The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2010)
A. Artusi, M. Joly-Parvex, G. Lucet, A. Ribes, and D. Pitzalis (Editors)

Exploration of large-scale museum artifacts through non-instrumented, location-based, multi-user interaction

X. Zabulis[1], D. Grammenos[1], T. Sarmis[1], K. Tzevanidis[1] and A. A. Argyros[1,2]

[1]Institute of Computer Science - FORTH Herakleion, Crete, Greece
[2] Department of Computer Science, University of Crete

Abstract

This paper presents a system that supports the exploration of digital representations of large-scale museum artifacts in through non-instrumented, location-based interaction. The system employs a state-of-the-art computer vision system, which localizes and tracks multiple visitors. The artifact is presented in a wall-sized projection screen and it is visually annotated with text and images according to the location as well as walkthrough trajectories of the tracked visitors. The system is evaluated in terms of computational performance, localization accuracy, tracking robustness and usability.

1. Introduction

In the past few years, museums worldwide have started exploring new ways for integrating interactive exhibits in their premises, moving beyond the typical "multimedia information kiosk" paradigm of the past, i.e. [BBBH05,HS06,SR09, RML06, KG08], in order to support constructive and engaging entertainment with the purpose of educating ("edutainment"). A main axis of effort is the enhancement of didactic information with captivating experiences employing as multimedia and immersive techniques but, most importantly, by supporting active user participation through natural interaction. In this context, this paper presents a multimedia system targeted at archeological and historical museums, which supports the exploration of large-scale artifacts in actual size (e.g. a wall-painting, a mosaic, a metope) through non-instrumented, location-based, multi-user interaction.

The system (see Fig. 1) is focused at the presentation of large scale artifacts, especially ones that are difficult for the public to access. For example, the artifact featured in the reported experiments is a wall painting which is located at the entrance of a tomb at a height of $2.5m$ and which due to preservation constraints the public can view only from a distance. Despite its large size ($\approx 5 \times 2m^2$) the artifact features an abundance of details and, thus, its actual-size observation is of significance. In this case, details cannot be easily observed, also due to the severely deteriorated state of the artifact.

Figure 1: *System overview. A display presents the artifact in its actual size and its content is updated based on the location and walkthrough trajectories of visitors, as estimated by a computer vision system. Based on these estimates the system presents graphical and textual annotations as well as a restored version of the artifact.*

The proposed system provides functionalities for visual and textual annotations on a digital representation of the artifact, based on individual user location and walkthrough tra-

jectory. In the presented case, such functionalities are utilized to indicate forms within the artifact that cannot be clearly observed, to present a restoration of the artifact and to providing information about the persons and activities illustrated in the painting. Furthermore, the system provides a means of supporting visits of various temporal lengths by detecting visitors who revisit its segments or spend a greater amount of time at some of them, thereby inferring their increased interest for the corresponding segment. The system serves also the goal of relieving the visitor from carrying printed material during the visit, providing detailed textual annotations for each segment of the artifact in the language selected by each user, individually.

The remainder of this paper is organized as follows. In Sec. 2, related work in non-instrumented, location-based interaction and person tracking is reviewed. In Sec. 3, the method utilized for person tracking is presented. In Sec. 4, the application mediating user interaction with the system and the response of the latter are described. In Sec. 5, the evaluation of the system in terms of computational performance and usability is reported. In Sec. 6 this work is summarized and directions for future work are provided.

2. Related work

2.1. Non-instrumented, location-based interaction

The term *non-instrumented* denotes that users do not have to carry any object pinpointing their location. This approach offers more natural interaction and has simple logistics (e.g. no need for a lending / return process), a fact that can be very important for a museum. One of the earliest examples of such interaction is *KidsRoom* [BID*99], an interactive playspace simulating a children's bedroom where young children are guided through an adventure story. In [LL06], a multiplayer game system was developed using one top-view camera where player motion is mapped to digital character 2D motion.

Another contemporary example are interactive floors (physical sensor-based, like *Magic Carpet* [PAHR97] or vision-based, *iGameFloor* [GIKN07]) which are mainly being used for playing games. In the domain of museum applications, the work in [KG08] explored three different ways for supporting location-based interaction: (a) a coarse grained passive infrared sensor, (b) pressure sensors embedded in the floor and a small staircase, and (c) camera tracking. In *Immersive Cinema* [Spa04], one ceiling-mounted camera is used to track a users position on a floor segmented in five areas. A different, but quite interesting approach was followed in [RML06], where a ceiling-mounted infra-red camera was employed in tracking user position and motion, which were subsequently combined into flocking behavior used to browse collections of photographs and texts.

The *proposed system* can present large scale images of artifacts, with which one or more visitors can concurrently interact, simply by walking around, thus effectively applying interaction techniques used by Magic Carpet [PAHR97] and iGameFloor [GIKN07] in a different application domain, also extending previous related approaches like [KG08] and [Spa04] through multi-user support and personalization.

2.2. Person tracking

Accurate, precise and real-time person localization is crucial in systems that monitor person motion. Though accuracy is successfully tackled by state-of-the-art systems, precision and real-time operation can be computationally demanding tasks to achieve. Multiview systems simplify the localization process, as opposed to monocular approaches [TLHD08, WSK*08], because they treat occlusions robustly. On the other hand, the large amount of visual data produces, respectively, large computational demands as well as bandwidth and latency issues as, typically, cameras are mounted on multiple computers. To meet real-time and precision constraints, the localization process is parallelized in multiple CPUs or GPUs, in modern systems.

Multiview localization systems produce a 3D reconstruction of the imaged persons to cope with occlusions, obtain an accurate representation of the dimensions of the imaged persons and register them to a map of the room. The methods in [KS06, MD03, RSCC08, FBLF08, LG09], employ multiple views and a planar homography constraint to map imaged persons to the ground plane. The system in [LBN08] utilizes a voxel grid to represent the 3D reconstruction and distributes computation in the GPUs of 4 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 representation capacity of the partial estimates is large. The system in [SS09] eliminates communication cost by centralizing computation but, this solution does not scale with the number of cameras (limited to 4), due to the constrained bandwidth of the computer's bus.

As in [FBLF08], the *proposed approach* utilizes a volumetric reconstruction of persons to increase localization robustness, but it does not require that the number of tracked persons is a priori known. Similarly to [LG09], we project voxels on the ground plane to increase the robustness of person detection, but instead of using the projection area, we also consider the volume occupied by it. Also, besides parallelizing reconstruction, we further accelerate computation by parallelizing operations such as radial distortion compensation and background subtraction in the GPU and optimizing communication across processing nodes.

3. Vision based person monitoring

A computer vision system localizes visitors and estimates their walkthrough trajectories. At each frame, the system

reconstructs in 3D the visitors from the synchronously acquired images. Due to the extrinsic calibration of cameras these reconstructions are registered to room coordinates. Furthermore, temporal correspondence of person locations, or tracking, estimates motion trajectories of persons in the room and compensates for localization and reconstruction errors. The frequent rate of this computation is important both for brisk system response but also for robust tracking. In the course of this work, we have applied the proposed system in increasingly large installations, varying in scene area, number of employed cameras, image resolution, and computational capacity. Our observation is that computational speedup is linearly related to the available computational resources and data capacity.

3.1. System setup

The system is installed in a $6 \times 6 \times 2.5m^3$ room, in which a $4.88 \times 1.83\,m^2$ dual backprojection display is installed at the wall opposite from its entrance. The display is implemented by two bright ($3000\,lm$) 1024×768 short-throw projectors, 2 stereo speakers and a projection screen. The projectors and speakers are connected to a conventional PC. The projectors are parallel and their projections partially overlap ($\approx 7\%$). In the overlapping image regions, the projected images are blended on the fly similarly to [RWF98], to attenuate the appearance of a "seam" between the projections. Additionally, there is an information kiosk and a stand with mobile phones. Mobile phones run a custom application that receives information about their holder's position, through *Bluetooth* communication.

The computer vision system includes 8 cameras (*Dragonfly*, *Point Gray Research*) and 2 computers. Cameras overlook the scene from the ceiling obtaining synchronized images. Cameras were synchronized by a timestamp-based software that utilizes a dedicated *FireWire* bus across computers and guarantees a maximum of $125\mu sec$ temporal discrepancy in images with the same timestamp. The cameras are connected and evenly distributed to the available computers. The software on the computers acquires the incoming synchronous groups of images. For each such group, an estimate of the locations of the visitors is computed. The computation is distributed across computers and parallelized on their GPUs. In a typical installation, 2 computers are equipped with a *Intel i920* quad-core CPU and a *nVidia GTX 275* programmable GPU, each, however we have performed experiments running the computation on 1, 2 and 4 computers. The two computers are connected by a $1\,Gbit$ *Ethernet* link.

On setup, cameras are automatically spatially calibrated by imaging a calibration target (a checkerboard) at multiple postures. Reference points on the target are recognized by the method in [SZA09] and passed on to a bundle adjustment procedure [LA09], estimating accurately the cameras' intrinsic and extrinsic parameters. In this way, estimated person locations can be corresponded to world coordinates.

3.2. Volumetric reconstruction

To locate persons in the scene, we perform a 3D reconstruction of the visitors in the imaged environment. The reconstruction is continuously computed in time and represented in a 3D matrix V that, in turn, represents the room volumetrically. The cells of this matrix represent cubes in space, otherwise *voxels*, and take the value of 0 or 1. In the reconstruction phase, the value of 1 is assigned at a voxel if the corresponding volume is *occupied* by a person and 0 otherwise.

The images acquired by each camera i, are initially corrected for lens distortion yielding image I_i on the GPU of the computer acquired. In images I_i and on the same GPU, persons are segmented from the "background" The result are binary images B_i which are utilized to estimate the volumetric occupancy of the visitors in the room. This is achieved by considering a voxel $\vec{x}$ as occupied if it projects on foreground regions in all images B_i, where $\vec{x}$ is visible from. This value can be computed independently for each voxel as $V(\vec{x}) = AND_i\,(B_i(P_i(\vec{x})))$, where $\vec{x}$ is the vector of 3D coordinates of the center of the voxel. In practice, a more lenient rule is applied, to compensate for local failures of background compensation and, thus, a voxel is considered as occupied if all except μ views concur that it projects in a foreground region, that is $V(\vec{x}) = \sum_i (B_i(P_i(\vec{x}))) > \mu$. Fig. 2, shows the resulting reconstruction for a challenging scene.

Although the computation is parallelized for each voxel $\vec{x}$, for each $\vec{x}$, all B_is are required. To perform the computation, images B_i are gathered to the same GPU. Image transmission cost is optimized by Run Length (RL) encoding of the transmitted images. The achieved framerate ($10 - 20\,Hz$, depending on the number of cameras, computers and utilized resolution), allows for the assumption of motion continuity during tracking.

Volumes of arbitrary size can be processed by partitioning V. In contrast, in [LBN08, SS09] V is processed at a single block and, thus, its dimensions are constrained by the GPU's memory capacity. This, sequential, part of the algorithm does not reduce the speedup obtained by parallelization in the GPU, as the GPUs processing nodes are less than the number of voxels.

3.3. Localization and tracking

In the volumetric representation (V) persons occur as 3D blobs, which could be merged if persons are very close or embracing, or if voxel resolution is coarse. Albeit such difficulties, a probabilistic tracker estimates motion trajectories of persons in time. In addition to the generic operation of the tracker, the system incorporates application-specific knowledge, related to person size, room limits, room entrances and

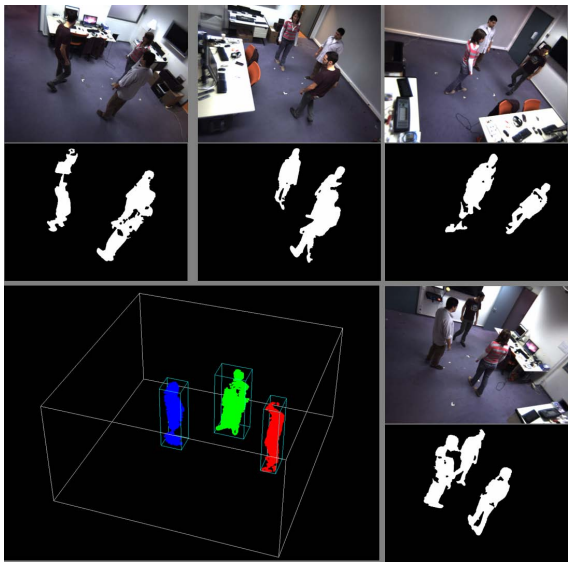

Figure 2: *Original images, background subtraction and volumetric reconstruction, for a scene imaged by* 4 *cameras. Persons are imaged against cluttered background and uneven illumination, resulting in inaccurate background subtraction. Persons occlude each other in all views. Reconstruction is not accurate, but sufficient for person tracking.*

exits. Even though tracking is a lightweight process, a part of it is parallelized in the GPU.

As in [KS06, MD03, RSCC08, FBLF08, LG09], a planar homography constraint is employed, mapping imaged persons to the ground plane. The occupied voxels of V are projected along the direction perpendicular to the floor. This results in a 2D buffer, F, registered to the ground plane and in which persons appear as "intensity blobs". Assuming that persons are standing, the highest intensities correspond to the projections of legs, torso and head. This projection is implemented on the GPU immediately after the computation of V. Image, F, is transferred to the computer's RAM. Pixels in F that exhibit small intensities are disregarded, typically corresponding to spurious volumes due to shadows or illumination artifacts. Thereby, only volumes of significant spatial extent are tracked.

The main difference with [LBN08], is that we parallelize on the background subtraction stage and transmit the RL encoded images B_i for the computation of V. This lasts significantly less than transmitting intermediate computations of V, as in [LBN08]. Another difference, in our case, is that increasing granularity of V does not increase communication cost, as does in [LBN08], because the transmission cost of images B_i is constant. The system in [AFM*06] exhibits a minimal communication cost, as it transmits silhouettes only and scales computation successfully to 11 dual-core CPUs. However, it can only achieve coarse, per-view parallelization as opposed to massive, per-voxel parallelization thereby requiring more computers.

Tracking, is achieved using the tracker in [AL04], which is efficiently implemented in CPU. The tracker is modified to track intensity blobs in F, rather than skin-colored blobs in color images for which it was originally formulated. This tracker is robust to transient localization failures, but most importantly, is designed to retain the tracking of blobs even if the occur merged for long temporal intervals. In this way, person tracking is successful even if subjects are very close to each other forming a single connected component in voxel space and in F (see Fig. 3). This is important for the system as often visitors tend to share a visit in companies or holding their hands, during intervals of the visit.

Two basic ingredients in the robustness of the tracking process are the high frame rate ($> 10Hz$) of operation and fine granularity of the volumetric representation ($1cm^3$). Achieving high frame rate casts blob motion in F smooth and continuous and, therefore, simpler to track. Fine granularity is important as proximate blobs will merge in F only if there are closer than voxel size.

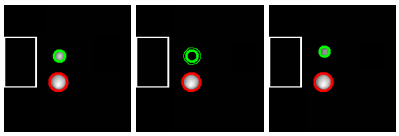

Figure 4: *Tracking example, showing the temporal evolution of F (left to right) with tracking results superimposed. The white rectangle marks the entrance of the room. The red and green circles correspond to two persons being tracked. Due to a reconstruction failure (i.e. change of illumination) a person is briefly lost from F, but its id is assigned to the closest re-appearing blob.*

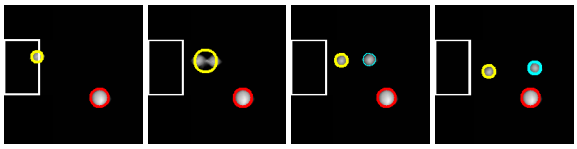

Figure 5: *Tracking example (left to right). A person is tracked (red circle) while another enters the room (yellow circle). When the blob under the yellow circle is split in two the tracker assigns a new id to the new blob, but only after this new blob persists for a sufficient time interval (represented as a thickening of of the cyan circle).*

Robustness enhancements consider environment constraints and were adopted after testing (see Sec. 5.2). A first constraint is imposed by the walls and entrance of the room. Persons are abandoned by the tracker when they disappear at the exit / entrance. Persons appearing at the same location are assigned with a new id. At the same time, if a person

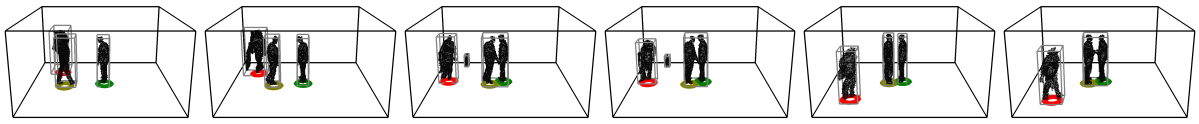

Figure 3: *Person localization and tracking. Estimated person location is marked with colored circles on the floor. Tracking is successful although in some frames visual hulls are merged. Occurring transiently, the spurious structure in frames 3 and 4 is disregarded by the tracker.*

is lost from tracking elsewhere than the exit (i.e. due to reconstruction error or lack of visibility), its id is assigned to the closest blob when it re-appears (Fig. 4 and 2^{nd} supplementary video). Furthermore, if blobs of significant extent abruptly appear and persist (i.e. $> 10\,sec$), tracking assigns a new id to them, inheriting the language preference from its parent. This incidence occurs after observing that some users may enter the room clustered together (Fig. 5 and 2^{nd} supplementary video).

3.4. Encapsulation

The complexity of camera control and synchronized image acquisition is encapsulated in a software platform. The platform supports communication of images and intermediate computation results across processing nodes, through a memory that is shared across computers. It is integrated with a middleware infrastructure to deliver output to the application layer. This enables integration of visual processes running on the platform through an API that supports a wide range of programming languages (*C/C++, .NET, Java, Python, Flash/ActionScript*). Through this middleware, output of visual processes is streamed to applications, as event notifications or peer-to-peer communication, hiding thus the details of the visual computation, network communication and data serialization. Such capabilities simplify the development of new visual processes and enable the integration of vision processes with the reasoning and actuating components of the interactive environment.

4. Application

An application runs on top of the computer vision infrastructure receiving changes in users' locations as event notifications and orchestrates system interaction, controlling the display and audio systems, according to the interaction scenario.

4.1. Description and functionality

In our installation, the display presents a wall painting located on the facade of the tomb of king Philip II in Vergina, Greece. The painting represents a hunting scene and it was found in a quite deteriorated state (Fig. 6, middle part of top image).

Visitors enter the room from an entrance opposite the display. The system assigns a unique id to each person entering the room. At the entrance an obstacle created by four queue posts guides visitors to move leftwards or rightwards to enter the room. As two help signs illustrate (Fig. 1), visitors entering the room from the right-hand side of the corridor are considered to be English-speaking, while those from the left-hand side, Greek-speaking. When at least one person is detected in the room, a soft piece of music starts to play.

The room is conceptually split in 5 zones perpendicular to the display, determined by an equal number of themes presented in the wall painting and corresponding to 5 vertical regions in it. Furthermore, the room is also split in 4 horizontal zones parallel to the wall painting, which are determined by their distance from it. Thus, a 5×4 grid is created, comprising 20 interaction slots. Fig. 6 bottom, presents an illustration of the grid, as rendered by the application in testing mode. In practice, the cells of the matrix in Fig. 6, are overlapping by 10%. This prevents the system from continuously alternating the content of the display, when a visitor standing at the border of a cell moves slightly in and out of its border.

The content presented on the display is determined by users' locations. The display renders images on the vertical slots which correspond to columns in the matrix of Fig.6. This content is relevant to the underlying theme, in the corresponding vertical slot of the display. The displayed visual content on each slot is also determined by the distance of the visitor to the display, quantized by the rows of the above matrix. In the current installation the system presents the painting in its current state at the farthest row. When a user is at the immediately closer row, an outline is superimposed on the corresponding vertical slot of the display, that intonates figures that cannot be clearly seen, due to deterioration of the fresco. In the closest two rows, the system presents a restored version of the painting. Additionally, a different textural annotation is presented at the bottom of the screen for each horizontal zone, that explains the displayed content in the corresponding vertical slot of the display.

4.2. Person tracking and system response

As the system tracks visitors, the textual components of the presented content are provided in the language selected by

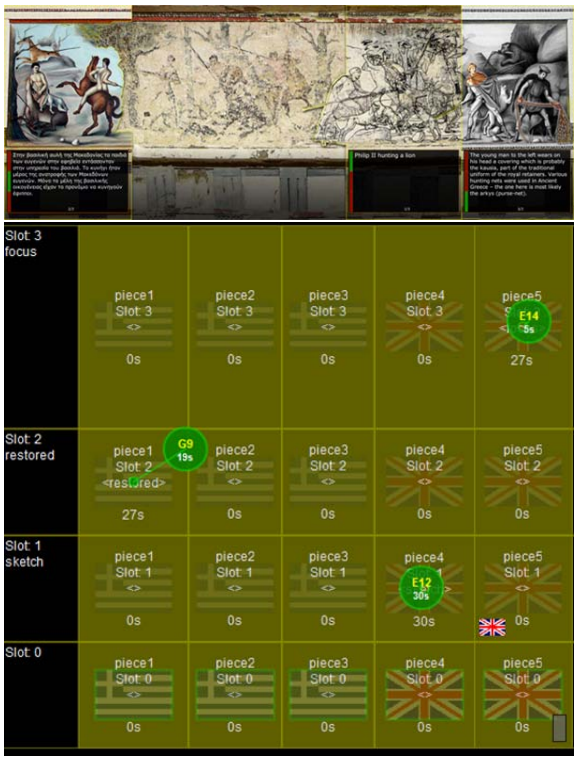

Figure 6: *Modulation of projected content (top) based on user locations (bottom). The respective wall painting vertical slot changes, depending on the row and column that the user is located at. In the example the 2^{nd} and 3^{rd} slot from the left are inactive, as users are located at columns 1, 4 and 5. The matrix 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. Three users are in the room, represented by green circles. A flag on each column shows the active language on each slot of the display.*

each user. Also, the system keeps track of the themes visited by each user and provides additional and more detailed information when a user revisits a theme. Several pieces of information are assigned to each theme, which are presented to the visitor when revisiting it or after a predefined time period has passed. The system keeps track of the "pages" of information that each user has seen, as well as of the time they have spent on each location and presents further textual details after a certain period of time.

When multiple visitors use the same vertical slot (are in the same column of the matrix in Fig. 6), the person standing closest to the wall determines the language of the presented textual annotation. When this person leaves, the next in line (if any) becomes the closest one to the wall.

An adjustment of system response concerned latency in performing updates on the display. In evaluation (Sec. 5.2) it was observed that users may rapidly cross the room, i.e. to join a friend, thereby causing an update of all intermediate themes in the display. Thus a minimum dwell time was adopted, for the establishment of control over a location.

4.3. Extensions

An editor is utilized to dynamically configure the content of the system and its active regions. Using this utility application, the administrator can replace the original and superimposed images as well as the graphical and textual annotations. Using the editor, new vertical slots in the presented image can be determined; the number of rows that the floor is tessellated and their width is automatically determined by the number and width of these slots, respectively. The "height" of the rows on the ground plane can be also configured (i.e. in Fig. 6 the row in front of the display is "taller" than the rest). The editor outputs all configurations in an XML file which is read by the application upon initiation.

Apart from location-sensing, the system also supports interaction using: (a) a kiosk and (b) mobile phones. The kiosk offers an overview of the wall painting, an introductory text and buttons for changing the user's language. When a visitor stands in front of the kiosk, all information is automatically presented in the visitor's language. Furthermore, the wall piece in front of which the visitor has spent most of the time is highlighted. Mobile phones are used as multimedia guides, presenting images and text, which can be read aloud, related to the visitor's position. In order to assign a mobile phone to a specific visitor, the visitor has to stand at a spot in the room denoted by a white X and press a selection button.

5. Experiments

5.1. Computational performance and robustness

System operation is demonstrated in the accompanying video. During development, the system was installed in increasingly large computational infrastructures. Representative cases are reported for the use of 4 or 8 cameras, to cover a $25\,m^2$ (Fig. 2) and a $36\,m^2$ (Fig. 1) room, while being supported by 1, 2 or 4 computers. The setup was also studied using lower resolution images, obtained by subsampling I_i.

Columns C1a, C2a, C3a, in Table 1 compare the performance of state-of-the-art implementations with the proposed one, in different system configurations, but for the same amount of computation. In the comparison, V was comprised of 2^{11} voxels and the scene was imaged by 8, 640×480 pixel cameras. In columns C1b, C2b, C3b image resolution was 320×240 pixels. The 1^{st} row marks the time required to process a frame, the 2^{nd} the amount of computational power utilized and the 3^{rd} the number of computers employed. The results indicate that the proposed approach improves state-of-the-art and that computational demands scale linearly with

the amount of computation. Most importantly, the proposed system requires less computers, therefore is simpler to install and more cost efficient. In the utilized installation the configuration of column C2a is employed, for voxels of $1\,cm^3$, a volume of $6 \times 6 \times 2m^3$ yielding a framerate of $\approx 15Hz$. The latency between a person's motion and the reception of the corresponding event is $\approx 140\,ms$ and localization accuracy is $\approx 4\,cm$.

Due to the use of the tracker the system is robust to transient reconstruction errors that may occur due to errors in background subtraction. Such errors occurred in an installation of the system which was installed in a room not isolated from exterior illumination (Fig. 2). Changes in the display and shadows do not cause temporally persistent errors in the reconstruction, as background subtraction is tuned to constantly update the background model.

Robustness has been also evaluated for the cases discussed in Sec. 3.3, indicating that the number of required cameras depends on the number of visitors. In the experiments, 4 cameras were adequate to disambiguate 3 persons even if clustered closely. Using 8 cameras, up to 7 persons were robustly tracked in challenging configurations, without errors. A system limitation is met when two or more embraced persons rotate; the result is a missasignment of tracking ids. Though this is rather improbable to occur, we plan to use color information for disambiguating such situations.

5.2. Usability

Due to the formative nature of our evaluations, we selected to use ethnographic field methods [BGMSW03], using a combination of the "observer participant" and "participant observer" approach. We avoided the use of video recording since, as our previous experience has shown, users tend to be more reluctant in freely exploring and experimenting with a system when knowing that they are being recorded. Participants were invited on an ad hoc basis, among people of all ages and cultural / educational background visiting (e.g., politicians, scientists, school classes) or working in our own facilities, including their families.

Up to now, more than 100 persons have participated in the usability evaluation. The robustness enhancements described in Sec. 3.3 stem from these experiments. Typically, 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 exhibits' characteristics to be assessed. For example, when, at earlier stages, 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 any instructions. During and after the sessions, the facilitator held free-form discussions with the participants eliciting their opinion and experience, identifying usability problems, as well as likes and dislikes. The facilitator kept a small notepad for taking notes. After the visitors have left, the two observers would discuss the session, keeping additional notes, often reenacting parts of it, in order to clarify or further explore some findings.

Language selection is a considerably challenging task for interactive exhibits, rarely addressed by previous efforts. For example, a kiosk (see Sec.4.3) was initially used as a means of language selection. In evaluation it was observed that it created both problems of visitor flow (people had to stand in line) and erratic behavior (multiple visitors standing too close during language selection). The current scheme of implicit selection was an improvement in terms of both usability and robustness.

Overall, the opinion of all participants about the exhibit ranged from positive to enthusiastic. In accordance with the findings of [KG08], people of all ages agreed that they would like to find similar systems in museums they visit. Usually, when visitors were first introduced to the exhibit there was a short exclamation and amusement phase, during which they seemed fascinated by the technology and tried to explore its capabilities but, interestingly, after that most of them spent considerable time exploring the exhibit's content. Over different installations we observed that the large size and luminous intensity contribute to the enhancement of visitor appreciation of the system.

6. Conclusion

This paper presents an interactive exhibit designed to support the exploration of large-scale artifacts in real-life size in museums. A 6-month period of iterative formative evaluations with a large, highly diverse, group of participants has shown that the exhibit achieved the goal of providing engaging and entertaining educational experiences to its users. After invitation, the system has been recently installed at the Archaeological Museum of Thessaloniki, Greece in a dedicated space within its premises, indicating its service of purpose.

The system is characterized by increased robustness in tracking persons in high framerate, and its reduced requirements in computational hardware. The requirements are linearly related to the the capacity of required computation, or otherwise, the spatial extent of the area to be covered and the number of cameras utilzed to cover this area. Due to the utilized middleware, the architecture of the system is flexible enough for the system to adapt to the availability of resources (few or abundant), either for larger installations or for cases of hardware failure where computational nodes may be fewer.

Plans for future work regard integration of the system with

Table 1: *Performance measurements and comparison (see text).*

	[LBN08]	[SS09]	[FMBR04]	C1a	C1b	C2a	C2b	C3a	C3b
ms	40	72	33.3	42	29.4	25	17.2	14	9.7
GFLOPS	1614	933	836	437	437	894	894	1788	1788
Computers	5	1	11	1	1	2	2	4	4

two pose estimation algorithms, developed by our laboratory, regarding the pose of visitor's head and that of a pointing device. The first is to provide the location of the room or display that each user is facing at. The second would facilitate the use of the system in guided tours, as the corresponding method would provide the information of which part of the display the guide is referring to. Further plans include robustness enhancements for larger and open area installations where illumination is not controlled.

Acknowledgements

This work was partially supported by the FORTH-ICS internal RTD Programme "Ambient Intelligence and Smart Environments".

References

[AFM*06] ALLARD J., FRANCO J., MENIER C., BOYER E., B. R.: The Grimage platform: A mixed reality environment for interactions. In *ICCVS* (2006). 4

[AL04] ARGYROS A., LOURAKIS M.: Real time tracking of multiple skin-colored objects with a possibly moving camera. In *ECCV* (2004), pp. 368–379. 4

[BBBH05] BANNON L., BENFORD S., BOWERS J., HEATH C.: Hybrid design creates innovative museum experiences. *Commun. ACM 48*, 3 (2005), 62–65. 1

[BGMSW03] BLOMBERG J., GIACOMI J., MOSHER A., SWENTON-WALL P.: Ethnographic field methods and their relation to design. In *Participatory design: Principles and practices* (2003), Lawrence Erlbaum Associates, pp. 123–155. 7

[BID*99] 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: Teleoperation Virtual Environments 8*, 4 (1999), 369–393. 2

[FBLF08] FLEURET F., BERCLAZ J., LENGAGNE R., FUA P.: Multicamera people tracking with a probabilistic occupancy map. *PAMI 30*, 2 (2008), 267–282. 2, 4

[FMBR04] FRANCO J., MENIER C., BOYER E., RAFFIN B.: A distributed approach for real time 3D modeling. In *CVPR Workshops* (2004), p. 31. 8

[GIKN07] GRONBAEK K., IVERSEN O., KORTBEK K., NIELSEN K. RAND AAGAARD L.: IGameFloor: a platform for co-located collaborative games. In *ACE: Advances in computer entertainment technology* (2007), ACM, pp. 64–71. 2

[HS06] HORNECKER E., STIFTER M.: Learning from interactive museum installations about interaction design for public settings. In *Australian conference on Computer-Human Interaction* (2006), pp. 135–142. 1

[KG08] KORTBEK K., GRONBAEK K.: Interactive spatial multimedia for communication of art in the physical museum space. In *ACM Multimedia* (2008), pp. 609–618. 1, 2, 7

[KS06] KHAN S., SHAH M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV* (2006), pp. 133–146. 2, 4

[LA09] LOURAKIS M., ARGYROS A.: SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software 36*, 1 (2009). 3

[LBN08] LADIKOS A., BENHIMANE S., NAVAB N.: Efficient visual hull computation for real-time 3d reconstruction using CUDA. In *CVPR Workshops* (2008), pp. 1–8. 2, 3, 4, 8

[LG09] LIEM M., GAVRILA D.: Multi-person tracking with overlapping cameras in complex, dynamic environments. In *BMVC* (2009). 2, 4

[LL06] LAAKSO S., LAAKSO M.: Design of a body-driven multiplayer game system. *Computer Entertainment 4* (2006), 7. 2

[MD03] MITTAL A., DAVIS L.: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. In *IJCV* (2003), pp. 189–203. 2, 4

[PAHR97] PARADISO J., ABLER C., HSIAO K., REYNOLDS M.: The magic carpet: physical sensing for immersive environments. In *Human factors in computing systems* (1997), pp. 277–278. 2

[RML06] ROBERTSON T., MANSFIELD T., LOKE L.: Designing an immersive environment for public use. In *Conference on Participatory design* (2006), pp. 31–40. 1, 2

[RSCC08] REDDY D., SANKARANARAYANAN A., CEVHER V., CHELLAPPA R.: Compressed sensing for multi-view tracking and 3-D voxel reconstruction. In *ICIP* (2008), pp. 221–224. 2, 4

[RWF98] RASKAR R., WELCH G., FUCHS H.: Seamless projection overlaps using image warping and intensity blending. In *Virtual Systems and Multimedia* (1998). 3

[Spa04] SPARACINO F.: Scenographies of the past and museums of the future: from the wunderkammer to body-driven interactive narrative spaces. In *ACM Multimedia* (2004), pp. 72–79. 2

[SR09] SNIBBE S. S., RAFFLE H. S.: Social immersive media: pursuing best practices for multi-user interactive camera/projector exhibits. In *Human factors in computing systems* (2009), pp. 1447–1456. 1

[SS09] SCHICK A., STIEFELHAGEN R.: Real-time GPU-based voxel carving with systematic occlusion handling. In *DAGM Symp. on Pattern Recognition* (2009), pp. 372–81. 2, 3, 8

[SZA09] SARMIS T., ZABULIS X., ARGYROS A. A.: *A checkerboard detection utility for intrinsic and extrinsic camera cluster calibration*. Tech. Rep. 397, FORTH-ICS, 2009. 3

[TLHD08] TRAN S., LIN Z., HARWOOD D., DAVIS L.: UMD VDT, an integration of detection and tracking methods for multiple human tracking. In *CLEAR* (2008). 2

[WSK*08] WU B., SINGH V., KUO C., ZHANG L., LEE S., NEVATIA R.: CLEAR'07 evaluation of usc human tracking system for surveillance videos. In *CLEAR* (2008), pp. 191–196. 2