Machine Vision and Applications (2013) 24:319–336
DOI 10.1007/s00138-012-0408-6

ORIGINAL PAPER

Multicamera human detection and tracking supporting natural interaction with large-scale displays

Xenophon Zabulis · Dimitris Grammenos · Thomas Sarmis · Konstantinos Tzevanidis · Pashalis Padeleris · Panagiotis Koutlemanis · Antonis A. Argyros

Received: 8 March 2011 / Revised: 9 January 2012 / Accepted: 17 January 2012 / Published online: 8 February 2012

Abstract This paper presents a computer vision system that supports non-instrumented, location-based interaction of multiple users with digital representations of large-scale artifacts. The proposed system is based on a camera network that observes multiple humans in front of a very large display. The acquired views are used to volumetrically reconstruct and track the humans robustly and in real time, even in crowded scenes and challenging human configurations. Given the frequent and accurate monitoring of humans in space and time, a dynamic and personalized textual/graphical annotation of the display can be achieved based on the location and the walk-through trajectory of each visitor. The proposed system has been successfully deployed in an archaeological museum, offering its visitors the capability to interact with and explore a digital representation of an ancient wall painting. This installation permits an extensive evaluation of the proposed system in terms of tracking robustness, computational performance and usability. Furthermore, it proves that computer vision technology can be effectively used to support non-instrumented interaction of humans with their environments in realistic settings.

Keywords Person localization · Person tracking · Camera network · Real-time volumetric reconstruction · Non-instrumented location-based interaction

X. Zabulis (✉) · D. Grammenos · T. Sarmis · K. Tzevanidis · P. Padeleris · P. Koutlemanis · A.A. Argyros
Institute of Computer Science, FORTH, N. Plastira 100, Vassilika Vouton, 700-13 Heraklion, Crete, Greece
e-mail: zabulis@ics.forth.gr

K. Tzevanidis · A.A. Argyros
Computer Science Department, University of Crete, Crete, Greece

1 Introduction

In the past few years, museums and expositions worldwide have started exploring new ways for integrating interactive exhibits [3,14,16,30,34], moving beyond the typical "multimedia information kiosk" paradigm. The motivation behind such efforts is to support constructive and engaging entertainment in education ("edutainment"). The enhancement of didactic content with captivating and immersive experiences and the support of active visitor participation plays central roles in the pedagogic value of such systems. In expositions, additional value can be gained through engaging, interactive informative sessions for multiple visitors. Towards this end, large-scale digital displays are typically employed to present large-scale artifacts in actual size or magnify smaller ones.

An interesting category of such displays are those offering non-instrumented, location-based interaction capabilities. In location-based interaction, the contents of the display are dynamically updated based on the position of the visitor relative to it. Location-based interaction is non-instrumented if visitors are not required to carry any device measuring or marking their location. Besides offering a much more natural interaction experience, this approach has simple logistics (e.g. no need for a lending–return process), which is of practical importance for a museum or exhibition. In this paper, the term large-scale refers to the size of the display whose dimensions are in the order of meters.

One of the earliest examples of such interaction is *KidsRoom* [6], an interactive playspace simulating a children's bedroom where young children are guided through an adventure story. In [17], a multiplayer game was developed using one top-view camera, mapping the estimated player 2D motion to that of a digital character. Another example are interactive floors (physical sensor-based like *Magic Carpet* [25], or vision-based like *iGameFloor* [12]), which are

mainly employed in games. In the domain of museum applications, the work in [16] explored three different ways of supporting location-based interaction: a coarse grained passive infrared sensor, pressure sensors embedded in the floor or small staircases, and vision-based tracking. In *Immersive Cinema* [35], one ceiling-mounted camera is used to track the position of a single visitor on a floor that is conceptually segmented in spatial regions. In [26], the same method was employed to track a single person in front of a large display and modulate the projected content according to his location. In [30], a ceiling-mounted infrared camera was employed to coarsely track human position and motion in groups. This information is subsequently exploited to derive a flocking behavior for browsing collections of photographs and texts.

The work presented in this paper is focused on the exploitation of computer vision towards the development of large-scale, multi-user digital displays with non-instrumented, enhanced interaction capabilities that are based on the location and walkthrough trajectories of users. We propose a multicamera vision system that localizes and tracks visitors in front of such displays (see Fig. 1). A camera network is employed to image the area in front of the display. The acquired images are used to compute a volumetric reconstruction of the visitors in real time that is employed to robustly track visitors, despite occlusions and challenging person arrangements.

The capability to maintain the identity of visitors over long periods of time is of key importance to the provision of personalized content. Thus, real-time, robust person tracking is crucial in systems that support such interaction. Monocular approaches to the problem [36,39] are based on the image cues such as color and silhouette shape and employ sophisticated tracking methods to cope with scene complexity. The method in [8] utilizes a binocular camera system and combines stereo, color, shape and face detection to improve tracking performance. More recently, it was shown that humans can be tracked effectively by relying on RGB-D data [33]. Still, all single-view approaches suffer from visibility limitations, due to the observation of the scene from a specific viewpoint. Multiview approaches simplify localization because they acquire information from diverse perspectives and treat occlusions systematically. On the other hand, the large amount of data to be processed induces large computational demands, which is typically addressed through parallel and distributed methods. Also, communication bandwidth issues are raised because the input of more than a handful of cameras needs to be distributed to multiple computers or bus channels.

Multiview human localization methods perform 3D reconstruction of the imaged persons to register them to a map of the workspace. The method in [37] fuses the results obtained by existing single-view tracking methods that are applied individually to each of the views. However, the limitations of single view methods in handling occlusions still affect the fused results. The methods in [10,15,20,23,28], employ multiple views and a planar homography constraint to map imaged persons to the ground plane. In [18], a voxel grid is utilized to represent the 3D reconstruction and computation is distributed in the GPUs of four computers. For each voxel, a partial estimate of its occupancy is obtained, transmitted centrally, and fused with the rest of estimates for this voxel. Communication cost is significant as the amount of data to be communicated is proportional to the number of voxels. The system in [32] eliminates the communication cost by mounting all cameras to a single computer and centralizing computation. This approach does not scale with the number of views which, in this case, is limited to four.

This work follows a multiview approach to person tracking and utilizes a volumetric 3D reconstruction of persons to increase localization robustness but it does not require that the number of tracked persons is a priori known, as in [10]. Computational efficiency is achieved by the proposed parallel and distributed computer architecture, whose implementation is based on the software platform in [42]. In this work, parallelization is not limited to 3D reconstruction but extends to other operations such as image rectification and background subtraction. The computation of pertinent volumetric data structures, such as volumetric occupancy and a mapping of this occupancy on the ground plane, is achieved in smaller execution time relatively to corresponding state-of-the-art systems. By optimizing execution time, more detailed spatial representations are obtained at larger framerates. We strive for efficient 3D reconstruction because this way, besides reducing system response time, tracking also becomes more robust. The reason is that as person motion becomes more densely sampled in time, it also becomes less ambiguous to track.

As a case study, we developed a computer vision system that was installed in front of a large display that is used for the interactive exploration of an ancient wall painting. The proposed system is used to present a personalized restoration of the artifact and provide detailed information on the persons and activities illustrated in the painting. The presented information depends on the location of individual visitors relative to the display, as well as on what information has already been presented to them. The proposed approach extends previous ones such as [16] and [35] through the support of multiple visitors and the provision of personalized content and interaction capabilities.

Several laboratory tests but also the real-life installation of the developed system at an archaeological museum provide the basis of the evaluation of the proposed approach in terms of tracking robustness, computational efficiency and usability. An early version of the supported cultural heritage application can be found in [40]. The obtained results show that the developed computer vision system can be used effectively to enhance user experience. The real-life deployment

of the system revealed another benefit that was not anticipated originally. The developed system provides an invaluable tool for acquiring statistics regarding the behavior of visitors over long periods of time. Based on this evidence, an informed redesign of an exhibit can be performed towards optimizing user experience.

The remainder of this paper is organized as follows. Section 2 presents the infrastructure that is necessary for the operation of the proposed system. Section 3 focuses on the core computer vision modules comprising the proposed vision system. In Sect. 4, the application mediating user interaction with the system is described. In Sect. 5, the evaluation of the system is presented. Finally, Sect. 6 summarizes this work and provides key directions for future work.

2 Infrastructure

A typical setup of the system involves a $4 \times 4.5 \times 2.5$ m room (see Fig. 2), in which a 4.2×1.58 m^2 dual backprojection display is mounted at the wall opposite from its entrance. Additional installations employed a 25 m^2 (Fig. 4) and a 36 m^2 (Fig. 1) terrain and correspondingly larger displays.

2.1 Multicamera vision system

The computer vision system employs a set of cameras that image the scene from multiple views. It also includes one or more computers that acquire the corresponding images, process them, and extract a spatial representation of the persons in the room. A middleware layer casts this representation available to the module that runs the application scenario.

In a typical setup, eight cameras (*Dragonfly*, *Point Gray Research*) are evenly distributed to two computers. The two computers are connected by a 1 Gbit *Ethernet* link. Each of them is equipped with an *Intel i920* quad-core CPU, and a *NVIDIA GTX 275* programmable GPU.

The cameras are placed evenly on the walls of the room and are mounted at the ceiling height, overlooking the workspace. As a background model will be acquired for each view, cameras that are opposite to the display are posed so as to avoid imaging a large portion of the screen. If this is not possible, a validity mask in the frame grabber discards the corresponding image region. The cameras are spatially calibrated by imaging a checkerboard calibration pattern at multiple postures. The employed calibration method [31] automatically detects the calibration pattern and extracts reference points. Camera calibration is further refined through bundle adjustment [21].

Camera synchronization is performed with the *MultiSynch* software [29] and is based on timestamps and a dedicated *FireWire* bus across computers. The technical complexity of synchronized image acquisition and communication is reduced by employing the software platform in [42]. This platform supports the communication of synchronized images and intermediate processing results across processing nodes that reside on the same or different computers, through a shared memory. Thus, when multiple computers are available, per view processing (i.e. preprocessing, background subtraction) is distributed at the computers hosting the cameras increasing computational speedup and reducing communication among computational nodes (see Sect. 3.1).

Finally, communication between visual processing and the application layer is facilitated by a custom middleware layer that hides the details of network and inter-process communication. This layer is accessible through an API compatible to a wide range of programming languages (*C/C++,.NET, Java, Python, Flash/ActionScript*). In this way, the application can be reprogrammed for different scenarios, independently of the visitor tracking infrastructure.

2.2 Display system

The employed display covers one of the room's walls and is created by combining the output of two bright

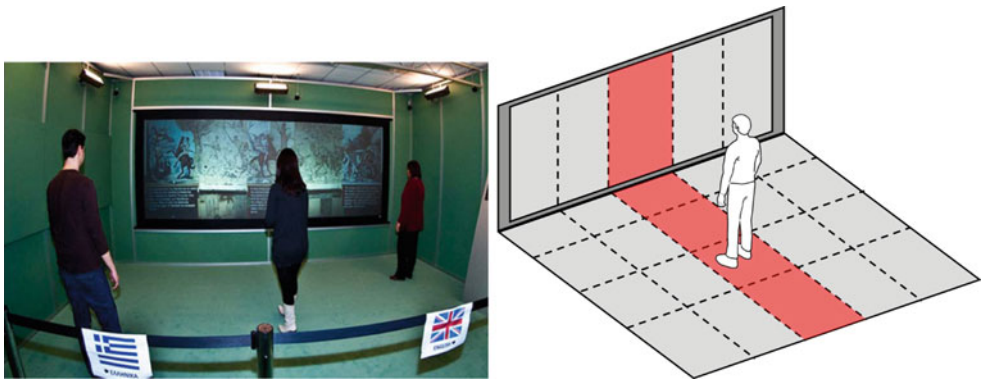


Fig. 1 *Left* visitors in front of a display that presents a large-scale artifact in actual size, while augmenting the display with personalized content in the form of visual and textual annotations. *Right* the display segment in front of each visitor is updated according to the visitor's distance from the display and his/her past walk-through trajectory

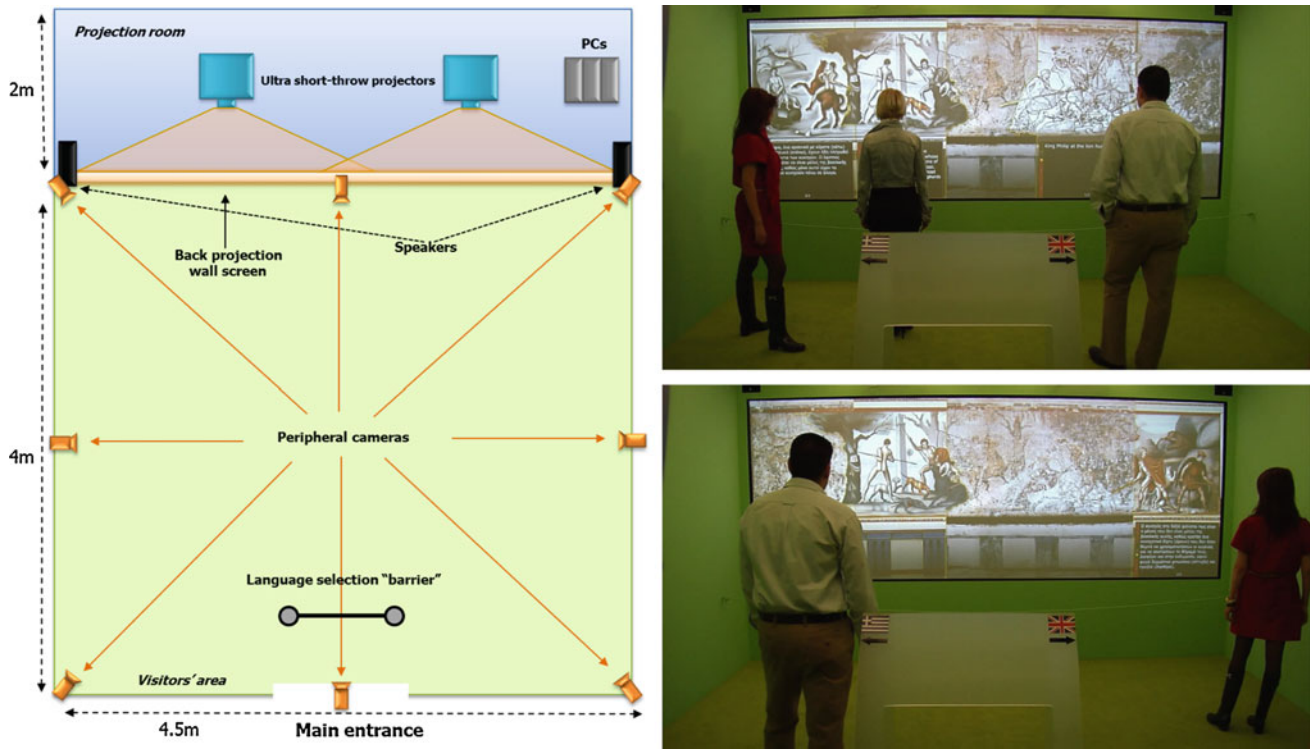


Fig. 2 Real-world installation at the Archaeological Museum of Thessaloniki, Greece. Floor plan (*left*) and two actual views (*right*). Due to the tracking of visitors, linguistic preferences are associated with each visitor, presenting textual information at the language of choice

Fig. 3 Combining two projections in a seamless display. Projected image before applying any adjustments (*left*), rectified projection (*right*)

(3000 lm) 1024 × 768 short-throw projectors. The projections are parallel and partially overlapping (≈10%). The reason for employing wide FOV projectors is to retain a small backprojection area. The disadvantage of this setup is that an intense projection distortion is induced, so, projected images need to be rectified. Vision-based projector lens calibration (i.e. as in [9]) requires an additional camera imaging the projection that complicates the hardware setup. To avoid this, the projections are fused by employing a piecewise linear transformation. More specifically, the projected image is divided into four overlapping quadrants. Geometrically, prior to projection, each quadrant is transformed through a homography determined by four control points. Visual markers in the overlapping regions indicate the corresponding control points that should, ideally, coincide. During system setup, a software tool allows the displacement of these control points across the display. By adjusting these markers to make them coincide, the four quadrants are stitched without any visible trace of a seam in the projection. Projections are photometrically blended in overlapping regions as in [27]. An example outcome of this process is shown in Fig. 3.

3 Volumetric human detection, localization and tracking

The core of the proposed system consists of a method for volumetric reconstruction of the humans in front of the large-scale display and the exploitation of this information for their accurate tracking in challenging situations. The rest of this section describes the proposed methods in more detail.

3.1 Scene reconstruction

Reconstruction employs the notion of the visual hull [19] that is computed based on a foreground detection process. The method computes the visual hull of all visitors on a voxel occupancy grid V. A voxel in V has the value of 1 if it is occupied by a person and 0 otherwise.

Each time a synchronized set of images is acquired (a multiframe), a 3D reconstruction of the scene is performed. This computation is divided in two stages. The first stage concerns preprocessing and foreground segmentation that is performed on each of the acquired images. This computation is distributed among the computers hosting the cameras. In the second stage, the binary images resulting from foreground detection are compressed and transmitted to a single computer which performs the volumetric reconstruction. Both stages are parallelized on the individual computers where they are performed, capitalizing on programmable GPU hardware.

3.1.1 Preprocessing and foreground detection

At each workstation, the images acquired by the hosted cameras are stored in RAM and processed locally, increasing the distribution of computational load. Each workstation holds a fixed-size buffer for every camera that is connected to it. Each buffer stores the captured frames after they have been converted from Bayer Tile to RGB format. The rate of storing images into buffers matches the camera's acquisition frame rate. Image data are stored together with their associated timestamps. To avoid buffer overflow as newer frames arrive, older frames are removed.

The image I_i from camera i is read into the GPU and transformed so that geometric distortions are canceled out based on the available calibration information. Foreground detection is performed also based on a GPU implementation of the method presented in [43] that parallelizes computation at the pixel level. This results in a binary image B_i. In contrast to [18,32] and aiming at efficiency, segmentation errors are not smoothed out by applying morphological filtering but taken into account later on. All B_is are gathered at a single computer for the next computational step. Before transmission, B_is are compressed using Run Length Encoding.

3.1.2 Volumetric reconstruction

To compute the occupancy grid V, images B_i are gathered at the GPU of a single computer. The value of each voxel of V is independently computed as follows. Let the 3D location $\vec{x}$ be called *potentially visible* from view i if its projection, $P_i(\vec{x})$, occurs within the field of view of view i. Then, if no errors would occur during the foreground segmentation, an occupied voxel $\vec{x}$ would project only to foreground regions of the views i' that it is potentially visible from. In this case, it would hold that:

$$s(\vec{x}) = \sum_{i'} (B_{i'}(P_{i'}(\vec{x}))) = \max(i'), \quad (1)$$

for all voxel centers $\vec{x}$ within the visual hull of the visitors, while for any other location, $s(\vec{x})$ would have a smaller value. To compensate for errors in foreground segmentation a more lenient rule is applied. A voxel is considered as occupied if it projects to a foreground region in all but μ views that it is potentially visible from. Thus, $V(\vec{x})$ is set to '1' if it holds that $s(\vec{x}) \geq max(i') - \mu$ and to '0' otherwise. By employing this rule, up to μ views may have a foreground detection error at pixels $B_{i'}(P_{i'}(\vec{x}))$. In some cases this might dilate the visual hull by a voxel. Considering the intended use of the visual hull computation, we found this inaccuracy to be acceptable in terms of person localization. Conversely, if this relaxed constraint is not employed, a failure to segment a person from the background in a single view could annihilate its 3D reconstruction.

The proposed methodology for 3D reconstruction and its implementation result in an efficient and scalable system. The achieved rate of updating V (10–30 Hz, depending on hardware configuration) allows for the assumption of motion continuity during tracking. At the same time, the system can support reconstructions of relatively high resolution, thereby facilitating robust disambiguation of visitors. A more detailed analysis of the implementation and its performance evaluation can be found in [38]. Figure 4 illustrates the resulting reconstruction for a case where the persons occlude each other in all views and failures of foreground detection are multiple.

3.2 Visitor detection and localization

Visitors are localized based on the information provided in V. As in [10,15,20,23,28], a 2D image F is formed from V that is aligned with the ground plane and V. Essentially, F is a 2D histogram: a pixel (or bin) in F counts the occupied voxels in V along a direction perpendicular to the ground plane. Persons appear in F as intense and size-dominant blobs, with their intensities and areas proportional to the volume they occupy. The sum of intensities within a blob is proportional to the occupied volume by the visual hull that gives rise to the blob. Localizing a blob in F is equivalent to registering the location of each human in the ground floor reference frame.

Depending on the number and the placement of the cameras, voxels of V are visible from a different number of cameras (see Fig. 5). In the corresponding regions, the signal in F becomes weaker as voxels are summarized along a smaller height. To compensate for such variations, the values in F are normalized as follows. A normalization map S of equal

Fig. 4 Original images I_i, foreground images B_i and volumetric reconstruction for a challenging scene imaged by four cameras. Persons occlude each other in all views and are imaged against cluttered background and uneven illumination, resulting in inaccurate foreground detection. Nevertheless, the obtained reconstruction accuracy is sufficient for robust person tracking

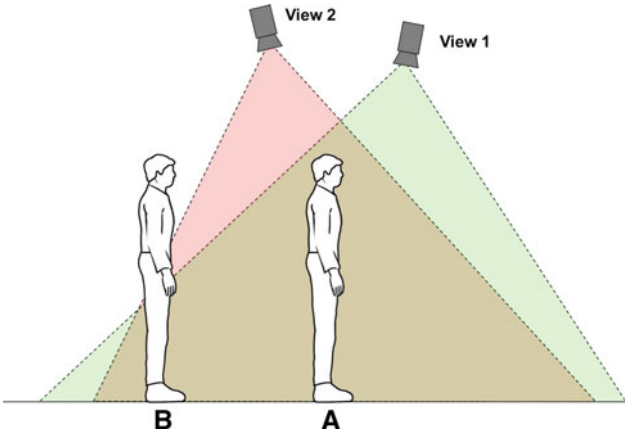


Fig. 5 Visual coverage and person tracking. When the visitor is at location A he gives rise to more occupied voxels than when in location B, as in B his torso is not visible by any view and is not reconstructed

dimensions to F is computed at initialization time. Each pixel in S summarizes the number of voxels along height which are potentially visible by μ or more views. At run time, the value of each pixel $\vec{q}$ in F is normalized as:

$$F_n(\vec{q}) = \begin{cases} F(\vec{q})/S(\vec{q}) & if\ \ S(\vec{q}) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Noise is initially suppressed in F_n by thresholding small values, followed by Gaussian smoothing. In the resulting image, conventional connected component labeling is performed to detect blobs. Blobs that correspond to very small volumes are filtered out, as they are typically due to reconstruction errors. Figure 6 demonstrates the steps of this operation. The detected blobs are directly localized as persons on the ground plane of the scene and their centroids and silhouettes are the measurements passed on to the tracking module (see Sect. 3.3).

3.3 Person tracking

A blob tracker is applied to F_n in order to identify blobs and associate them with individual visitors. More specifically, the tracker in [2] is modified to track intensity blobs in F_n, rather than skin-colored blobs in color images for which it was originally developed. This tracker may track a potentially varying number of targets and is robust to transient localization failures. Even more importantly, it is designed

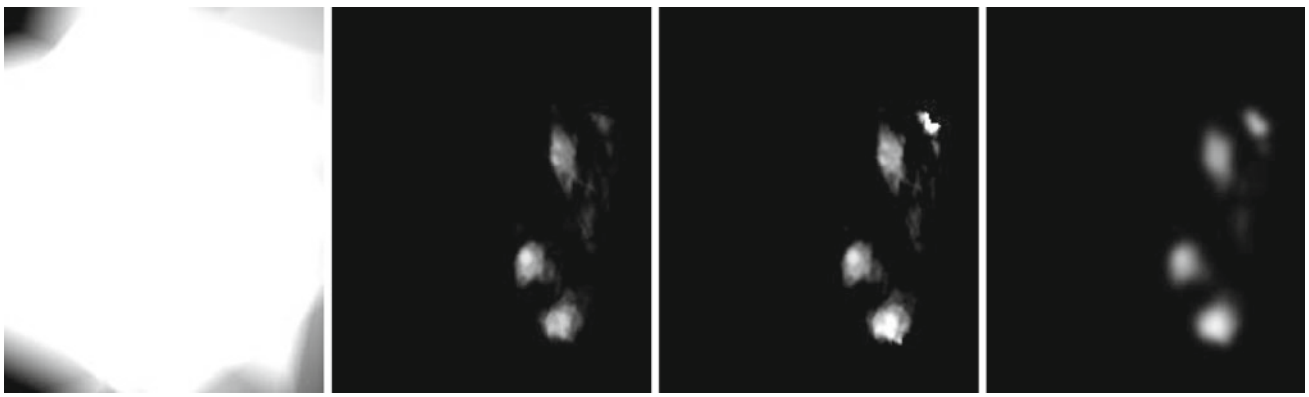


Fig. 6 Visibility based normalization of F. From *left* to *right*, normalization map S, F, F_n and F_n after thresholding and Gaussian smoothing. The normalization of blobs at regions of relatively lower visibility (*top* and *bottom right*) increases tracking robustness

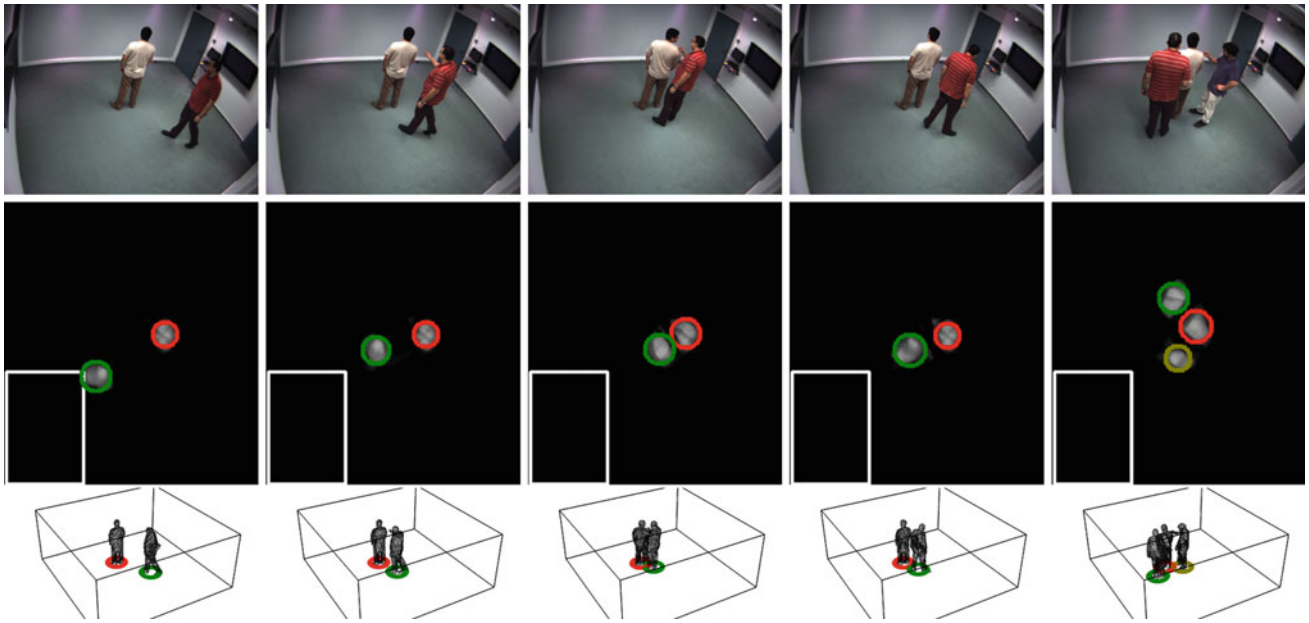


Fig. 7 Person localization and tracking. Original images from one out of eight views (*top row*), blob tracking in F_n (*middle row*), and 3D reconstruction (*bottom row*), for frames 150, 156, 164, 179, and 475 of a sequence (*left–right*). Tracking results are rendered as circles, superimposed on F_n and on the ground plane of the 3D reconstruction. *Circle colors* correspond to track ids. In the example, first two and then three persons produce a single connected component in V. Tracking retains the correct id for all persons although the visual hulls of certain individuals are merged (color figure online)

to retain the tracking of blobs even if they occur merged for long temporal intervals. In this way, person tracking is robust even if visitors come close enough to give rise to a single connected component in V and F_n (see Fig. 7). The output of the tracker is a list of person ids and the associated user location estimates. This is transmitted to the application through the use of a custom middleware layer.

The robustness of tracking has been further enhanced by exploiting contextual information relevant to the particular application domain. First, a person may appear and disappear in the scene only at the entrance/exit of the room. Hence, the id of a tracked person that is lost elsewhere, is not immediately removed from further consideration. If, later on, a blob that is not already associated with the id of another tracked person appears close-by, then the id of the temporarily lost person is assigned to this blob (Fig. 8). Second, when visitors enter the room clustered together (e.g., holding hands, a child being carried by a parent, etc), the tracker assigns to them a single identity. If later on the corresponding blob splits into different ones, then the tracker assigns to them separate identities, which, nevertheless, inherit the properties (i.e., history, language) of the id from which both originated (see Fig. 9). To avoid false positive person detections due to transient reconstruction artifacts, the presence of new blobs is required to show sufficient temporal persistence (i.e. 2–3 s) to be assigned with an id.

The robustness of tracking also improves with increasing the density in temporal sampling, as well as with increasing the spatial resolution of the reconstruction. Observing the scene at a high framerate (>10 Hz) casts blob motion in F_n smooth and continuous and, thus, supports the unambiguous tracking of blobs. Fine granularity ($1\ cm^3$) in occupancy estimation is important as proximate blobs will merge in F_n only if they occur closer than voxel size. Therefore, the computational efficiency achieved by the reconstruction approach in Sect. 3.1 is not only important to brisk system response but also to tracking robustness.

4 Interactive application

In a prototype installation, the developed system was used for the multi-user interactive exploration of the "Wall-painting of the Royal Hunt". This painting was found in the tomb of Philip II at Vergina, ancient Aigai. To date, this is the largest ancient Greek painting that has been found, its length exceeding 5.5 m. It is of great archaeological significance and widely admired as a rare masterpiece of ancient Greek art. It is in a quite deteriorated state and depicts ten hunters chasing five different animals in a complex landscape.

The digital representation of the painting is conceptually separated in five zones perpendicular to the display,

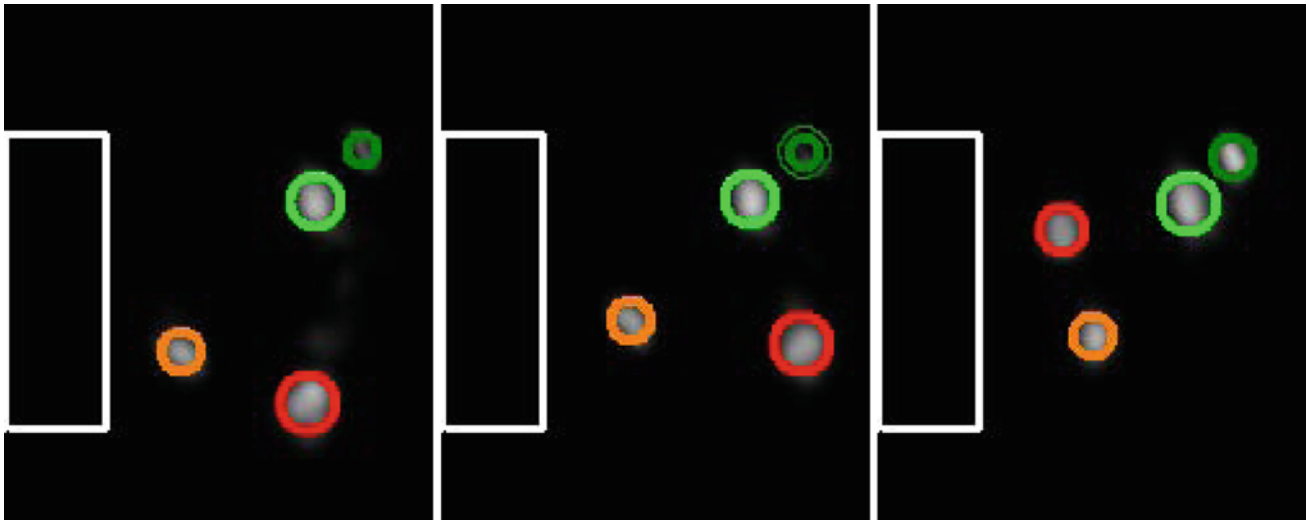


Fig. 8 Tracking example, showing the temporal evolution of F_n (*left–right*) with tracking results superimposed. The *white rectangle* marks the entrance of the room and circles correspond to tracked blobs. The visitor marked with a *green circle* (*left image*) is temporarily lost by the tracker (*middle image*). This is signified by an additional circle. However, in a subsequent frame (*right image*) the visitor is again visible and the system reassigns the same id to the closest re-appearing blob

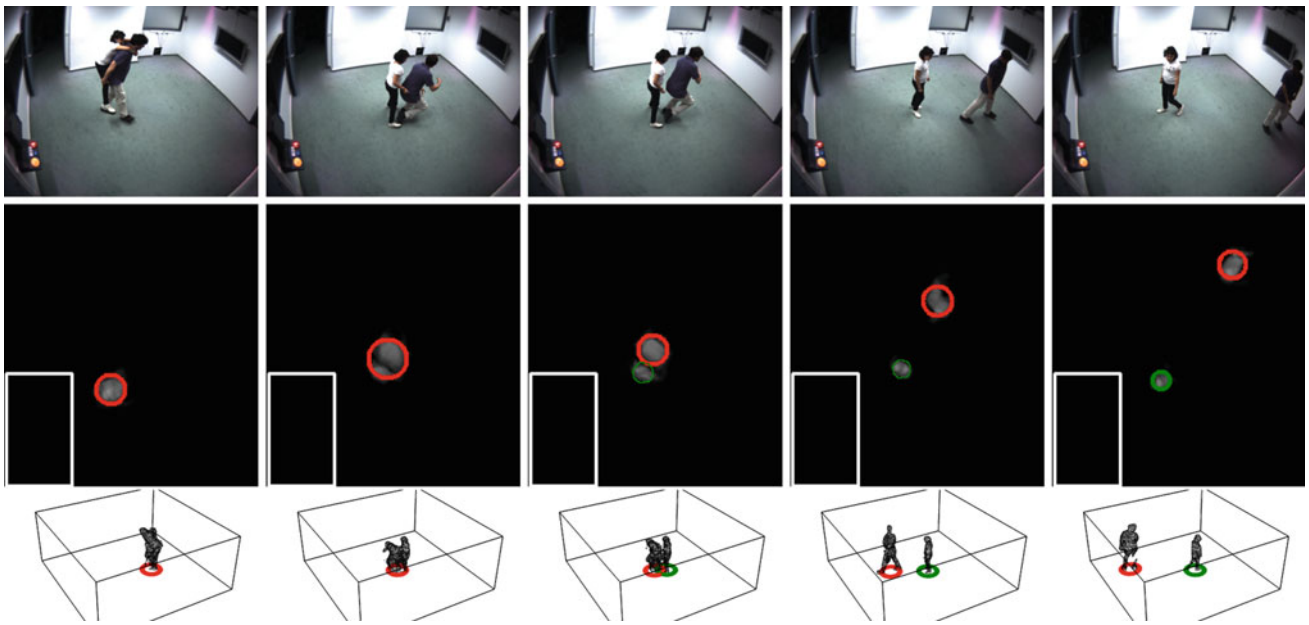


Fig. 9 Tracking example (*left–right*). Two persons enter the room clustered together, forming a single connected component in F_n and V. *Top row* original images from a system camera, out of a total of eight views. *Middle row* blob tracking in F_n. *Bottom row* 3D reconstruction. Tracking results are also rendered in the 3D reconstruction as circles on the ground plane. When the blob under the red circle is split into two, the tracker assigns a new id to the new blob, but only after this new blob persists for a sufficient time interval. This is represented as a thickening of the *green* circle, in the images of the middle row (color figure online)

based on a semantic interpretation of the themes that appear in the painting. The room is also conceptually separated in four rows parallel to the display, which correspond to different distances of observation. Figure 10 presents an illustration of the resulting grid. To prevent from continuous alternation in the case of visitor lingering across the boundary of a cell of the grid, the cell size is assumed magnified by 10% comprising a grid of partially overlapping slots. At the room entrance, signs guide English-speaking visitors to enter the room by moving rightwards and Greek-speaking visitors to enter the room by moving leftwards. The corresponding textual annotation for each distance is presented at the bottom of the screen (see Fig. 11).

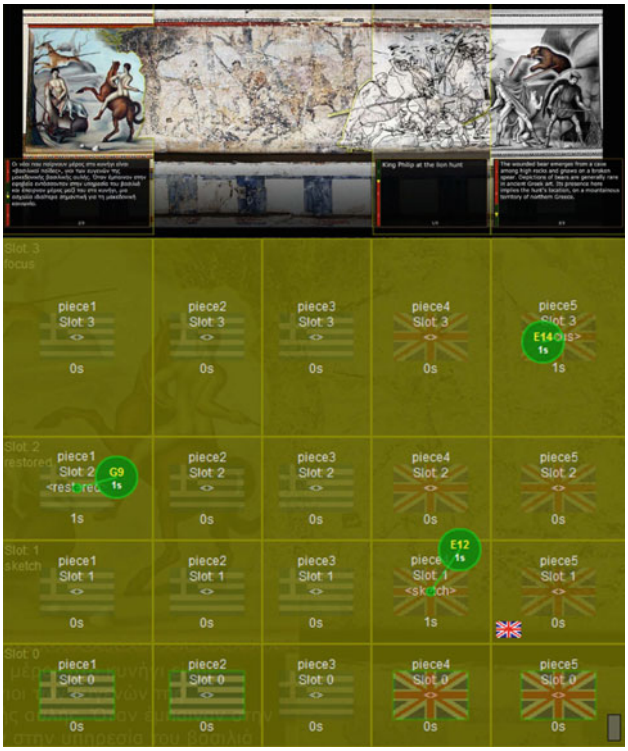


Fig. 10 Modulation of projected content (*top*) based on visitor locations (*bottom*). The *top image* shows the content presented on the display, for the arrangement of visitors that is illustrated at the *green* grid at the bottom. The grid at the *bottom* corresponds to the ground floor of the room and illustrates the spatial extent of the regions and the state of the application. Matrix cells represent the 5×4 interaction slots and circles visitor locations. In this example, the 2nd and 3rd cells from the left are inactive, as visitors are located at columns 1, 4 and 5 (color figure online)

Within the context of a zone, the presented content is varied graphically and conceptually according to the distance of observation. When a visitor enters a zone, the presented content matches the viewing distance. The visitor has the capability to explore the corresponding theme by stepping back to get a more abstract view or step closer and focus on the details of the exhibit. This capability is communicated to visitors through real-time visual feedback (see Fig. 11), which helps them perceive the distance they have to walk in order to move to the adjacent row. When idle and upon visitor entrance, the system presents the wall-painting in its current state (Fig. 11a). As one or more visitors approach the display, graphical outlines are superimposed to the corresponding region(s) of the display reviving the deteriorated forms (Fig. 11b). In the next row, the system presents a fully restored version of the painting (Fig. 11c). In the closest row, the restored version is grayed out, and a specific detail is highlighted, using a combination of color and animation (Fig. 11d).

When multiple visitors stand in the same zone (i.e. in the same column of the grid in Fig. 10), the person closest to the display determines the content of the presented textual annotation. When this person leaves, the next in line (if any) becomes the closest one to the display. An adjustment of system response introduced after evaluation (Sect. 5.3), concerned latency in performing updates on the display. It was observed that visitors may rapidly cross the room, e.g., to join a friend or move to a specific point of interest, thereby causing an update of all intermediate display segments. Thus a minimum dwell time was adopted, for granting control over a location.

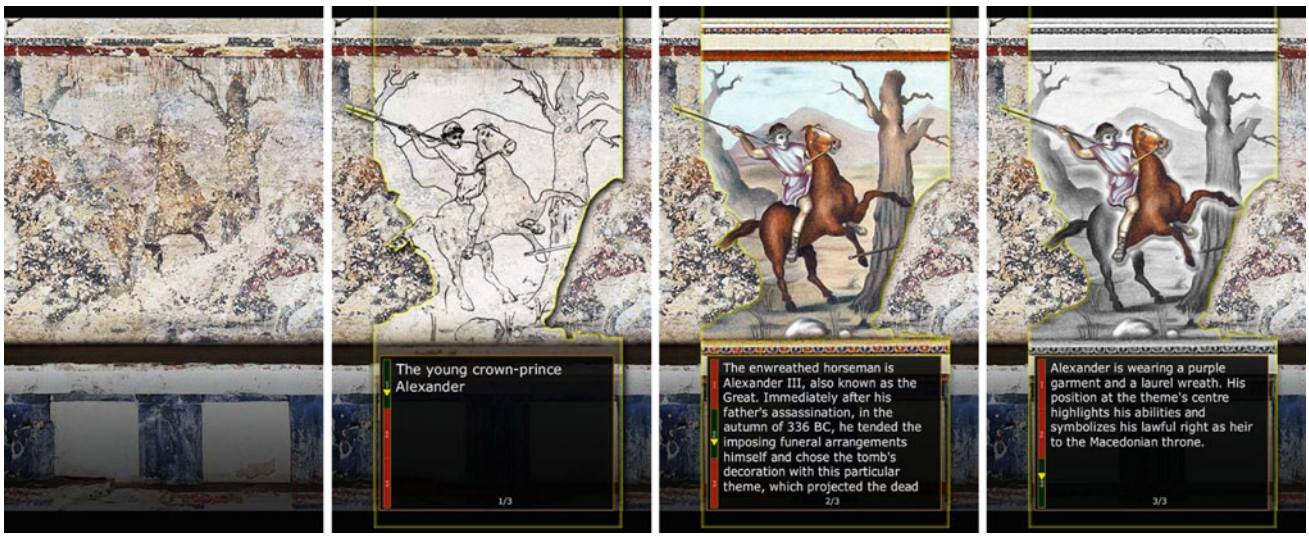


Fig. 11 Graphical presentation of a vertical segment of the display based on its distance from the visitor. *Left* to *right*: **a** initial state, **b** intonated outlines, **c** restoration, **d** detail highlight. Next to the textual information, *three rectangles* and a *yellow arrow* are presented. The *triangle* represents the visitor's distance from the wall, while the rectangles represent the three available rows of information. The active slot is presented in *green color*, while the inactive ones in *red*. The triangular marker is updated in real-time providing visual feedback to the visitor (color figure online)

By tracking visitors and assigning a unique identifier to each one, the system also retains attributes for each corresponding blob in F_n. Using an attribute for the language, which is set upon visitor entrance, the textual components of the presented content are provided in the language selected by each visitor. Similarly, the system keeps track of the slots visited by each user and provides additional and more detailed information when a visitor has viewed all content in a zone and revisits it. Several "pages" of information are assigned to each slot of a zone and presented when a visitor revisits the slot or after a predefined time period has passed; correspondingly, the system keeps track of the pages each visitor has seen. Finally, the time each visitor has spent in each slot is logged in order to gather and visualize statistical data concerning the visits.

5 Evaluation

In this section, we present the results of experiments evaluating the proposed system from three perspectives. First, the computer vision system is evaluated with respect to its accuracy and precision in visitor tracking. Second, the vision system is benchmarked as to its computational performance. Third, the integrated application is evaluated in terms of usability.

During this evaluation, some of the experiments were performed based on a laboratory prototype. However, most of the testing and evaluation was performed through the installation of a final prototype at the Archaeological Museum of Thessaloniki, Greece. This prototype, "Macrographia", is part of a permanent exhibition of prototypical interactive systems with themes drawn from ancient Macedonia. This installation resulted in a series of in situ experiments with visitors and museum staff in real-life conditions. Demonstration videos of system operation and information about the exhibition are available online [22].

5.1 Visitor tracking

Two experiments were performed to evaluate the computer vision system in terms of tracking accuracy and robustness. The first experiment was performed by employing the laboratory prototype, where the recording of data and the subsequent annotation of these data with ground truth was possible. The second experiment was based on field data acquired at the museum installation. This experiment inquires the efficacy of tracking from the application perspective, i.e. whether the visitor encountered a system failure or not. An inquiring parameter in the above experiments was also a quantification of the limitations of the approach and setup.

5.1.1 Laboratory experiments

The laboratory setup is essentially identical to that of the museum installation (see Fig. 3). During the experiment, the system was storing the acquired images (achieved at a rate of 11 fps) which enabled the subsequent annotation of these data with ground truth.

System precision was quantified using the MOTP metric [4] on datasets where visitors were instructed to walk at marked spots on the floor, whose locations were a priori measured. For three and five persons, the localization error was in the order of 2–4 cm, respectively. In effect, voxel size is sufficiently small with respect to localization requirements. On the other hand, although precision is sufficient, small fluctuations of visitor location estimation upon a cell boundary are possible and are addressed at the application layer (see Sect.4).

Tracking accuracy was quantified with the MOTA metric [4] on four datasets of increasing complexity. Initially, a baseline dataset $D0$ was recorded where a single visitor enters the setup, visits practically all of its locations, and exits. In the $D1$ dataset (662 frames, 1181 tracked objects), two visitors walk together and then one of them walks at a corner of the room that constitutes a blind spot (not sufficiently visible to be reconstructed). The other visitor repeatedly walks towards and away from him. In the $D2$ dataset (904 frames, 1,909 tracked objects), four persons visit the installation and perform typical walkthroughs that visitors perform in the museum installation. Often, a visitor stands at a corner of a room while another passes closeby or is in contact, thus increasing the possibilities of a tracking mismatch. Finally, the $D3$ dataset (1,011 frames, 4,945 tracked objects) is overly challenging as seven persons move rapidly, getting in contact sometimes altogether and constantly occluding each other.[1] We included this overly difficult data set in order to demonstrate the performance of the proposed human tracking approach in situations that go beyond a typical museum scenario.

Based on the ground truth annotations, we measured the tracker's misses, false positives and mismatches in all frames of datasets $D0$–$D3$. A distance threshold of 0.5 m indicated whether a tracking error would be considered as a tracking miss rather than tracking imprecision.

Figure 12 shows characteristic images from these datasets and Table 1 reports the obtained MOTA values, as well as tracking misses, mismatches, and false positives. In $D0$, no misses, false positives, or mismatches occurred, even when the visitor entered and exited blind spots. Hence MOTA was 1 for this dataset. Regarding $D1$ and $D2$, although the tracker is able to correctly assign the ids, there exist occasions where

[1] Datasets $D1$, $D2$ and $D3$ can be downloaded at http://www.ics.forth.gr/~zabulis/tracking.html

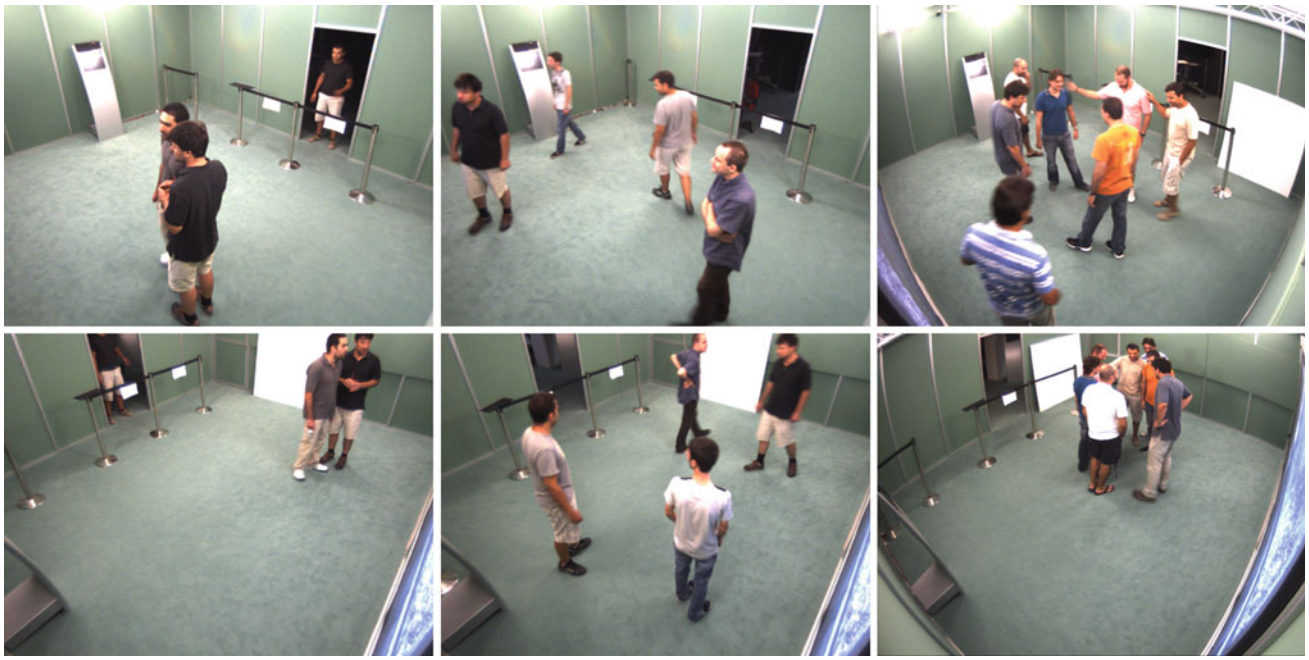


Fig. 12 Characteristic snapshots from datasets $D1$, $D2$ and $D3$ (*left*, *middle* and *right column*, respectively)

Table 1 Quantitative evaluation of tracking performance for the ground truth datasets (see text)

Dataset	MOTA	Miss	Mismatch	False positive
$D0$	1.000	0.000	0.000	0.000
$D1$	0.935	0.000	0.000	0.643
$D2$	0.986	0.007	0.020	0.002
$D3$	0.712	0.149	0.019	0.147

lack of visibility produces false-positive detections (assignment of a tracking hypothesis to a non-existing person in the scene). These false-positives are mainly due to a volumetric artifact occurring at the dynamic occlusion of an area, which has been studied in [13]. Due to the adequate number of views these errors are rare and mainly occur when visitors are almost in contact with a wall. In this case, a spurious blob occurs between the visitor and the wall (see [13]) and inherits the attributes of its parent (see Sect. 3.3); the blob disappears when the user leaves the location. These false positives do not affect the system's response, as such a falsely-detected "visitor" has the same location and preferences with the original user. The increased complexity of $D3$ results in lower MOTA values for this dataset. In addition to the above error, in datasets with multiple persons, mismatches occur when visitors are lost by the tracker in a low visibility region of the scene and are, then, mismatched with another close-by visitor. In this case, mismatches in id tracking of blobs case apparent errors in system response.

An interesting case is encountered when visitors enter the room clustered together, which initially evokes a single blob in the tracker's representation, F_n. The blobs resulting after the split of such blobs inherit the attributes of the parent blob (see Sect. 3.3) and, thus, corresponding users obtain the intended system response. Though not affecting system performance, this tracking failure reduces the MOTA score, as during the presence of a single blob, misses are counted.

5.1.2 Experiments with field data

Another quantitative evaluation of the human detection and tracking performance was obtained based on the museum installation of the prototype. The experiments were performed after more than three months of continuous operation of the installation. A period of 186 days (4 February 2011–18 August 2011) of normal museum operation was considered. During this period, 9,873 museum visitors visited "Macrographia", with the average visit duration being 165 s. A significant percentage of the visitors were school students in organized visits that gave rise to a crowded workspace and challenging visitor configurations. Overall, visitors ranged considerably in their nationality, gender, education and, most importantly, age. Museum visitors were unaware of the execution of this experiment. In that sense, the experimental conditions constitute an unbiased, real-life field test. To automate the process of data gathering and assessment, during this test the behavior of the tracker was logged. Then, we considered the following failure cases:

Table 2 Quantitative assessment of detection and tracking errors (see text)

Error type	Count	Percentage
A	513	5.1
B	566	5.7
C	1,197	12.1

- *T*ype A errors: The first occurrence of a visitor id is observed in the inner part of the workspace and not in the entrance; i.e., a detected visitor never enters the room.
- *T*ype B errors: The last occurrence of a visitor id is observed in the inner part of the workspace and not in the exit (i.e., a detected visitor never exits the workspace).
- *T*ype C errors: A visitor is observed to perform unlikely motions between two frames. Assuming a maximum visitor speed, such situations are attributed to tracking failures, i.e. misses.

Note that the case of a visitor appearing and disappearing in the workspace without ever entering and exiting it (spurious visitor detections) are already counted in type A and B errors. Table 2 summarizes the errors detected in this experiment.

Though marked as tracking failures, in practice, we have observed that less than a third of the instances of the above errors lead to failures to the user experience. For example, type *A* errors include blob split that occurs when two persons enter the room together (see Sect. 3.3) without necessarily implying an application error. In addition, even though a single person may be initially detected further than the exit, the system recovers in most cases by assigning the linguistic preference according to the closest entrance point. Similarly, type *B* errors are counted even for cases of visitors exiting the installation together, where a tracking loss has no consequences to the users' visit. Finally, type *C* errors are mainly encountered when a visitor is transiently lost by the tracker but eventually kept tracked by the system. As the system recovers from such errors, they either have no effect in system behavior or experienced as a transient system delay.

Furthermore, in Fig. 13, the number of these errors is plot as a function of the number of visitors within the workspace at the time instance that the error occurred. It can be verified that most errors occur when the workspace is crowded. Errors for six and seven visitors are fewer, as usually, up to five persons enter the room. Overall, human detection and tracking performs very well, especially if someone considers the challenging experimental conditions. The tests also demonstrate that four cameras are sufficient to disambiguate three persons even if clustered closely, while with eight cameras, up to seven visitors can be adequately tracked. As the number of required cameras scales with the number of visitors, it is important that the required computational resources scale in a linear relation, a fact that is verified in Sect. 5.2.

Finally, as an indirect, yet vivid indication of the human localization accuracy, Fig. 14 (top) shows a color-encoded histogram of visitor localization results. More specifically, the map shown in this figure corresponds to the floor map of the workspace. Warm colors correspond to workspace locations where visitors spent most of their time during the full 186 days field experiment. The structure of this map bears a close resemblance to the defined application grid (see Fig. 10).

The logging of such data provided an additional visualization tool of visitor behavior. This is quite important to the museum as these data can be used to optimize the presentation and, consequently, the user experience.

5.2 Computational performance

To evaluate the computational performance and the scalability of the proposed approach, we have considered several system configurations. This performance evaluation is focused on the volumetric 3D reconstruction because this accounts for more than 95% of the total system computational load.

Table 3 summarizes the execution time of the proposed 3D volumetric reconstruction method in four different configurations. The reported figures correspond to the time required to process a multiframe (in milliseconds) as a function of the number of employed computers and image resolution. The results confirm that execution time on the GPU scales well

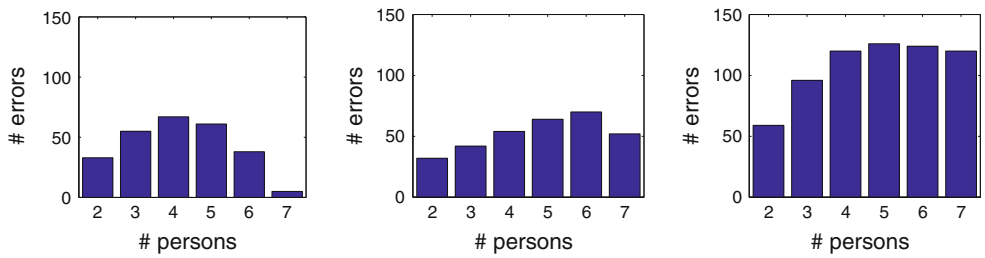


Fig. 13 Number of errors as a function of the number of persons in the workspace when the error occurs. Histograms from *left* to *right* correspond to error types A, B, and C, respectively (see text)

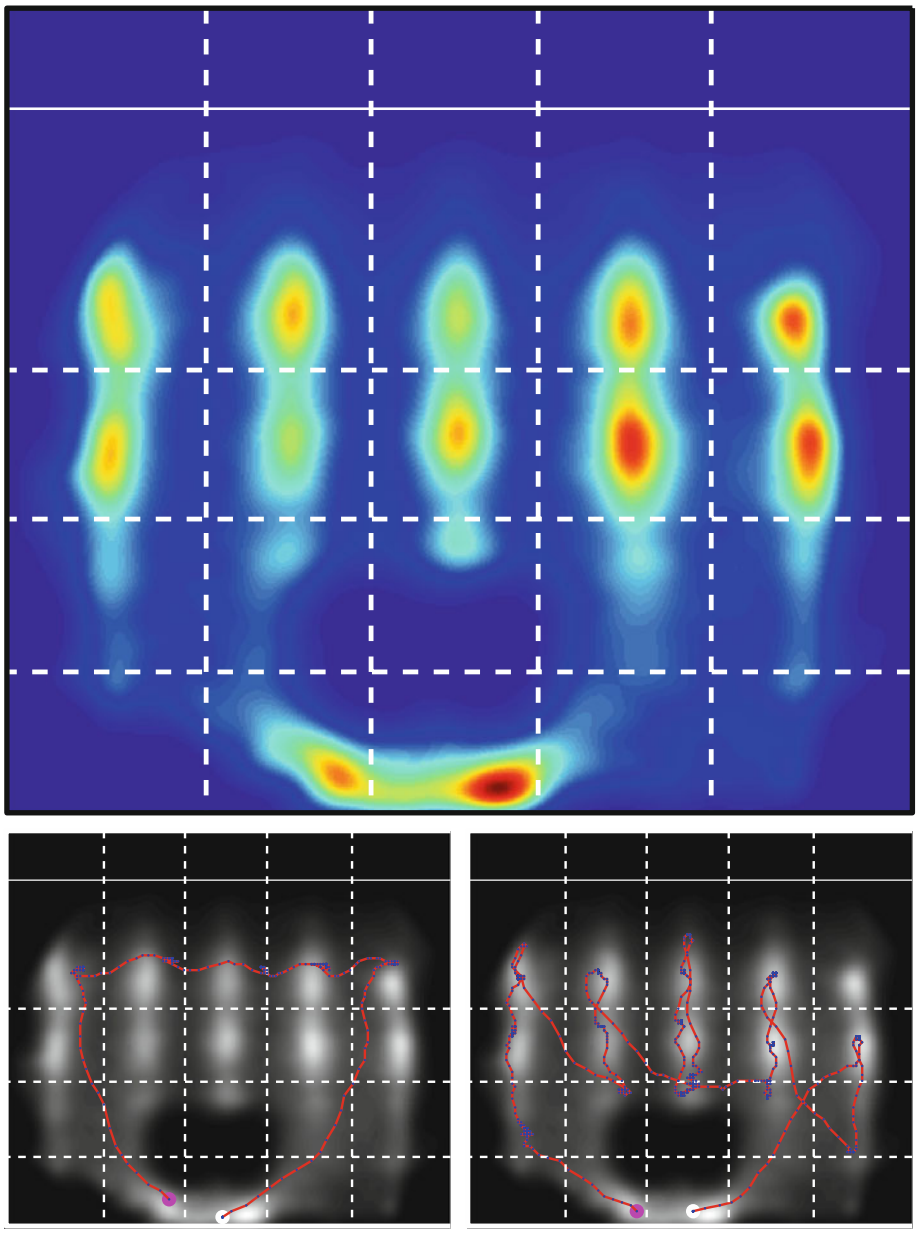


Fig. 14 *Top* A view of workspace "hot-spots" during the 186-day field test. Warm colors correspond to the most visited workspace locations. *Dashed lines* indicate the limits of the workspace slots and zones. The display is on the top boundary of the figure, while the entrance on its bottom boundary. The thin continuous line near the top of the figure, indicates the closest distance at which visitors are allowed to approach the display. *Bottom* sample visitor trajectories, superimposed upon a grayscale version of the above image (color figure online)

Table 3 Computational time (in milliseconds) for volumetric reconstruction as a function of image resolution and number of employed computers

	1	2	4
320 × 240	29.4	17.2	9.7
640 × 480	42	25	14

Table 4 Comparative evaluation of volumetric reconstruction versus state of the art

	[18]	[32]	[11]	Proposed
ms	40	72	33.3	25
G FLOPS	1,614	933	836	894
Computers	5	1	11	2

with the number of computers and with image resolution. Table 4 evaluates the performance of the proposed volumetric reconstruction method in comparison with other state of the art methods. In all cases, the voxel space was composed of 2^{11} voxels, eight views were considered and the resolution of each input image was equal to 640 × 480 pixels. The three rows of the table show the time required to process a frame (in milliseconds), the amount of computational power utilized and the number of computers employed. The results indicate that the proposed approach improves the state-of-the-art

by being more efficient while requiring less computational resources.

The significant differences in performance are due to several reasons. To compute $V(\vec{x})$ all B_is are gathered over the local network on a single computer. In contrast to [18], we parallelize the background subtraction stage and transmit the Run Length encoded images B_i for the computation of V. This costs significantly less than the transmission of intermediate computation results for all $\vec{x}$ in V as in [18]. The system in [1] exhibits a minimal communication cost, as it transmits silhouettes. However, it parallelizes computation only per-view, as opposed to massive, per-voxel, parallelization. Additionally, its implementation is more complex and requires CPU processing. Thus, multiple computers (11 dual-core processors for 8 views) are required to achieve a satisfactory framerate. In the proposed implementation, volumes of arbitrary size and resolution can be processed by partitioning V and, thus, computation of wide areas can be achieved even with modest computational resources. In [18,32] V is processed as a single block, thus, the dimensions of the voxel space are constrained by the GPU's memory capacity.

In the employed museum installation, a volume of $6 \times 6 \times 2$ m was covered with voxels of 1 cm^3, yielding a framerate of $\approx$ 15 Hz. The latency between a person's motion and the reception of the corresponding event was $\approx$ 140 ms.

5.3 Usability evaluation

The usability of the developed computer vision system and application was extensively evaluated both in laboratory conditions but also in an actual museum setting.

5.3.1 Formative evaluation

Laboratory evaluation was a continuous and ongoing process aiming at testing successive versions of the developed prototypes. Several of the enhancements (i.e. as those in Sects. 3.2 and 3.3) were based on observations from this evaluation. During the development of the system, due to the formative nature of this type of evaluations, we selected to use ethnographic field methods [5], using a combination of the "observer participant" and "participant observer" approach. Participants were invited on an ad hoc basis, were naive to the experimental hypotheses, and exhibited a large diversity in age, educational, and cultural background. During a 6-months period, more than 200 persons have participated in the informal laboratory usability evaluation. Evaluation sessions involved a facilitator accompanying the visitors, acting as a "guide" and another, distant observer discretely present in the exhibition space. Since there were numerous evaluation sessions, alternative approaches were used, depending on the characteristics to be assessed. When the interactive behavior of the exhibit was tested, the facilitator would first provide a short demonstration to the participants and then invite them to try it for themselves. Alternatively, when ease of use and understandability were assessed, the facilitator would prompt participants to freely explore the exhibit without further instructions. During and after the sessions, the facilitator held discussions with the participants eliciting their opinion and experience, identifying usability problems, as well as preferences. At the end of each session, the two observers would discuss about it, often reenacting parts of it, in order to clarify or further explore the findings.

5.3.2 Summative evaluation

Using the final version of the system, summative evaluation sessions took place both at the premises of ICS-FORTH, as well as at the Archaeological Museum of Thessaloniki. For the purposes of evaluating the experience of the users with the system, a 13-item attitude Likert scale questionnaire was created which was based on Brook's System Usability Scale (SUS) questionnaire [7]. This type of questionnaire was chosen because its questions cover most aspects of usability such as system complexity, learnability, likeability, and effectiveness. The questions of the original questionnaire were adapted to fit our application and three more questions were added to measure specific aspects of user experience and satisfaction that were of particular interest. The questionnaire is divided in both positively and negatively stated questions. The participants marked each question with a number from 1 (strong disagreement) to 5 (strong agreement). The scoring of the positive questions is produced by taking the number that the user marked and subtract 1 from it, while the scoring of the negative questions is produced by subtracting the marked number from 5. This way, the best score that any question can get regardless of whether it is positive or negative, is 4 points. The sum of the points of each questionnaire is then normalized to come up with a percentage score as an indication of the overall system usability. The questionnaire also included a part for collecting background information about the respondents.

Details on the usability tests performed at the two different installations are provided in the following sections.

Evaluation at ICS-FORTH

The laboratory room that housed the usability evaluation was set-up in a way to resemble as close as possible the actual museum exhibit where this application is currently being housed. The size of the room was only slightly bigger than the actual museum room. In addition, one camera was set up in one of the corners of the room to record each session and the comments made by the participants during the evaluation.

Twenty two volunteers participated of which 13 were males and 9 were females. The average age of the participants was 31.7 years old, the youngest being 18 and the oldest 41 years old. 17 participants stated having a "high" level of computer experience, 5 "Intermediate" level, 3 "Low" level, and 2 "No" computer skills at all. Also, no one's studies or profession was related to archeology.

Before each session the evaluator explained to the participants the purpose of the testing and the process that was going to be followed. After that, the participants were asked to sign a participation consent form and to fill out a short questionnaire with personal information. The participants were then asked to enter the room in groups of two or three, one of them usually assuming the role of the English speaking visitor. No specific instructions were given to the participants as to how the Macrographia system actually worked or displayed the information. The purpose of not giving detailed instructions was to examine if the users were able to understand how the system actually worked and how to retrieve the information that corresponded to each section of the Macrographia. The participants were asked to use the "Think-Aloud" method [24] and express their thoughts and comments freely.

Evaluation at the Archaeological Museum of Thessaloniki

Twenty two questionnaires were filled-out by 15 visitors and 7 guards at the Archaeological Museum of Thessaloniki. Of those, 11 were males and 11 were females. Their average age was 34.6 years, the youngest being 21 and the oldest 56 years old. Four of them were close or over 50. 17 participants marked that their level of computer expertise was "Intermediate", 4 participants marked "High", and only 2 marked "Low". 13 participants had a background in archeology either in their profession or studies.

In contrast to the in-house evaluation sessions, the museum evaluation was "not administered". Museum visitors were asked by one of the guards before leaving the area, if they were willing to fill in the questionnaire, without any further help or guidance in order to maximize the objectivity of the results.

5.3.3 Evaluation results

The detailed quantitative results for each separate statement from both set-ups are presented in Tables 5 and 6. The overall usability of the system was rated high in both studies: 82.8% in-house and 80.8% at the museum.

The questionnaire results were also supported by data collected by analyzing the recorded sessions or interviewing the participants for the in-house evaluation, and by hand-written comments of the museum visitors. Overall, the comments that were made by the all participants were highly majority positive. Most of them were impressed with the system's ability to track accurately their position in the room and display the information in the language that was chosen. Even though very little instruction was given to them before they entered the room, they all managed to understand that their movement was tracked and that the information changed according to their distance from the screen. As a result, all users were able to read all the information that was presented in each section of the Macrographia. They also offered some suggestions on how to improve it. For example, a few users suggested that it would be better if the text font size changes from larger to smaller as the user approaches the display.

Finally, a number of positive comments were provided in written on the questionnaires such as: "The interactive system is very interesting. It helped me a lot to get important knowledge and I believe it is extremely useful", "Congratulation for this effort, I wish that all museums had rooms like this. Excellent interactive guidance", "The wider use of this system could help children to get in touch with Greek antiquity in an easy way. Excellent!!!" and "I liked it! It is very useful to have such kind of electronic material in museums, it helps the visitor understand better.".

Table 5 Evaluation results (average scores and standard deviation) for the positive statements *(best score is 4 = totally agree)*

Positive questions	In-house	Museum
1. I liked it	3.45 (1.05)	3.45 (0.59)
3. It was easy to use	3.59 (0.51)	3.50 (0.74)
5. It responded correctly to my actions	3.18 (0.95)	3.23 (1.02)
7. It was fun	3.50 (1.05)	3.27 (0.82)
9. The content was of high quality	3.55 (0.80)	2.86 (1.08)
11. It helped me learn information about the"Wall-painting of the Royal Hunt"	3.64 (0.72)	3.41 (0.67)
13. The concurrent presence of multiple visitors did not cause any problems	3.23 (0.87)	3.09 (0.75)

Table 6 Evaluation results (average scores and standard deviation) for the negative statements *(best score is 4 = totally disagree)*

Negative questions	In-house	Museum
2. I would not recommend it to my friends	3.59 (0.85)	3.64 (0.58)
4. It was hard learning how to use it	2.73 (1.42)	3.45 (1.10)
6. It did not work as I expected	3.18 (1.09)	2.77 (1.41)
8. I got confused while using it	3.14 (1.35)	3.36 (1.22)
10. It did not respond timely to my actions	2.91 (1.37)	2.45 (1.50)
12. Using such a system does not benefit my experience while visiting the museum	3.45 (0.96)	3.50 (1.05)

6 Summary

This paper presented the use of computer vision towards supporting multi-user, location based and non-instrumental

exploration of large-scale artifacts. A 6-month period of evaluations in laboratory conditions and in a museum setting reveals that the developed system achieved the goal of providing engaging and entertaining educational experiences to its visitors. The system is capable of tracking persons robustly and at high framerate. The requirements for processing power scale well with the amount of data to be processed, or otherwise, with the spatial extent of the area to be covered and the number of the employed cameras. Additionally, the system adapts to the availability of resources, either for larger installations or for cases of hardware failure where computational nodes may be fewer.

One key direction for future work concerns the ability of the system to operate in less constrained environments. Towards achieving open area installations and to cope with uncontrollable illumination changes and backgrounds, we are experimenting with the incorporation of depth from stereo techniques for 3D reconstruction. In addition, visitor detection and tracking is going to be enhanced by exploiting color besides the use of purely geometric information. Probably, the most important future research direction is the incorporation of additional, vision based extracted information regarding the visitors and their posture/behavior in the observed workspace. Currently, the interaction is based on the location and the walk-though trajectories of visitors. Clearly, the real-time reconstruction of the visitors provides an abundance of useful and directly exploitable information. As a concrete example, head pose estimation methods that rely on this kind of reconstruction [41] can be used to determine where each visitor is facing at, and therefore, enable more intricate visitor-exhibit interaction.

Acknowledgments This work was partially supported by the FORTH-ICS internal RTD Programme "Ambient Intelligence and Smart Environments" and by the IST-FP7-IP-215821 project GRASP. The restoration of the wall-painting was provided by George Miltsakakis. Authors wish to thank the Archaeological Museum of Thessaloniki for hosting the developed system in their premises, as part of the permanent exhibition "Macedonia: from fragments to pixels". Authors thank Michalis Sifakis for fruitful conversations and the formulation of the archaeological content, Antonios Katzourakis for the graphical designs, Panagiotis Panagiotidis for technical support, and Ilia Adami for her substantial contribution in planning, setting-up and conducting the usability evaluation studies as well as in the analysis of the evaluation results.

References

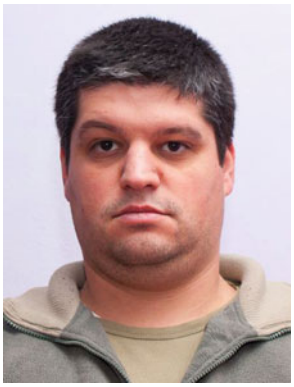
1. Allard, J., Franco, J., Menier, C., Boyer, E., Raffin, B.: The GrImage platform: a mixed reality environment for interactions. In: IEEE International Conference on Computer Vision Systems, p. 46 (2006)
2. Argyros, A.A., Lourakis, M.I.A.: Real time tracking of multiple skin-colored objects with a possibly moving camera. In: European Conference on Computer Vision, pp. 368–379 (2004)
3. Bannon, L., Benford, S., Bowers, J., Heath, C.: Hybrid design creates innovative museum experiences. Commun ACM **48**(3), 62–65 (2005)
4. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP J Image Video Process (2008)
5. Blomberg, J., Giacomi, J., Mosher, A., Swenton-Wall, P.: Ethnographic field methods and their relation to design. In: Participatory design: Principles and practices, pp. 123–155. Lawrence Erlbaum Associates (2003)
6. Bobick, A., Intille, S., Davis, J., Baird, F., Pinhanez, C., Campbell, L., Ivanov, Y., Schutte, A., Wilson, A.: The KidsRoom: a perceptually-based interactive and immersive story environment. Presence Teleoper Virtual Environ **8**(4), 369–393 (1999)
7. Brooke, J.: SUS: a quick and dirty usability scale. pp. 189–194. Taylor and Francis, UK (1996)
8. Darrell, T., Gordon, G., Harville, M., Woodfill, J.: Integrated person tracking using stereo, color, and pattern detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 601–609 (1998)
9. Falcao, G., Hurtos, N., Massich, J., Fofi, D.: Projector-camera calibration toolbox (2009). http://code.google.com/p/procamcalib
10. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans Pattern Anal Mach Intell **30**(2), 267–282 (2008)
11. Franco, J., Menier, C., Boyer, E., Raffin, B.: A distributed approach for real time 3D modeling. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, p. 31 (2004)
12. Gronbaek, K., Iversen, O., Kortbek, K., Nielsen, L., Rand Aagaard, K.: IGameFloor: a platform for co-located collaborative games. In: ACE: Advances in Computer Entertainment Technology, pp. 64–71 (2007)
13. Haro, G., Pardàs, M.: Shape from incomplete silhouettes based on the reprojection error. Image Vis Comput **28**(9), 1354–1368 (2010)
14. Hornecker, E., Stifter, M.: Learning from interactive museum installations about interaction design for public settings. In: Australian conference on Computer-Human Interaction, pp. 135–142. Sydney, Australia (2006)
15. Khan, S., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: European Conference on Computer Vision, pp. 133–146 (2006)
16. Kortbek, K., Gronbaek, K.: Interactive spatial multimedia for communication of art in the physical museum space. In: ACM Multimedia, pp. 609–618 (2008)
17. Laakso, S., Laakso, M.: Design of a body-driven multiplayer game system. Comput Entertain **4**, 7 (2006)
18. Ladikos, A., Benhimane, S., Navab, N.: Efficient visual hull computation for real-time 3d reconstruction using CUDA. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2008)
19. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Trans Pattern Anal Mach Intell **16**(2), 150–162 (1994)
20. Liem, M., Gavrila, D.M.: Multi-person tracking with overlapping cameras in complex, dynamic environments. In: British Machine Vision Conference (2009)
21. Lourakis, M.I.A., Argyros, A.A.: SBA: A software package for generic sparse bundle adjustment. ACM Trans Math Softw **36**(1) (2009)
22. Macedonia: from fragments to pixels (2010). http://www.makedonopixels.org. Demonstration video: http://www.makedonopixels.org/videos.php?c=8&sub_c=7&l=e
23. Mittal, A., Davis, L.: M2tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. Int J Comput Vis 189–203 (2003)

24. Nielsen, J.: Usability Engineering, chapter Thinking Aloud. pp. 195–199. Academic Press, san diego (1993)
25. Paradiso, J., Abler, C., Hsiao, K., Reynolds, M.: The magic carpet: physical sensing for immersive environments. In: Human factors in computing systems, pp. 277–278 (1997)
26. Pietroni, E., Antinucci, F.: The rule confirmation: virtual experience among the characters of Giotto's work. In: International Symposium on Virtual Reality, Archaeology and Cultural Heritage (2010)
27. Raskar, R., Welch, G., Fuchs, H.: Seamless projection overlaps using image warping and intensity blending. In: Virtual Systems and Multimedia (1998)
28. Reddy, D., Sankaranarayanan, A., Cevher, V., Chellappa, R.: Compressed sensing for multi-view tracking and 3-D voxel reconstruction. In: IEEE International Conference on Image Processing, pp. 221–224 (2008)
29. Point Grey Research. MultiSync. http://www.ptgrey.com/products/multisync
30. Robertson, T., Mansfield, T., Loke, L.: Designing an immersive environment for public use. In: Conference on Participatory design, pp. 31–40 (2006)
31. Sarmis, T., Zabulis, X., Argyros, A.A.: A checkerboard detection utility for intrinsic and extrinsic camera cluster calibration. Technical Report 397, FORTH-ICS (2009)
32. Schick, A., Stiefelhagen, R.: Real-time GPU-based voxel carving with systematic occlusion handling. In: DAGM Symposium on Pattern Recognition, pp. 372–81 (2009)
33. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: EEE Conference on Computer Vision and Pattern Recognition (2011)
34. Snibbe, S., Raffle, H.: Social immersive media: pursuing best practices for multi-user interactive camera/projector exhibits. In: Human factors in computing systems, pp. 1447–1456 (2009)
35. Sparacino, F.: Scenographies of the past and museums of the future: from the wunderkammer to body-driven interactive narrative spaces. In: ACM Multimedia, pp. 72–79 (2004)
36. Tran, S., Lin, Z., Harwood, D., Davis, L.: UMD VDT, an integration of detection and tracking methods for multiple human tracking. In: Multimodal Technologies for Perception of Humans. Lecture Notes in Computer Science, vol. 4625/2008, pp. 179–190. Springer, Berlin (2008)
37. Tyagi, A., Keck, M., Davis, J., Potamianos, G.: Kernel-based 3d tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
38. Tzevanidis, K., Zabulis, X., Sarmis, T., Koutlemanis, P., Kyriazis, N., Argyros, A.: From multiple views to textured 3d meshes: a gpu-powered approach. In: European Conference on Computer Vision Workshops, pp. 5–11 (2010)
39. Wu, B., Singh, V., Kuo, C., Zhang, L., Lee, S., Nevatia, R.: CLEAR'07 evaluation of usc human tracking system for surveillance videos. In: Multimodal Technologies for Perception of Humans. Lecture Notes in Computer Science, vol. 4625/2008, pp. 191–196. Springer, Berlin (2008)
40. Zabulis, X., Grammenos, D., Sarmis, T., Tzevanidis, K., Argyros, A.A.: Exploration of large-scale museum artifacts through non-instrumented, location-based, multi-user interaction. In: International Symposium on Virtual Reality, Archaeology and Cultural Heritage (2010)
41. Zabulis, X., Sarmis, T., Argyros, A.A.: 3d head pose estimation from multiple distant views. In: British Machine Vision Conference (2009)
42. Zabulis, X., Sarmis, T., Tzevanidis, K., Koutlemanis, P., Grammenos, D., Argyros, A.A.: A platform for monitoring aspects of human presence in real-time. In: International Symposium on Visual Computing (2010)
43. Zivkovic, Z.: Improved adaptive Gaussian mixture model for background subtraction. In: International Conference on Pattern Recognition, pp. 28–31 (2004)

Author Biographies

Xenophon Zabulis is a Researcher at the Institute of Computer Science - Foundation for Research and Technology, Hellas. He received his B.Sc., M.Sc. and Ph.D. degrees in Computer Science from the University of Crete, Greece, in 1996, 1998 and 2001, respectively. From 2001 until 2003 he has worked as a Postdoctoral Fellow at the Computer and Information Science Department, at the interdisciplinary General Robotics, Automation, Sensing and Perception laboratory and at the Institute for Research in Cognitive Science, both at the University of Pennsylvania, USA. In addition, during 2004 to 2007, he has worked as a Research Fellow at the Institute of Informatics and Telematics, Centre of Research and Technology Hellas, Greece. His research interests are in the following areas: stereo and multiple-view vision, real-time 3D reconstruction of static and dynamic scenes, person localization and head pose estimation, camera networks, medical image processing, human stereo vision, and applications of computer vision in Ambient Intelligence environments.

Dimitris Grammenos is a Principal Researcher at the Institute of Computer Science (ICS) of the Foundation for Research and Technology - Hellas (FORTH) and the lead interaction designer of the Human-Computer Interaction (HCI) Laboratory, specializing in the domains of Universal Design, Universal Access and Ambient Intelligence Environments. Over the past 17 years, he has designed the user interface of numerous interactive applications using a large variety of technologies (e.g., desktop computers, virtual reality, multimedia, mobile location-aware handheld computers, ambient intelligence) in diverse application domains (e.g., assistive technology, tools for working with guidelines, education, entertainment, serious games, healthcare). Currently, as a member of the Ambient Intelligence (AmI) Programme of FORTH-ICS, he is heavily involved in the design and development of novel, human-centred, AmI environments and interactive applications. In this context, he has designed (and often developed) several AmI-related applications and systems and is particularly interested in the design of (universally accessible) applications and systems for museums and public spaces, education, entertainment and their cross-section, and in the shaping of novel interaction concepts and techniques for Ambient Intelligence Environments.



Thomas Sarmis has graduated from Applied Information Technology and Multimedia Department of TEI of Crete. He received his M.Sc. in Computer Science from University of Crete. He has worked as Research and Development engineer in Computational Vision and Robotics Laboratory at the Institute of Computer Science - Foundation of Research and Technology, Hellas. His research interests are in the following areas: multi-camera image acquisition and processing, on-line 3D reconstruction, object tracking and applications of computer vision.

Konstantinos Tzevanidis is a Software Engineer specialized in the area of Computer Vision. He received his B.Sc. degree in Computer Science from the Informatics School of Aristotle University of Thessaloniki while working on transferring experience in multi-agent domains. He also holds a M.Sc. degree in Computer Vision received from the Computer Science Dept. of the University of Crete for his work on GPU powered multi-view 3D reconstruction. His interests include GPU computing, 3D graphics, parallel processing, machine learning. He has been a member of the Computational Vision and Robotics Laboratory of ICS FORTH. Currently he is focusing on the development of vision and computer graphics applications.

Pashalis Padeleris received his B.Sc. degree in Computer Science from the University of Crete in 2004. He has a long experience in designing and building systems providing public information over mobile devices and information points. As a member of the Computational Vision and Robotics Laboratory of the ICS-FORTH he has worked on problems including 3D scene reconstruction and real time people/object tracking.

Panagiotis Koutlemanis received his B.Sc. degree in Music Technology and Acoustics from the Technological Educational Institute of Crete, with a major in virtual reality using binaural audio, in 2008. He has worked as a developer for the Technological Educational Institute of Crete from 2005 to 2008. Since 2009, he has been working as a developer at the Institute of Computer Science - Foundation for Research and Technology, Hellas (FORTH), participating in research programs in the field of Ambient Intelligence.

Antonis A. Argyros is an Associate Professor at the Computer Science Department, University of Crete (CSD-UoC) and a Researcher at the Institute of Computer Science - FORTH, in Heraklion, Crete, Greece. He received B.Sc. (1989) and M.Sc. degrees (1992) in Computer Science, both from the CSD-UoC. On July 1996, he completed his Ph.D. on visual motion analysis at the same Department. He has been a postdoctoral fellow at the Computational Vision and Active Perception Laboratory, KTH, Sweden. Antonis Argyros is an area editor for the Computer Vision and Image Understanding (CVIU) Journal, member of the Editorial Board of the IET Image Processing Journal and a General Chair of ECCV'2010. He is also a member of the Executive Committee of the European Consortium for Informatics and Mathematics (ERCIM). The research interests of Antonis fall in the areas of computer vision with emphasis on tracking, human gesture and posture recognition, 3D reconstruction and omnidirectional vision. He is also interested in applications of computational vision in the fields of robotics and smart environments.