

# Optimal Camera Selection in Vision Networks through Shape Approximation

Marleen Morbee <sup>#1</sup>, Linda Tessens <sup>#1</sup>, Huang Lee <sup>\*2</sup>, Wilfried Philips <sup>#1</sup>, Hamid Aghajan <sup>\*2</sup>

<sup>#</sup> *TELIN-IP1-IBBT, Ghent University*

*Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium*

<sup>1</sup>{marleen.morbee, linda.tessens, philips}@telin.UGent.be

<sup>\*</sup> *Wireless Sensor Network Lab, Department of Electrical Engineering, Stanford University*

*350 Serra Mall, David Packard room 318, Stanford, CA 94305, USA*

<sup>2</sup>{huanglee, aghajan}@stanford.edu

**Abstract**—Within a camera network, the contribution of a camera to the observation of a scene depends on its viewpoint and on the scene configuration. This is a dynamic property, as the scene content is subject to change over time. An automatic selection of a subset of cameras that significantly contributes to the desired observation of a scene can be of great value for the reduction of the amount of transmitted or stored image data.

In this work, we propose low data rate schemes to select from a vision network a subset of cameras that allows for an efficient observation of the persons present in a scene. We also experimentally investigate to what degree computational efficiency and low data rates trade off quality of reconstructed 3D shapes.

## I. INTRODUCTION

In many applications, the deployment of a camera network provides substantial advantages over a single fixed viewpoint camera. E.g. in scene monitoring, camera networks can alleviate occlusion problems ; in gesture recognition, cues coming from different viewpoints can lead to a more robust decision.

Camera networks increase not only the amount of data available for further storage, observation or processing, but also the redundancy within this data. It is therefore beneficial, and from a practical point of view often necessary, to have a system that can fully exploit the additional information available in the network, while simultaneously keeping the redundancy under control. A possible way of achieving this is by selecting a limited number of cameras or views from the network and transmitting and/or storing only these. Thus the amount of data that requires transmission, further processing, storage or observation is greatly reduced and resources are saved. The recent introduction of “smart cameras” with on-board image processing and communication hardware allows for a distributed implementation of such a selection algorithm, hereby reducing the required communication bandwidth and spreading the computational burden. This is beneficial to the scalability of the system and can even allow the cameras to work wirelessly.

Viewpoint selection has been extensively studied in the context of computer graphics and robot navigation [1]–[3]. These systems require an accurate model of the observed shape(s). So called next best view systems [4], [5] deal with the problem of selecting the best view of a scene given some

already selected view(s). All these systems have difficulties coping with the background present in natural scenes as they were all designed for artificial circumstances.

A topic very much related to viewpoint selection is camera allocation within vision networks. Chu *et al.* and Zhao *et al.* studied this problem in a more general context of sensor networks [6], [7]. However, devising the necessary sensor and target models is far from straightforward in the case of camera networks. In [8], [9], cameras within a network are tasked in order to minimize the number of active cameras [8] while determining the occupied space in the scene [9].

In this work, our interest lies in the observation of persons in a 3D scene. We propose two low data rate algorithms that select a limited number of cameras from a network in a content adaptive manner, such that this subset constitutes the most complete view of the scene possible for the given number of selected cameras. We compare the performance of these methods to their theoretically optimal counterpart, which requires high data rates. The viability of the methods is proven on extensive video data from natural scenes obtained with a network of ten cameras.

The remainder of this paper is organized as follows. In Section II, we elaborate on the setup of the system for which we devise our methods and on the assumptions we make. In Section III, we describe the proposed algorithms in detail. The performance of these methods is discussed in Section IV and conclusions are presented in the last section.

## II. SYSTEM SETUP AND NOTATIONS

A scheme of the system setup is depicted in the bottom-right corner of Figure 4. The system consists of multiple smart camera sensors that observe a room with one or more persons inside. The camera sensors are battery powered and can communicate with each other through wireless channels. The cameras’ positions and orientations are fixed and calibrated. A base station is deployed to receive the observations from all camera sensors and is responsible for coordinating all sensors in the network. We assume that there are  $N$  camera sensors in the network. They are denoted by  $C_i$  for  $i = 1, \dots, N$ . The complete collection of cameras is the set  $\mathbf{C} = \{C_1, \dots, C_N\}$  where  $|\mathbf{C}| = N$ .

Our goal is to select a subset of cameras denoted as  $\mathbf{S} \subseteq \mathbf{C}$  where  $|\mathbf{S}| = n \leq N$ , and only the selected cameras will send their image data to the base station. These images should comprise the most complete view of the scene possible for the given number of selected cameras. The subset  $\mathbf{S}$  contains two types of cameras:

- The key camera: the camera with the view that contributes most to the desired observation of a scene at a certain time instant. The key camera is indicated by  $K$ .
- The helper cameras: cameras with views that complement the selected key view, such that the total selected view subset constitutes a significantly more efficient scene representation than the totality of the available views. The  $n - 1$  helper cameras are indicated by  $W_k$  where  $k = 1, \dots, n - 1$ .

The remaining  $N - n$  cameras do not send any image data. The image captured by the  $i$ -th camera at a certain time instant  $t$  is denoted by  $\mathbf{X}_i(t)$ . Note that  $\mathbf{S} = \{K\} \cup \{W_1, \dots, W_{n-1}\}$ .

### III. ALGORITHM

In this section, we will discuss the details of the key camera and helper camera selection algorithms.

#### A. Key Camera Selection

To assign the role of key camera  $K$  in the camera network, we run the following algorithms on each of the smart cameras  $C_i$  in the network (see Figure 1). In a first step, we segment the foreground (FG)  $\mathbf{F}_i(t)$  and the background (BG)  $\mathbf{B}_i(t)$  of the frames  $\mathbf{X}_i(t)$  using the method of [10]. Then, we detect the frontal faces in the FG regions of the frame with the object detector that was initially proposed by Viola *et al* [11] and then improved by Lienhart *et al* [12]. At each time instant  $t$ , the face detector returns the following values:  $f_i(t)$  and  $Q_i^l(t)$  ( $l = 1, \dots, f_i(t)$ ).  $f_i(t)$  is the number of faces detected in the frame  $\mathbf{X}_i(t)$ .  $Q_i^l(t)$  is a measure of the quality of the  $l^{\text{th}}$  detected face. The lower this measure, the less certain the detection. In our implementation, we assume that the number of windows that have passed all classification stages and that constitute a detected face is such a measure.

To deal with spurious face detections and to obtain smoothness over time, the decision on the key camera for time instant  $t$  not only depends on the current face detection output, but also on the previous observations. For each camera  $C_i$ , this temporal filtering is implemented as an exponentially weighted moving average (for  $t \geq 2$ ):

$$Q_i^s(t) = \alpha \sum_{l=1}^{f_i(t)} Q_i^l(t) + (1 - \alpha) \sum_{l=1}^{f_i(t-1)} Q_i^l(t-1) \quad (1)$$

where  $Q_i^s(t)$  is the smoothed face detection output of camera  $C_i$  at time instant  $t$  and  $\alpha$  is a constant between 0 and 1 that determines the importance of previous observations. Then, the key camera at time instants  $t \geq 2$  is

$$K(t) = \underset{C_i}{\operatorname{argmax}} Q_i^s(t) \quad (2)$$

In the remainder of this paper, we will leave out the time variable  $t$ , in order not to overload the notations.

#### B. Helper Camera Selection

In this section, we present three approaches to select among the remaining  $N - 1$  cameras those  $n - 1$  helper cameras  $\{W_1, \dots, W_{n-1}\}$  that add most information to the image data coming from the already selected key camera.

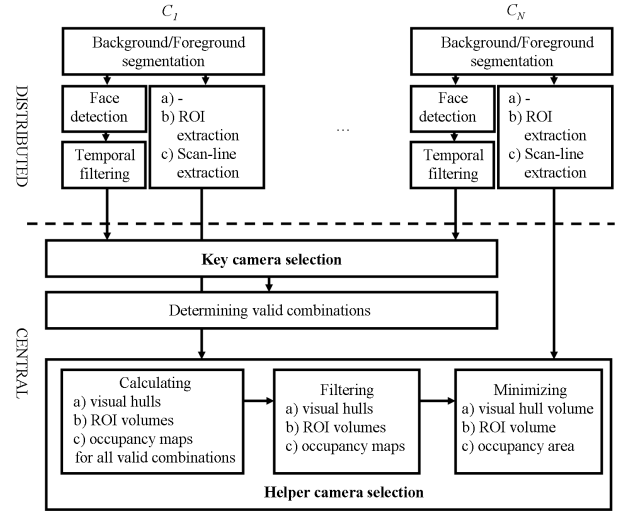


Fig. 1. Block diagram for camera selection based on: a) minimal visual hull volume criterion, b) minimal ROI volume criterion, c) minimal occupancy area criterion.

1) *Minimal Visual Hull Volume Criterion*: This approach is depicted in Figure 1 (track a). All the smart cameras  $C_i$  ( $i = 1, \dots, N$ ) send, without further processing, their segmented FG silhouettes  $\mathbf{F}_i$  obtained with the BG/FG segmentation presented in Section III-A to the central base station. Then, at the base station we determine all the valid candidate subsets  $\hat{\mathbf{S}} \subseteq \mathbf{C}$ , for which  $|\hat{\mathbf{S}}| = n$  and  $K \in \hat{\mathbf{S}}$ , with  $K$  determined as in Equation (2). For all these candidate subsets, we reconstruct the visual hull  $\mathbf{H}_{\hat{\mathbf{S}}}^{\text{hull}}$  from the silhouette images  $\mathbf{F}_i$  (see Fig. 2a) using the shape-from-silhouette technique [13]. More precisely, within a cuboid-shaped volume  $V^3$  in the 3D space of the observed room

$$V^3 = [X_1, X_2) \times [Y_1, Y_2) \times [Z_1, Z_2) \subset \mathbb{N}^3, \quad (3)$$

$\mathbf{H}_{\hat{\mathbf{S}}}^{\text{hull}}(\mathbf{j})$ , with  $\mathbf{j} \in V^3$ , assumes value 0 when the voxel  $\mathbf{j}$  is observed as empty by at least one of the cameras from the set  $\hat{\mathbf{S}}$ . This is the case when it is part of the reprojected BG region from at least one of the cameras of the subset. All other voxels have value 1.

The resulting visual hulls can contain *shadow* volumes. These are parts of the visual hull that do not represent real objects but result from an insufficient number of used cameras  $n$ . To remove these, we filter  $\mathbf{H}_{\hat{\mathbf{S}}}^{\text{hull}}$  by multiplying its voxels with the corresponding voxels of a visual hull  $\mathbf{H}_{\mathbf{C}}^{\text{hull, filt}}$  which we obtain in the following way. From the ideal visual hull  $\mathbf{H}_{\mathbf{C}}^{\text{hull}}$ , reconstructed from the complete set of cameras  $\mathbf{C}$ , we extract sections parallel to the ground plane and we perform a morphological dilation operation on each section. In this way, we ensure that we base our camera selection only on the reconstructed shapes of objects that are also detected when all

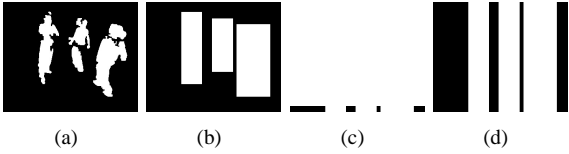


Fig. 2. Used silhouette images: (a) BG/FG segmentation ( $\mathbf{F}_i$  and  $\mathbf{B}_i$ ), (b) ROI extraction ( $\mathbf{F}_{i,ROI}$  and  $\mathbf{B}_{i,ROI}$ ), (c) scan-line (d) column-wise extended scan-line ( $\mathbf{F}_{i,sc}$  and  $\mathbf{B}_{i,sc}$ )

cameras  $N$  are active and the influence of shadow volumes is minimized.

We consider the hulls  $\mathbf{H}_{\hat{S}}^{\text{hull}}$  to be approximations of the actual shape of the objects present in the scene. As we wish to select the set of cameras with the most complete view on the scene, we choose the ones that allow for the best approximation. That is why in a final step, we select from all candidate subsets the one that yields the visual hull with the smallest volume (or in other words, the least voxels with value 1) within  $\mathbf{H}_{\mathbf{C}}^{\text{hull,flt}}$ :

$$\mathbf{S}_n^{\text{hull}} = \underset{\forall \hat{S}}{\operatorname{argmin}} \sum_{\forall \mathbf{j} \in V^3} \mathbf{H}_{\hat{S}}^{\text{hull}}(\mathbf{j}) \mathbf{H}_{\mathbf{C}}^{\text{hull,flt}}(\mathbf{j}). \quad (4)$$

This approach provides the best performance but requires large data transmissions (entire silhouettes) between sensor nodes and base station. Therefore, the performance of this approach is used only as a baseline for comparison with the two other proposed methods of Section III-B1 and III-B2.

2) *Minimal ROI Volume Criterion*: This approach is depicted in Figure 1 (track b) ) and has similarities with approach III-B1. The difference is that now at each smart camera, we extract the bounding boxes of the segmented FG silhouettes obtained with the BG/FG segmentation presented in Section III-A. We will call these bounding boxes the Regions of Interest (ROIs). They are an approximated version of  $\mathbf{F}_i$ , and therefore we will denote them by  $\mathbf{F}_{i,ROI}$ . The corresponding BG region is denoted by  $\mathbf{B}_{i,ROI}$ . Each camera sends the upper left and bottom right vertices of the ROIs to the base station. Then, as in method III-B1, we determine at the base station all the valid candidate subsets  $\hat{S}$ , for which  $|\hat{S}| = n$  and  $K \in \hat{S}$ , with  $K$  determined as in Equation (2). For all these candidate subsets, we reconstruct the 3D ROI volume  $\mathbf{H}_{\hat{S}}^{\text{ROI}}$  in 3D space from the approximate silhouette images  $\mathbf{F}_{i,ROI}$  (see Fig. 2b).

Again, as explained in Section III-B1, this ROI volume needs to be filtered to remove *shadow* parts. The filtering visual hull  $\mathbf{H}_{\hat{S}}^{\text{ROI,flt}}$  is a dilated version of the ideal ROI volume  $\mathbf{H}_{\hat{S}}^{\text{ROI}}$ , reconstructed from the complete set of cameras  $\mathbf{C}$ . The 3D ROI volume  $\mathbf{H}_{\hat{S}}^{\text{ROI}}$  are again approximations of the actual shape of the objects present in the scene, albeit cruder than the approximations in Section III-B1. To select the set of cameras with the most complete view on the scene, we choose the subset that yields the minimal reconstructed ROI volume within  $\mathbf{H}_{\mathbf{C}}^{\text{ROI,flt}}$ :

$$\mathbf{S}_n^{\text{ROI}} = \underset{\forall \hat{S}}{\operatorname{argmin}} \sum_{\forall \mathbf{j} \in V^3} \mathbf{H}_{\hat{S}}^{\text{ROI}}(\mathbf{j}) \mathbf{H}_{\mathbf{C}}^{\text{ROI,flt}}(\mathbf{j}). \quad (5)$$

3) *Minimal Occupancy Area Criterion*: This approach is depicted in Figure 1 (track c) ). In this method, we extract

at each camera the scan-line from the silhouettes in the scene. This scan-line is the projection of the 2D-foreground silhouettes  $\mathbf{F}_i$  to a 1D-line (see Figure 2c). All the cameras send their (run-length coded) scan-lines to the base station. At the base station, a list of all possible candidate subsets  $\hat{S}$  is made based on the key camera selection (as for method III-B1 and III-B2). Then, for each camera in a particular candidate subset, we column-wise extend the scan-lines to a 2D image, such that we get a rough approximation of the original backgrounds extracted at the sensor nodes (see Figure 2d). This approximation will be denoted by  $\mathbf{B}_{i,sc}$ . From all the  $\mathbf{B}_{i,sc}$  for the different cameras  $C_i \in \hat{S}$ , we then calculate an occupancy map  $\mathbf{O}_{\hat{S}}$ . This is a 2D raster image, uniformly distributed in a plane  $P^2$  horizontal to the ground floor of our observed 3D scene (or the defined voxel volume  $V^3$ ):

$$P^2 = [X_1, X_2] \times [Y_1, Y_2] \subset \mathbb{N}^2 \quad \text{and} \quad Z = c. \quad (6)$$

It is obtained by intersecting the visual hull reconstructed from the silhouette images  $\mathbf{B}_{i,sc}$  with the plane  $Z = c$  [14].

As in Sections III-B1 and III-B2, this occupancy map is then filtered to remove *shadow* areas. The filtering occupancy map is a dilated version  $\mathbf{O}_{\hat{S}}^{\text{flt}}$  of the ideal occupancy map  $\mathbf{O}_{\mathbf{C}}$ , reconstructed from the complete set of cameras  $\mathbf{C}$ . Also here, we consider that the occupancy maps  $\mathbf{O}_{\hat{S}}$  are (very crude) shape approximations of the objects in the scene. The subset that yields the minimal occupied area is assumed to provide the most complete view on the scene. More specifically,

$$\mathbf{S}_n^{\text{area}} = \underset{\forall \hat{S}}{\operatorname{argmin}} \sum_{\forall \mathbf{j} \in P^2} \mathbf{O}_{\hat{S}}(\mathbf{j}) \mathbf{O}_{\mathbf{C}}^{\text{flt}}(\mathbf{j}). \quad (7)$$

#### IV. PERFORMANCE EVALUATION

As mentioned previously, we evaluate the performance of the proposed helper camera selection algorithms of Section III-B2 and III-B3 by comparing them to the benchmark method of Section III-B1. We consider two aspects: accuracy and communication overhead. Each aspect is discussed in the following subsections.

Experimental data to test the methods on was recorded using  $N = 10$  cameras. Five of these were Logitech QuickCam Pro 5000 cameras and the five others Logitech QuickCam Sphere MP. The cameras were calibrated using the method for multi-camera self calibration of [15]. Sequences were recorded at 5 frames per second and at a resolution of  $352 \times 288$ . Only the starting points of the recordings were synchronized. The parameters of the BG/FG segmentation and the face detection are summarized in Table I. We allowed the BG/FG segmentation algorithm to build up its BG model during 30 frames at the start of each sequence. These first 30 frames of each sequence are not considered in the experiments in this section. The weighting factor of the temporal filtering  $\alpha$  was set to 0.05. The structuring element for the dilation to obtain the filters  $\mathbf{H}_{\mathbf{C}}^{\text{hull,flt}}$ ,  $\mathbf{H}_{\mathbf{C}}^{\text{ROI,flt}}$ ,  $\mathbf{O}_{\mathbf{C}}^{\text{flt}}$  (see Section III-B) and  $\mathbf{H}_{\mathbf{C}}^{\text{flt}}$  (see later, in Section IV-A) is a square of  $3 \times 3$  with the origin at its center and the dilation is performed in five iterations. The voxel volume  $V^3$  was  $[0, 200] \times [0, 100] \times [0, 50] \subset \mathbb{N}^3$ , where

each voxel is a cube with edges of  $0.04m$ . The plane  $P^2$  is the plane in the voxel volume at  $Z = 1.29m$ . We discuss results for subsets of both 3 and 6 cameras.

TABLE I  
PARAMETERS OF THE BG/FG SEGMENTATION AND THE FACE DETECTION.

Parameters BG/FG Segmentation			
$L$ (color comp.)	128	$L$ (color co-occ.)	64
$N_1$ (color comp.)	15	$N_1$ (color co-occ.)	25
$N_2$ (color comp.)	25	$N_2$ (color co-occ.)	40
$\alpha_1$	0.1	$\alpha_2$	0.005
$\alpha_3$	0.1	$T$	0.9
$\delta$	2	MINAREA	15.0
UPDATE_TRESH	0.5		
Parameters Face Detection			
scale factor		1.10	
min. number (-) of neighbors		2	
min. window size		$5 \times 5$	
classifier training window size		$20 \times 20$	

### A. Accuracy Evaluation

In our experiments we distinguish between four scenarios, depending on the number of people in the scene. The higher this number, the harder the selection problem.

We will refer to the benchmark visual hull, reconstructed from the whole set  $\mathbf{C}$  of ten available cameras, by  $\mathbf{H}_C$ . The dilated version of  $\mathbf{H}_C$  is indicated by  $\mathbf{H}_C^{\text{filt}}$  and will be used later on in this paragraph.

To evaluate the accuracy of the proposed selection methods, we calculate the visual hull reconstructed from the *foreground silhouettes*  $\mathbf{F}_i$  of the selected camera subset  $\mathbf{S}_n^{\text{method}}$ , where  $\text{method} \in \{\text{hull, ROI, area}\}$  and  $\mathbf{S}_n^{\text{method}}$  as in Equations (4), (5) and (7). We will denote these hulls by  $\mathbf{H}_{\mathbf{S}_n^{\text{method}}}^{\text{hull}}$ . Note that for the minimal ROI volume and minimal occupancy area methods, these FG silhouettes  $\mathbf{F}_i$  are *not* available in the actual method, only their approximate versions  $\mathbf{F}_i^{\text{ROI}}$  and  $\mathbf{F}_i^{\text{area}}$ . From these hulls, we determine at each time instant the number of voxels  $d_n^{\text{method}}$  that are different between the hull reconstructed from the selected subset and the benchmark hull  $\mathbf{H}_C$ :

$$d_n^{\text{method}} = \sum_{\forall \mathbf{j} \in V^3} \left[ \left( \mathbf{H}_C^{\text{filt}}(\mathbf{j}) \mathbf{H}_{\mathbf{S}_n^{\text{method}}}^{\text{hull}}(\mathbf{j}) \right) - \mathbf{H}_C(\mathbf{j}) \right] \quad (8)$$

where  $\text{method} \in \{\text{hull, ROI, area}\}$ . Note that for the calculation of this difference we only take into account differences within  $\mathbf{H}_C^{\text{filt}}$ . This filtering is needed in order to focus on the interesting objects, without having the disturbing influence of *shadow* volumes (as described in Section III-B). At the same time, due to the dilation operation, we still consider the whole object as reconstructed by the subset. The amount of extra volume in the filtered reconstructed hull (or in other words  $d_n^{\text{method}}$ ) gives us an insight in how well the selected subset observes the objects in the scene from all sides. Indeed,

TABLE II  
NUMBER OF BITS REQUIRED

Criterion	Visual Hull	Visual Hull (Compressed)	ROI	Occupancy Area
Required bits	$l \times w$	$l \times w \times \rho$	4MB	2MB
5 objects	101376	4055	180	90
10 objects	101376	4055	360	180

the more voxels are observed as empty around the object of interest, the less redundant the views from the selected cameras are.

In Table III, we compare for the three methods (hull, ROI, area) the mean value of the number of different voxels  $d_n^{\text{method}}$  over all frames of the sequences with a certain scenario, both for  $n = 3$  and  $n = 6$ . The lower this number, the higher the quality of the observation with the selected camera subset. The number of frames available per scenario is indicated in the second column, and the average voxel volume of  $\mathbf{H}_C$  in the third column. The results confirm the expectations. First of all, we observe that the visual hull volume criterion yields better results than both the minimal ROI volume criterion and the minimal occupancy area method. However, at the same time we should note that the communication overhead for the visual hull volume criterion is much larger than for the other two methods, as will be discussed in Section IV-B. Also, we can see from Table III that the minimal ROI volume criterion performs slightly better than the minimal occupancy area method. The differences between the latter ones become, relatively speaking, smaller when the number of persons in the scene increases. This is due to the fact that when more persons walk around in the scene, silhouettes merge and the bounding boxes become less accurate and more similar to the column-wise extended scan-lines.

As an illustration, we show in Figure 3 the selection performance when selecting  $n = 3$  cameras from 10. For a representative sequence of each scenario we plot per frame the volume (in number of voxels) of the visual hull from all the candidate subsets  $\hat{\mathbf{S}}$  contained within  $\mathbf{H}_C^{\text{filt}}$  (green dotted lines). As a reference, for each frame the number of voxels of the benchmark visual hull  $\mathbf{H}_C$  is also indicated (solid magenta line). The number of voxels in the filtered hulls  $\mathbf{H}_C^{\text{filt}}, \mathbf{H}_{\mathbf{S}_n^{\text{hull}}}^{\text{hull}}, \mathbf{H}_C^{\text{filt}}, \mathbf{H}_{\mathbf{S}_n^{\text{ROI}}}^{\text{hull}}$  and  $\mathbf{H}_C^{\text{filt}}, \mathbf{H}_{\mathbf{S}_n^{\text{area}}}^{\text{hull}}$  per frame are drawn as the thicker lines (resp. solid blue, dash-dotted black with round markers and dashed red). This graph confirms that the minimal visual hull volume method selects, for a given key camera and subset size, the best possible subset of cameras for visual hull reconstruction. Indeed the curve corresponding to this method is the lower envelope of the curves of the valid subsets. The curves of the minimal ROI volume and occupied area methods are mostly very near to the one from the ideal method, although occasionally important differences occur. In these cases, the curves still follow the trend of the ideal method, and as expected, on average the ROI volume method deviates less from the ideal than the occupancy area method.

Figure 4 shows a visual example of the selection of  $n = 6$  cameras from 10 for a 3-persons scene. We display the views of all the cameras  $C_1, \dots, C_{10}$ . To give an insight into the system setup, we depicted in the bottom-right corner a top view of the scene, which indicates the relative positions of the ten cameras and the three persons in the scene. The selected key camera is  $C_1$  and is marked by a magenta bounding box. The detected face is indicated by a red circle. Due to this current face detection and previous face detections, this camera

TABLE III

MEAN VOXEL DIFFERENCE  $d$  (EQUATION (8)) FOR THE THREE SUBSET SELECTION METHODS (SECTION III-B1, III-B2 AND III-B3) FOR FOUR DIFFERENT SCENARIOS. IN THE SECOND COLUMN WE INDICATE THE TOTAL NUMBER OF FRAMES OVER WHICH THE AVERAGE IS CALCULATED. THE AVERAGE VOXEL VOLUME OF  $\mathbf{H}_C$  IS SHOWN IN THE THIRD COLUMN. COLUMNS 3-5 ARE THE RESULTS FOR  $n = 3$  AND COLUMNS 6-8 FOR  $n = 6$ .

Scenario	# frames	$\sum_{\forall \mathbf{j} \in V^3} \mathbf{H}_C(\mathbf{j})$	$d_3^{\text{hull}}$	$d_3^{\text{ROI}}$	$d_3^{\text{area}}$	$d_6^{\text{hull}}$	$d_6^{\text{ROI}}$	$d_6^{\text{area}}$
1 persons	1629	615.61	583.68	960.97	1204.29	21.61	189.17	348.39
2 persons	2213	2450.99	1823.97	2976.72	3365.90	99.17	555.43	763.40
3 persons	826	4584.35	4335.62	7267.32	7349.20	292.88	1374.98	1326.16
4 persons	290	8079.53	6200.01	9156.85	10036.74	327.99	1230.80	1356.97

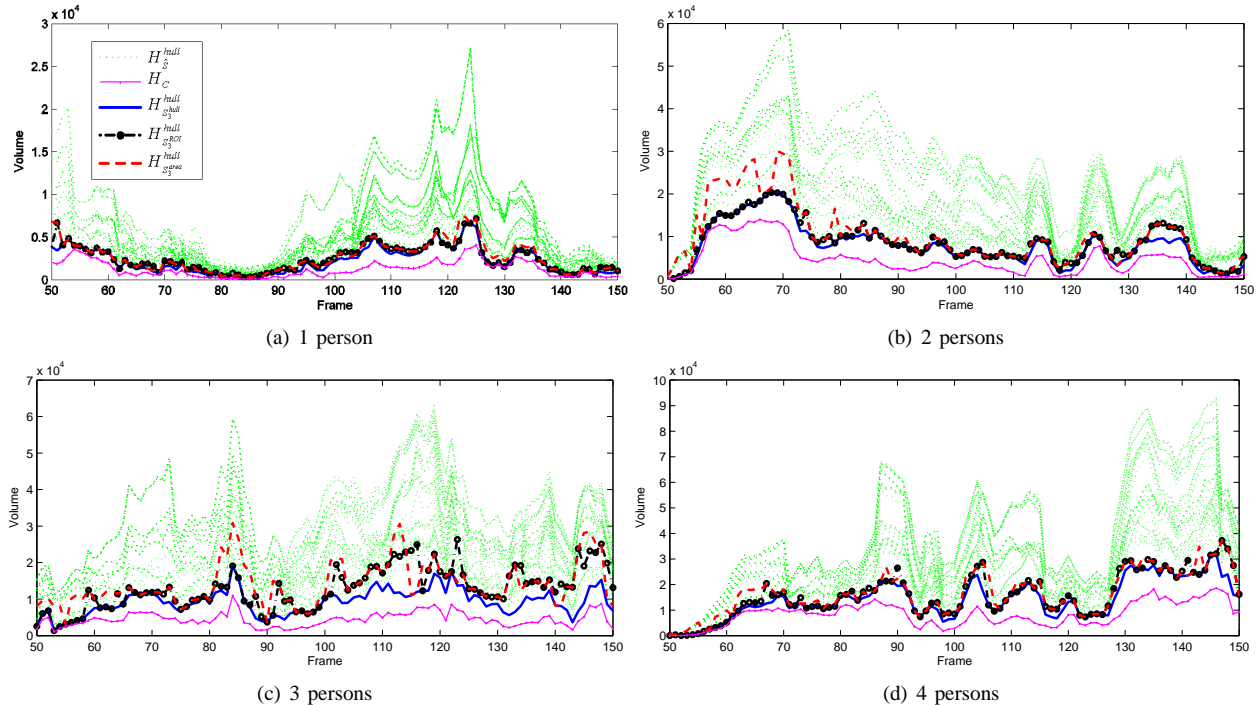


Fig. 3. Selection performance for 100 frames of a representative sequence of each scenario. The number of cameras in the subset is  $n = 3$ . For each frame, we plotted the volume (in number of voxels) of the visual hulls  $\mathbf{S}$  (green dotted lines),  $\mathbf{H}_C$  (solid magenta line),  $\mathbf{H}_C^{\text{hull}}$  (solid blue line),  $\mathbf{H}_C^{\text{ROI}}$  (dash-dotted black line with round markers) and  $\mathbf{H}_C^{\text{area}}$  (dashed red line).

was chosen to be the key camera (according to Equation (2)). The helper cameras are  $C_3$ ,  $C_5$ ,  $C_8$ ,  $C_9$  and  $C_{10}$ , and are marked by a cyan bounding box. We can observe from the displayed views that the selected subset gives us a complete view of the persons, and that the non-selected cameras add rather redundant information.

### B. Communication Overhead

In this section we show how many bits are required for communication for the three approaches of Section III-B. For the minimal visual hull method, each camera sensor sends a silhouette image to the base station. The number of bits required to represent a silhouette depends on the image size, which is  $l$  by  $w$ . Therefore the number of required bits is  $l \times w$ . This number can be further minimized by compression. We assume that a Lempel-Ziv-Welch code is used and the average compression rate  $\rho = 0.04$ . So, the required number of bits after compression is given by  $l \times w \times \rho$ .

For the minimal ROI volume method, we need to send positions of two corners on the diagonal of the ROI. Since the image length is  $l$ , we need at most  $M = \lceil \log_2 l \rceil$  bits to describe the position of a corner on one axis.

We also assume that there are at most  $B$  distinguishable objects on an image frame. Therefore, the number of required bits for transmission is  $4MB$ . As for the minimal occupancy area method, since the foreground is projected to one axis, the required amount of bits is half of the minimal ROI volume method and is given by  $2MB$ .

We will now give some numerical examples to compare the required bits among the three methods. For image size, we assume that  $l = 352$  by  $w = 288$ . So, we have  $M = 9$ .

For the minimal visual hull method, the number of bits is fixed and does not depend on the number of objects. As for the minimal ROI volume and the minimal occupancy area methods, we need the number of objects in the room. We assume that there are at most 10 objects, and give the maximum number of required bits. However, in a realistic situation objects may occlude each other, so the number of objects visible in each image frame is usually smaller than the number of objects in the room. The cases when  $B = 5$  and  $B = 10$  are listed in Table II. It can be observed that the proposed methods significantly decrease the communication overhead.

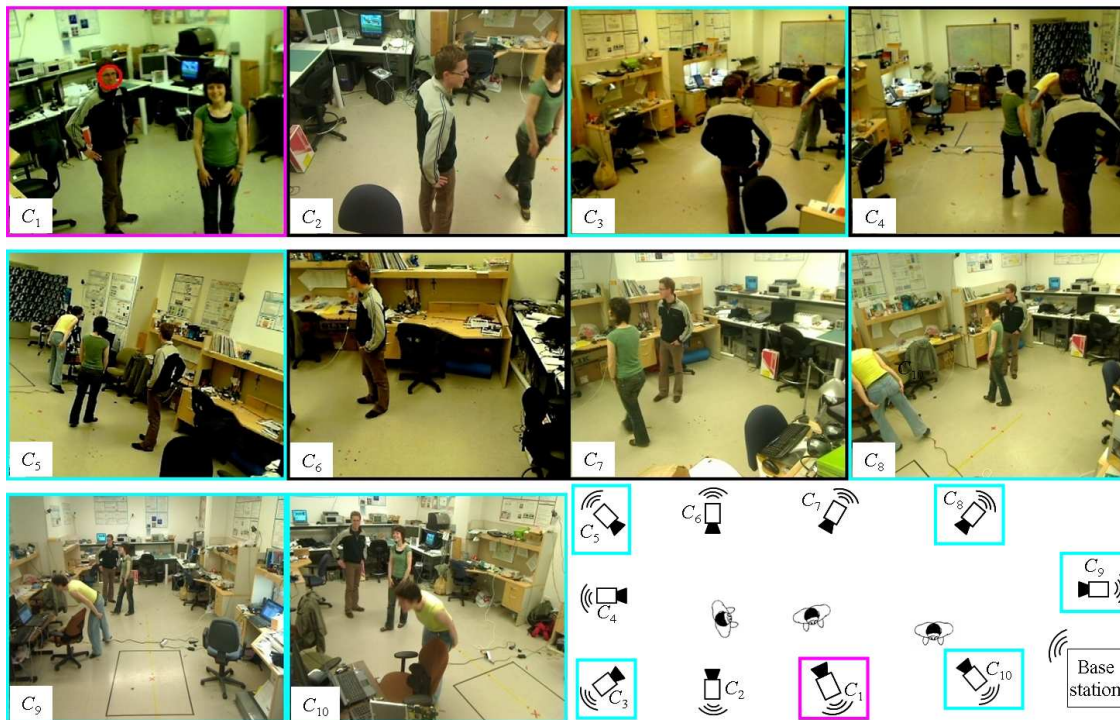


Fig. 4. Example of the selection of 6 out of 10 cameras for a 3-persons scene. The views of the 10 cameras ( $C_1, \dots, C_{10}$ ) are shown. In the bottom-right corner, we depicted a top view of the scene which shows its geometry and the positions of the cameras and persons. The selected key camera is  $C_1$  (marked by a magenta bounding box). The helper cameras are  $C_3, C_5, C_8, C_9$  and  $C_{10}$  (marked by a cyan bounding box).

## V. CONCLUSIONS

In this paper we have presented two methods for selecting within a camera network a subset of views that constitutes a significantly more efficient scene representation than the totality of the available views. These systems allow to drastically reduce the data output of camera networks, while preserving the richer information content they produce.

In a first, distributed processing phase of the algorithms, the smart cameras in the network each transmit a small amount of data to a central base station. In a second, central phase, the base station uses this information to approximate the shapes present in the scene. The subset that allows for the best shape approximation is then selected.

We have tested both methods on an extensive set of realistic sequences. From these we can conclude that the minimal ROI volume method selects subsets that constitute a slightly more complete view of the scene. However, the minimal occupied area method achieves nearly the same selection quality, while requiring less data for communication.

## REFERENCES

- [1] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich, "Viewpoint selection using viewpoint entropy," in *VMV '01: Proceedings of the Vision Modeling and Visualization Conference 2001*. Aka GmbH, 2001, pp. 273–280.
- [2] T. Kamada and S. Kawai, "Simple method for computing general position in displaying three-dimensional objects," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 1, pp. 43–56, 1988.
- [3] D. R. Roberts and A. D. Marshall, "Viewpoint selection for complete surface coverage of three dimensional objects," in *Proc. of the British Machine Vision Conference (BMVC)*, Southampton, England, 1998.
- [4] N. A. Massios and R. B. Fisher, "A best next view selection algorithm incorporating a quality criterion," in *Proc. of the British Machine Vision Conference (BMVC)*, Southampton, England, 1998.
- [5] L. Wong, C. Dumont, and M. Abidi, "Next best view system in a 3-d object modeling task," 1999. [Online]. Available: [citeseer.ist.psu.edu/wong99next.html](http://citeseer.ist.psu.edu/wong99next.html)
- [6] M. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks," *International Journal of High Performance Computing Applications*, vol. 16, no. 3, pp. 293–313, 2002.
- [7] F. Zhao, J. Shin, and J. Reich, "Information-driven dynamic sensor collaboration," *Signal Processing Magazine, IEEE*, vol. 19, no. 2, pp. 61–72, Mar 2002.
- [8] T. Matsui, H. Matsuo, and A. Iwata, "Dynamic camera allocation method based on constraint satisfaction and cooperative search," in *Proceedings of 2nd International Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, vol. 8, 2001, pp. 955–964.
- [9] D. Yang, J. Shin, A. O. Ercan, and L. Guibas, "Sensor tasking for occupancy reasoning in a camera network," in *Proc. of IEEE/ICST 1st Workshop on Broadband Advanced Sensor Networks*, 2004.
- [10] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*. New York, NY, USA: ACM, 2003, pp. 2–10.
- [11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," 2001. [Online]. Available: <http://citeseer.ist.psu.edu/663084.html>
- [12] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," vol. 1, 2002, pp. I-900–I-903 vol.1. [Online]. Available: <http://dx.doi.org/10.1109/ICIP.2002.1038171>
- [13] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 150–162, 1994.
- [14] A. Hoover and B. D. Olsen, "A real-time occupancy map from multiple video streams," in *International Conference on Robotics and Automation (ICRA)*, 1999, pp. 2261–2266.
- [15] T. Svoboda, D. Martinec, and T. Pajdla, "A convenient multi-camera self-calibration for virtual environments," *PRESENCE: Teleoperators and Virtual Environments*, vol. 14, no. 4, pp. 407–422, August 2005.