Machine Vision and Applications
DOI 10.1007/s00138-014-0602-9

ORIGINAL PAPER

Shape from interaction

Damien Michel · Xenophon Zabulis · Antonis A. Argyros

Received: 5 June 2013 / Revised: 27 December 2013 / Accepted: 19 February 2014

Abstract We present "shape from interaction" (*SfI*), an approach to the problem of acquiring 3D representations of rigid objects through observing the activity of a human who handles a tool. *SfI* relies on the fact that two rigid objects cannot share the same physical space. The 3D reconstruction of the unknown object is achieved by tracking the known 3D tool and by carving out the space it occupies as a function of time. Due to this indirection, *SfI* reconstructs rigid objects regardless of their material and appearance properties and proves particularly useful for the cases of textureless, transparent, translucent, refractive and specular objects for which there exists no practical vision-based 3D reconstruction method. Additionally, object concavities that are not directly observable can also be reconstructed. The 3D tracking of the tool is formulated as an optimization problem that is solved based on visual input acquired by a multicamera system. Experimental results from a prototype implementation of *SfI* support qualitatively and quantitatively the effectiveness of the proposed approach.

Keywords 3D reconstruction · Object tracking · 3D pose estimation · Transparent objects

Electronic supplementary material The online version of this article (doi:10.1007/s00138-014-0602-9) contains supplementary material, which is available to authorized users.

D. Michel · X. Zabulis · A. A. Argyros (✉)
Institute of Computer Science, FORTH, Heraklion, Greece
e-mail: argyros@ics.forth.gr
URL: http://www.ics.forth.gr/argyros

A. A. Argyros
Computer Science Department, University of Crete, Heraklion, Greece

1 Introduction

The automatic, vision-based estimation of accurate 3D object models is a fundamental problem of great theoretical importance and practical significance. Conventional passive 3D reconstruction methods such as binocular or multiview stereo, structure from motion, and shape from shading (see [41] for a review) typically assume Lambertian surfaces and become very inaccurate when this assumption is violated. Similar is the case for photometric stereo techniques [48]. Multiview methods treat deviations from Lambertian behavior as outliers and avoid reconstructing them [28] or suppress the pronunciation of such effects through optical filtering (i.e., polarization for specularities [34]). Though very accurate in the reconstruction of Lambertian objects, structured light methods [8] perform poorly when applied to objects with specular, refractive or translucent surfaces. The reason is that these methods are based on the detection of a projected illumination pattern upon the surface of interest, which is expected to appear clearly and distorted only due to the shape of the surface. When the surface exhibits any of the above properties this assumption does not hold, making hard or even impossible to detect the projected pattern and to reconstruct the surface.

Our approach to the problem of vision-based 3D reconstruction (see Fig. 1) is based on the fact that two rigid objects cannot share the same physical space. The observation of the interaction of a human handling a known tool with an unknown 3D object provides a wealth of constraints that permit the full recovery of the 3D shape of the unknown object. Thus, the problem of 3D reconstruction of an unknown object is essentially transformed into a problem of tracking its interaction with a known one.

The proposed idea is similar in spirit to that of digitizing shapes using a touch probe. This has been used quite

Published online: 01 April 2014

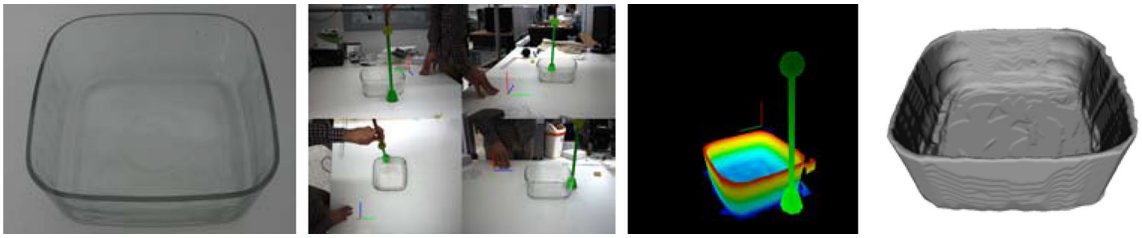

Fig. 1 Shape from interaction (*SfI*): the interaction of a known tool with an object of unknown 3D structure is observed. The accurate and occlusions-tolerant tracking of the tool provides constraints that are enough to reconstruct densely the unknown object. Thus, *SfI* recovers 3D models of objects that because of their physical properties (non-textured, transparent, translucent, specular, highly concave, etc) cannot be reconstructed by existing vision-based methods

extensively in industry by the so-called coordinate measurement machines (CMMs). A CMM is a device for measuring the geometrical 3D shape of an object. Current CMMs use probes mounted on mechanically tracked arms. As such, they are expensive and have a small operational workspace. We demonstrate that we can develop CMMs that, being vision-based, can be accurate, cost-effective and operational in large workspaces. A vision-based method with high-level similarities to the proposed one is presented in [26] in the context of augmented reality applications. However, that method depends solely on tip tracking as opposed to the robust and occlusions tolerant 3D tracking of various tools that we propose in this paper.

By delegating the 3D reconstruction of an unknown object to the 3D tracking of an known one, most of the material and appearance properties of the unknown object become irrelevant. This way, the proposed methodology is not only able to handle ordinary textured objects but also non-textured and even transparent, translucent, refractive and highly non-convex ones.

In the remainder of this paper, after presenting related work, we describe in detail an instantiation of the *SfI* methodology as well as its qualitative and quantitative evaluation. Experimental results confirm that *SfI* reconstructs in 3D objects that are impossible to obtain by any other existing vision-based 3D reconstruction technique.

2 Related work

A common classification of 3D reconstruction methods is defined based on the visual cue they employ. The term *shape-from-X* is utilized to refer to *X* as the visual cue and the process on which reconstruction is based. For example, shape-from-shading [49] and shape-from-texture [3] are approaches for single-view 3D reconstruction based on the corresponding cues. When more viewpoints are available and feature correspondences among views can be established, structure- and shape-from-motion [15] and shape-from-stereo approaches [45] are typically used. To ease the establishment of feature correspondences and to increase accuracy, structured light can be projected upon the scene [44].

The above methods assume the approximate Lambertian behavior of surfaces and, thereby, encounter difficulties for transparent, refractive, and translucent objects for which this assumption does not hold. Recently transparent and specular objects' reconstruction has gained attention in the literature. A comprehensive review can be found in [19]. The common ground amongst all of the existing approaches is their tight coupling to the type and individual properties of the surface to be reconstructed.

Methods targeting specifically the reconstruction of specular surfaces can be classified into three classes. The first aims at the reconstruction of intensely reflective (e.g. mirror) surfaces and is based on the projection of known patterns upon them; their shape is computed from the apparent distortion that this pattern undergoes when projected onto the surface [46]. In certain cases [9] the pattern itself and the process of projecting it onto the object can be as simple as casting shadows of a wand upon the surface of an object in a purposeful manner. Some methods pursue the tracking of reflected features upon the surface [1], thus employing the apparent distortion of the optical flow field instead of a calibrated pattern. Existing analyses of the phenomenon [43] do not reduce the requirements for careful setup of the reconstruction target and the projection apparatus. Another class of methods utilize surface specularities induced by one or more light sources to reconstruct them explicitly. A complicating factor in the reconstruction process is that the 3D location of the light source(s) needs to be calibrated. Methods in this class have been used to measure directly the structure through the appearance [20,35,50] while some emphasize in the recovery of surface details [10]. Finally, a third class of methods relax the use of a calibrated light source but still require the use of a calibration pattern [25].

The reconstruction of translucent surfaces by active illumination methods encounters similar difficulties, as the projected pattern is typically blurred leading to inaccurate results. A method that targets specifically such surfaces [11]

filters out subsurface scattering through polarization and phase shifting of the actively projected pattern.

The reconstruction of solid refractive and transparent objects is even more challenging than the above cases and, to date, there exist no generic reconstruction approaches for such objects. As in the case of specular surfaces, some methods rely on the apparent distortion [6] of a projected pattern to recover the shape of such objects, as well as the tracking of refracted features [2]. Another group of methods is based on the imaging of calibration targets after they have been refracted through the objects and avails 3D coordinates on the surface of these objects [5,25]. The method in [23] utilizes a known and controlled environment and employs differences between a conventional camera and a range camera to detect and reconstruct transparent objects. In [47], reconstruction of transparent and refractive objects is achieved by employing light-field probes encoding light ray positions and angles in varying intensities and colors. Using non-visible radiation, the method in [14] achieves reconstruction of transparent objects but requires special hardware. The method in [32] measures surface shapes of transparent objects by using a polarizing filter.

The above review reveals that the 3D reconstruction of transparent, translucent, refractive and specular objects is an open research problem. In this work we propose a new approach that treats the 3D reconstruction of each and all the above object categories robustly and in a unified manner.

3 Overview of the proposed method

The proposed method assumes a multicamera system consisting of n synchronized and calibrated cameras with projection matrices P_c, $1 \leq c \leq n$. Cameras are placed so that they all overlook the scene containing an object O of unknown 3D structure S_O that needs to be reconstructed. It is also assumed that the 3D position of O is roughly known in the sense that its actual volume lies inside a parallelepiped V of known dimensions and position. O is not allowed to move relative to the cameras. At time t, the cameras acquire a *multiframe* M_t consisting of n images, i.e. $M_t = \{I_1, I_2, \ldots, I_n\}$.

The proposed method also considers a tool T of known 3D structure S_T. The 3D position and orientation of T is tracked in a temporal sequence of multiframes. Provided that this can be achieved, S_T can be registered in the 3D scene observed by the cameras. Thus, the space S_T^t that this occupies at time t can be estimated. The fundamental idea behind *SfI* is that as each moment in time, T and O cannot share the same physical space. More specifically, it holds that

$$\forall t,\ S_T^t \cap S_O^t = \emptyset. \tag{1}$$

Under this assumption, *SfI* suggests that at time t the 3D structure S_O^t of O can be approximated with

$$S_O^t = V - \bigcup_{1 \leq i \leq t} S_T^i = S_O^{t-1} - S_T^t, \tag{2}$$

where S_T^i denotes the space occupied by T at time i and $S_O^0 = V$. To minimize the time required for *SfI*, it is important that T purposefully interacts with O. The tracking of the tool needs to be robust in occlusions, because as T interacts with O, considerable portions of it might not be visible by some (or even by any) of the cameras. Additionally, tracking needs to be accurate, because if some volume is incorrectly carved off the objects' shape, the resulting error cannot be recovered at a later stage.

The notion of space carving is central to the proposed approach. Space carving was proposed in [24] and makes use of photoconsistency as a means to decide whether a voxel is occupied by matter or not. Voxels that are not photoconsistent, are "carved out" during 3D reconstruction. In the same general spirit, we carve volumes that are occupied by the tool. Evidently, the criteria based on which space carving is performed are completely different. Moreover, in photoconsistency-based space carving, the manipulation of a voxel (either maintaining it to the reconstruction or carving it) requires that this is visible from at least two cameras. This is not a requirement in *SfI*.

Another approach that requires some interaction by the user is [16] which provides coarse models of spaces through the volume that dynamic entities such as walking persons occupy in a scene. The method relies on accurate background segmentation in order to reconstruct these entities through their visual hulls.

We next present the class of tools that we have considered (Sect. 4), the way that those are tracked by the employed multicamera system (Sect. 5) and how S_O is reconstructed by observing the interaction of T with O (Sect. 6).

4 Tool design

The tools we have considered consist of three parts, a wand W, a sphere S and an effector E. Figure 2 provides a close-up view of the tool and a collection of effectors, with one of them (large dark sphere) attached to the wand.

The wand W is an elongated cylinder. The sphere S is attached close to one of the endpoints of W so that the 3D main axis of W passes through the center of S. The effector E may have any rotationally symmetric shape. E is attached to one of the endpoints of W at a known distance from the center of S. Moreover, its main axis of rotational symmetry coincides to that of W. Both W and S are painted on individual, highly discriminant colors. In our experiments, the effectors used were modelled as geometric primitives whose shape parameters were measured manually.

The presented design facilitates the 3D detection and tracking of the tool in the case of severe occlusions and/or

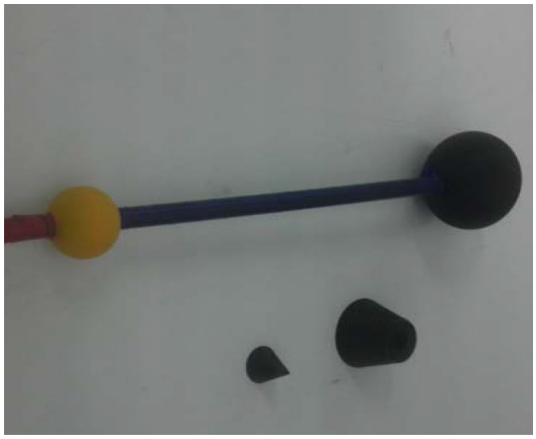

Fig. 2 A closeup view of the tool used in this work

clutter. The spherical part of the tool provides strong evidence on the 3D location of T. This is important because even if the endpoint of the tool that carries the effector is totally occluded (e.g., the effector is inside a concave object as shown in Fig. 6b) an accurate estimation of its correct 3D location and pose is still feasible. The elongated wand determines the 3D orientation of the tool. The rotational symmetry of the effector and of the wand guarantees that the space occupied by the tool is invariant to its rotation around the corresponding axis. This makes it possible to estimate the space occupied by the tool even in situations (see Fig. 6a) where the effector is totally occluded. In this case, the tool has been inserted in the object in a way that the effector is totally invisible. Now consider different rotations of the wand around its main axis. All those rotations are indistinguishable because, essentially, they change nothing in the acquired images. If the effector is not rotationally symmetric, a different part of the voxel space would have been removed in the (invisible) concave part of the object. But since the rotations of the wand are unobservable, the part of the voxel space that needs to be carved cannot be decided. On the contrary, if the effector is rotationally symmetric and its axis of symmetry coincides with the main axis of the wand, all possible rotations of the wand result in the same voxels to be carved. Essentially, the rotational symmetry of the effector is required to guarantee that carving is performed correctly in the case that the effector is severely occluded.

Another solution to the same problem might have been to use a rotationally asymmetric wand. Theoretically, if the wand is asymmetric, even if the effector is invisible, it is possible to estimate correctly its pose. Practically, in order to avoid ambiguities and to achieve high accuracy, this would require a tool that is far from being a typical wand. As such, it would be counter-intuitive for the user and difficult to manipulate. For this reason, we decided to use a rotationally symmetric effector attached to a rotationally symmetric wand.

In what follows, the 3D position of the tool is that of the center of S. The orientation of the tool is the orientation of W which, due to the rotational symmetry, is a 2D vector. Thus, the 3D pose of the tool is encoded as a 5D vector.

5 Tool tracking as an optimization problem

The 3D tracking of a tool T is treated as an optimization problem whose objective function quantifies the similarity between the appearance of hypothesized instances of T and its actual visual observations. Optimization (maximization of the objective function) is performed through particle swarm optimization (PSO) [22]. PSO has been employed successfully in the past to other 3D pose estimation problems such as human body tracking [21], multicamera-based hand pose estimation [36] and 3D head pose estimation based on depth data [40].

The input to the proposed method is a multiframe acquired by the employed multicamera system. Given the characteristic colors of the wand W and the sphere S parts of the tool, simple color-based segmentation suffices to isolate them in the images of the multiframe. Each tool pose hypothesis h is represented as a vector of five parameters. Tool 3D tracking is formulated as the problem of estimating the tool 3D pose h^* that maximizes the similarity between synthesized views of T in that pose and actual visual observations. Towards quantifying this similarity, we employ simple rendering techniques to produce comparable color maps for a given tool pose hypothesis. An appropriate objective function is then formulated and a variant of PSO is employed to optimize it. The result h^* of this optimization process is the output of the method for the given multiframe. Temporal continuity is exploited to track the tool in a sequence of frames. The remainder of this section describes these algorithmic steps in more detail.

5.1 Observing a tool

For each image I of a multiframe M, a wand color map o_W^I and a sphere color map o_S^I is computed using a variant of the skin color detection method presented in [4]. As a convention, the label of 1 indicates presence and the label of 0 indicates the absence of the tool in the corresponding maps. For each image I, the maps o_W^I and o_S^I constitute its observation cues. Note that the effector part of the tool does not contribute to the observation model.

5.2 Rendering a tool

Given a tool 3D pose hypothesis h we can render the complete 3D model of the tool in the view of the cameras of

the employed multiframe. Given the simplicity of the 3D primitives synthesizing the model, this process can be accelerated by analytically determining the image areas where tool primitives are expected to project. More specifically, we approximate the projection of the sphere with a circle and the projection of the cylinder with a trapezoid region. Thus, for each image I of a multiframe M, a rendered wand color map r_W^I and a rendered sphere color map $r_S^{I,h}$ is computed. Maps $r_W^{I,h}$ and $r_S^{I,h}$ are directly, pixelwise comparable to their observation counterparts o_W^I and o_S^I.

5.3 Evaluating a tool hypothesis

The proposed method establishes a measure quantifying the compatibility of a given tool pose h to the actual camera-based observations of multiframe M. This is based on the computation of an affinity function $A(h, M)$ which measures the similarity between wand and sphere maps computed in M and the wand and sphere maps that are rendered for h:

$$A(M, h) = \sum_{I \in M} F(I, h), \tag{3}$$

where F is defined as

$$F(I,h) = \alpha \frac{\left| o_W^I \wedge r_W^{I,h} \right|}{\left| o_W^I \vee r_W^{I,h} \right| + \epsilon} + (1-\alpha) \frac{\left| o_S^I \wedge r_S^{I,h} \right|}{\left| o_S^I \vee r_S^{I,h} \right| + \epsilon}. \tag{4}$$

In Eq. (4), o_W^I, o_S^I, $r_W^{I,h}$, $r_S^{I,h}$ are as defined in Sects. 5.1 and 5.2, respectively. Function $|\cdot|$ represents the cardinality of a set and the logical operators are applied to the binary maps in a pixelwise manner. A small term ϵ is added to the denominators of Eq. (4) to avoid divisions by zero. Parameter α can be used to tune the relative importance of the otherwise normalized contributions of W and S in the objective function. The value of $\alpha = 0.5$ was used in all our experiments.

Figure 3 illustrates the intuition behind the specific selection of the objective function. This figure shows the actual wand and sphere segmentation in a particular frame (yellow sphere, blue wand with solid lines) as well as the projection of a tool hypothesis on the same frame. The first term of the right hand side of Eq. (4) is proportional to the area of intersection of the sphere observation with the sphere hypothesis (area of region A), normalized by the areas of the union of regions A, B and C. Similarly, the second term of the right hand side of Eq. (4) is proportional to the area of intersection of the wand observation with the wand hypothesis (area of region D), normalized by the area of the union of regions D, E and F. Thus, in this particular example, the objective function is equal to

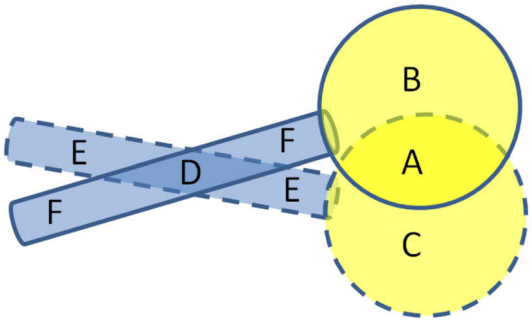

Fig. 3 The image projection of a tool hypothesis (*blue* wand and *yellow sphere with dashed lines*) in relation to the actual tool observation (*blue* wand and *yellow sphere with solid lines*). Regions A, B, C, D, E and F are defined. The areas of these regions determine the score that needs to be maximized by the optimization process so that the 3D pose of the tool can be recovered

$$F(I,h) = \frac{1}{2}\left(\frac{|A|}{|A|+|B|+|C|+e}\right) + \frac{1}{2}\left(\frac{|D|}{|D|+|E|+|F|+e}\right). \tag{5}$$

It can be verified that if the projection of the tool hypothesis is perfectly aligned with the actual observation, the value of the objective function is equal to 1. On the contrary, the objective function is equal to 0 when the area covered by the projection of the tool hypothesis is disjoint to that of its observation.

It should also be noted that Eq. (3) gathers contributions from all views to form the final value of the objective function.

5.4 Tool 3D pose estimation through PSO

PSO [22] is a stochastic, evolutionary algorithm that optimizes an objective function through the evolution of particles of a population (i.e., candidate solutions) that lie in its parameter space. The particles evolve in generations (i.e., runs) according to rules that emulate social interaction.

At a generation k, every particle has a position x_k and a velocity v_k. P_k stores the position at which the particle achieved, up to generation k, the best value of the objective function. G_k represents the best position encountered across all particles of the swarm, i.e., the global optimum up to generation k. The equations that re-estimate each particles' velocity and position in every generation k are

$$v_{k+1} = w(v_k + c_1 r_1 (P_k - x_k) + c_2 r_2 (G_k - x_k)) \tag{6}$$

and

$$x_{k+1} = x_k + v_{k+1}, \tag{7}$$

where w is a constant *constriction factor* [12]. In Eq. (6), c_1 is the so-called *cognitive component*, c_2 is the *social component* and r_1, r_2 are random samples of a uniform distribution in

the range $[0 \ldots 1]$. Finally, $c_1 + c_2 > 4$ must hold [12]. In all performed experiments the values $c_1 = 2.8$, $c_2 = 1.3$ and $w = 2/\left|2 - \psi - \sqrt{\psi^2 - 4\psi}\right|$ with $\psi = c_1 + c_2$ were used.

The particles are initialized at random positions and zero velocities. Each dimension of the parameter space is bounded in some range. Special treatment is required if, during position update, a velocity component forces a particle to move to a point outside this space. We adopted the "nearest point" method [18] according to which, if the velocity of a particle forces it to move to a point p_o outside the bounds of the parameter space, that particle moves to the point p_b inside the bounds that minimizes the distance $|p_o - p_b|$.

In this work, PSO operates on the $5D$ tool pose parameter space (see Sect.4). This means that each particle of the swarm corresponds to a tool pose hypothesis and the population is a set of candidate 3D tool poses hypothesized for a single multiframe. The objective function to be maximized is $A(M, h)$ [Eq. (3)]. For each time t, the optimization is executed for a fixed number of generations. After all generations have evolved, the best hypothesis h^* is dubbed as the solution for this time step. Essentially, PSO searches the 5D space of tool poses. This is achieved by a systematic process that defines candidate solutions based on Eq. (6) which are evaluated based on Eq. (3). PSO has been applied with great success to a number of high dimensional optimization problems in computer vision such as head pose estimation [40] and hand articulations tracking [27,37–39].

The process of tracking the tool requires the solution of a sequence of optimization problems, one for each acquired multiframe. By exploiting temporal continuity, the solution over the previous frame is used to generate the initial population for the optimization problem for the current frame (see also Sect.5.5). More specifically, the first member of the population for the current frame is the solution for the previous frame. The rest of the population consists of perturbations of that solution. The variance of these perturbations is experimentally determined based on the maximum pose variation that can be observed between two consecutive multiframes at a given image acquisition frame rate.

In all conducted experiments, optimization run for 12 particles and 40 generations.

5.5 Data-driven tool pose hypotheses

The above described PSO formulation tracks the 5D global pose of the tool by searching for it in a neighbourhood of the estimation performed in the previous time step. However, for the estimation of the pose of the tool in the first frame no such previous estimation exists. Therefore, the ranges of parameters to be searched is rather broad. To ease pose detection in the first frame but also in order to enrich PSO particles at each frame based on bottom-up evidence, the following strategy has been adopted. First, 3D line (i.e., wand) candidates are identified. To do this, we employ a technique [17], where a 3D line is estimated based on its projections on at least two cameras of a calibrated multicamera system. The recovered 3D lines form hypotheses for the 3D orientation of the wand.

Additionally, we seek for candidate 3D positions of the sphere. First, in each frame of the multiframe, we estimate the centroid of the 2D blob corresponding to the projection of the sphere in this frame. This is computed as follows. The oriented contour of each such 2D blob in each o_S^I is considered. Each pixel of the contour votes in a Hough transform-like manner for a half line where the projection of the 3D center of the sphere is constrained to lie on. The starting point of this half line is the contour point itself, and the direction is perpendicular to the estimated slope of the contour at this point, in the direction where more blob pixels lie. The point that receives the largest number of votes is assumed to be the center of the circle. By triangulating sphere centroid projections in pairs of images of the multiframe, we obtain hypotheses for the 3D position of the sphere.

It is important that in order to form a hypothesis for either the wand orientation or the sphere center, only two of its observations are required. Thus, hypotheses can be formulated even in the case of heavy occlusions.

The wand orientation and sphere center hypotheses are used to synthesize particles that are considered in a particular frame. The synthesis of wand poses and ball centroids cannot be performed with their simple concatenation because this does not lead necessarily to plausible tool pose hypothesis. As an example, the estimated 3D direction of the wand might not correspond to a line passing through the sphere center. For this reason, to form a particle, we consider the 3D orientation of the wand and the projection of candidate sphere centers on that orientation.

In our experiments, all particles in the first frame are formed based on this bottom up evidence. In subsequent multiframes, those are reduced to one third of the population. The rest of the particles consist of perturbations of the solution sought in the previous multiframe, as described in Sect. 5.4.

6 3D reconstruction

Given the tracking of the 3D pose of the tool, the space S_T^t that occupies in time t can be estimated. This estimation includes not only the wand W and the sphere S but also the effector E which is a known 3D shape attached to one of W's endpoints. S_T^t is carved off the solid parallelepiped volume V that originally approximates the 3D shape of the unknown object. In our experiments, a voxel of 0.9^3 mm^3 was used. Assuming that the user purposefully scans the unknown object O with the tool, and due to Eq. (1), it is expected that

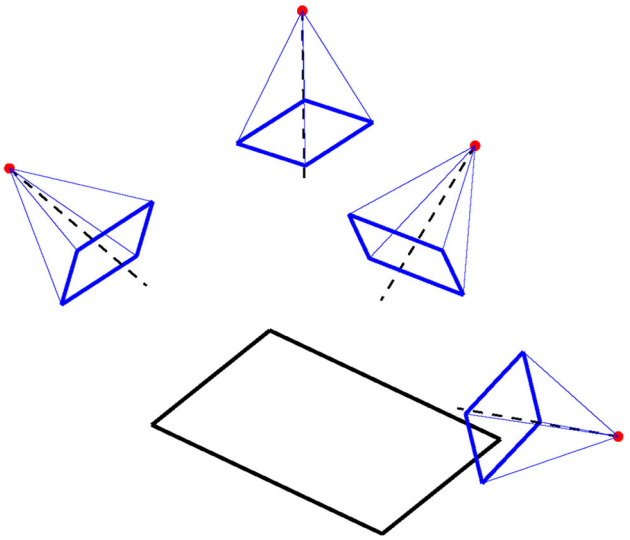

Fig. 4 The geometry of the four-cameras system used in the *SfI* experiments of this work

at the end of this process the evolution of V will be the true shape S_O of the unknown 3D object O.

During carving with *SfI*, the voxel space may become fragmented. Connected components labelling of voxels based on six-connectedness is performed. The largest connected component in terms of voxel count that is connected to the table is maintained. The rest are automatically carved out and removed from further consideration.

To safeguard for the case of relatively low frame rates, interpolation is performed between successive tool 3D pose estimations and the interpolated volumes are also carved. The amount of interpolation depends on the distances of the recovered wand endpoints. Additionally, the value of the objective function [Eq. (3)] lies in the range $[0 \ldots 1]$ with values close to 1 signifying very good agreement between hypothesis and observations. Thus, this is used as a confidence measure of the accuracy of tool pose estimation. If this value is lower than a predefined, experimentally set threshold, the carving of the reconstructed tool 3D volume is not performed.

7 Experimental evaluation

The proposed *SfI* framework has been validated quantitatively and qualitatively and has been compared with a number of standard 3D reconstruction methods. In order to do so, we employed a setup of four cameras (see Fig. 4). The cameras were calibrated intrinsically and extrinsically using [42].

A number of objects with diverse material, appearance and geometric properties has been collected to form a dataset. Table 1 summarizes the objects used and the properties that make their 3D reconstruction a challenging task. Sample views of these objects can be seen in the first column of Fig. 5.

Table 1 Characteristics of the objects used in experiments

Object	CON	NT	TP	TL	SP	REF
Statue	√	X	X	X	X	X
Spray	√	X	X	X	X	X
Doll	√	√	X	X	X	X
Mug	√	√	X	X	√	X
Mirror 1	X	√	X	X	√	X
Mirror 2	√	√	X	X	√	X
Bowl	√	√	√	X	X	X
Ashtray	√	√	√	√	√	√
Flower	√	√	√	√	√	√

From left to right CONcave, No Texture, TransParent, TransLucent, SPecular, REFractive

7.1 Qualitative comparative evaluation

We employed the proposed *SfI* method to reconstruct in 3D all objects listed in Table 1. Example videos of the whole process together with intermediate results is presented in the supplemental video material accompanying this submission and at http://youtu.be/uZQQkGTk6-k. For each object, we also attempted 3D reconstruction with the following three standard methods.

Stereo Stereo reconstruction was based on the plane sweeping method [13]. The photoconsistency metric employed was the modified normalized cross correlation [33]. The resulting depth map obtained from plane sweeping was median filtered for noise suppression. Additionally, small holes were filled through morphological filtering. The resulting depth map was transformed into a mesh of triangles, by taking into account the neighbourhood relationships of pixels.

Visual hull The multiframes that were employed for *SfI* were background subtracted utilizing the method in [51], yielding four binary foreground masks. The volume occupied by the object was approximated by its visual hull as a volumetric occupancy grid, as in [30]. This space was then smoothed with a Gaussian filter to suppress voxelization artifacts and, thereafter, the visual hull was extracted as its 0-isosurface, using [29]. For each multiframe, the visual hull is encoded as a mesh of triangles. In all experiments, a voxel was a cube with a side length of 2 mm.

Kinect The *Microsoft Kinect* sensor [31] was employed to obtain a third set of reconstructions. The acquired depth map was represented as a surface by a triangle mesh, which was obtained as in the case of stereo. The final 3D mesh was smoothed with a Gaussian 3D kernel to attenuate noise and depth quantization effects.

Fig. 5 3D reconstruction results for the objects listed in Table 1. Results. From *left to right SfI*, Kinect, visual hull, stereo. From *top to bottom* statue, spray, doll, mug, mirror 1, mirror 2, bowl, ashtray, flower. *N/A* denotes that the extracted foreground masks in individual views were inconsistent and gave rise to a null 3D reconstruction

For each object, Fig. 5 shows a view of the reconstruction obtained by each of the employed methods. Animated views of all reconstructed models for all object/method combinations can be viewed in the supplemental material accompanying this submission. From the visual inspection of these results, a number of conclusions can be drawn. Stereo reconstructions of non-textured objects are sparse. Visual hull performs better in textureless objects. Still, concavities cannot be recovered. Additionally, if foreground/background segmentation fails (i.e., white object in a white background, etc) visual hull fails to deliver reasonable 3D models. The same happens in the case of transparent and translucent objects. Kinect is not affected by the lack of texture. Still, it fails completely in the case of objects made of mirror, glass or crystal. As can be verified in the last five rows of Fig. 5 and in the supplemental material, Kinect-based reconstruction of such objects returns a hole in the reconstructed table. Finally, *SfI* manages to provide reasonable approximations of the 3D shapes of all these objects without making any assumptions regarding their material properties or their appearance. As an indicator of the accuracy by which the effector pose is estimated, it has to be noted that the bowl has been reconstructed without any holes, although the thickness of its walls does not exceed 5 mm. It can also be verified that for the objects made of mirror, glass or crystal, *SfI* is the only method that produces usable results.

It should be noted that in all *SfI*-based experiments the tracking of the pose of the tool had to deal with a number of challenging situations such as severe occlusions of the tool, refracted views of the effector, existence of multiple wand hypotheses due to mirror surfaces, distorted views of the wand, etc. Representative such situations are illustrated in Fig. 6.

7.2 Quantitative evaluation of 3D tool tracking

Besides the qualitative evaluation of *SfI* in comparison to other 3D reconstruction approaches, experiments were performed to assess the accuracy of tool 3D pose recovery as well as that of *SfI*. To assess quantitatively the performance of tool tracking, we performed the following experiment. A typical data CD was rigidly attached to the surface of a table. The conic effector was attached to the tool. Then a human manipulated the tool so that the endpoint of the effector was always touching the circumference of the CD. The tool and, consequently, the endpoint of the effector was tracked in 3D. The recovered 3D points were then fitted to a 3D circle. The average distance of the recovered 3D points from the center of the circle was measured to be 61 mm which should be compared to the 60 mm of the actual CD radius. The average distance of the recovered 3D points from the disc circumference was 0.96 mm with a standard deviation of 0.73 mm. The *precision* [45] of the method is 1.98 mm, in the sense that 90 % of the distances are below that value. The *accuracy* of the method [45] is 69.1 %, i.e., this is the percentage of points reconstructed within a minimum distance of 1.25 mm from the circumference of the CD.

In order to assess quantitatively the whole *SfI* framework and 3D structure estimation, we performed quantitative experiments with the mirror, spray and mug objects for which detailed, ground truth 3D models were available. More specifically, after reconstructing these objects with *SfI*, we registered the computed models to the ground truth ones through ICP [7] and then measured the average distance of the recovered points to the ground truth, their standard deviation, as well as the related accuracy and precision. Table 2 summarizes the obtained results. As it can be verified, *SfI* manages to deliver satisfactory 3D reconstruction results.

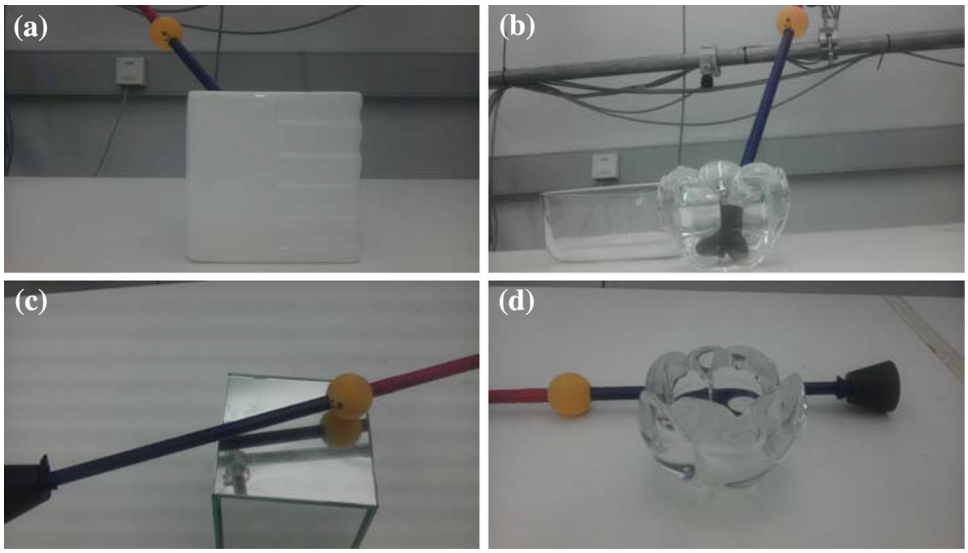

Fig. 6 Challenges for tool 3D pose tracking. **a** Severely occluded tool, **b** refracted view of the effector, **c** multiple wand hypotheses due to mirror surfaces, **d** distorted view of the wand due to occlusion caused by a crystal object of complex geometry

Table 2 Quantitative evaluation of *SfI*-based 3D reconstruction of the mirror, mug and spray objects

Object	Mean dist. (mm)	Std (mm)	Accuracy (mm)	Precision (%)
Mirror	0.44	0.43	1.00	94
Mug	1.09	0.79	2.17	66
Spray	1.53	1.42	3.14	54

7.3 Practical considerations and computational performance issues

The obtained experimental results demonstrate that the accuracy achieved by *SfI* is a function of the structural complexity of the object and not of its material or appearance properties. Additionally, it is very cheap and effective to have a selection of different end-effectors matching most of the basic shapes.

3D shape acquisition based on *SfI* differs considerably to the standard CMM practice. At each moment in time, CMMs sample a single point, but it is guaranteed that this point belongs to the surface of the reconstructed object. In contrast, *SfI* carves a volume that depends on the actual shape of the tool. However, at a certain moment in time, it is not known which points of the carved space belong to the object and which are not. It is only through visual inspection that the user may decide that the model that has been reconstructed so far is an acceptable approximation of the true shape of the object. Thus, depending on the goal of 3D model acquisition, the shape of the object and the required accuracy, the one or the other method might be preferable.

Computational performance is important because if *SfI* runs online, interaction is more natural and the user may get visual feedback that is very valuable in deciding "what to carve next" and in determining the appropriate subsequent manipulation of the tool. Regarding the reported experiments, the "statue", being the most complex object, required 2,687 frames or ∼15 min. Similarly, the "flower" required 1,068 frames, the "ashtray" 882, the "bowl" 911, the "hollow mirror box" 890 and the "mirror box" 520. This effectively means that simple objects can be carved in <5 min. Our current CPU-based implementation runs at an improved frame rate of 8 fps on a computer equipped with a 8-core Intel i7 CPU at 3 GHz and 4 GB RAM.

Still, it is expected that further computational performance optimization will have a significant impact on 3D reconstruction accuracy, too. Fortunately, there is ample room for such optimizations, because the computational time per frame is dominated by data parallel operations that can be speeded up by exploiting GPU processing. It is expected that reaching truly real time performance is feasible. This is because several time-consuming operations are inherently and highly data parallel. As an example, the evaluation of Eq. (3) for different PSO particles is totally independent. Additionally, the evaluation of this function involves data independent pixel-wise logical operations and summations that can be implemented very efficiently by exploiting the GPU of contemporary graphics cards. Thus, computations can be accelerated dramatically by employing this type of hardware.

8 Summary

In this paper, we proposed a new approach to the problem of vision-based 3D reconstruction of rigid objects. The approach is termed "shape from interaction" because it achieves 3D reconstruction by monitoring the interaction of a known tool with the object to be reconstructed. Tracking is formulated as an optimization problem that seeks for the 3D position and orientation of the tool that is mostly

compatible with visual observations obtained by a multicamera system. Thus, it becomes possible to estimate the space that the tool occupies at each moment in time accurately, and with high tolerance to occlusions and other optical effects. The estimated tool volume is carved continuously from a parallelepiped workspace. The remainder of this process constitutes the structure of the unknown object. A series of qualitative and quantitative experiments demonstrates that the robustness and the tolerance in occlusions and clutter are translated to accuracy in *SfI*-based 3D model acquisition. *SfI* proves particularly useful in estimating 3D models of objects that due to their material and reflectance properties are difficult or even impossible to obtain with other vision methods.

Acknowledgments This work was partially supported by the EU IST-FP7-IP-288533 project RoboHow.Cog.

References

1. Adato, Y., Vasilyev, Y., Ben-Shahar, O., Zickler, T.: Toward a theory of shape from specular flow. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
2. Agarwal, S., Mallick, S., Kriegman, D., Belongie, S.: On refractive optical flow. In: European Conference on Computer Vision, pp. 483–494 (2004)
3. Aloimonos, Y.: Shape from texture. Biol. Cybern. **58**(5), 345–360 (1988)
4. Argyros, A.A., Lourakis, M.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: European Conference on Computer Vision, pp. 368–379. Springer, New York (2004)
5. Atcheson, B., Ihrke, I., Bradley, D., Heidrich, W., Magnor, M., Seidel, H.-P.: Imaging and 3D tomographic reconstruction of time-varying inhomogeneous refractive index fields. Tech. Rep. University of British Columbia, UBC CS TR-2007-06 (2007)
6. Ben-Ezra, M., Nayar, S.: What does motion reveal about transparency? In: IEEE International Conference on Computer Vision (2003)
7. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 239–256 (1992)
8. Blais, F.: Review of 20 years of range sensor development. J. Electron. Imaging **13**, 231 (2004)
9. Bouguet, J., Perona, P.: 3D photography on your desk. In: IEEE International Conference on Computer Vision, pp. 43–50 (1998)
10. Chen, T., Goesele, M., Seidel, H.: Mesostructure from specularity. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1825–1832 (2006)
11. Chen, T., Lensch, H., Fuchs, C., Seidel, H., Informatik, M.: Polarization and phaseshifting for 3D scanning of translucent objects. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
12. Clerc, M., Kennedy, J.: The particle swarm—explosion, stability, and convergence in a multidimensional complex space. Trans. Evol. Comput. **6**(1), 58–73 (2002)
13. Collins, R.: A space-sweep approach to true multi-image matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 358–363 (1996)
14. Eren, G., Aubreton, O., Meriaudeau, F., Secades, S., Fofi, D., Naskali, T., Truchetet, F., Ercil, A.: Scanning from heating: 3D shape estimation of transparent objects from local surface heating. Opt. Express **17**(14), 11457–11468 (2009)
15. Faugeras, O., Luong, Q., Papadopoulos, T.: The Geometry of Multiple Images. MIT Press, Cambridge (2001)
16. Guan, L., Franco, J., Pollefeys, M.: Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction. Int. J. Comput. Vis. **90**(3), 283–303 (2010)
17. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN: 0521540518
18. Helwig, S., Wanka, R.: Particle swarm optimization in high-dimensional bounded search spaces. In: Proceedings of IEEE Swarm Intelligence Symposium, pp. 198–205 (2007)
19. Ihrke, I., Kutulakos, K., Lensch, H., Magnor, M., Heidrich, W.: Transparent and specular object reconstruction. Comput. Graph. Forum **29**(8), 2400–2426 (2010)
20. Ikeuchi, K.: Determining surface orientations of specular surfaces by using the photometric stereo method. IEEE Trans. Pattern Anal. Mach. Intell. **3**(6), 661–669 (1981)
21. John, V., Ivekovic, S., Trucco, E.: Articulated human motion tracking with HPSO. In: Proceedings of International Conference on Computer Vision Theory and Applications (2009)
22. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, vol. 4. pp. 1942–1948 (1995)
23. Klank, U., Carton, D., Beetz, M.: Transparent object detection and reconstruction on a mobile platform. In: IEEE International Conference on Robotics and Automation, pp. 5971–5978 (2011)
24. Kutulakos, K., Seitz, S.: A theory of shape by space carving. Int. J. Comput. Vis. **38**, 307–314 (2000)
25. Kutulakos, K., Steger, E.: A theory of refractive and specular 3D shape by light-path triangulation. In: IEEE International Conference on Computer Vision, pp. 1448–1455 (2005)
26. Kutulakos, K., Vallino, J.: Calibration-free augmented reality. IEEE Trans. Vis. Comput. Graph. **4**(1), 1–20 (1998)
27. Kyriazis, N., Argyros, A.A.: Physically plausible 3d scene tracking: the single actor hypothesis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9–16 (2013)
28. Li, Y., Lin, S., Lu, H., Kang, S., Shum, H.: Multibaseline stereo in the presence of specular reflections. In: International Conference on Pattern Recognition, pp. 573–576 (2002)
29. Lorensen, W., Cline, H.: Marching cubes: a high resolution 3D surface construction algorithm. In: SIGGRAPH, pp. 163–169 (1987)
30. Matsuyama, T., Wu, X., Takai, T., Nobuhara, S.: Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video. CVIU **96**(3), 1077–3142 (2004)
31. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–12 (2012)
32. Miyazaki, D., Kagesawa, M., Ikeuchi, K.: Transparent surface modeling from a pair of polarization images. IEEE Trans. IEEE Trans. Pattern Anal. Mach. Intell. **26**(1), 73–82 (2004)
33. Moravec, H.: Towards automatic visual obstacle avoidance. In: International Joint Conference on Artificial Intelligence, pp. 584–594 (1977)
34. Nayar, S., Fang, X., Boult, T.: Removal of specularities using color and polarization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 583–590 (1993)
35. Nayar, S., Sanderson, A., Weiss, L., Simon, D.: Specular surface inspection using structured highlight and gaussian images. IEEE Trans. Robot. Autom. **6**(2), 208–218 (1990)
36. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Markerless and efficient 26-DOF hand pose recovery. In: Asian Conference on Computer Vision, pp. 744–757 (2010)
37. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: British Machine Vision Conference, Dundee (2011)

38. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: IEEE International Conference on Computer Vision, pp. 2088–2095 (2011b)
39. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
40. Padeleris, P., Zabulis, X., Argyros, A.: Head pose estimation on depth data based on particle swarm optimization. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Workshop on Human Activity Understanding from 3D Data (HAU3D) (2012)
41. Remondino, F., El-Hakim, S.: Image-based 3D modelling: a review. Photogramm. Rec. **21**, 269–291 (2006)
42. Sarmis, T., Zabulis, X., Argyros, A.A.: A checkerboard detection utility for intrinsic and extrinsic camera cluster calibration. Tech. Rep. 397, FORTH-ICS (2009)
43. Savarese, S., Chen, M., Perona, P.: Local shape from mirror reflections. Int. J. Comput. Vis. **64**(1), 31–67 (2005)
44. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 195–202 (2003)
45. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 519–528 (2006)
46. Tarini, M., Lensch, H., Goesele, M., Seidel, H.: 3D acquisition of mirroring objects using striped patterns. Graph. Models **67**(4), 233–259 (2005)
47. Wetzstein, G., Roodnick, D., Raskar, R., Heidrich, W.: Refractive shape from light field distortion. In: IEEE International Conference on Computer Vision (2011)
48. Woodham, R.: Photometric method for determining surface orientation from multiple images. Opt. Eng. **19**(1), 513–531 (1980)
49. Zhang, R., Tsai, P., Cryer, J., Shah, M.: Shape from shading: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **21**(8), 690–706 (1999)
50. Zheng, J., Murata, A.: Acquiring a complete 3D model from specular motion under the illumination of circular-shaped light sources. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 913–920 (2000)
51. Zivkovic, Z.: Improved adaptive Gaussian mixture model for background subtraction. In: International Conference on Pattern Recognition, pp. 28–31 (2004)

Damien Michel received his M.Sc. degree in computational and electrical engineering in 2006 from the Ecole Nationale Superieure de l' Electronique et de ses Applications (ENSEA), one of the "grandes ecoles" in Paris. Since then, he has been working as a Research and Development Engineer at the Computational Vision and Robotics Laboratory (CVRL), of the Institute of Computer Science, FORTH. His previous works include shape matching, object tracking and 3D reconstruction, as well as the development of several computer vision techniques at the service of Ambient Intelligence applications. His current research interests are focused on human body and human hand posture recognition and tracking.

Xenophon Zabulis is a Researcher at the Institute of Computer Science, Foundation for Research and Technology, Hellas. He received the B.A., M.S. 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 2004–2007, 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, camera networks, and applications of computer vision in Ambient Intelligence environments.

Antonis A. Argyros is a Professor of Computer Science 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. In these areas he has (co-)authored more than 100 papers in scientific journals and conference proceedings.