PRINCIPAL VIEW DETERMINATION FOR CAMERA SELECTION IN DISTRIBUTED SMART CAMERA NETWORKS

Linda Tessens [#1*], *Marleen Morbee* [#1], *Huang Lee* [*2], *Wilfried Philips* [#1] *and Hamid Aghajan* [*2]

[#] TELIN-IPI-IBBT
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
[1] {linda.tessens,marleen.morbee,philips}@telin.UGent.be

[*] *Wireless Sensor Networks Lab*
Department of Electrical Engineering, Stanford University
350 Serra Mall, David Packard room 318, Stanford, CA 94305, USA
[2] {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 and the camera configuration might not be fixed, e.g. in a mobile network. In this work, we address the problem of effectively determining the principle viewpoint within a network, i.e. the view that contributes most to the desired observation of a scene. This selection is based on the information from each camera's observations of persons in a scene, and only low data rate information is required to be sent over wireless channels since the image frames are first locally processed by each sensor node before transmission. The principal view, complemented with one or more helper views, constitutes a significantly more efficient scene representation than the totality of the available views. This is of great value for the reduction of the amount of image data that needs to be stored or transmitted over the network.

Index Terms— principal view ; camera selection ; smart camera ; distributed processing

1. INTRODUCTION

For many purposes, 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. In applications such as for example human behavior observation and person identification, the main information content of the joint network observation can be summarized into one principle view at each time instant. Transmitting, processing, storing and/or observing only this one view results in substantial resources savings. Depending on the acceptable information loss associated with this data reduction, one view might not suffice. In this case, a limited number of additional views can be selected to complete the principal one in order to obtain the desired observation.

The recent introduction of "smart cameras" with on-board image processing and communication hardware allows to distributedly extract from the captured images the necessary observations for principle view determination and for the selection of additional views, thus eliminating the need to collect the image data at a central point. This diminishes the required communication bandwidth within the network, which allows the cameras to work wirelessly, and spreads the computational burden over the nodes, resulting in a scalable system.

Viewpoint selection has been extensively studied in the context of computer graphics and robot navigation [1, 2]. These methods require an accurate model of the observed shape(s) and have difficulties coping with the background present in natural scenes, as they were all designed for artificial circumstances.

The allocation of resources in vision networks has been studied in [3], where resources are reconfigured based on sensor coverage and geometry, sensor allocation policies, and the dynamic processes in the environment, and in [4], where a distributed system is presented in which the video transmis-

[*]Linda Tessens is supported by the Flanders Fund for Research (FWO).

978-1-4244-2665-2/08/$25.00

sion parameters are adapted according to available resources. In [5], an integrated system with the functionality of human tracking, active camera control, face recognition, and speaker recognition is proposed.

Algorithms for automatically selecting a subset of cameras within a network have been designed for several purposes. In [6] and [7], the authors investigate camera selection within wireless vision networks of battery-powered nodes under lifetime constraints for user-specified viewpoint synthesis. 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 previous work, we have developed a method for selecting a limited number of cameras from a network in a content adaptive manner, such that this subset constitutes the most complete view on the shape of the objects in the scene possible for the given number of selected cameras [10]. None of these systems focuses on the problem of selecting one best view from all available cameras.

More directly related to this work is [11], where a single camera collects key frames of people in surveillance video based on face detections. Another related topic is the allocation of tasks to the best suited cameras within a multi-sensor surveillance system [12, 13].

In this work, our interest lies in the determination of one view that contributes most to the observation of the persons in a 3D scene in a vision network. We propose a low data rate method that is designed to be implemented in a distributed way on smart cameras. Experimental results on extensive video data from natural scenes show that the algorithm automatically selects a view that is also selected by a human observer in a high number of cases for a limited number of people in the scene. Furthermore, we describe an algorithm that selects, starting from the already selected principal view, a number of additional views and we demonstrate that the final selected subset of cameras (principal view together with a desired number of additional views) gives a complete view of the shape of the objects in the scene given the number of selected cameras.

The remainder of this paper is organized as follows. In Section 2, we elaborate on the setup of the system for which we devise our methods and on the assumptions we make. In Section 3, we describe the proposed principal view determination algorithm in detail. In Section 4, we apply the new method in the camera selection system of [10]. The performance of the method is discussed in Section 5 and conclusions are presented in the last section.

2. SYSTEM SETUP AND NOTATIONS

The system we consider consists of multiple smart camera sensors that observe a room with persons inside. A scheme of the system setup is depicted in the bottom-right corner of Figure 6. The smart camera sensors are battery powered and can communicate with each other through wireless channels. Their positions and orientations are fixed and calibrated. If calibration data is available at each time instant, the proposed algorithm can also be applied in a mobile camera network. A base station is deployed to receive the observations from the camera sensors and is responsible for coordinating all sensors in the network. The cameras are denoted by C_i for $i = 1, \ldots, N$, with N the total number of sensors. The complete collection of cameras is the set $\mathbf{C} = \{C_1, \ldots, C_N\}$ where $|\mathbf{C}| = N$. The image captured by the i-th camera at a certain time instant t is denoted by $\mathbf{X}_i(t)$. The different persons, or more general objects, are denoted by O_j for $j = 1, \ldots, L$, with L the total number of objects in the scene.

We assume that all cameras can exchange information with the base station. Each camera first processes the observed images locally, and then sends its processed information to the base station. The base station determines which camera contributes most to the desired observation of the objects in the scene at the current time instant and selects it as the key or principal camera K. The selection result is then broadcast to all the camera sensors, and only the selected camera will later send its image to the base station.

Although the transmission of an image is now delayed by the time it takes the base station to make and communicate its selection decision, the huge time gain resulting from not having to transmit complete images from all nodes ensures that the observation frequency of this system can be considerably higher than that of one without principal view determination. To reduce the time of the principal view determination, it is important to lower the time needed for the base station to collect the input data from the nodes. This is why it is of paramount importance that the nodes send only small amounts of *processed* information as input for the principal view determination algorithm.

Note that it is possible to augment the observation frequency by not running the principal view determination algorithm for every frame but applying a selection decision to several frames. Also, one could determine the principal view at a certain time instant based on the observation data of a previous time instant. These frame rate increasing strategies will have an impact on the accuracy of the principal view selection as necessary switches of the principal view will be delayed.

3. PRINCIPAL VIEW DETERMINATION

In this section, we will discuss the details of the principal view determination algorithm. The system diagram is depicted in Fig. 1.

3.1. Distributed Processes

In this phase of the algorithm, the nodes process the observed images to yield only the information necessary for the base station to determine the principal view. The lower the amount

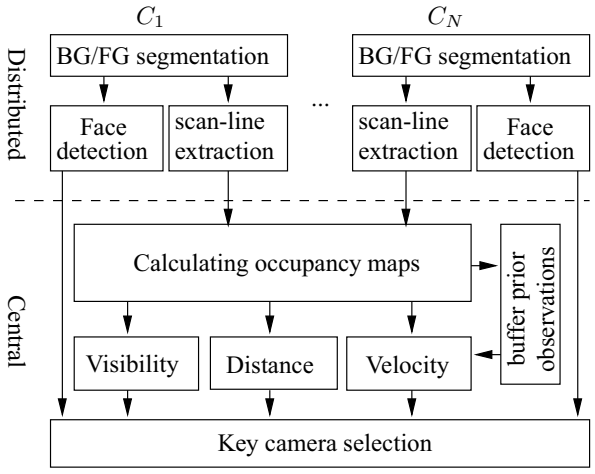


Fig. 1. Block diagram for principal view determination.

of data that needs to be transmitted, the quicker this decision can be made and the higher the achievable observation frequency.

Each smart camera C_i independently runs the following algorithms on its image $\mathbf{X}_i(t)$ captured at a certain time instant t. 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 [14]. Then, we detect the frontal faces in the foreground regions of the frame with the object detector that was initially proposed by Viola *et al* [15] and then improved by Lienhart *et al* [16]. At each time instant t, the face detector returns the following values: $f_i(t)$ and $Q_i^l(t)$ ($l = 1, \ldots, f_i(t)$). $f_i(t)$ 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^{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. The face detection measures $Q_i^l(t)$ of all detected faces are added into one general score

$$Q_i(t) = \sum_{l=1}^{f_i(t)} Q_i^l(t),$$

which is sent to the base station. With proper quantization if needed, this score can be represented by at most one byte.

Additionally, we project at each camera C_i the segmented foreground $\mathbf{F}_i(t)$ to a 1D-line horizontal line. This line is called scan-line and an example is shown in Figure 2b. All cameras send their (run-length coded) scan-lines to the base station.

If we assume there are at most B distinguishable objects in an image frame and the amount of bits needed to encode start and end point of each object on the scan-line is at most $2\lceil \log_2 w \rceil$ bits, where w is the image width, then the payload of this transmission can be approximated by $2\lceil \log_2 w \rceil B$.

Fig. 2. Example of (a) background/foreground segmentation ($\mathbf{B}_i$ and $\mathbf{F}_i$), (b) scan-line, and (c) column-wise extended scan-line ($\mathbf{B}_{i,\text{sc}}$ and $\mathbf{F}_{i,\text{sc}}$).

For example, for an image with width=352 and for 5 detected objects, this number amounts to 90 bits. Together with the output of the face detector, only 98 bits require transmission.

In the remainder of this paper, we will leave out the time variable t when talking about the observations and processing of the current time instant, in order not to overload the notations. We will again use the time variable when previous observations are taken into account.

3.2. Central Processes

At the base station, we extend each received scan-line in a column-wise manner to a 2D image, such that we get a rough approximation of the original background B_i and foreground F_i extracted at the sensor nodes (see Figure 2c). These approximations are denoted by $\mathbf{B}_{i,\text{sc}}$ and $\mathbf{F}_{i,\text{sc}}$.

We then calculate an occupancy map $\mathbf{O_C}$ in the following way. Using the shape-from-silhouette technique [17] we reconstruct the visual hull $\mathbf{H}_\mathbf{C}^{sc}$ from the background approximations $\mathbf{B}_{i,\text{sc}}$ of all cameras $C_i \in \mathbf{C}$. More precisely, within a cuboid-shaped volume V^3 in the 3D space of the observed room

$$V^3 = [X_1, X_2) \times [Y_1, Y_2) \times [Z_1, Z_2) \subset \mathbb{N}^3, \qquad (1)$$

$\mathbf{H}_\mathbf{C}^{sc}(\mathbf{j})$, 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. This is the case when it is part of the reprojected BG region from at least one of the cameras. All other voxels have value 1. Let us assume that the z-axis is perpendicular to the ground plane. Intersecting this visual hull with the plane

$$z = c \qquad (2)$$

yields us the desired occupancy map $\mathbf{O_C}$ [18]. This is a 2D raster image, uniformly distributed in the plane P^2 parallel to the ground floor (at $z = c$) of our observed 3D scene (or the defined voxel volume V^3):

$$P^2 = [X_1, X_2) \times [Y_1, Y_2) \subset \mathbb{N}^2 \quad \text{and} \quad z = c. \qquad (3)$$

Note that in a practical implementation, $\mathbf{O_C}$ can be directly obtained from the scan-lines, without actually reconstructing either $\mathbf{B}_{i,\text{sc}}$ nor $\mathbf{H}_\mathbf{C}^{sc}$. We consider the occupancy

map as a (very crude) shape approximation of the object in the scene.

From this shape approximation, we extract a number of cues that play a role in determining the principal view in the network. These cues are the position and velocity of each detected object O_j, with $j = 1 \dots L$ and L the total number of objects encountered in the occupancy map. Velocities are determined by calculating the distance covered by each object from the previous to the current frame. Armed with this information and with the output $Q_i(t)$ of the face detector on each camera, we can now assess the suitability of each camera to be assigned the role of key camera.

We propose different factors to determine this suitability.

- The *visibility* ν_{ij} of each object O_j in the view of camera C_i: This measure takes on value 1 if the center of mass of the object lies within the viewing range of the camera and 0 otherwise. The viewing range of each camera is determined from the calibration data.

- The *moving direction* of each object O_j *relative to the viewing direction* $\mathbf{\Psi}_i$ of camera C_i: With $\mathbf{V}_j$ the velocity of object O_j, negative values of $G_{ij} = \mathbf{V}_j \cdot \mathbf{\Psi}_i$ (with $\cdot$ denoting the scalar product between two vectors) indicate that the object is moving towards the camera. In this work, we assume that an observed person's body is oriented in the direction of his or her movement. As we wish to obtain frontal views of the observed persons, we introduce a binary value γ_{ij} which is 1 when G_{ij} is negative and 0 otherwise.

- The *distance* D_{ij} between the center of mass of object O_j and the camera center of C_i: This distance is normalized by dividing it by the maximal possible distance between an object in the observed space and a camera center. If we denote the z-coordinate (in V^3) of the camera center which is at the greatest height above the ground plane by Z_C^{max}, we approximate this distance by

$$D_{\max} = \sqrt{(Z_C^{\max} - c)^2 + (X_1 - X_2)^2 + (Y_1 - Y_2)^2} \tag{4}$$

  with X_1, X_2, Y_1, Y_2 as in Equation (1) and c as in Equation (2). To avoid evaluating square roots, we always work with the square of distances. If an observed person's body is oriented towards a camera - which can be ascertained when the velocity vector points in this direction or if the person's face is detected - a small distance between camera and object is desirable.

- The *speed* $\|\mathbf{V}_j\|$ at which each object O_j is moving: If this speed is very small, we assume that we cannot conclude anything about the body orientation of the observed person as (s)he might be standing still or rotating around his or her axis. The binary value σ_j indicates if the speed exceeds a certain threshold K_S, in which case $\sigma_j = 1$. Otherwise $\sigma_j = 0$. This measure is camera independent.

- The output Q_i of the *face detector* on each camera C_i: As the face detection score of each camera is the sum of the scores of all faces detected in its view, it is not linked to a particular object.

We summarize these factors into a score for each camera C_i:

$$S_i = K_Q Q_i + \sum_{j=1}^{L} \nu_{ij}\gamma_{ij}\sigma_j \left(-K_G G_{ij} + K_D(1 - \frac{D_{ij}^2}{D_{max}^2}) \right), \tag{5}$$

where K_Q, K_G and K_D are tuning parameters that weight the contribution of each factor. In the current system, these parameters have been optimized experimentally and then fixed. The dynamic adaptation of their value (e.g. based on modeling of the scene dynamics) would make the algorithm more universally applicable and flexible. This will be the subject of further research. Note that this score can be zero if no faces are detected in any of the cameras and if the observations extracted from the occupancy map are inconclusive. The latter case occurs

- if no objects are visible in any of the cameras,
- if all objects move away from the cameras in which they are visible, or
- if all objects move at speeds below the threshold K_S.

To obtain smoothness over time, the decision on the key camera for time instant t not only depends on the current observations, but also on those obtained at previous time instants. The default choice for the key camera $K(t)$ at time instant t is the previous key camera $K(t-1)$. It is then possible for the cameras to place a request to take over the role of key camera. The camera placing the request (the requester) is denoted by $R_{\text{key}}(t)$, with $R_{\text{key}}(t) \in \mathbf{C} \cup \{\text{NRQ}\}$. If all scores are zero, no request is placed and $R_{\text{key}}(t) = \text{NRQ}$. Otherwise, the camera with the highest score $S_i(t)$ at time instant t places the request :

$$R_{\text{key}}(t) = \begin{cases} \text{NRQ} & \text{if } \forall C_i, S_i(t) = 0 \\ \underset{C_i}{\text{argmax}}\, S_i(t) & \text{otherwise} \end{cases}. \tag{6}$$

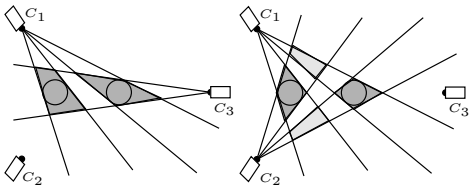
This request is granted if the same camera also placed a request at time instant $t-1$ or if the current key camera has not placed a request during the past T frames. In this way, excessive switching between cameras that are equally suitable to provide the principle view is averted, while simultaneously avoiding that alternating requests from such cameras prevents the role from being passed on to a more suitable camera than the current key camera. Also, the delay for a necessary switch of principal view is limited to one frame at most.

The algorithm is summarized in Algorithm 1.

Algorithm 1 Principal View Determination

Input: $S_i, i = 1 \dots N$ (the scores from the different cameras)
Output: $K(t)$ (the key camera for this time instant)

$K(t) \leftarrow K(t-1)$
$S_{max} \leftarrow 0$ and $R_{key}(t) = \text{NRQ}$
for $i = 1$ to N **do**
  if $S_i > S_{max}$ **then**
    $S_{max} \leftarrow S_i$
    $R_{key}(t) \leftarrow C_i$
  end if
end for
if $R_{key}(t) \neq \text{NRQ}$ **then**
  if $R_{key}(t) = R_{key}(t-1)$ **then**
    $K(t) \leftarrow R_i(t)$
  end if
  if $K(t) \notin \{R_{key}(t), \dots, R_{key}(t-T)\}$ **then**
    $K(t) \leftarrow R_i(t)$
  end if
end if



(a) Occupancy map when C_1 and C_3 are selected.

(b) Occupancy map when C_1 and C_2 are selected.

Fig. 3. Occupancy maps when specific cameras are selected. Detected objects are marked in gray. The shadow regions in (b), marked in light gray, are filtered out with the knowledge of the occupancy map calculated from all three cameras. As the dark gray area around the circular objects is smaller in (b) than in (a), C_2 adds more shape information than C_3.

4. APPLICATION IN CAMERA SELECTION

If the information provided by the principal key camera is not sufficient to have the desired observation of the scene, one can decide to select additional views that complement the view of the key camera. These additional cameras are called helper cameras and are indicated by W_k where $k = 1, \dots, n-1$, with n the total number of selected views. The main idea of this section is to choose from the remaining $N-1$ cameras those helper cameras that add most information about the shape to the image data coming from the already selected key camera K (see Figure 3). We denote the final selected camera subset by $\mathbf{S}_n = \{K\} \cup \{W_1, \dots, W_{n-1}\}$. This subset constitutes a significantly more efficient scene representation than the totality of the available views.

Our approach consists of the following steps. First, the base station determines all the valid candidate subsets $\hat{\mathbf{S}} \subset \mathbf{C}$, for which $|\hat{\mathbf{S}}| = n$ and $K \in \hat{\mathbf{S}}$, with K determined as in Section 3. For the principal view selection, all the cameras had already sent their scan-lines to the base station. For each candidate subset $\hat{\mathbf{S}}$, we now use the scan-lines from only $C_i \in \hat{\mathbf{S}}$ to reconstruct the occupancy map for that candidate subset, with the method described in 3.2.

Subsequently, this occupancy map is filtered to remove *shadow* areas. These are parts of the occupancy map that do not represent real objects but result from an insufficient number of used cameras n (see figure 3). The filtering occupancy map is a dilated version $\mathbf{O}_{\mathbf{C}}^{\text{filt}}$ of the ideal occupancy map $\mathbf{O}_{\mathbf{C}}$, reconstructed from the complete set of cameras $\mathbf{C}$ (as calculated in Section 3.2). In this way, we ensure that we base our camera selection only on the reconstructed shapes of objects that are also detected when all N cameras are active and the influence of shadow areas is minimized. As in Section 3.2, we consider the occupancy maps $\mathbf{O}_{\hat{\mathbf{S}}}$ as (very crude) 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 shape of the objects in the scene, we choose the cameras that allow for the best approximation. This is why in a final step, we select from all candidate subsets the subset that yields the minimal occupied area:

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

5. RESULTS

In this section, we assess the performance of the principal view determination and the selection of additional views. To evaluate the quality of the principal view selected by the method of Section 3, we use sequences labeled by human observers as a benchmark (Section 5.1). To illustrate the benefit of using this principal view as a starting point for the selection of additional cameras to complete the observation of the shape of the objects in a scene, we compare the accuracy of this selection result to the accuracy of selections made when no key camera is provided as a starting point and when only the face detection score is used to select the principal view (Section 5.2).

Experimental data to test the method on was recorded with a camera network set up as described in Section 2, in which a maximum of 4 persons were present. We believe that the proposed system provides a framework that is sufficiently general to also be able to cope with drastically other scenarios such as outdoor and crowded scenes and a wider coverage area. Indeed, if needed, the background-foreground segmentation in Fig. 1 can easily be replaced by a change detection algorithm that is robust to outdoor conditions or can e.g. detect only abnormal behavior in crowded scenes. Such experiments will be performed and assessed in future work.

In this work, we have recorded indoor scenes with $N = 10$ cameras, of which five 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 [19]. Sequences were recorded at 5 frames per second and at a CIF resolution (352×288). Only the starting points of the recordings were synchronized. The parameters of the BG/FG segmentation and the face detection are summarized in Table 1. 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 structuring element for the dilation to obtain the filters $\mathbf{O}_{\mathbf{C}}^{\text{filt}}$ (see Section 4) and $\mathbf{H}_{\mathbf{C}}^{\text{filt}}$ (see later, in Section 5.2) is a square of 3×3 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 occupancy map is the intersection of the voxel volume at $z = 1.29m$. The tuning parameters in Equation (6) are set to $K_Q = 1$, $K_G = 2$ and $K_D = 1$. The threshold for the speed is $K_S = 0.08m$/frame. The parameter T is set to 4. We discuss results for subsets of both 3 and 6 cameras.

Table 1. Parameters of the BG/FG segmentation and the face detection.

Parameters BG/FG Segmentation			
L (color comp.)	128	α_1	0.1
N_1 (color comp.)	15	α_2	0.005
N_2 (color comp.)	25	α_3	0.1
L (color co-occ.)	64	δ	2
N_1 (color co-occ.)	25	T	0.9
N_2 (color co-occ.)	40	MINAREA	15.0
UPDATE_TRESH	0.5		

Parameters Face Detection	
scale factor	1.10
min. number (-1) of neighbors	2
min. window size	5×5
classifier training window size	20×20

5.1. Principal View Quality

Which view provides the best observation of a person is in many cases not clearly defined, even for a human observer (see for example Fig. 4). For this reason, at each time instant up to three views can get the label of being a view that provides a good observation of the persons in a scene. If the scene is empty, none of the views is labeled as principal view.

In our experiments we distinguish between four scenarios, depending on the number of people in the scene.

Table 2 indicates the percentage of frames in which the view selected by the method of Section 3 was labeled as a principal one by a human observer. The total number of labeled frames is indicated in the second column. Table 2 also shows the hit rate of the key camera selection method of [10], in which only the face detection score is used to select the principal view.

Fig. 4. Example of ambiguity when choosing the best observation of the person. Both images display a nearly frontal view of the person. In the left one, the face is tilted somewhat more towards the camera, while in the right view the person appears slightly bigger. Both views can therefore be considered equivalent.

Table 2. Percentage of frames in which the view selected by the method of Section 3 and [10] was labeled as a principal one by a human observer.

Scenario	# frames	Key as in [10]	Proposed
1 persons	316	0.46	0.70
2persons	297	0.51	0.71
3 persons	376	0.50	0.55
4 persons	262	0.51	0.53

From these numbers, we can observe that the proposed principal view selection method achieves a good hit rate for small numbers of people in the scene, or in other words, that it very often selects the view which also a human observer would judge as providing a good observation of the persons in a scene. For more persons, the hit rate drops. In these cases, determining the principal view becomes more ambiguous, as more than one camera might have a good frontal view of different persons. Selecting more than one key view would then be appropriate. This will be the subject of further research. Comparing column four to column three, we can also conclude that including knowledge about position and velocity of the observed objects in the principle view determination provides a powerful means to boost the hit rate.

A demo video illustrating principal view selection in a network of 10 cameras can be found online at [20].

5.2. Application in Camera Selection

To evaluate the accuracy of the camera selection method of Section 4, we reconstruct the visual hull for each frame from the foreground *silhouettes* $\mathbf{F}_i$ of the selected camera subset $\mathbf{S}_n$, with $\mathbf{S}_n$ as in Equation (7). We will denote this hull by $\mathbf{H}_{\mathbf{S}_n}^{\text{silh}}$. Note that these FG silhouettes $\mathbf{F}_i$ are *not* available

at the base station nor used in the actual method, only their approximate versions $\mathbf{F}_{i,\mathrm{sc}}$. From these hulls, we determine at each time instant the number of voxels d_n that are different between the hull reconstructed from the selected subset and a benchmark hull. This benchmark visual hull is reconstructed from the whole set $\mathbf{C}$ of ten available cameras and will be denoted by $\mathbf{H_C}$. For the calculation of this difference we only take into account differences within $\mathbf{H}_\mathbf{C}^{\mathrm{filt}}$, which is the dilated version of $\mathbf{H_C}$:

$$d_n = \sum_{\forall \mathbf{j} \in V^3} \left[\left(\mathbf{H}_\mathbf{C}^{\mathrm{filt}}(\mathbf{j}) \mathbf{H}_{\mathbf{S}_n}^{\mathrm{silh}}(\mathbf{j}) \right) - \mathbf{H_C}(\mathbf{j}) \right]. \qquad (8)$$

Filtering with $\mathbf{H}_\mathbf{C}^{\mathrm{filt}}$ is needed in order to focus on the interesting objects, without having the disturbing influence of *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. These shadow volumes can be seen as the 3D versions of the shadow areas described in Section 4. 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 within the filtered reconstructed hull (or in other words d_n) gives us an insight in how well the selected subset observes the objects in the scene from all sides. Indeed, the more voxels are observed as empty around the object of interest, the less redundant the views from the selected cameras are.

To assess the impact of the principal view determination on the performance of the camera selection algorithm of Section 4, we compare its accuracy with the one obtained when the key camera is assigned based only on the face detection scores, as in [10], and when cameras are selected with no prior assignment of a key camera. In this last case all possible combinations of n out of N cameras are valid and the camera subset that leads to the occupancy map with the smallest area is guaranteed to be found. This is not the case when a key camera is chosen, as it might exclude the "optimal" subset from the valid combinations. Note that the lack of prior key camera assignment drastically increases the computational burden of the algorithm and eliminates the guarantee that the view that contributes most to the desired observation of the scene is selected. Accuracies for the camera selection methods with a different or no principal view assignment are calculated in the same way as described above for the method of Section 4.

In Table 3, we compare for the three methods (no key camera assigned, key camera assigned as in [10], proposed method of Section 3) the mean value of the number of different voxels d_n 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 as a reference.

First of all, we observe that the proposed method yields similar results as camera selection without prior key camera assignment. Occasionally, it even outperforms that method. This is possible because the occupied area is only an approximation of the shape of the people present in the scene. The subset of cameras that minimizes the occupancy area does not necessarily lead to the solution that gives the best visual hull. Also, we can see from Table 3 that camera selection with the principal view determined as proposed in Section 3 outperforms selection where the choice of the key camera is only based on the face detection scores [10].

As an illustration, we show in Figure 5 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 possible subsets $\check{\mathbf{S}} \in \mathbf{C}$, with $|\check{\mathbf{S}}| = n$, contained within $\mathbf{H}_\mathbf{C}^{\mathrm{filt}}$ (green dotted lines). Note that there are more possible subsets $\check{\mathbf{S}}$ than there are candidate subsets $\hat{\mathbf{S}}$ (as introduced in Section 4), since for the latter also holds $K \in \hat{\mathbf{S}}$. 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 $\mathbf{H}_\mathbf{C}^{\mathrm{filt}} \mathbf{H}_{\mathbf{S}_3}^{\mathrm{silh}}$[1] per frame are drawn as the thicker lines. The solid blue line indicates the camera selection when no key camera is assigned. The dash-dotted black line with round markers is the subset selection with the key camera selection based on face detection only (as in [10]). The dotted red line with triangular markers is the camera subset selection with the principal camera selection proposed in Section 3. This graph visualizes that, regardless of how the principle view is determined, the camera selection method of Section 4 selects from all possible subsets one that is always close to the best possible subset. Indeed, the curves of all methods are close to the lower envelope of the curves of all possible subsets. A second observation is that the curves corresponding to camera selection without prior key camera assignment and with principal view determination as in Section 3 mostly coincide and that both methods lead to lower visual hull volumes than the same selection method but with the principal view determined as in [10].

Figure 6 shows a visual example of the selection of $n = 3$ cameras from 10 using the proposed method. We display the views of all the cameras $C_1, \ldots, C_{10}$. 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 person in the scene. The selected key camera is marked by a magenta bounding box (C_3). This camera was chosen to be the key camera by the principal view determination method of Section 3. The helper cameras are marked by a cyan bounding box and are selected as explained in Section 4 (C_8 and C_{10}). We can observe from the displayed views that the selected principle view contributes most to the observation of the person, while the helper cameras complete the observation. The non-selected cameras add only redundant information. Note that the person sitting at the desk in camera view C_{10} is operating the start and the end of the cap-

[1] The multiplications here are performed between corresponding voxels.

Table 3. Mean voxel difference d_n (Equation (8)) for the three key camera assignment methods (no key camera assigned, key camera assigned as in [10], proposed method of Section 3) 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			d_6		
			No key	Key as in [10]	Proposed method	No key	Key as in [10]	Proposed method
1 persons	1629	598.93	842.70	1204.29	894.42	298.60	348.29	310.36
2 persons	2213	2450.99	2945.63	3365.90	2755.85	1095.09	755.97	750.32
3 persons	826	4584.35	5364.68	7349.20	5816.58	1476.38	1326.16	1306.34
4 persons	290	8079.53	7630.01	10036.74	8917.38	1361.51	1356.97	1415.02

turing of the sequence. This person is immobile during the whole sequence and is assumed to be background.

A demo video illustrating the application of principal view selection in camera selection can be found online at [20].

6. CONCLUSION

In this paper, we presented a method to determine which camera in a smart camera network has the best frontal view of the persons in a scene. The algorithm consists of two types of processes. The *distributed* processes run on the smart cameras themselves and strongly reduce the amount of data that needs to be sent over the network to the base station to a couple of tens of bits per node. At the base station the *central* principal view selection takes place. In order to choose an appropriate key camera, this algorithm takes into account the number of faces detected by each of the cameras, and the velocity and positions of the objects relative to the viewing direction and viewing angle of the cameras. If one view does not suffice to obtain the desired information about the objects in the scene, one can decide to select additional views that complement the view of the key camera. The method we use for this is based on shape approximation. The subset that is expected to approximate the shape most accurately, is chosen to be the optimal one.

Experimental results on human-labeled sequences show that the selected principal view is equal to the view selected by a human observer in a high number of cases for a limited number of people in the scene. Also, we showed that this view together with the additional views give a good approximation of the shape compared to the best achievable shape approximation with the selected amount of cameras. Additionally, it is shown that the principal key camera selection is a good starting point for the selection of additional views, since it greatly reduces the computational complexity, while still allowing the reconstruction of the shape of the objects to be almost as accurate as in the optimal subset selection case (without principal view determination).

The proposed system provides a framework that is sufficiently general to handle different scene environments and scenarios with only minor adaptations. In future work we will investigate the replacement of the background-foreground segmentation by a change detection algorithm to operate in outdoor conditions and in crowded circumstances, and we will consider wide area setups.

Acknowledgement

The authors would like to thank everyone in the test video sequences and all reviewers for their instructive comments.

7. REFERENCES

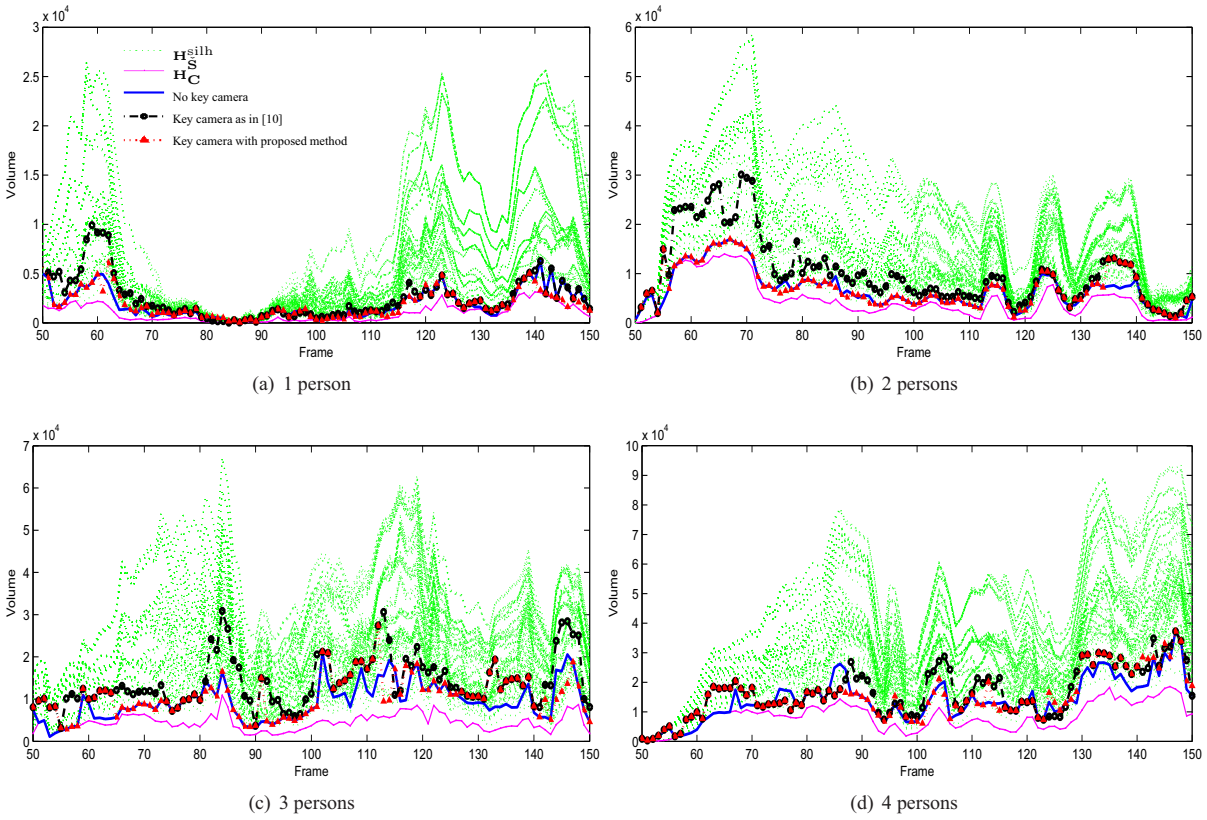
[1] Pere-Pau Vázquez, Miquel Feixas, Mateu Sbert, and Wolfgang Heidrich, "Viewpoint selection using viewpoint entropy," in *Proceedings of the Vision Modeling and Visualization Conference 2001*, 2001, pp. 273–280.

[2] D. R. Roberts and A. David Marshall, "Viewpoint selection for complete surface coverage of three dimensional objects," in *Proc. of the British Machine Vision Conference (BMVC)*, Southampton, England, 1998.

[3] D. Karuppiah, R. Grupen, A. Hanson, and E. Riseman, "Smart resource reconfiguration by exploiting dynamics in perceptual tasks," in *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems)*, 2005, pp. 1513 – 1519.

[4] Christian Micheloni, Marco Lestuzzi, and Gian Luca Foresti, "Adaptive video communication for an intelligent distributed system: Tuning sensors parameters for surveillance purposes," *Machine Vision and Applications*, 2007, available online.

[5] Mohan Trivedi, Kohsia Huang, and Ivana Mikic, "Intelligent environments and active camera networks," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Nashville, TN, USA, 2000, vol. 2, pp. 804 – 809.



(a) 1 person (b) 2 persons

(c) 3 persons (d) 4 persons

Fig. 5. 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 reconstructed from all possible subsets $\check{\mathbf{S}}$ (green dotted lines), the benchmark hull $\mathbf{H_C}$ (solid magenta line) and the hull $\mathbf{H}_{\mathbf{C}}^{\text{filt}}\mathbf{H}_{\mathbf{S}_3}^{\text{silh}}$ from the subsets selected with different principal view selection strategies. The lower this volume, the less redundant the selected views. One can see that from all possible subsets that might be selected, all methods always pick one that is close to the best possible subset (the one that leads to a visual hull volume closest to the benchmark hull $\mathbf{H_C}$). The curves corresponding to camera selection without prior key camera assignment (solid blue line) and with principal view determination as in Section 3 (dotted red line with triangular markers) mostly coincide and both methods lead to lower visual hull volumes than the same selection method but with the principal view determined as in [10] (dash-dotted black line with round markers).

[6] Stanislava Soro and Wendi B. Heinzelman, "Camera selection in visual sensor networks with occluding objects," in *Proceedings of ACM/IEEE First International Conference on Distributed and Smart Cameras (ICDSC)*, Vienna, Austria, 2007.

[7] G. Sharma C. Yu, S. Soro and W. Heinzelman, "Lifetime-distortion trade-off in image sensor networks," in *Proceedings of International Conference on Image Processing (ICIP)*, San Antonio, Texas, USA, September 2007, vol. V, pp. 129–132.

[8] Toshihiro Matsui, Hiroshi Matsuo, and Akira 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*, 2001, vol. 8, pp. 955–964.

[9] Danny Yang, Jaewon Shin, Ali O. Ercan, and Leonidas Guibas, "Sensor tasking for occupancy reasoning in a camera network," in *Proc. of IEEE/ICST Workshop on Broadband Advanced Sensor Networks*, 2004.

[10] Marleen Morbee, Linda Tessens, Huang Lee, Wilfried Philips, and Hamid Aghajan, "Optimal camera selection in vision networks for shape approximation," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Cairns, Queensland, Australia, Oktober 2008.

[11] Rogerio Feris, Ying-Li Tian, and Arun Hampapur, "Capturing people in surveillance video," in *Pro-*

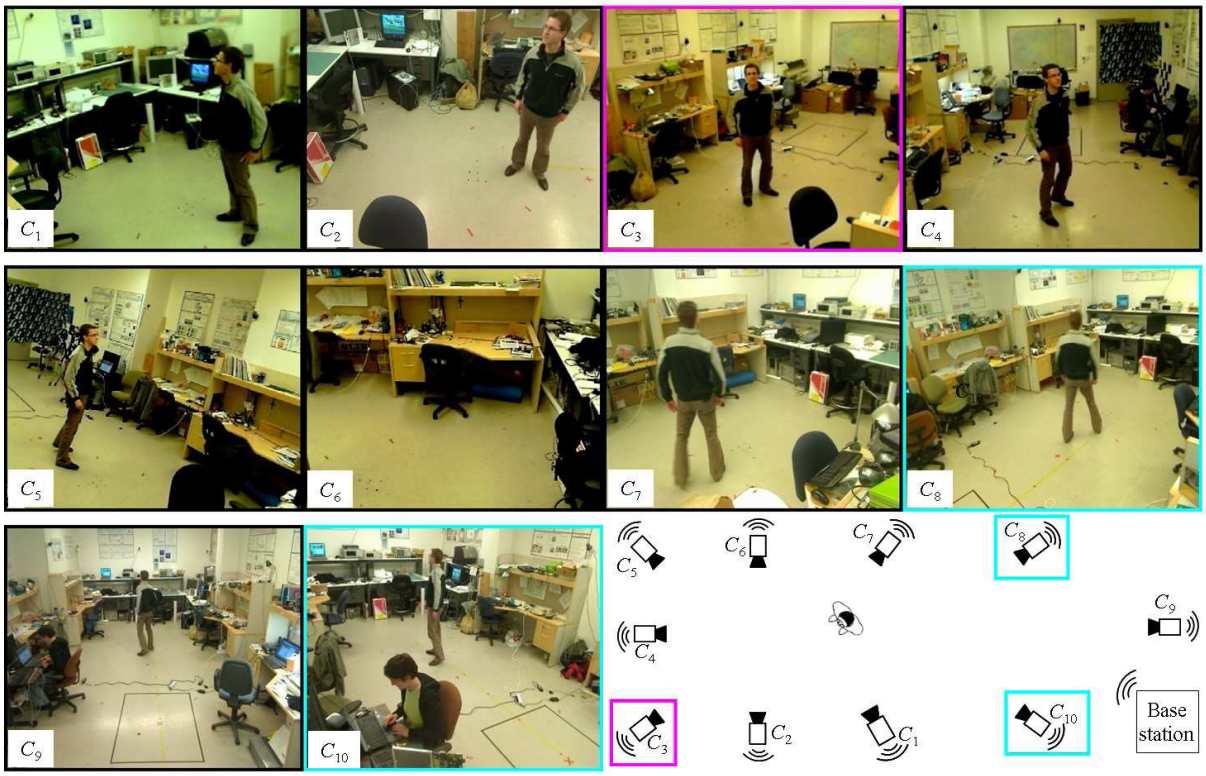


Fig. 6. Example of the selection of 3 out of 10 cameras. The views of the 10 cameras ($C_1, \ldots, C_{10}$) are shown. In the bottom-right corner, we depicted a top view of the scene which shows its geometry and the positions of the cameras and person. The selected key camera (C_3) is marked by a magenta bounding box and the helper cameras (C_8 and C_10) by a cyan bounding box.

ceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, United States, 2007.

[12] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance," in *Proceedings of the IEEE*, October 2001, vol. 89(10), pp. 1456–1477.

[13] M. Bramberger, B. Rinner, and H. Schwabach, "A method for dynamic allocation of tasks in clusters of embedded smart cameras," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, Waikoloa, HI, United States, 2005, vol. 3, pp. 2595 – 2600.

[14] Liyuan Li, Weimin Huang, Irene Y. H. Gu, and Qi Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, pp. 2–10, ACM.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," 2001.

[16] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, vol. 1, pp. I–900–I–903 vol.1.

[17] 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.

[18] Adam Hoover and Bent David Olsen, "A real-time occupancy map from multiple video streams," in *International Conference on Robotics and Automation (ICRA)*, 1999, pp. 2261–2266.

[19] Tomáš Svoboda, Daniel Martinec, and Tomáš Pajdla, "A convenient multi-camera self-calibration for virtual environments," *PRESENCE: Teleoperators and Virtual Environments*, vol. 14, no. 4, pp. 407–422, August 2005.

[20] "Principal view determination demos," `http://telin.ugent.be/ipi/drupal/cameraSelection`.