14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

COOPERATIVE BACKGROUND MODELLING USING MULTIPLE CAMERAS TOWARDS HUMAN DETECTION IN SMART-ROOMS

J.L. Landabaso, M. Pardas

Image Processing Group, Technical University of Catalonia, Barcelona, Spain
{jl,montse}@gps.tsc.upc.edu

ABSTRACT

In multi-camera systems for human detection and analysis, Shape-from-Silhouette (SfS) is a common approach taken to reconstruct the Visual Hull, i.e. the 3D-Shape, of the bodies. The reconstructed 3D-Shape is later used in 3D-trackers and body fitting techniques. 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. Silhouettes are usually extracted using a foreground classification process, which is performed independently in each camera view. In this paper we present a novel approach in which 2D-foreground classification is achieved in 3D accordance in a Bayesian framework. In our approach, instead of classifying images and reconstructing the volume later, we simultaneously reconstruct and classify in the 3D space. Furthermore, the classification of the 3D space is used to obtain a more accurate model of the 2D-background.

1. INTRODUCTION

Background modelling has been one of the fields where computer vision has had a major impact, driving successful deployment of several visual surveillance/behavior modelling systems. Advances in the field have led to multi-camera systems designed to build volumetric models from a set of silhouettes (foreground masks), in what is called Shape-from-Silhouette (SfS) [1, 10, 11].

In traditional SfS, a calibrated [18] set of cameras must be placed around the scene of interest, and pixels in all cameras must be provided as either part of the foreground (ϕ) or background (β). Each of the foreground camera point defines a ray in the scene space that intersects 3D entities at some unknown depth along this ray; the union of these visual rays for all points in the silhouette defines a generalized cone within which the entities must lie. The 3D-Shape defined by the intersection of all the cones encloses all the foreground entities of the scene.

SfS plays an important role in smart-room environments, where fast 3D-Shape reconstructions are used as the building blocks of 3D-trackers[9] and body fitting techniques[2, 3, 7]. Usually, a foreground separation process is performed at each camera view. Then, the 3D-foreground scene is discretized into voxels, making use of the voxel-based SfS approach. Finally, foreground voxels are grouped into meaningful blobs and temporally tracked in 3D, preventing the difficulties of occlusions in 2D trackers. Alternatively, foreground voxels are used to fit models of the human body.

We propose a 2D-background modelling and foreground classification scheme to be used in this context. While all the state-of-the-art techniques independently learn the 2D-background models in each view, in our approach, the 2D-background models are learned using evidence from all the cameras in a Bayesian framework. Note that this is quite different from the standard approach, since 2D classification is performed using the redundancy of all the cameras. Thus, the system advantages is twofold; better 2D foreground regions are extracted and, therefore more accurate 3D foreground volumetric models of the humans in the scene can be obtained.

The paper is structured as follows. In the next section, the foreground and background models used in 2D are discussed, with emphasis on the segmentation as a classification process. Section 3 is devoted to discussion of the voxel-based SfS approach, including the approaches taken by some practitioners. These techniques first classify the pixels into foreground and background for all cameras and then reconstruct the 3D volumes using the previous classification result. In section 4, two approaches for reconstructing the 3D shapes using the 2D probabilistic models cooperatively are discussed. In the first approach, 2D-models are considered to be error-free, and in the second one, an error model is introduced. Section 5 presents some experimental results. The paper concludes in section 6 with a discussion of future research direction.

2. 2D MODELS

In this section we describe the probabilistic fore/background models that are estimated for the pixels in a certain view independently of the other pixels in the rest of views. This is the common procedure that can be used when working with only one camera.

In order to obtain accurate 2D-segmentations using a Bayesian framework, 2D-models should fulfill the following requirements:

- Should not only provide fore/background classifications, but also their probabilities.
- Foreground segmentations should not be performed as an exception to the background model.
- Finally, the models should be adaptive to slow lighting changes so that they can cope with fluctuations of the light such as daylight changes in a room with windows, or when a beamer is being used, etc.

2.1 Single-Class Adaptive Background Models

We adopt a single-class statistical model for modelling the background colour of a pixel $\mathbf{x}$ (indicating its spatial coordinates), given observations of its colour value $\mathbf{I}(\mathbf{x})$ across time. For this purpose, we use a Gaussian probability density function. Gaussians have been previously proposed in [5, 15, 17], among others, to ensure that the cameras thermal noise does not produce classification errors. Some of these works[5, 15], adopt multi-class models to model repetitive background, such as in weaving flags, or moving trees. However, a single-class model is enough in our approach, since our system is being developed to operate in a scene that consists of a relatively static situation:

$$G_{\mathbf{x}}(\mathbf{I}(\mathbf{x})) = \frac{1}{(2\pi)^{3/2}\sqrt{|\mathbf{\Sigma}_{\mathbf{x}}|}} e^{-\frac{1}{2}(\mathbf{I}(\mathbf{x})-\mu_{\mathbf{x}})^T \mathbf{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{I}(\mathbf{x})-\mu_{\mathbf{x}})}, \quad (1)$$

corresponding to the Gaussian that models the colour of the background process of pixel $\mathbf{x}$, and where pixel colour val-

ues ($\mathbf{I}(\mathbf{x})$) are expressed as a vector of three dimensions in the RGB colour space. Often it is assumed that the covariance matrix is diagonal with R, G and B sharing the same variances: $\mathbf{\Sigma_x} = \sigma_{\mathbf{x}}^2 \cdot \mathbf{Id_{3\times 3}}$.

Similarly as in[15], model adaptation is implemented as a low pass filter procedure. Thus, once the pixel value has been classified into the background, the model is adapted as follows

$$\begin{aligned}\mu_{\mathbf{x}}[t] =&(1-\rho)\mu_{\mathbf{x}}[t-1] + \rho\mathbf{I}(\mathbf{x}) \\ \sigma_{\mathbf{x}}^2[t] =&(1-\rho)\sigma_{\mathbf{x}}^2[t-1]+ \\ &+\rho(\mathbf{I}_{\mathbf{x}}(\mathbf{x})-\mu_{\mathbf{x}}[t])^T(\mathbf{I}_{\mathbf{x}}(\mathbf{x})-\mu_{\mathbf{x}}[t])\end{aligned} \quad (2)$$

where ρ is the adaptation learning rate: $\rho \propto G_{\mathbf{x}}(\mathbf{I}(\mathbf{x}))$.

However, our approach differs from[15] in an important way. In[15], background classification is performed when the pixel value falls within 2.5 standard deviations of the mean of the Gaussian. Otherwise, it is classified as foreground. Our approach differs in that the foreground is not classified as an exception to the background model. Instead, we prefer to express the problem in a Bayesian form. To do so, first we need to model the foreground process.

2.2 Uniform Foreground Model

The foreground process can be modelled using histograms, Gaussians or any other *pdf*. However, we simply use a uniform *pdf* to model the foreground process in each pixel, which is in fact the probabilistic extension of classifying a foreground pixel as an exception to the model, as discussed in[13].

Since a pixel admits 256^3 colours in the RGB colour space, we model its *pdf* as

$$U_{\mathbf{x}}(\mathbf{I}(\mathbf{x})) = \frac{1}{256^3} \quad (3)$$

2.3 2D Fore/Background Classification

Once that the foreground and background likelihoods of a pixel have been introduced, and assuming that we have some knowledge of foreground and background prior probabilities, $P(\phi)$ and $P(\beta)$[1], respectively, we are now in position to further discuss how the 2D-classification process can be done.

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

$$P(\phi|\mathbf{I}(\mathbf{x})) = \frac{P(\phi)p(\mathbf{I}(\mathbf{x})|\phi)}{p(\mathbf{I}(\mathbf{x}))}, \quad (4)$$

In order to compute (4), the unconditional joint probability density ($p(\mathbf{I}(\mathbf{x}))$) can be expressed in terms of the conditional distributions as

$$p(\mathbf{I}(\mathbf{x})) = P(\phi)p(\mathbf{I}(\mathbf{x})|\phi) + P(\beta)p(\mathbf{I}(\mathbf{x})|\beta) \quad (5)$$

Then, in the case of the models described in the previous section, (4) is

$$P(\phi|\mathbf{I}(\mathbf{x})) = \frac{P(\phi)\frac{1}{256^3}}{P(\phi)\frac{1}{256^3} + P(\beta)G_{\mathbf{x}}(\mathbf{I}(\mathbf{x}))}, \quad (6)$$

and $P(\beta|\mathbf{I}(\mathbf{x})) = 1 - P(\phi|\mathbf{I}(\mathbf{x}))$.

[1]Foreground and background priors depend on the application. However, approximate values can be easily obtained for each application by manually segmenting the foreground in some images, and averaging the number of segmented points over the total.

Thus, a pixel is classified into the foreground class using maximum a posteriori if $P(\phi|\mathbf{I}(\mathbf{x})) > \frac{1}{2}$ is satisfied. Alternatively, the following test can also be used

$$P(\phi)P(\mathbf{I}(\mathbf{x})|\phi) > P(\beta)P(\mathbf{I}(\mathbf{x})|\beta), \quad (7)$$

which is faster, since the denominator in (4) does not have to be computed.

Note that, in practice (see Fig. 1), this is very similar to the approach previously described[15] consisting in determining background when a pixel value falls within 2.5 standard deviations of the mean of the Gaussian.

Finally, the Gaussian model is adapted using (2), when the pixel is classified into the background. In our cooperative approach, the Gaussians will be updated only when the corresponding projected 3D-Shape, built using multi-camera information, has been classified as background.

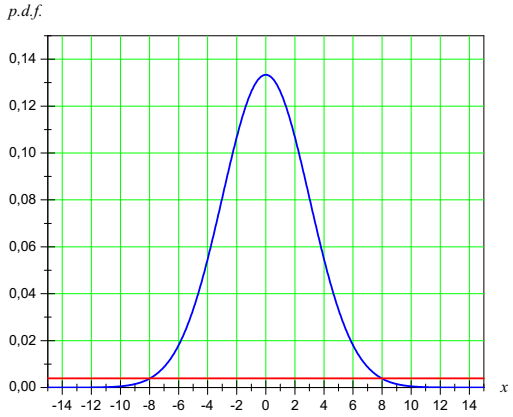


Figure 1: Probability density functions of a 1D-Gaussian, with $\sigma = 3$, and a uniform ($\frac{1}{256}$) distribution, assuming equiprobable $P(\phi) = P(\beta)$. Note that, for this value of σ, the point of cross of the distributions is very close to 2.5σ.

3. VOXEL-BASED SHAPE FROM SILHOUETTE

In order to present the proposed cooperative classification method, first we need to introduce the voxel-based SfS approach, which will be later extended to a probabilistic framework.

For each camera view, a fore/background classification is done first in all the pixels as described in section 2. In the literature, the resulting binary image is often referred as the silhouette of a view. Silhouettes are used in the voxel-based SfS algorithm as follows.

Given a bounded volume, we can divide it into voxels. Voxels are then projected into all the camera views to test (using a Projection Test) if all the projections are contained within a silhouette. Note that in this approach 3D classification is just an outcome of the 2D segmentation process.

For the sake of simplicity we consider a simple Projection Test which is passed if the pixel corresponding to the projection of the center of the voxel belongs to a silhouette. Other more robust tests can be found in [7].

The voxel-based SfS approach for any Projection Test is detailed in Algorithm 1, where MINC is used to refer to the minimum number of visual cones intersection required to reconstruct the voxel.

Algorithm 1 Voxel-based SfS algorithm

Require: Silhouettes: $S(c)$, a Projection Test Function: $PT_c(voxel, Silhouette)$

for all *voxel* **do**
  $in \leftarrow 0$
  $voxel \leftarrow$ Background
  for all cameras $(c = 1, \cdots, C)$ **do**
    if $PT_c(voxel, S(c))$ is *passed* **then**
      $in \leftarrow in + 1$
    end if
  end for
  if $in \geq$ MINC **then**
    $voxel \leftarrow$ Foreground
  end if
end for

In conventional voxel-based SfS, MINC $= C$ (see Algorithm 1). However, some practitioners [14, 9] prefer setting MINC $= C - P$, where P is the number of acceptable 2D-misses among the set C of cameras. Although single misses do not block the reconstruction in this approach, the resulting 3D-Shape is larger than the real Visual Hull for requiring fewer visual cones intersections.

4. COOPERATIVE BACKGROUND MODELLING

In this section, we propose a Bayesian method for classifying the voxels making use of the 2D-fore/background models of all the views. As will be shown further on, in this new approach, pixel models cooperate to classify the voxelized scene. In compensation, they benefit from the 3D-classification since they are only updated when the foreground voxels do not project over them.

We adapt the notation in the rest of the text so that $\mathbf{x}, \mathbf{I}$ and $G_{\mathbf{x}}(\mathbf{I}(\mathbf{x}))$ can be referred to each one of the C views: $\mathbf{x_i}, \mathbf{I}_i$ and $G_{i,\mathbf{x_i}}(\mathbf{I}_i(\mathbf{x_i}))$. In the following, we first consider error-free 2D-models, and later in the section, we adapt an outlier model to the 2D-models.

4.1 Probabilistic voxel classification

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 $\mathbf{v}$ 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 [6].

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

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

And prior probabilities are:

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

where a detected voxel, *i.e.* a voxel of the 3D-Shape, belongs to super class Γ_0, with P_S prior probability[2]. 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:

$$
P(\Gamma_j|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C})) = \max(P(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C}))) \tag{8}
$$

Assuming here and in the rest of the text that the super classes are conditionally independent, and using the Bayes theorem:

$$
P(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}), \cdots, \mathbf{I}_C(\mathbf{x_C})) = \frac{P(\Gamma_k)\prod_{i=1}^{C} p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_k)}{p(\mathbf{I}_1(\mathbf{x_1}))\cdots p(\mathbf{I}_C(\mathbf{x_C}))}, \tag{9}
$$

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

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

Substituting (9) into (8) we finally obtain the decision rule

Choose Γ_j *if:*

$$
P(\Gamma_j)\prod_{i=1}^{C} p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_j[i]) = \max P(\Gamma_k)\prod_{i=1}^{C} p(\mathbf{I}_i(\mathbf{x_i})|\Gamma_k[i]) \tag{10}
$$

Or in terms of a posteriori probabilities

Choose Γ_j *if:*

$$
P(\Gamma_j)\prod_{i=1}^{C} \frac{P(\Gamma_j[i]|\mathbf{I}_i(\mathbf{x_i}))}{P(\Gamma_j[i])} = \max P(\Gamma_k)\prod_{i=1}^{C} \frac{P(\Gamma_k[i]|\mathbf{I}_i(\mathbf{x_i}))}{P(\Gamma_k[i])} \tag{11}
$$

which is equivalent to

Choose Γ_j *if:*

$$
P(\Gamma_j)^{1-C}\prod_{i=1}^{C} P(\Gamma_j|\mathbf{I}_i(\mathbf{x_i})) = \max P(\Gamma_k)^{1-C}\prod_{i=1}^{C} P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})), \tag{12}
$$

[2] The prior probability of detecting a foreground voxel can be simply obtained by computing the detected voxel / total voxel occupancy ratio using conventional SfS, for instance. $P(\phi)$ and $P(\beta)$ are obtained from P_S: $P(\phi) = \sqrt[C]{P_S}$ and $P(\beta) = 1 - P(\phi)$.

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

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

where $P(\mathbf{I}_2(\mathbf{x_2}))$ can be computed using (5).

Both (10) and (12) decide the most probable super class. However (10) can be used to obtain faster classification, even though the probabilities are not explicitly computed.

4.2 Discussion

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. Miss-classifications 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.

In order to prevent such type of errors, we can force the classifiers not to deviate from the prior probabilities. This can be done with two different interpretations of the problem: (1) considering an outlier model in the 2D-models [12]; and (2) assuming that $P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) = P(\Gamma_k)(1+\delta_{ki})$ [8]. Both interpretations are discussed in the following.

4.3 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

$$P'(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) = eP(\Gamma_k) + (1-e)P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) \quad (13)$$

and then,

$$P'(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = \prod_{i=1}^{C}(eP(\Gamma_k) + (1-e)P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))) \quad (14)$$

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

$$P'(\Gamma_k|\mathbf{I}_1(\mathbf{x_1}),\cdots,\mathbf{I}_C(\mathbf{x_C})) = (eP(\Gamma_k))^C + + (eP(\Gamma_k))^{C-1}(1-e)\sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) + O((1-e)^2) \quad (15)$$

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

Choose Γ_j *if:*

$$(eP(\Gamma_j))^C + (eP(\Gamma_j))^{C-1}(1-e)\sum_{i=1}^{C}P(\Gamma_j|\mathbf{I}_i(\mathbf{x_i})) = \max\left((eP(\Gamma_k))^C + (eP(\Gamma_k))^{C-1}(1-e)\sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))\right) \quad (16)$$

4.4 Probabilistic voxel classification considering non-deviated posteriors

A similar result to (16) can be obtained expressing a posteriori probabilities as

$$P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})) = P(\Gamma_k)(1+\delta_{ki}), \quad (17)$$

where $\delta_{ki} \ll 1$. This expression assumes that the a posteriori probabilities computed by the respective classifiers will not deviate dramatically from the prior probabilities [8].

Substituting (17) into (12), and neglecting terms of second and higher order we obtain

Choose Γ_j *if:*

$$(1-C)P(\Gamma_j) + \sum_{i=1}^{C}P(\Gamma_j|\mathbf{I}_i(\mathbf{x_i})) = \max\left((1-C)P(\Gamma_k) + \sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i}))\right) \quad (18)$$

Note that both interpretations described in (16) and (18) convert the product $(\prod_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})))$ in (12) into a sum $(\sum_{i=1}^{C}P(\Gamma_k|\mathbf{I}_i(\mathbf{x_i})))$. Interestingly, this is the probabilistic justification of the previously described approach taken by practitioners in voxel-based SfS, consisting in letting a voxel reconstruction with only a partial sum of $C-P$ foreground projections, instead of requiring total intersection.

4.5 Implementation issues

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 $P(\Gamma_0|\mathbf{I}_i(\mathbf{x_i}), i = 1,\cdots,C)$ and set a threshold on the probability of the 3D-Shape. The probability of the 3D-Shape $(\mathrm{P}(\Gamma_0))$ can be obtained using (9) when working with reliable 2D-models, or with (14) 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. 2(a) and (c)). 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.

However, it has to be remarked that the most reliable classification, with a Bayesian justification, is done using (10), when considering error-free 2D-models and (16) or (18), 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.

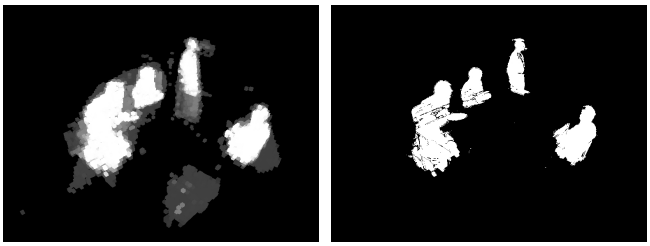
Once the voxels have been classified with any of the previously discussed procedures, the resulting foreground voxels are projected to all the views, and Gaussians are adapted using (2) in all those pixels which do not belong to the projected 3D-Shape.

5. RESULTS

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 artefacts, 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. 2 shows an example in a certain view and instant. 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 (14), 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.

(a) (b)



(c) (d)

Figure 2: The original image is show in (a). Picture (b), shows the foreground segmentation using conventional classification as explained in section 2. 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.

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 improvement of the models and silhouettes will be very important when combining them with other processing modules which make use of the segmented foreground as their input, such as face detectors or 2D gesture analysis modules.

6. CONCLUSION AND FUTURE WORK

In this paper, 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. Some of the future works include adapting the presented Bayesian framework to other SfS approaches[4, 16] which are able to detect some 2D-classification errors based on the geometric constraints of Visual Hull (testing correspondences of the frontier points).

Acknowledgments

This material is based upon work partially supported by the E.U. under the NoE MUSCLE FP6-507752 and by the Spanish Ministry of Ed. under action ACERCA TEC2004-01914.

REFERENCES

[1] B. G. Baumgart. *Geometric modeling for computer vision.* PhD thesis, 1974. 1

[2] C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Fusion of multiple viewpoint information towards 3d face robust orientation detection. In *IEEE International Conference on Image Processing (ICIP)*, Genoa, Italy, 2005. 1

[3] C. Canton-Ferrer, J. R. Casas, M. Tekalp, and M. Pardàs. Projective Kalman Filter: Multiocular Tracking of 3D Locations Towards Scene Understanding. In *Multimodal Interaction and Related Machine Learning Algorithms (MLMI05)*, Edinburgh, UK, 2005. 1

[4] K. Forbes, A. Voigt, and N. Bodika. Using silhouette consistency constraints to build 3d models. In *Proceedings of the Fourteenth Annual South African Workshop on Pattern Recognition, 2003.* PRASA, 2003. 5

[5] N. Friedman and S. J. Russell. Image segmentation in video sequences: A probabilistic approach. In *UAI*, pages 175–181, 1997. 1

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521540518, second edition, 2004. 3

[7] J.-Y. Bouguet K. M. Cheung, T. Kanade and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, volume 2, pages 714 – 720, June 2000. 1, 2

[8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239, 1998. 4

[9] J.L. Landabaso and M. Pardàs. Foreground regions extraction and characterization towards real-time object tracking. In *Proceedings of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI '05)*, 2005. 1, 3

[10] A. Laurentini. The visual hull: A new tool for contour-based image understanding. *Proc. Seventh Scandinavian Comperence on Image Processing*, pages 993–1002, 1991. 1

[11] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994. 1

[12] T. Minka. The 'summation hack' as an outlier model, 2003. Available from: http://www.stat.cmu.edu/minka/papers/minka-summation.pdf. 4

[13] W. P. Power and J. A. Schoonees. Understanding background mixture models for foreground segmentation. *Image and Vision Computing New Zealand*, 2002. 2

[14] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, pages 345–353, 2000. 3

[15] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22(8):747–757, 2000. 1, 2

[16] K.-Y. K. Wong. *Structure and Motion from Silhouettes.* PhD thesis, Department of Engineering, University of Cambridge, 2001. 5

[17] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In *FG*, pages 51–59, 1996. 1

[18] Z. Zhang. A flexible new technique for camera calibration. technical report msr-tr-98-71, microsoft research, 1998. 1