False alarms correspond to erroneous foreground detections, while misses correspond to erroneous background detections.

These errors in the silhouettes can be due to different causes: regular noise and non-Gaussian systematic errors. The first type of error is because of the cameras thermal noise. It produces isolated background pixels within the foreground silhouettes and foreground pixels within the background. The second one often consists in large regions missed or falsely detected due to the arrangement of the scene or limitations of the foreground segmentation technique. Systematic misses in a view often occur when, for instance, foreground objects have similar colours and texture to their counterparts in the background. Systematic misses can also be due to background structures, occluding the foreground objects in some views. Analogously, specular reflections can form large areas of falsely detected foreground pixels.

Some approaches oriented to eliminate gaussian noise classifying voxels as shape or background using cooperatively the information from the multiple cameras before extracting the 2D silhouettes have also been developed in the past. In [30] an algorithm based on graph cuts determines the 3D shape with lowest cost (smoothest shape consistent with the observations). In this case, the 2D silhouettes are not explicitly computed. In [8] the shape-from-silhouette problem is restated as a sensor fusion problem, providing each pixel from each camera with a forward sensor formulation which models the pixel observation responses to the voxel occupancies in the scene. However, in none of these approaches the 3D map was used to feed back the background models in the images.

Other approaches involve the usage of voxel-based reconstructions to reduce the probability of voxel miss-classification. In [4], Cheung et al. propose an algorithm called SPOT in this direction. SPOT achieves lower voxel miss-classification rate compared to other SfS algorithms that use naive projection tests such as testing only one point per voxel and view or testing all the pixels within the projection of the voxel.

Proposed 3D foreground segmentation

We propose in this paper to use a new Bayesian 3D activity detection technique that better exploits the redundancy present in a multi-camera environment by using the 3D maps to feed back the background models in the images. A first approach to this system was propose in [17].

The Bayesian framework proposed gives a unified manner to interact between the planar and the volumetric detection subsystems. In addition, an outlier model in the probabilistic framework prevents noisy pixel observations from propagating to the rest of the views through a bogus reconstruction. However, when large systematic errors occur in the 2D detection plane, then outlier models simply cannot help.

Errors in the planar detection task often produce a set of incompatible foreground planar regions in the sense that they cannot be globally explained as the projection of the detected 3D volume. This is a problem with significant implications that is not considered in most of current Shape from Silhouette approaches. Current 3D reconstruction methods simply assume that errors do not occur in the 2D plane. Instead, we use a new three-dimensional foreground detection scheme ([18], [19]) that is able to correct errors in 2D planar detections by checking the consistency between 3D foreground detections and the set of corresponding 2D foreground regions. The technique allows to obtain accurate 3D models that provide the most reasonable explanation of a 3D detection based on 2D observations. In this paper we propose to incorporate back the reassigned classifications into our Bayesian framework so that the 3D probabilistic map reflects them. In this way, we can deal with both gaussian and systematic errors.

This paper is structured as follows. Next Section presents our Bayesian framework for the planar foreground detection task. Section III is devoted to the 3D-Shape reconstruction techniques. We describe in Section IV the interaction between the cooperative Bayesian method proposed and the 3D reconstruction considering silhouette inconsistencies. Some results with real-world sequences are shown in Section V and, finally, Section VI provides some conclusions.

II. Bayesian Foreground Segmentation

A. Background and Foreground Models

A very fast method to learn and update a representation of the background of the imaging scene is to model the background color at each pixel location fitting a Gaussian function [33]. A more elaborated method consists in using a Mixture of Gaussians (MoG) to model the background process at each pixel [11], [31]. This is very similar to the previous method. But, in addition, an MoG is also able to model a background scene that is constantly changing along the time such as in raining situations, waving flags, water, etc. Finally, it is possible to obtain better approximations of the background process at each pixel by learning a smooth continuous version

of the histogram obtained from the last number of observed values in the same pixel location. This can be achieved by summing one Gaussian centered at a pixel value for each sample that is observed in the same location along the time [7].

All the mentioned models above share that they use a pdf function to represent the background. The MAP-based foreground segmentation scheme that is presented in this section may be used with any models which are expressed as a pdf, including the models described above. Besides, the 2D foreground detection can be adjusted according to the higher level applications needed. For instance, a decision needs to be taken regarding the incorporation into the background of the foreground objects that remain static for a long period of time. If the application needs to consider these static objects as background (for instance in parking lots), then the model of the pixels needs to be updated even if a pixel is considered as foreground, in such a way that when a given value for a given foreground pixel has been observed for a long time, this information is incorporated into the background model. For other applications it is better not to include these objects into the background. This is the case of the smart-room scenarios, where people can remain static for a long period, but we wish to continue to track them. In this case, background models should not be updated with foreground values' information. This is the option taken in the experiments performed in this paper. But the methodology presented can be easily extended to the aforementioned situation and also to new techniques which consider three classes of objects: background, moving objects and static objects [9].

In order to make use of a maximum a posteriori setting, apart from the background model, a foreground model must also exist. The main problem is how to obtain a reliable characterization of the foreground process of a pixel. The foreground entities of an image are those which are in a prominent place in a scene, due to the fact that they are constantly moving. Therefore, it is difficult to obtain a foreground characterization by inspection of a single pixel location, without using global information of what is happening at the whole image level.

Several approaches have been proposed in the literature that try to characterize complete foreground entities (the so-called blobs in the literature) [24], [23], [11], [7], [25], [15], [21]. Basically, in these approaches each foreground entity is characterized by means of a complex model that takes into account the geometrical properties of the entity. The fundamental premise of these methods is that a tracker is employed so that the foreground models of each entity can be correctly updated along the time. Finally, in a per-pixel MAP setting, the blob-based foreground

models must be mapped to each pixel before performing the foreground segmentation. A MAP setting has been used in [25], [15], [21], among others. However, in the model update stage of these works, the update process is taken as a separate task, not making use of the classification probabilities that MAP provides.

We think it is important to provide also a method which allows using MAP without requiring a complex setup (that is, the estimation of the pdf of the foreground relying on the tracking) to obtain foreground models. To do so, we propose using a uniform *pdf* to model the foreground process at each pixel. A MAP setting, even with this naive foreground characterization, provides better results than a classification based only on background models.

First off, let's assume that we don't have any clue about the foreground process in the scene. We can, however, consider that in images with D channels, each pixel in the image has values in $\mathbf{D} = \{0, \cdots, 255\}^D$. Then, without more information about the foreground entities, we can assume that the likelihood of observing one of the values in $\mathbf{D}$, given that it belongs to a foreground process is

$$p(\mathbf{I_x}|\text{foreground}) = \frac{1}{256^D}, \tag{1}$$

where $\mathbf{I_x}$ denotes the value of a certain pixel $\mathbf{x}$ in the image.

B. MAP Classification

Several pixel foreground/background classification settings have been proposed in the past. We will provide here the development of the MAP based classifiers, their error probabilities and principal characteristics. The model update part will be covered in the next sub-section. As will be shown, in our proposal model maintenance is partially built on the classification procedure reviewed here. Thus, it is important to provide first a solid foundation of the classification methods in order to introduce the update scheme later.

A per-pixel probabilistic foreground and background classification setting involves two classes: foreground, denoted with symbol ϕ, and background β. The classification task can be solved by choosing for each pixel the most probable class, i.e., that one with the highest probability.

Therefore, a pixel is classified into foreground if

$$\phi = \underset{c=\{\phi,\beta\}}{\operatorname{argmax}} P(c|\mathbf{I_x}), \tag{2}$$

and analogously, a pixel is considered to belong to the background stochastic process if

$$\beta = \underset{c=\{\phi,\beta\}}{\operatorname{argmax}} P(c|\mathbf{I_x}). \tag{3}$$

If we assume that the classes of each pixel are independent, then a pixel can be classified as foreground if

$$\phi = \underset{c=\{\phi,\beta\}}{\operatorname{argmax}} \frac{P(c)p(\mathbf{I_x}|c)}{p(\mathbf{I_x})} = \underset{c=\{\phi,\beta\}}{\operatorname{argmax}} \frac{P(c)p(\mathbf{I_x}|c)}{P(\phi)p(\mathbf{I_x}|\phi) + P(\beta)p(\mathbf{I_x}|\beta)}, \tag{4}$$

which completes the maximum a posteriori setting.

1) Classification Probabilities: Assuming that we are using the uniform foreground model in (1), we can derive the foreground and background probabilities for each pixel,

$$P(\phi|\mathbf{I_x}) = \frac{P(\phi)p(\mathbf{I_x}|\phi)}{P(\phi)p(\mathbf{I_x}|\phi) + P(\beta)p(\mathbf{I_x}|\beta)} = \frac{P(\phi)\frac{1}{256^D}}{P(\phi)\frac{1}{256^D} + P(\beta)p(\mathbf{I_x}|\beta)} \tag{5}$$

$$P(\beta|\mathbf{I_x}) = \frac{P(\beta)p(\mathbf{I_x}|\beta)}{P(\phi)p(\mathbf{I_x}|\phi) + P(\beta)p(\mathbf{I_x}|\beta)} = \frac{P(\beta)p(\mathbf{I_x}|\beta)}{P(\phi)\frac{1}{256^D} + P(\beta)p(\mathbf{I_x}|\beta)}. \tag{6}$$

For the sake of simplicity, let us consider a simple background model, using only one Gaussian to represent a single background mode. Inherently, this model assumes that the background is static, without moving leaves in a tree or waving flags, water and so on. The model is:

$$\mathbf{G_x}(\mathbf{I_x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma_x}|}} e^{-\frac{1}{2}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{\Sigma_x^{-1}}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})}. \tag{7}$$

which leads to

$$P(\phi|\mathbf{I_x}) = \frac{P(\phi)\frac{1}{256^D}}{P(\phi)\frac{1}{256^D} + P(\beta)\frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma_x}|}} e^{-\frac{1}{2}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{\Sigma_x^{-1}}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})}} \tag{8}$$

$$P(\beta|\mathbf{I_x}) = \frac{P(\beta)\frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma_x}|}} e^{-\frac{1}{2}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{\Sigma_x^{-1}}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})}}{P(\phi)\frac{1}{256^D} + P(\beta)\frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma_x}|}} e^{-\frac{1}{2}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{\Sigma_x^{-1}}(\mathbf{I_x}-\boldsymbol{\mu}_\mathbf{x})}}. \tag{9}$$

2) Classification Error: In an exception-to-background segmentation setting it is impossible to obtain a measure of the probability of a given classification because only the background class is available. On the contrary, we have shown that an MAP setting provides the posterior probabilities of each class. Moreover, it is also possible to obtain a measure of classification reliability in terms of the error probabilities.

Note that it is important to know the segmentation error rate to inform to the subsequent parts of the system that make use of the classifications. For instance, the error rate of a foreground segmentation scheme is critical in the 3D reconstruction module that we present in Section III.

In order to present the formulation of the error rate of a foreground segmentation scheme, let us assume that we are using the foreground and background models in (1) and (7), respectively.

There are many sources of stochastic fluctuation. Suppose an observation $\mathbf{I_x}$ is made leading to a decision $\hat{c}$. We can summarize average performance in terms of a confusion matrix, $P(\hat{c}|c)$, which for the detection task is a 2×2 array representing the hit (correct detections): $P(\hat{\phi}|\phi)$, false positive, also known as false alarm: $P(\hat{\phi}|\beta)$, false negative (miss): $P(\hat{\beta}|\phi)$, and correct rejection rates: $P(\hat{\beta}|\beta)$:

$$\begin{pmatrix} P(\hat{\phi}|\phi) & P(\hat{\phi}|\beta) \\ P(\hat{\beta}|\phi) & P(\hat{\beta}|\beta) \end{pmatrix}. \tag{10}$$

Obviously, an ideal system would maximize the frequency of hits and rejections while minimizing the frequency of misses and false alarms. If the probabilistic distributions are known, the best decision one can make is choosing one class or another according to (4). For example, consider the foreground detection task using the models depicted in the figure and assume that we do not have information about the priors. In that case, the cross of both distributions occurs at

$$\frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} e^{-\left(\frac{\mathbf{I}_{\mathbf{x}}^{\diamond}}{\sqrt{2}\sigma_{\mathbf{x}}}\right)^2} = \frac{1}{256}, \tag{11}$$

where $\mathbf{I}_{\mathbf{x}}^{\diamond}$ is used to represent $\mathbf{I_x} - \boldsymbol{\mu}$. Computing $\mathbf{I}_{\mathbf{x}}^{\diamond}$ and integrating the foreground and background likelihood functions in the corresponding intervals, we obtain the following confusion matrix for the MAP setting

$$\begin{pmatrix} P(\hat{\phi}|\phi) & P(\hat{\phi}|\beta) \\ P(\hat{\beta}|\phi) & P(\hat{\beta}|\beta) \end{pmatrix}_{\text{MAP}} = \begin{pmatrix} 1 - \frac{2\sqrt{2}\sigma_{\mathbf{x}}}{256}\sqrt{-\ln\left(\frac{\sqrt{2\pi}\sigma_{\mathbf{x}}}{256}\right)} & 1 - \operatorname{erf}\left(\sqrt{-\ln\left(\frac{\sqrt{2\pi}\sigma_{\mathbf{x}}}{256}\right)}\right) \\ \frac{2\sqrt{2}\sigma_{\mathbf{x}}}{256}\sqrt{-\ln\left(\frac{\sqrt{2\pi}\sigma_{\mathbf{x}}}{256}\right)} & \operatorname{erf}\left(\sqrt{-\ln\left(\frac{\sqrt{2\pi}\sigma_{\mathbf{x}}}{256}\right)}\right) \end{pmatrix}.$$

C. Model Update

Once the classification determination has been made, it is important to update the foreground and background models by including the most recent input. There exist many different methods to update the models. The most important bit is how to seize the entry information, that is, how to measure the relevance of the observations when updating the models. For instance, in the approach of Stauffer and Grimson [31] (S&G from now on), they propose to adapt the learning rate of background models to the likelihood that a pixel value belongs to the model, without further Bayesian justification.

Instead, we propose using a Bayesian scheme. The method we advocate is centered on the Expectation Maximization approach [5]. Following, we provide the parameters update equations using the Expectation Maximization (EM) algorithm. We have derived the equations for learning the parameters of an MoG background model, considering a uniform foreground model. This is the main difference with respect to the other approaches, that simply maximize the background model without considering the foreground process [14], [24] . Using an MoG in the background makes it possible to compare EM to the update scheme of the S&G method. In addition, the derivation of the single Gaussian background model can be simply obtained as a particularization of the MoG case. Moreover, we will show that the update step of the EM method is directly tied to the MAP classification described in Section II-B, closing the Bayesian classification-update loop.

1) EM of a Gaussian Mixture and Uniform Functions: The mixture density parameter estimation problem is probably one of the most widely used applications of the EM algorithm.

In this case, we are assuming the following probabilistic background model:

$$p(\mathbf{I_x}|\beta) = \text{MoG}_\mathbf{x}(\mathbf{I_x}) = \sum_{k=1}^{K} w_{\mathbf{x},k}\text{G}_{\mathbf{x},k}(\mathbf{I_x}) = \sum_{k=1}^{K} \frac{w_{\mathbf{x},k}}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}_{\mathbf{x},k}|}} e^{-\frac{1}{2}(\mathbf{I_x}-\boldsymbol{\mu}_{\mathbf{x},k})^T\boldsymbol{\Sigma}_{\mathbf{x},k}^{-1}(\mathbf{I_x}-\boldsymbol{\mu}_{\mathbf{x},k})}, \tag{12}$$

where K is the total number of Gaussians used in each pixel, and where $w_{\mathbf{x},k}$ is the prior probability that a background pixel is represented by a certain mode k of the mixture ($\sum_{k=1}^{K} w_{\mathbf{x},k} = 1$). These priors are often referred as the weights of the Gaussians. Also note that the means and covariances are indexed with respect to a Gaussian k of the MoG in $\mathbf{x}$: $\boldsymbol{\Sigma}_{\mathbf{x},k}$ and $\boldsymbol{\mu}_{\mathbf{x},k}$.

The foreground model we are considering in the derivation of the EM method is a uniform function. Therefore, the likelihood function for a certain pixel $\mathbf{x}$ is:

$$p(\mathbf{I_x}|\theta) = P(\beta)\text{MoG}_\mathbf{x}(\mathbf{I_x}) + P(\phi)\frac{1}{256^D}, \tag{13}$$

where the model parameters (θ) to estimate for the pixel are:

$$\theta = \left\{\boldsymbol{\mu}_{\mathbf{x},1}, \cdots, \boldsymbol{\mu}_{\mathbf{x},K}, \boldsymbol{\Sigma}_{\mathbf{x},1}, \cdots, \boldsymbol{\Sigma}_{\mathbf{x},K}, w_{\mathbf{x},1}, \cdots, w_{\mathbf{x},K}\right\}. \tag{14}$$

These parameters are iteratively estimated using a number of statistically independent observations $\mathbf{I_x}[1], \cdots, \mathbf{I_x}[M]$ at pixel location $\mathbf{x}$ drawn from either Gaussians $\text{G}_{\mathbf{x},1}, \cdots, \text{G}_{\mathbf{x},K}$ or

from the uniform function. Developing the Expectation Maximization procedures, leads to the following update parameters in the n-th iteration:

$$
\begin{aligned}
P(y_i = k|\theta)^\star &= \frac{\sum_{i=1}^{M} P(y_i = k|\mathbf{I}_\mathbf{x}[i], \theta_n)}{M} \\
\mu^\star_{\mathbf{x},k} &= \frac{\sum_{i=1}^{M} P(y_i = k|\mathbf{I}_\mathbf{x}[i], \theta_n)\mathbf{I}_\mathbf{x}[i]}{\sum_{i=1}^{M} P(y_i = k|\mathbf{I}_\mathbf{x}[i], \theta_n)} \\
\mathbf{\Sigma}^\star_{\mathbf{x},k} &= \frac{\sum_{i=1}^{M} P(y_i = k|\mathbf{I}_\mathbf{x}[i], \theta_n)(\mathbf{I}_\mathbf{x}[i] - \mu^\star_{\mathbf{x},k})(\mathbf{I}_\mathbf{x}[i] - \mu^\star_{\mathbf{x},k})^T}{\sum_{i=1}^{M} P(y_i = k|\mathbf{I}_\mathbf{x}[i], \theta_n)},
\end{aligned} \tag{15}
$$

where it has to be noted that priors $P(y_i = k|\theta)$ and $w_{\mathbf{x},k}$ relation is $w_{\mathbf{x},k} = \frac{P(y_i=k|\theta)}{P(\beta)}$.

Note that one of the best characteristics of the approach of Stauffer and Grimson is that the background model is updated instead of fully recomputed at any time, which makes their algorithm very fast. However, these update equations assume a fixed number of observations M. In other words, EM, as presented so far, is inherently offline. It requires multiple passes through the data set. As the data set grows, so does the computation per iteration of EM. This limitation is a common limitation of the EM algorithm. A practical online implementation that is capable of foreground segmentation of each frame as it is acquired has to re-estimate all the parameters incrementally from each new sample.

In the following we adapt the derivations of offline EM to online EM considering both a foreground and background model. The main idea behind an online update scheme is that last observation M is used to feed the last iteration n:

$$
\begin{aligned}
&P(y_i = k|\theta)[t] = P(y_i = k|\theta)[t-1] + \alpha\left(P(y_i = k|\mathbf{I}_\mathbf{x}[t], \theta_t) - P(y_i = k|\theta)[t-1]\right) \\
&\mu_{\mathbf{x},k}[t] = \mu_{\mathbf{x},k}[t-1] + \alpha P(y_i = k|\mathbf{I}_\mathbf{x}[t], \theta_t)\left(\frac{\mathbf{I}_\mathbf{x}[t] - \mu_{\mathbf{x},k}[t-1]}{P(y_i = k|\theta)[t]}\right) \\
&\mathbf{\Sigma}_{\mathbf{x},k}[t] = \mathbf{\Sigma}_{\mathbf{x},k}[t-1] + \\
&\qquad + \alpha P(y_i = k|\mathbf{I}_\mathbf{x}[t], \theta_t)\left(\frac{(\mathbf{I}_\mathbf{x}[t] - \mu_{\mathbf{x},k}[t-1])(\mathbf{I}_\mathbf{x}[t] - \mu_{\mathbf{x},k}[t-1])^T - \mathbf{\Sigma}_{\mathbf{x},k}[t-1]}{P(y_i = k|\theta)[t]}\right).
\end{aligned} \tag{16}
$$

And if only one Gaussian per pixel is used:

$$
\begin{aligned}
&\mu_{\mathbf{x},k}[t] = \mu_\mathbf{x}[t-1] + \frac{\alpha P(\beta|\mathbf{I}_\mathbf{x}[t], \theta_t)}{P(\beta)}\left(\mathbf{I}_\mathbf{x}[t] - \mu_{\mathbf{x},k}[t-1]\right) \\
&\mathbf{\Sigma}_{\mathbf{x},k}[t] = \mathbf{\Sigma}_{\mathbf{x},k}[t-1] + \\
&\qquad + \frac{\alpha P(\beta|\mathbf{I}_\mathbf{x}[t], \theta_t)}{P(\beta)}\left((\mathbf{I}_\mathbf{x}[t] - \mu_\mathbf{x}[t-1])(\mathbf{I}_\mathbf{x}[t] - \mu_\mathbf{x}[t-1])^T - \mathbf{\Sigma}_\mathbf{x}[t-1]\right),
\end{aligned} \tag{17}
$$

where $P(\beta)$ is a constant prior value.

Similar equations have been derived in the past [14], [24], [32]. However, the main difference with our derivation is that we have included a foreground model into the likelihood of the pixel, permitting us to obtain the parameters that maximize the complete likelihood. The background models are updated using all the color values that are observed along the time and the system is in charge of automatically weighting the contributions of each sample. Our system first determines the background probability of an observation (using MAP), and then the probability of a certain mode, assuming that the system has observed a color value corresponding to the background. Thus, implicitly, the process of pixel model update is making use of the MAP classification setting. This permits obtaining better background models and, therefore, also better classifications.

The advantages of the planar foreground detection scheme can be outlined as follows:

- Only those observations with high background probability contribute more. That is, the speed of adaptation of the background models is proportional to the certainty of the background observation. Contrarily, in the systems that only maximize the likelihood of the background, all uncertain observations that are erroneously classified as background contribute to wrongly update a model *at full speed*.
- The system supports MoGs for the background models, among other pdfs. Note that MoGs have been successfully used many times in the past in other frameworks. Thus, the system has been built over robust building blocks which have been adapted and improved in a Bayesian way.
- The parameters used in our system are more intuitive than those needed in classical approaches such as [14]. In fact, in our case, we only need to choose the priors of foreground and background and the number of Gaussians that will be used. In the approach of S&G one has to decide the number of Gaussians, the threshold T, corresponding to the minimum prior probability that the background is in the scene and the adaptation speed α. And in practical scenarios, the user also needs to find a suitable value for the adaptation learning rate $\rho_{\mathbf{x},k}$, which is usually fixed to a constant value.
- The system is not only able to classify pixel observations but to provide the probability of a given classification. These probabilistic values will be used in Section III-B to build three-dimensional probabilistic maps. Moreover, the system also provides its probabilities of false

alarm and miss, which will be used in Section III-C to evaluate the error probabilities of the volumetric reconstruction method that employs foreground segmentations from multiple views.

- Finally, background models are updated proportionally to the probability that an observation belongs to the background. This correlation between model update and pixel probability can be exploited by using other more-informed probabilistic sources information that are external to the pixel process. This finding will be used to bridge the gap between the planar and volumetric foreground detection tasks, unifying the algorithms presented in this section and in the next one.

III. 3D-Shape Reconstruction Using Multiple Cameras

A. *Shape from Silhouette*

As it has been commented in the Introduction many multi-camera 3D foreground detection systems are based on 2D foreground segmentation techniques that make use of 2D background modeling. SfS is the approach taken in most of these systems. As mentioned, it consists on building the Visual Hull from the set of binarized foreground masks (the silhouettes). The Visual Hull is constructed back-projecting the silhouettes into the scene and creating the 3D intersection area, or projecting every voxel of the scene into the images and testing if their projection is contained in the 2D foreground region. Only if all the projections are within a foreground region, a voxel will be considered as Shape.

The construction of the 3D Shape from the 2D foreground region has some inherent advantages:

- Accurate 3D location information of the entities of interest
- Most of the 2D occlusion problems between foreground objects are solved, because they do not occur in the 3D space
- Many 2D foreground detection errors due to noise or shadows are corrected, because they only appear in one of the image projections, and thus are not reconstructed when performing the 3D intersection.

Regarding the shadows, let us mention that shadows and highlights are an important problem when dealing with 2D foreground segmentation. Let us consider for instance the shadows cast by the foreground objects. They produce pixel values which do not correspond to the learnt

background model and are thus detected as foreground. Actually, shadow removal algorithms are usually incorporated in the background subtraction/modeling step. Several studies have been carried out to extract cues from the background reference images/models and use them to identify if a pixel is a cast shadow/highlight pixel or not. Prati et al. have presented an in-depth survey of these algorithms [27]. There are two main sets of works that incorporate these extracted cues, including the use of color and texture information to find chrominance and texture similarities between the background representation and the incoming frame. In [34] we proposed to use a combination of the two and correct the remaining classification errors using images prior to the shadow-removal process where shapes are still well defined to assist blob reconstruction.

However, when Shape from Silhouette is performed, shadows will only be reconstructed if they are seen from all camera views. This can be observed in the results of Fig. 3 and Fig. 5, where some shadows can be observed in the 2D only Segmentation, and they are not reconstructed in any of the SfS approaches.

B. Probabilistic Voxel Classification

In the standard SfS procedure, two kinds of errors have been observed: errors due to regular noise and what we have called systematic errors. To deal with the first kind of errors we propose here to simultaneously reconstruct and classify the 3D space, instead of previously classifying images in 2D and reconstructing the volume later, as it is the case of regular SfS approaches that use the 2D foreground binary masks. This is achieved by obtaining 3D probabilistic information from 2D probabilistic maps and then projecting back 3D probabilities to each view. Projected probabilities are then used to update the two-dimensional background models. The scheme permits to obtain better 3D foreground detections that in turn also permit to obtain better planar foreground detections. In the state-of-the-art SfS approaches that employ 2D binary masks extracted after background modeling, the models in each view are always temporally maintained along the time using evidence only from that view. In our approach the 2D models are updated using evidence from all the cameras in a Bayesian framework. Thus, both 2D and 3D classifications are performed using existing information in all views along the time. Finally, the proposed technique permits setting external probabilistic information or foreground and background priors to 3D regions instead of a more tedious 2D prior setting. Indeed, 2D regions represent some part of the viewing scene at different depths and this makes it difficult

to apply 2D priors. However, in a 3D Bayesian framework, 3D priors do not suffer from this limitation since occlusions are inherent to the 2D space. Apart from setting 3D priors, the scheme permits incorporating external 3D probabilistic information to the 3D map. In the next section we will propose a reconstruction method that is able to obtain the volume that minimizes the probability of volumetric misclassification. We will show how to incorporate this information back into the Bayesian 3D map.

A 3D probabilistic representation is only possible when the two-dimensional foreground classifications of the scenes can be probabilistically justified. The proposed approach takes the MAP scheme for 2D foreground classification as its building block.

A pixel position $\mathbf{x}$ and an image view $\mathbf{I}$ will be referred to each one of the C views: $\mathbf{x_i}$ and $\mathbf{I}_i$. For the sake of simplicity, we particularize the pixels' background models to the simple case of a single Gaussian per pixel $G_{i,\mathbf{x_i}}(\mathbf{I}_i(\mathbf{x_i}))$ (corresponding to view i, pixel $\mathbf{x_i}$)) and we employ a uniform function as the foreground distribution. However, any other foreground and background likelihood functions can be used without any problems whatsoever.

In the following, we first consider error-free 2D models, and next we adapt an outlier model to the 2D models.

1) Probabilistic Voxel Classification Considering Error-Free 2D Models: Voxel-based SfS can be thought as a classification problem. Consider a pattern recognition problem where, in a certain view $\mathbf{I}_i$, a voxel in location $\mathbf{v}$ is assigned to one of the two classes ϕ (2D-foreground), or β (2D-background), given a measurement $\mathbf{I}_i(\mathbf{x_i})$, corresponding to the pixel value of the projected voxel: $\mathbf{v} \rightarrow \mathbf{x_i}$, in camera i [12][2].

Now, let us represent with super classes $(\Gamma_0, \cdots, \Gamma_K)$ all possible combinations of 2D-

[2]By taking only 1 pixel per view for each voxel we are implicitly considering a very simple, though common, projection test.

fore/background detections in all views ($i = 1, \cdots, C$):

$$\begin{aligned}
\Gamma_0 &= \{ \quad \phi, \quad \phi, \quad \phi, \quad \cdots, \quad \phi \quad \} \\
&\vdots \\
\Gamma_j &= \{ \, \Gamma_j[1], \Gamma_j[2], \Gamma_j[3], \, \cdots, \, \Gamma_j[C] \, \} \\
&\vdots \\
\Gamma_C &= \{ \quad \beta, \quad \beta, \quad \phi, \quad \cdots, \quad \phi \quad \} \\
&\vdots \\
\Gamma_K &= \{ \quad \beta, \quad \beta, \quad \beta, \quad \cdots, \quad \beta \quad \}
\end{aligned}$$

with the following prior probabilities

$$\begin{aligned}
P(\Gamma_0) &= P(\phi)P(\phi)\cdots P(\phi) = P(\phi)^C = P_S \\
P(\Gamma_1) &= P(\beta)P(\phi)\cdots P(\phi) = P(\beta)P(\phi)^{C-1} \\
&\vdots \\
P(\Gamma_K) &= P(\beta)P(\beta)\cdots P(\beta) = P(\beta)^C,
\end{aligned}$$

where a voxel classified as foreground, i.e., a voxel of the 3D-Shape, belongs to super class Γ_0, with P_S prior probability[3]. Contrarily, an undetected voxel, i.e., a voxel of the 3D background, belongs to any of the other super classes ($\Gamma_{k\neq 0}$), since voxels are not detected when *at least* one projected voxel ($\mathbf{x_i}$) is not classified as a foreground pixel. The total number of 3D background super classes is $K = \sum_{i=1}^{C} \binom{C}{i}$.

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

$$P(\Gamma_j|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C})) = \max(P(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C}))). \tag{18}$$

If the cameras are positioned over a short baseline, then the views have high correlation between them. However, it is reasonable to assume that the camera views are statistically independent among them in environments with a scatter of cameras around the scene, which

[3]The prior probability of detecting a foreground voxel can be simply obtained by computing the ratio $\frac{\text{detected voxel}}{\text{total voxel occupancy}}$ using conventional SfS, for instance. $P(\phi)$ and $P(\beta)$ are obtained from P_S: $P(\phi) = \sqrt[C]{P_S}$ and $P(\beta) = 1 - P(\phi)$. Priors can also be set according to the particularities of each set-up, setting low foreground priors where activity is unlikely, for instance.

is the case considered here. Thus, assuming here and in the rest of the section that the super classes are conditionally independent, and using the Bayes theorem:

$$P(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = \frac{P(\Gamma_k)\prod_{i=1}^{C} p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_k)}{p(\mathbf{I}_1(\mathbf{x_1}))\cdots p(\mathbf{I}_C(\mathbf{x_C}))}, \tag{19}$$

where $p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_k)$ is the likelihood of the observation, given a certain super class. For instance, given $\Gamma_2 = \{\phi, \beta, \phi, \cdots, \phi\}$, likelihoods $p(\mathbf{I}_1(\mathbf{x_1}))$ and $p(\mathbf{I}_2(\mathbf{x_2}))$ are

$$p(\mathbf{I}_1(\mathbf{x_1})|\Gamma_2) = p(\mathbf{I}_1(\mathbf{x_1})|\Gamma_2[1]) = p(\mathbf{I}_1(\mathbf{x_1})|\phi) = \frac{1}{256^3}$$
$$p(\mathbf{I}_2(\mathbf{x_2})|\Gamma_2) = p(\mathbf{I}_2(\mathbf{x_2})|\Gamma_2[2]) = p(\mathbf{I}_2(\mathbf{x_2})|\beta) = G_{2,\mathbf{x_2}}(\mathbf{I}_2(\mathbf{x_2})).$$

Substituting (19) into (18) we finally obtain the decision rule

$$\Gamma_j = \underset{\Gamma_k}{\operatorname{argmax}}\, P(\Gamma_k)\prod_{i=1}^{C} p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_k[i]). \tag{20}$$

Or in terms of a posteriori probabilities

$$\Gamma_j = \underset{\Gamma_k}{\operatorname{argmax}}\, P(\Gamma_k)\prod_{i=1}^{C} \frac{P(\Gamma_k[i]|\mathbf{I}_i(\mathbf{x_i}))}{P(\Gamma_k[i])}, \tag{21}$$

which is equivalent to

$$\Gamma_j = \underset{\Gamma_k}{\operatorname{argmax}}\, P(\Gamma_k)^{1-C}\prod_{i=1}^{C} P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})), \tag{22}$$

where $P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))$ is the probability of a super class, given a certain observation. For instance, given $\mathbf{I}_2(\mathbf{x_2})$, the probability of super class $P(\Gamma_{C+1})$ is

$$\begin{aligned} P(\Gamma_{C+1}|\mathbf{I}_2(\mathbf{x_2})) &= P(\beta)P(\beta|\mathbf{I}_2(\mathbf{x_2}))P(\phi)^{C-2} \\ &= P(\beta)\frac{P(\beta)G_{2,\mathbf{x_2}}(\mathbf{I}_2(\mathbf{x_2}))}{p(\mathbf{I}_2(\mathbf{x_2}))}P(\phi)^{C-2}, \end{aligned}$$

where $p(\mathbf{I}_2(\mathbf{x_2}))$ is the unconditional joint distribution of pixel $\mathbf{x_2}$ in view $\mathbf{I}_2$.

Both (20) and (22) decide the most probable super class. However (20) can be used to obtain faster classifications, even though the probabilities are not explicitly computed.

Note that the decision rule is very strict in the sense that a single misclassification in a view inhibits a correct interpretation of the process occurred. Misclassifications are specially sensible in the case of super class Γ_0, since a single misdetection of a ϕ class will let a erroneous 3D background detection. On the contrary, misclassifications in a 3D background super class often

will lead to another 3D background super class, which is not a severe problem. A more in-depth analysis of this unbiased behavior to error types is given in the following section.

In order to prevent such type of errors, we can force the classifiers not to deviate from the prior probabilities. This can be done considering an outlier model in the 2D models [1].

2) Probabilistic Voxel Classification Considering Outliers in the 2D Model: If we consider that the 2D model has an associated probability of outlier e, then we can use the prior probability when the model fails

$$P'(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) = eP(\Gamma_k) + (1-e)P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})), \tag{23}$$

and then,

$$P'(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = \prod_{i=1}^{C}\left(eP(\Gamma_k) + (1-e)P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))\right). \tag{24}$$

A Taylor expansion in f around 0, after replacing variables $f = (1-e)$, gives

$$P'(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = (eP(\Gamma_k))^C + \\ + (eP(\Gamma_k))^{C-1}(1-e)\sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) + O((1-e)^2). \tag{25}$$

If e is close to 1, then only the first two terms matter. This is a rather strong assumption but it may be satisfied when observed data is highly ambiguous.

Under this assumption, super class Γ_j is chosen using the following decision rule

$$\Gamma_j = \underset{\Gamma_k}{\operatorname{argmax}}\left((eP(\Gamma_k))^C + (eP(\Gamma_k))^{C-1}(1-e)\sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))\right). \tag{26}$$

3) 2D Model Update: Once the voxels have been classified with any of the previously discussed procedures, the resulting voxels probabilities are projected to all the views. Note that when the probabilities are projected, special care has to be taken so that pixels are assigned the highest foreground probability value among all voxels whose projection belongs to the pixel. Additionally, the corresponding foreground probability that is projected from 3D to 2D has to be adapted to the change of dimensionality:

$$P(\phi_i|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = \sqrt[C]{P(\Gamma_0|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C}))}, \tag{27}$$

assuming that all the views contributed to the voxel with identical probabilities. This probability can be used to update the 2D background models described in Section II-C. In the 2D MAP setting described in the mentioned section, background models are updated according to their background probabilities ($P(\beta|\mathbf{I_x})$). The Bayesian setting of both approaches let us easily incorporate this 3D extra probabilistic information to the models update process by redefining the $P(\beta|\mathbf{I_x})$ as follows:

$$P'(\beta|\mathbf{I_x}) = P(2D)P(\beta|\mathbf{I_x}) + (1 - P(2D))(1 - P(\phi_i|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C}))), \tag{28}$$

where $P(2D)$ is a design parameter (a prior) that determines the influence that 3D information has into the 2D model update process (a value of $P(2D) = 0.5$ has proved to work well in our experiments).

Projecting back 3D probabilities permits to update 2D background models with higher precision. In Section II we proved that, based on EM, the background models should be updated proportionally to the probability that an observation belongs to the background. Thus, the equations derived here are important to provide more robust learning speeds based on the information acquired from multiple cameras. Note that better adaptation speeds also permit to obtain better 2D background models. In this scheme, the background models are constantly updated making use of the redundancy present in a multi-camera system. In addition, the framework presented here can be extended to incorporate other 3D probabilistic values obtained using other techniques. In this regard, the method developed in the following section will be used to refine the 3D map obtained with the presented method, leading to even better 2D/3D foreground detections.

4) System Implementation: When using a large number of cameras, the class of maximum probability has to be found in a large search-space (K), and computational costs may be too high for certain applications. If this is the case, one can compute the probability of foreground in a voxel, that is the 3D shape probability, $P(\Gamma_0|\mathbf{I}_i(\mathbf{x_i}), i = 1, \cdots, C)$ and set a threshold on this probability. The probability of the 3D-Shape (P(Γ_0)) can be obtained using (19) when working with reliable 2D-models, or with (24) when considering a certain probability of outliers (e) in the 2D-models.

Threshold selection is performed only once per each different type of working environment. The threshold can be simply obtained by inspection of original image confronted to the projected

probabilities (see Fig. 1(a) and (c)). Similarly as discussed in Section III-B.3, note that when the probabilities of the 3D-Shape are projected, special care has to be taken so that pixels are assigned the highest probability value among all voxels whose projection belongs to the pixel. Note that this threshold can be set with very high precision, since probabilities are numbers defined in $\mathbb{R}$. On the contrary, in classical SfS, thresholds have to be set in the realm of integer numbers $\mathbb{Z}$, i.e., one has to decide the minimum number of foreground projections in 2D that form a voxel in 3D.

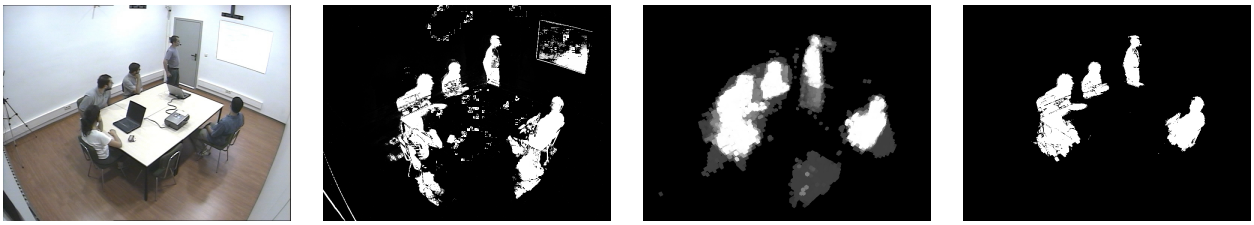
Finally, it has to be remarked that the most reliable classification, with a Bayesian justification, is done using (20), when considering error-free 2D models and (26) when considering an error model. The drawback is that the probabilities of all the 3D background super classes, which we are not interested in, will have to computed.

5) Results: The proposed scheme 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. Our system has dealt with all these problems successfully, improving the results of conventional 2D segmentators and standard SfS reconstruction methods.

Fig. 1 shows an example in a certain view and instant. In this example, we have used foreground priors equal to 0 in those regions which are within $0.4m$ of the walls. The original image (a) can be compared to the resulting mask after performing a conventional 2D foreground segmentation in (b) and a cooperative 2D foreground segmentation in (d). In the example, the outlier model in (24), without further simplifications is used. In this example, we have used $e = 0.5$. The classification is performed setting a threshold to the probability of 3D-foreground by inspection of (c), as discussed in the previous section.

Inspection of silhouettes (b) and (d) shows that the 2D models learned in the cooperative approach are clearly better than those which are learned using a single-view approach.

The Bayesian setting presented has proved to work well. However there are a set of problems which this technique cannot alleviate. When errors are systematic, that is, some areas which are completely missed in certain views consistently along the time, then the technique is not able to detect the problem. In the next section we present an approach in this direction. The integration of both methods is discussed in Section IV.



(a) (b) (c) (d)

Fig. 1. 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.

Finally, in Section II we showed the close relation between the pixel model update and pixel background probability. This relation can be exploited by using more-informed probabilistic sources information that are external to the pixel process. In this section, we have proposed to use the projection of probabilistic 3D maps that are created form 2D views. This has allowed to bridge the gap between the planar and volumetric foreground detection tasks. Moreover, the framework can be extended to incorporate external 3D probabilistic values which are obtained using other set of techniques. In this respect, the method developed in the following section will be used to refine the 3D map obtained with the presented method, leading to even better 2D/3D foreground detections.

C. 3D-Shape Reconstruction Considering Geometric Constraints

The previous subsection has presented an approach to improve 2D and 3D detections in the SfS framework, by performing a joint 2D-3D classification. However, we have not dealt with systematic errors, such us those due to occlusions in one or more of the views or wrong detections in one or more of the views due for instance to a similar color of the object and its background model. In order to deal with this kind of errors, we proposed in [18], [19] to examine in more detail the concept of the Visual Hull, and how the 2D errors are propagated to the 3D reconstruction.

Regarding occlusions, two kind of occlusions of the foreground objects are possible. The first one is produced by other foreground objects. In this case, Shape from Silhouette algorithms work correctly, because the voxels occluded in a given view project into foreground pixels, although

this projection corresponds to the occluding object. The second one is produced by background objects of the scene, such as tables or other furniture placed in the middle of the scene. In this case, a classical Shape from Silhouette, or the probabilistic one proposed in Section III-B, will not reconstruct the corresponding voxels. However, geometric considerations can be used to improve the situation when the number of occluded views is limited.

The concept of VH is strongly linked to the one of silhouettes' consistency: A set of silhouettes is consistent if there exists at least one volume which exactly explains the complete set of silhouettes, and the VH is the maximal volume among the possible ones. If the silhouettes are not consistent, then it does not exist an object silhouette-equivalent, that is, the VH does not exist. Total consistency hardly ever happens in realistic scenarios due to inaccurate calibration or noisy silhouettes caused by errors during the 2D detection process. In spite of that, most SfS methods have been designed in the past assuming that the silhouettes are consistent, thus reconstructing only the part of the volume which projects consistently in all the silhouettes, i.e., the volume where the visual cones intersect, without further considerations.

Our proposal in [18], [19] is to use a shape reconstruction method based on the silhouette consistency principle. Our system validates the regions in the silhouettes which are consistent in all the projections and adjusts the regions which are not, dealing with 2D errors, i.e., misses (foreground voxels detected as background) and false alarms (background voxels detected as foreground), in an unbiased way. By contrast, other SfS systems usually treat differently the 2D errors on the basis of their type.

In classical SfS, a false alarm in a view does not contribute to a false alarm in 3D unless the visual cone that is erroneously created intersects simultaneously with other $C-1$ visual cones, where C is the total number of cameras. If the intersection is produced, then the volumetric points corresponding to the intersection are wrongly reconstructed. Since the reconstructed shape is consistent because its projection in all the views matches with the silhouettes, then the 2D false alarm is undetectable. However, the shape is not reconstructed in the parts of the volume where at least one of the erroneous visual cones does not intersect simultaneously with other $C-1$ visual cones. This is the most typical case in scenarios where the major part of the volume is unoccupied. In such case, the cones produced by 2D false alarms do not intersect with visual cones from the rest of cameras, then 2D false alarms are inconsistent with the reconstructed shape, allowing their detection as we will show in the following.

Contrarily, a miss in a view inhibits the simultaneous intersection of C visual cones in 3D, leading to an ineluctable miss in the shape. This makes the SfS algorithm highly sensitive to this type of errors, whereas 2D false alarms do not produce erroneous reconstructions in most of the cases. 2D misses can also be indirectly detectable, since the projection of the incomplete Visual Hull reconstructed will not match with the rest of correct silhouettes.

Apart from the methods mentioned in the previous section for dealing with the 3D reconstruction errors, another approach which is specifically oriented to counteract this asymmetry between misses and false alarms, is to require the intersection of at least $C - P$ visual cones to allow a reconstruction, where P is the number of acceptable misses among the set C of cameras. Although single misses do not block the reconstruction in this approach, the resulting shape is larger than the real Visual Hull for requiring fewer intersections of visual cones. A drawback of this approach is that larger hulls are reconstructed either if the silhouettes are consistent or not.

We will use multi camera consistency constraints for detecting systematic errors. We use a fast technique for estimating that part of the volume which projects inconsistently and propose a criteria for classifying it either as part of the shape or not by minimizing the probability of voxel misclassification. Our approach is voxel-based and can be used to correct errors from any Shape from Silhouette technique, from the standard ones to those which were proposed to minimize the effects of noise in the foreground detection [30], [8]. In particular, we will develop its application in the context of the Cooperative 2D-3D framework developed in last section.

1) Shape from Inconsistent Silhouette (SfIS): In Shape from Inconsistent Silhouette (SfIS), the VH is reconstructed using SfS methods and corrected later with those parts of the volume which were not correctly classified. 3D misclassifications can be detected by examining the inconsistent regions of the silhouettes. To detect inconsistent regions, one can project back the VH and test whether the projections match with the generative silhouettes. Then, the shape can be reconstructed using a different criterion when there are parts of the volume (*Inconsistent Volume:IV*) which project to inconsistent regions in the silhouettes (*Inconsistent Silhouettes:ISs*).

As we showed in [18], the main problem to solve will be how to choose the minimum number of inconsistent intersections ($T^\star$) that have to be produced so that it can be determined that a part of the Shape was missed during the reconstruction process.

The optimal threshold $T^\star$ has to be such that if $\mathcal{I} \geq T^\star$, the voxel is better explained as Shape

than Background:

$$\begin{aligned} \mathfrak{I} \geq T^{\star} \quad &\Rightarrow \quad \text{decide } \textit{Shape} \\ \mathfrak{I} < T^{\star} \quad &\Rightarrow \quad \text{decide } \textit{Background}, \end{aligned} \tag{29}$$

where $\mathfrak{I}$ corresponds to the number of inconsistent foreground projections.

In order to find $T^{\star}$, we only have to express the probability of voxel misclassification for any $P(Err_{3D}[T])$ so that $T^{\star}$ is that one which minimizes it:

$$T^{\star} = \underset{T}{\operatorname{argmin}}\, P(Err_{3D}[T]), \tag{30}$$

which is a function of the false alarm and miss probabilities that we derived in Section II.

Since voxel classification errors may be due to either false alarms or misses, the probability that a voxel is misclassified is:

$$P(Err_{3D}) = P_B P(FA_{3D}) + P_S P(M_{3D}), \tag{31}$$

where P_B and P_S are prior probabilities of a voxel forming part of the Background or Shape, respectively[4], and $P(FA_{3D})$ and $P(M_{3D})$ correspond to the probabilities of false alarm and miss in a voxel. They can be computed, as shown in [18], as a function of the threshold T, being $\mathcal{O}$ the number of consistent foreground projections of a voxel, as:

$$P(FA_{3D}) = \sum_{i=max(T,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(FA_{2D})^i (1 - P(FA_{2D}))^{C-i}, \tag{32}$$

corresponding to the summation of all possible combinations that trigger a false alarm in a voxel, and assuming equiprobable $P_i(FA_{2D}) = P(FA_{2D})$ in all views (i)

$$P(M_{3D}) = \sum_{i=max(C-\mathcal{O}-T+1,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i}, \tag{33}$$

where $P(M_{2D})$ corresponds to the probability that the projection test has not been passed by error, and assuming equiprobable $P_i(M_{2D}) = P(M_{2D})$ in all views (i).

SfIS can be very fast, once the optimal thresholds have been computed for each possible case of occlusion and stored in a lookup table (LUT). Real-time operation of SfIS can be achieved when using it in combination with fast projection tests. Often, the One Pixel Projection Test is used for being fast and simple since it simply consists in projecting the point in the center of a

[4]Priors P_S and $P_B = 1 - P_S$ can be simply obtained by computing the detected/total voxel occupancy ratio, for instance.

voxel into a pixel for each camera views. However, LUTs cannot be used when probabilities of 2D miss and false alarm of the projection test change over time ($P(FA_{Pix}(t))$ and $P(M_{Pix}(t))$). For example, when a mixture of Gaussians is used to model the Background, the probabilities of miss and false alarm of the pixels depend on the variances of the Gaussians, which are constantly changing over time. Under these circumstances, it is important to have a fast search strategy that can compute the optimal thresholds on-line. The computation of this thresholds on-line has been introduced in [19].

IV. A Unified Cooperative-SfIS Bayesian Framework

In the previous sections we have described the probabilistic methods for obtaining 2D silhouettes and a cooperative framework that allowed obtaining 3D classifications using 2D probabilities. In addition, the bases for 2D model update using probabilistic 3D information were also established. In the last section, we have reviewed a tool that allows to reclassify an initial volumetric estimate making use of the geometrical constraints of the problem. The last question to be solved is then, how to incorporate the information generated by these geometrical constraints to the integrated 2D-3D Bayesian framework described. In fact, both Bayesian and geometrical approaches can cooperate to the benefit of the system.

To obtain a volumetric estimate using geometrical constraints in our 2D-3D Bayesian framework, first, a set of probabilistic pixel models are created for each image in the set-up. Pixel models can be used for Bayesian classification of 2D silhouettes, which can be later employed for obtaining 3D reconstructions. In addition, it is possible to maintain those pixel models with new observations so that their posterior probabilities are always maximum.

This 2D probabilistic information is incorporated to the SfIS approach as follows. Once the pixel models have been estimated, an initial version of the SfS is obtained using the SfS cooperative approach described. The cooperative approach makes use of the two-dimensional probabilistic methods previously mentioned to obtain foreground and background probabilities for each voxel. These probabilities are then used to classify all the voxels. Contrarily to the classical SfS approach, in the cooperative approach, 2D probabilistic values are transferred to 3D and used there to classify the volume.

Once a volumetric Shape has been classified, then it is projected back to each one of the views and compared there with a set of silhouettes that are temporarily classified using only the

pixel models. The inconsistencies between Shape and Silhouettes are determined and then the SfIS algorithm is applied as usual so that a refined 3D model is obtained.

The last step corresponds to the process in which the images' pixels models are updated. Note that in Bayesian 2D foreground segmentation, the models' adaptation speed varies according to the probabilities of each class, as shown in Section II. In the cooperative approach, the voxel probabilities are projected and used as the adaptation speed of the pixels models. However, at this point, SfIS provides an extra source of information which can be incorporated to the probabilistic voxel representation obtained with the cooperative SfS approach. Those voxels which SfIS reclassifies are therefore reassigned with prior fixed probabilistic values (0.9^C for a 3D miss and 0.1^C for a 3D false detection).

Finally, the 2D pixel models are updated using (16), after projecting the reassigned 3D probabilities with the projection rules defined. The main steps of the presented approach are summarized in the algorithm below.

Algorithm 1 Cooperative SfIS algorithm

Require: Video Sequences for each camera: $VS(camera)$

for all frames **do**
  for all camera **do**
    Compute 2D Probabilities using (8) and (9)
  end for
  Compute 3D Probabilities from 2D Probabilities P(Γ_0) using (19) or (24)
  Binarize 3D Probabilities
  Perform SfIS (algorithm details available in [19])
  Assign foreground priors (0.9^C) for the voxels reclassified as shape with SfIS
  Obtain final reconstruction using MAP over 3D Probabilities using (26)
  Update 2D Models employing 3D probabilities using (28)
end for

Incorporating SfIS to the cooperative Bayesian framework clearly improves the overall system accuracy. Indeed, it is important not to update models with wrong observation values due to erroneous classifications. SfIS helps in detecting some of these errors in the silhouettes making

use of the existing redundancy in a multi-camera setup.

This integration of probabilistic and geometrical information compiles all the different aspects that this paper has addressed. First, the 2D Bayesian approach is used, then the 3D cooperative background learning is employed to obtain a preliminary set of 3D probabilistic values. This initial volumetric probabilistic representation is refined using SfIS and, finally, the 2D models are updated using our unified Bayesian framework.

The reconstructed 3D Shapes provide detailed information of the moving entities of the scene. This information can be used for the applications of higher semantic content. Some of the applications are the computation of the trajectories of the moving entities in the scene and the extraction of activity information. It can also be the input information for a human motion analysis system using body models [3]. By using these higher level systems, the reconstructed entities can be separated in humans or objects and their motion can be determined with more precision.

V. RESULTS

The experiments have been performed using a low-resolution volume, employing voxels with edge size $2.5\ cm$, therefore prioritizing fast 3D detections over a more accurate Shape.

Since we are using real-world images with imprecise calibration, we have opted to indirectly evaluate the performance of the reconstruction methods. To do so, we have compared the projection of the 3D volumes with a set of five manually classified silhouettes from images that have been randomly selected in a video sequence. These manually labeled silhouettes are the ground truth.

Four different techniques have been compared. The first technique is a version of SfS where a voxel is not classified as Shape if there are *more than one* views where its projection test fails. We identify this method as *SfS* $C-1$ *intersections* in our experiments. The second evaluated technique is traditional SfS. Third and fourth tested methods are SfIS and cooperative SfIS, respectively. In this experiment, we have always employed the One Pixel Projection Test in all the methods for a fair comparison.

The pixel models employed for 2D classifications are a single Gaussian per pixel for the background and a uniform distribution for the foreground. 2D classifications are obtained using MAP and the models are updated using EM. The pixels models classification and update steps

were described in detail in Section II. In this experiment, the models adaptation speed corresponds to the probability after MAP of 2D models except in the cooperative SfIS approach that uses the projection of the 3D probabilistic representation described in Section III.

For visual inspection purposes, we present two figures (Fig. 2 and Fig. 3) with results corresponding to different times and camera views of a scene[5]. These tests were performed in the smart-room of the UPC. In the first row of images for all the camera groups, the original view and some intermediate results are presented and, in the second row, the projections of the Shapes obtained with the methods under evaluation are shown.

In Fig. 2, the images corresponding to camera 2 in frame number 175 are shown. Note that the 2D only segmentation (2nd column, 1st row) -not using 3D redundancy information- has failed due to the similar colors of the person in the foreground and the clutter in the background. However, see that the projected voxel probabilities using the cooperative SfIS approach (3rd column, 1st row) do correct these errors and, therefore, 2D segmentations using the cooperative approach are more precise (4th column, 1st row).

Similar problems are observable in Fig. 3. The figure corresponds to frame 650 and shows three out of the five camera views used in all the methods. Note that 2D misses in a view are transferred to the rest of views in the SfS approach. The SfS $C-1$ approach does not propagate 2D misses but incorporates many false alarms conducing to larger Shapes and silhouettes' projections. As it can be observed from the images, SfIS is a good approach for not propagating 2D misses as well as for not incorporating many false alarms. The cooperative SfIS approach behaves even better than SfIS because it informs the 2D models when an error is made and, thus, the pixels models are updated with a more-informed strategy.

Quantitative results of this experiment are presented in Table I. These results have been obtained by averaging the number of 2D false alarms, 2D correct detections and 2D misses over a set of projected reconstructions. These projections correspond to the five views where the silhouettes were manually labeled to be the ground truth, as previously commented. The

[5]Due to the great number of images resulting from the methods compared, the number of camera views and the large time interval evaluated, it is not possible to show here the complete set of results. However, the video sequences with all the evaluated methods at all the cameras views can be obtained in http://gps-tsc.upc.es/imatge/_Montse/Coop_SFIS.html

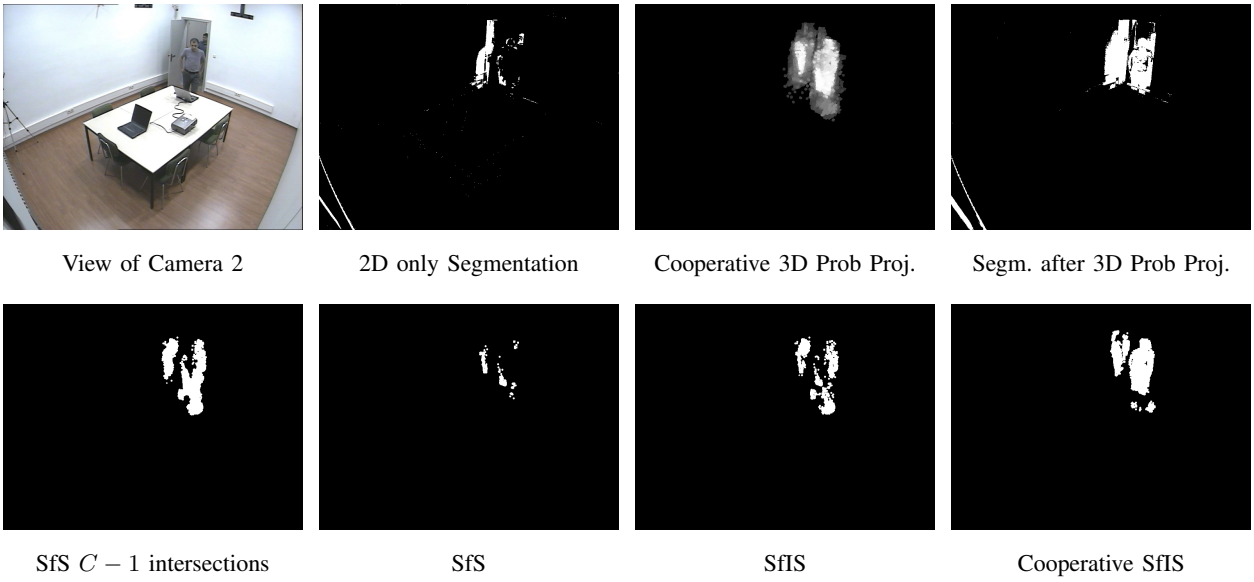


Fig. 2. Silhouettes and 3D volumetric projections corresponding to frame 175 with different techniques using the One Pixel Projection Test.

verification measures that have been used are defined as follows:

$$\text{Recall} = \frac{\#\text{correct Shape detections}}{\#\text{correct Shape detections} + \#\text{misses}}$$
$$\text{Precision} = \frac{\#\text{correct Shape detections}}{\#\text{correct Shape detections} + \#\text{false Shape detections}}. \quad (34)$$

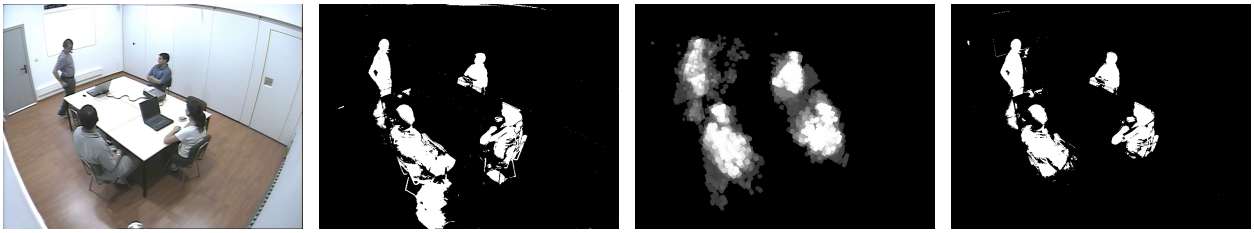
In order to combine precision and recall, we employ the F-measure, also known as the harmonic mean of precision and recall. The F-measure is the measure that we use to evaluate the overall performance of the system:

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (35)$$

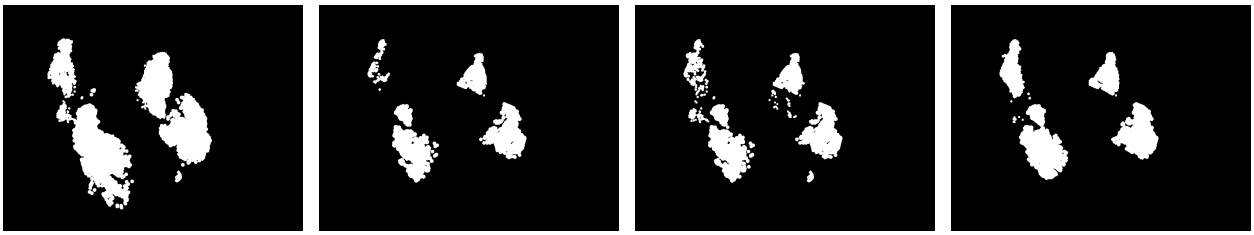
To sum up, recall measures how well the classifier detects voxels that form part of the Shape and precision measures how well it weeds out the voxels in the background. A well balanced system should have high, similar values of both recall and precision.

Some interesting conclusions can be extracted from the table.

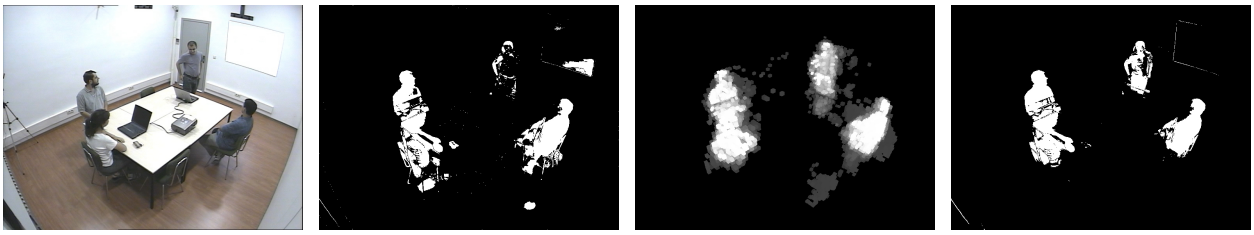
Note that the SfS $C-1$ approach has a highest recall rate. Indeed, it also has a very large number of false alarms and, therefore, a poor precision rate, but it is a good method if we want to be sure to detect the foreground voxels when they exist.



View of Camera 1 | 2D only Segmentation | Cooperative 3D Prob Proj. | Segm. after 3D Prob Proj.

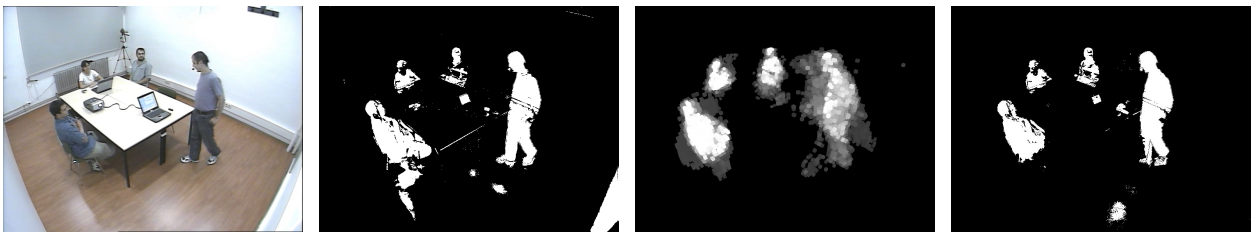


SfS $C-1$ intersections | SfS | SfIS | Cooperative SfIS

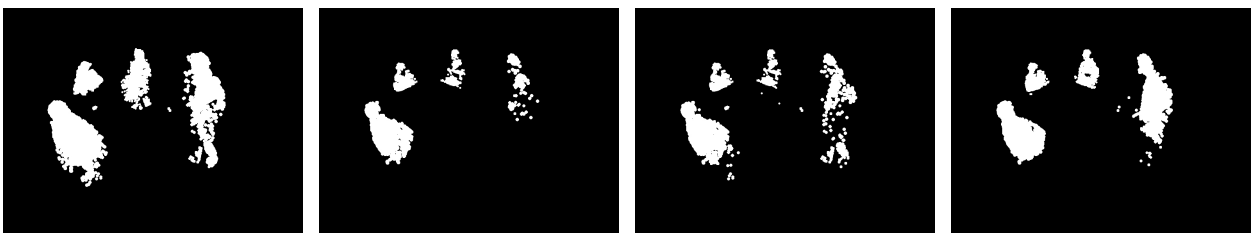


View of Camera 2 | 2D only Segmentation | Cooperative 3D Prob Proj. | Segm. after 3D Prob Proj.

SfS $C-1$ intersections | SfS | SfIS | Cooperative SfIS



View of Camera 3 | 2D only Segmentation | Cooperative 3D Prob Proj. | Segm. after 3D Prob Proj.



SfS $C-1$ intersections | SfS | SfIS | Cooperative SfIS

Fig. 3. Silhouettes and 3D volumetric projections corresponding to frame 650 with different techniques using the One Pixel Projection Test.

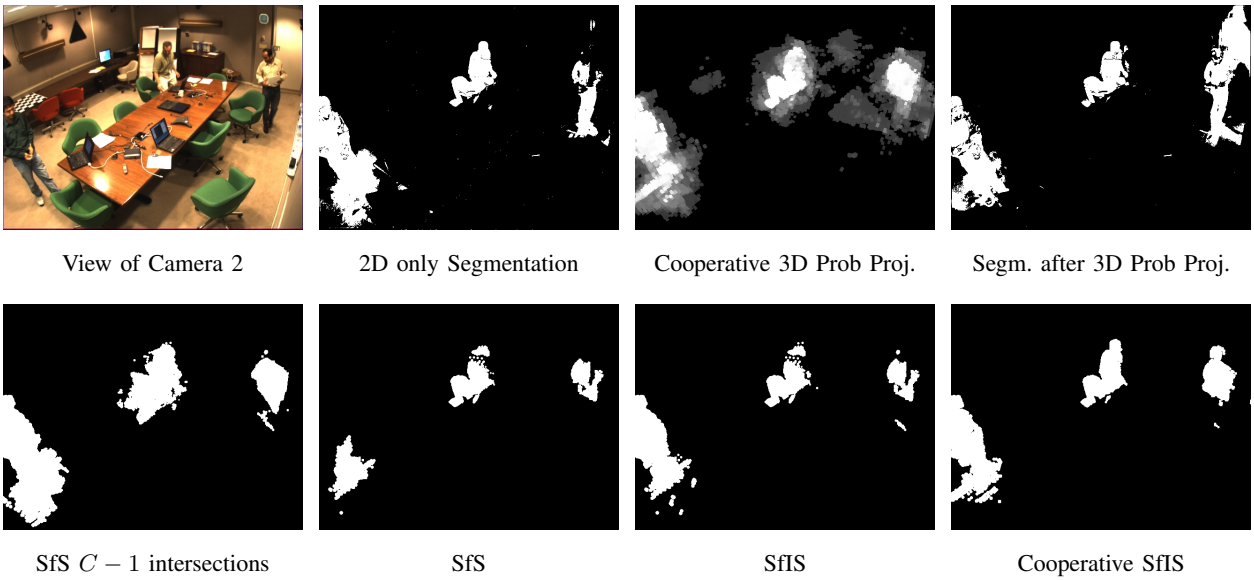


Fig. 4. Silhouettes and 3D volumetric projections corresponding to one frame captured in the smart room of IBM Czech Republic research labs. Results obtained for precision, recall and F-measure are 0.48, 0.93, 0.63 for SfS C-1, 0.62, 0.46, 0.53 for SfS, 0.69, 0.77, 0.73 for SfIS and 0.67, 0.86, 0.75 for the cooperative SfIS approach, respectively.

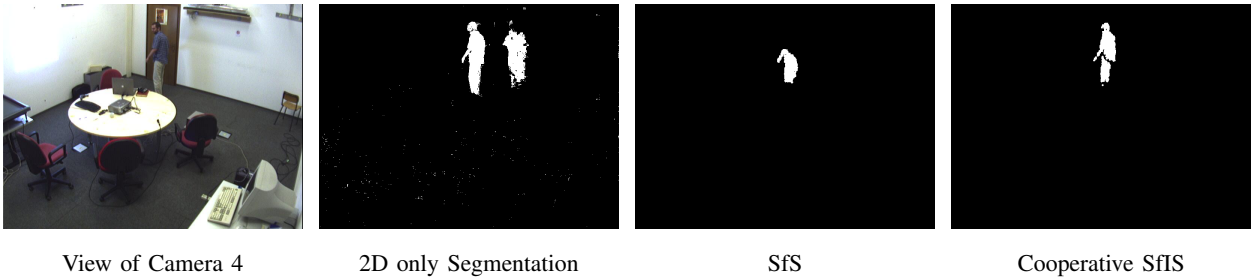


Fig. 5. Silhouettes and 3D volumetric projections corresponding to one frame captured in the smart room of the Istituto Trentino di Cultura (ITC) research labs. Results obtained for precision, recall and F-measure are 0.85, 0.40, 0.54 for SfS and 0.93, 0.71, 0.80 for the cooperative SfIS approach, respectively.

In contrast, traditional SfS is very precise, even more than SfIS. Note that SfS detects fewer voxels but it is very good at asserting that those voxels form part of the Shape.

SfIS and cooperative SfIS are the most balanced methods. They have high precision and recall rates and their F-measures are the best. SfIS improves when combined with the cooperative Bayesian framework since the 3D information is continuously flowing to 2D in a probabilistic manner.

Similar considerations can be made from the set of images presented in Fig. 4 and Fig. 5,

TABLE I

SYSTEM EVALUATION THROUGH THE PROJECTION OF 3D RECONSTRUCTIONS IN VIDEO SEQUENCES

	Ground truth	SfS ($C-1$ int.)	SfS	SfIS	Coop. SfIS
# Correct foreground det.	32270	**27471**	15023	20445	25061
# False alarms	0	29760	**5077**	7529	6872
# Misses	0	**4808**	13256	11834	7218
Recall	1	**0.85**	0.53	0.63	0.77
Precision	1	0.48	0.75	0.73	**0.78**
F-measure : $\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$	1	0.62	0.62	0.68	**0.77**

captured in other scenarios. The figures are presented together with the values for precision, recall and F-measure for different reconstruction methods.

In conclusion, the cooperative SfIS approach definitely has the best F-measure of them, as simple visual inspection of the images confirms and it is the method that performs best when operating with video sequences.

VI. CONCLUSION

In this paper we have proposed a unified framework for the planar and volumetric foreground detection tasks. The proposed framework operates in a collaborative manner by transferring more-informed 3D probabilistic information back to the planar detectors in each image that, in turn, also become more precise.

The cooperative planar-volumetric detector has been achieved by first extending planar detection techniques to a Bayesian framework. We have set the bases for Maximum a Posteriori classification and introduced model update equations based on Expectation Maximization. This shows that for proper Bayesian update, new background observations have to be incorporated to the background model proportionally to the probability of the background class. This has permitted us to employ the update scheme of the planar activity detectors to include higher-level probabilistic information obtained by other means. In this way, we have developed a new framework in which 3D probabilistic information is attained from 2D probabilistic maps and then projected back to each one of the original views to be used as the update speed of the two-dimensional background models. Finally, we have reexamined the volumetric activity detection

task. Following a different line of thought, we have studied the coherence between planar and volumetric detectors and incorporated a novel technique, called Shape form Inconsistent Silhouette (SfIS). Basically, SfIS is able to reclassify some of the initial volumetric detections so that the misclassification error is minimized. SfIS can be used to extract better volumes by minimizing the effects that inconsistencies have over the reconstructed Shape. In addition, SfIS can also be used to recover errors in the silhouettes from more-informed decisions made at the volume level, where individual detections at the 2D level are compared for consistency. Finally, reassigned classifications can be introduced back into our Bayesian cooperative framework providing better long-term video activity detections in both the 2D and 3D domains.

REFERENCES

[1] Thomas Minka August. The 'summation hack' as an outlier model. Technical report, Department of Statistics, Carnegie Mellon University, 2003. Available from: http://www.stat.cmu.edu/minka/papers/minka-summation.pdf.

[2] Bruce G. Baumgart. *Geometric Modeling for Computer Vision*. PhD thesis, CS Department, Stanford University, October 1974. AIM-249, STAN-CS-74-463.

[3] C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Exploiting structural hierarchy in articulated objects towards robust motion capture. *Submitted to 5th Conference on Articulated Motion and Deformable Objects MLMI*, 2008.

[4] Kong Man Cheung, Takeo Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 714 – 720. IEEE Computer Society, June 2000.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[6] C. R. Dyer. Volumetric scene reconstruction from multiple views. In *Foundations of Image Understanding*, pages 469–489. Kluwer, 2001.

[7] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, Sept 1999.

[8] Jean-Sébastien Franco and Edmond Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, October 2005.

[9] J. Gallego, M. Pardàs, and J.L. Landabaso. Segmentation and tracking of static and moving objects in video surveillance scenarios. *Submtted to IEEE International Conference on Image Processing*, 2008.

[10] Jack Goldfeather, Jeff P M Hultquist, and Henry Fuchs. Fast constructive-solid geometry display in the pixel-powers graphics system. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 107–116, New York, NY, USA, 1986. ACM Press.

[11] Haritaoglu, D. Harwood, and L. Davis. W4: Real time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2000.

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[13] T. Horpraset, D. Harwood, and L. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, 1999.

[14] P. Kaewtrakulpong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Proceedings of 2nd European Workshop on Advanced Video Based Surveillance Systems*, volume 5308, 2001.

[15] S. Khan and M. Shah. Tracking people in presence of occlusion. In *Proceedings of Asian Conference on Computer Vision*, 2000.

[16] J. L. Landabaso and M. Pardàs. Foreground regions extraction and characterization towards real-time object tracking. In *Proceedings of Multimodal Interaction and Related Machine Learning Algorithms*, Lecture Notes in Computer Science. Springer, 2005.

[17] J. L. Landabaso and M. Pardàs. Cooperative background modelling using multiple cameras towards human detection in smart-rooms (invited paper). In *Proceedings of European Signal Processing Conference*, 2006.

[18] J. L. Landabaso, M. Pardàs, and J.R. Casas. Reconstruction of 3D shapes considering inconsistent 2D silhouettes. In *Proceedings of International Conference on Image Processing*. IEEE Computer Society, 2006.

[19] J. L. Landabaso, M. Pardàs, and J.R. Casas. Shape from Inconsistent Silhouette. *Accepted for publication in Journal of Computer Vision and Image Understanding*, 2008.

[20] A. Laurentini. The Visual Hull: A new tool for contour-based image understanding. In *Proceedings of Seventh Scandinavian Comperence on Image Processing*, pages 993–1002, 1991.

[21] Liyuan Li, Weimin Huang, Irene Y. H. Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[22] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 369–374, New York, NY, USA, 2000. ACM Press.

[23] Stephen J. McKenna, Sumer Jabri, Zoran Duric, Azriel Rosenfeld, and Harry Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, 2000.

[24] Stephen J. McKenna, Yogesh Raja, and Shaogang Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3-4):225–231, 1999.

[25] Anurag Mittal and Larry S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *Proceedings of European Conference on Computer Vision*, pages 18–36, London, UK, 2002. Springer-Verlag.

[26] Saied Moezzi, Arun Katkere, Don Y. Kuramura, and Ramesh Jain. Reality modeling and visualization from multiple video sequences. *IEEE Computer Graphics and Applications*, 16(6):58–63, 1996.

[27] A. Prati, I. Mikic, MM Trivedi, and R. Cucchiara. Detecting moving shadows: algorithms and evaluation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):918–923, 2003.

[28] Ari Rappoport and Steven Spitz. Interactive boolean operations for conceptual design of 3-d solids. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 269–278, New York, NY, USA, 1997. ACM Press.

[29] Greg Slabaugh, Bruce Culbertson, Thomas Malzbender, and Ron Shafer. A survey of methods for volumetric scene reconstruction from photographs. In *International Workshop on Volume Graphics*, Stony Brook, New York, June 2001.

[30] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *Proceedings of Computer Vision and Pattern Recognition*, pages 345–353. IEEE Computer Society, 2000.

[31] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[32] H. G. C. Traven. A neural network approach to statistical pattern classification by'semiparametric' estimation of probability density functions. *IEEE Transactions on Neural Networks*, 2(3):366–377, May 1991.

[33] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[34] L.Q. Xu, JL Landabaso, and M. Pardas. Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.