Contributions to Robust Multi-view 3D Action Recognition

Luis Díaz Más

Departamento de Informática y Análisis Numérico

University of Córdoba

A thesis submitted for the degree of

Computer Science Doctor (PhD)

2012 Septiembre

TITULO: *CONTRIBUTIONS TO ROBUST MULTI-VIEW 3D ACTION RECOGNITION*

AUTOR: *LUIS DÍAZ MÁS*

© Edita: Servicio de Publicaciones de la Universidad de Córdoba.
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TÍTULO DE LA TESIS: Contributions to Robust Multi-View 3D Action Recognition
DOCTORANDO/A: D. Luis Díaz Mas.

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

D. Francisco José Madrid Cuevas, Doctor en Informática y D. Rafael Muñoz Salinas, Doctor en Informática, ambos adscritos al grupo de investigación "Aplicaciones de la Visión Artificial" de la Universidad de Córdoba, como directores de la tesis titulada "Contributions to Robust Multi-View 3D Action Recognition" desarrollada por el doctorando D. Luis Díaz Mas

INFORMAN

que la presente tesis doctoral ha sido realiza como fruto de la investigación realizada por el doctorando en el seno del grupo de investigación "Aplicaciones de la Visión Artificial" y dentro del programa de doctorado "*Técnicas Avanzadas de Análisis y Control de Sistemas*" de la Universidad de Córdoba, siguiendo los criterios de calidad exigidos para este tipo de trabajo. En su investigación, alumno ha profundizado en el campo del reconocimiento de acciones humanas utilizando Visión por Computador. Como fruto de su investigación el alumno ha contribuido al citado campo de conocimiento con tres propuestas originales que han originado sendas publicaciones en revistas indexadas:

- Luis Díaz-Mas, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Three-dimensional action recognition using volume integrals. Pattern Analysis and Applications, pages 1–10, September 2011.
- Luis Díaz-Mas, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Shape from silhouette using Dempster–Shafer theory. Pattern Recognition, 43(6):2119–2131, June 2010.
- Luis Díaz-Mas, Francisco José Madrid-Cuevas, Rafael Muñoz-Salinas, Angel Carmona-Poyato, and Rafael Medina-Carnicer. An octree-based method for shape from inconsistent silhouettes. Pattern Recognition, 45:3245–3255, March 2012.

Así mismo, el trabajo realizado por el doctorando también ha propiciado una aportación al campo del seguimiento de objetivos que ha originado una publicación en revista indexada:

- Rafael Muñoz-Salinas, Enrique Yeguas-Bolivar, Luis Díaz-Mas, and Rafael Medina-Carnicer. Shape from pairwise silhouettes for plan-view map generation. Image and Vision Computing, In press, February 2012.

Por todo esto, los directores de la tesis creen que tanto el doctorando como su trabajo reúnen las condiciones necesarias exigidas a este tipo de trabajo y autorizan la presentación de la tesis doctoral.

Córdoba, 25 de julio de 2012

Fdo.:F.J. Madrid Cuevas

Fdo.: R. Muñoz Salinas

Contributions to Robust Multi-view 3D Action Recognition

Luis Díaz Más

Departamento de Informática y Análisis Numérico
University of Córdoba

A thesis submitted for the degree of
Computer Science Doctor (PhD)

2012 Septiembre

Abstract

This thesis focus on human action recognition using volumetric reconstructions obtained from multiple monocular cameras. The problem of action recognition has been addressed using different approaches, both in the 2D and 3D domains, and using one or multiple views. However, the development of robust recognition methods, independent from the view employed, remains an open problem.

Multi-view approaches allow to exploit 3D information to improve the recognition performance. Nevertheless, manipulating the large amount of information of 3D representations poses a major problem. As a consequence, standard dimensionality reduction techniques must be applied prior to the use of machine learning approaches. The first contribution of this work is a new descriptor of volumetric information that can be further reduced using standard Dimensionality Reduction techniques in both holistic and sequential recognition approaches. However, the descriptor itself reduces the amount of data up to an order of magnitude (compared to previous descriptors) without affecting to the classification performance.

The descriptor represents the volumetric information obtained by SfS techniques. However, this family of techniques are highly influenced by errors in the segmentation process (e.g., undersegmentation causes false negatives in the reconstructed volumes) so that the recognition performance is highly affected by this first step. The second contribution of this work is a new SfS technique (named SfSDS) that employs the Dempster-Shafer theory to fuse evidences provided by multiple cameras. The central idea is to consider the relative position between cameras so as to deal with inconsistent silhouettes and obtain robust volumetric reconstructions.

The basic SfS technique still have a main drawback, it requires the whole volume to be analized in order to obtain the reconstruction. On the other hand, octree-based

representations allows to save memory and time employing a dynamic tree structure where only occupied nodes are stored. Nevertheless, applying the SfS method to octree-based representations is not straightforward. The final contribution of this work is a method for generating octrees using our proposed SfSDS technique so as to obtain robust and compact volumetric representations.

Contributions to Robust Multi-view 3D Action Recognition

Luis Díaz Más

Departamento de Informática y Análisis Numérico
University of Córdoba

A thesis submitted for the degree of
Computer Science Doctor (PhD)

2012 Septiembre

Resumen

Esta tesis se centra en el reconocimiento de acciones humanas usando reconstrucciones volumétricas obtenidas a partir de múltiples cámaras monoculares. El problema del reconocimiento de acciones ha sido tratado usando diferentes enfoques, en los dominios 2D y 3D, y usando una o varias vistas. No obstante, el desarrollo de métodos de reconocimiento robustos, independientes de la vista empleada, sigue siendo un problema abierto.

Los enfoques multi-vista permiten explotar la información 3D para mejorar el rendimiento del reconocimiento. Sin embargo, manipular las grandes cantidades de información de las representaciones 3D plantea un importante problema. Como consecuencia, deben ser aplicadas técnicas estándar de reducción de dimensionalidad con anterioridad al uso de propuestas de aprendizaje. La primera contribución de este trabajo es un nuevo descriptor de información volumétrica que puede ser posteriormente reducido mediante técnicas estándar de reducción de dimensionalidad en los enfoques de reconocimiento holísticos y secuenciales. El descriptor, por si mismo, reduce la cantidad de datos hasta en un orden de magnitud (en comparación con descriptores previos) sin afectar al rendimiento de clasificación.

El descriptor representa la información volumétrica obtenida en técnicas SfS. Sin embargo, esta familia de técnicas está altamente influenciada por los errores en el proceso de segmentación (p.e., una sub-segmentación causa falsos negativos en los volúmenes reconstruidos) de forma que el rendimiento del reconocimiento está significativamente afectado por este primer paso. La segunda contribución de este trabajo es una nueva técnica SfS (denominada SfSDS) que emplea la teoría de Dempster-Shafer para fusionar evidencias proporcionadas por múltiples cámaras. La idea central consiste en

considerar la posición relativa entre cámaras de forma que se traten las inconsistencias en las siluetas y se obtenga reconstrucciones volumétricas robustas.

La técnica SfS básica sigue teniendo un inconveniente principal; requiere que el volumen completo sea analizado para obtener la reconstrucción. Por otro lado, las representaciones basadas en octrees permiten salvar memoria y tiempo empleando una estructura de árbol dinámica donde sólo se almacenan los nodos ocupados. No obstante, la aplicación del método SfS a representaciones basadas en octrees no es directa. La contribución final de este trabajo es un método para la generación de octrees usando nuestra técnica SfSDS propuesta de forma que se obtengan representaciones volumétricas robustas y compactas.

Esta tesis está dedicada a mi padre, quien me contagió su pasión por la ciencia y me dio los consejos necesarios para vivir la vida y conseguir mis metas profesionales. También está dedicada a mi madre, quien me enseñó el significado de la palabra "paciencia" tras soportar en muchas ocasiones mis cambios de humos en los momentos de trabajo más duro. Y finalmente a todas aquellas personas, compañeros de trabajo, amigos y familiares, que han compartidos buenos momentos conmigo desde que empecé mi carrera investigadora.

Agradecimientos

Me gustaría transmitir mis agradecimientos en primer lugar a mis directores, los doctores Francisco José Madrid Cuevas y Rafael Muñoz Salinas. Siempre tuvieron un hueco en su tiempo para atender mis dudas, pude aprender de ellos diferentes buenas cualidades que todo investigador debería tener, y vivimos buenos momentos juntos durante los tres años que permanecí en el grupo de investigación.

También agradecerle al Dr. Rafael Medina Carnicer la oportunidad que me brindó en su día para colaborar con el grupo de investigación AVA. Desde el primer momento se preocupó de conseguir la financiación necesaria para poder realizar mi doctorado en buenas condiciones.

Agradezco a todos los familiares y amigos el apoyo y comprensión mostrada durante este tiempo. No todo en la vida consiste en trabajar, y sin los momentos de diversión y tranquilidad pasados con ellos habría sido imposible llegar a culminar este trabajo.

Por último, no puedo olvidarme del trabajo desinteresado de millones de personas alrededor del mundo colaborando con el software libre. Todo el trabajo desarrollado durante estos años ha sido realizado sobre plataformas GNU/Linux y todos los documentos escritos con LaTeX. El mundo es un lugar mejor gracias a ambos proyectos.

Contents

Glossary

Background subtraction Process in which the background of the scene is segmented in a specific image giving as result usually two type of elements: *background* pixels and *shape* or *foreground* ones.

BBA *Basic Belief Assignment*: Assignment of belief to facts defined in the Dempter-Shafer probability framework.

F-Measure Common measure in the information retrieval field which represent the weighted harmonic mean of *precision* and *recall*.

FN *False Negative*: a non-expected negative value in the information retrieval field.

FP *False Positive*: a non-expected positive value in the information retrieval field.

GT *Ground Truth*: previous measurements of properties which act as reference model against to produce comparisons.

HMM *Hidden Markov model*: can be considered as the simplest dynamic Bayesian network. Its internal states are not directly visible, but output, dependent on the state, is visible. It is highly used in temporal pattern recognition such as speech, handwriting, gesture recognition, etc.

LDA *Linear Discriminant Analysis*: Dimensional reduction technique used in machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events.

MHV *Motion History Volume*: 3D action spatio-temporal descriptor for action recognition that comprises a whole action into a single voxelset.

OAR *Occupation Area Rate*: The percentage inside a area of an image which is occupied by foreground silhouettes.

Octree Tree-based structures for representing 3D data that are recursively divided into eight cubes of the same size until a desired resolution is reached.

PCA *Principal Components Analysis*: Dimensional reduction technique used in machine learning that transform data to a new coordinate system such that greatest variances by any projection of input data lie on first new coordinates.

Precision Common measure in the information retrieval field that indicates the percentage of detected TP in relation to the total number of positives detected.

Projection test Process in *SfS* algorithms which evaluates the final state of a *voxel*.

Recall Common measure in the information retrieval field that indicates the percentage of detected TP in relation to the TP in the GT.

SfIS *Shape from Inconsistent Silhouettes*: Technique for reconstructing 3D volumes from 2D silhouette images considering inconsistencies.

SfS *Shape from Silhouettes*: Technique for reconstructing 3D volumes from 2D silhouette images.

SfSDS *Shape from Silhouettes using Dempster-Shafer theory*: Technique for reconstructing 3D volumes from 2D silhouette images taking into account inconsistencies and relative positions between views. It was proposed in one of the contributions presented in this thesis.

SVM *Support Vector Machine*: Machine learning technique used for classification and regression analysis. It could be generalized as a non-probabilistic binary linear classifier.

TN *True Negative*: an expected negative value in the information retrieval field.

TP *True Positive*: an expected true value in the information retrieval field.

VH *Visual Hull*: is the closest 3D solid equivalent to the real object that explains the silhouettes extracted in different views.

VI *Volume Integral*: A spatio-temporal descriptor for action recognition that comprises a specific frame or a whole action into three planes that maximize the action's discriminability. It was proposed in one of the contributions presented in this thesis.

Voxel It can be considered the 3D version of a pixel (or volumetric pixel) and represents a volume element in a space usually divided as a regular grid in the three dimensional space.

Voxel set Regular grid of voxels in the three dimensional space that represents the existing foreground objects in a scene.

Part I

Introduction

1

Introduction

"Things which we see are not by themselves what we see ... It remains completely unknown to us what the objects may be by themselves and apart from the receptivity of our senses. We know nothing by our manner of perceiving them."

Immanuel Kant

Computer vision is gradually becoming a known term to the general public. The reduction of costs, miniaturization and emergence of new visual devices bring this technology as a day-to-day element in our lives. We can see how video consoles, smart phones and operative systems start to incorporate vision-based devices and image processing algorithms as their human-computer interfaces. Furthermore, the efficiency and robustness of vision-based systems are increasing in the last years so that we can find applications using this technology in several other fields such as: surveillance, medicine, aerospace, agriculture, etc.

Along with the evolution of computer vision, the interest of understanding automatically the activities of people has increased. This interest is motivated by a lot of promising applications both offline and online, which can be roughly classified in three categories: surveillance, control and analysis (40). *Surveillance applications* cover some of the classical problems related with automatically monitoring and understanding the behaviour of crowds in airports (53), subways (12), public spaces like malls or in front of shop windows (57), etc. Also, special attention has been focus on smaller places like houses or rooms for many diverse purposes such as care-dependent assistance (46). In

control applications the estimated motion or pose parameters can be used to develop complex human-computer interfaces such as in EyeToy (50) or Kinect (39). They have also been used in the film industry and video games for modelling the appearance and movements of avatars. Finally, *analysis applications* includes automatic medical diagnostic, analysis and optimization of athletes performances and annotation of videos (37) or content-based retrieval (9, 38, 52), among other.

As already seen, action recognition is a wide area of research covering many applications which imply the need of different degrees of complexity. Several taxonomies has been proposed to structure the concept of action recognition, and the more recurrently adopted is in (40): primitive action, action and activity. A *primitive action* is an atomic movement that could be described at the level of a limb ("forehand", "backhand", "kick with left leg", etc.). An *action* consists of several primitive actions and it describes a full-body movement, probably cyclic ("run", "stand-up"). Finally, *activities* contain a number of consecutive actions and give meaning to the movements being performed, e.g., "110m hurdles" is an activity containing the actions "start", "jump" and "run". This work focus mainly on the concept of action (Fig. 1.1).

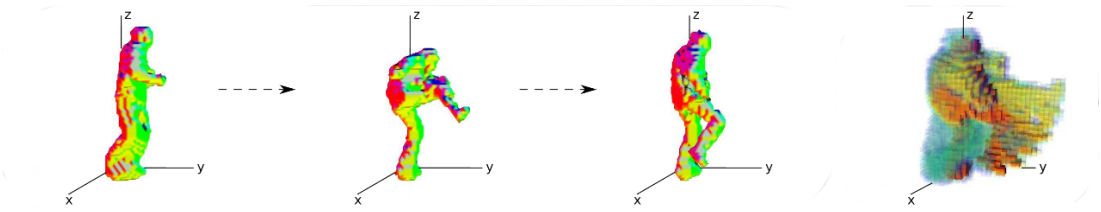

Figure 1.1: Three frames of the primitive action "kick" and its final volumetric representation.

In order to make a machine able to understand an action using computer vision, video sequences from one or several cameras must be provided to it. Then, a spatio-temporal descriptor must be developed trying to minimize the intra-variance existing between actions in the same class performed by different persons (42). Such descriptor could use internal features, external or a mixture of both. External features focus on the contours, shapes and movements while internal ones are focused on regional properties such as color or textures. In any case, the selected features for the descriptors must be as invariant as possible to scale, translation, rotation, and other possible transformations (25).

The accuracy level obtained in action recognition is highly influenced by the information sources selected. There exist precise approaches based in markers but the need of external elements for increasing the robustness of the algorithms is a deterrent for many applications. Initial works in markerless action recognition used 2D approaches with or without explicit shape models. When model are not employed, movements are described in terms of low-level 2D features from regions of interest (8, 44). Model based approaches, however, describe the human body as a kinematic tree, consisting of segments that are linked by joints. Every joint has a number of Degrees of Freedom (DOF) indicating in its possible movements and the configuration of all the model's joints represents a pose. Two dimentional models are suitable for motion parallel to the image plane and are sometimes used for gait analysis. When explicit a priori knowledge is available, human models are usually employed to segment, track and label the different body parts (3, 4, 30, 31, 36, 43, 49, 51). Two-dimensional approaches have several limitations, though, such as self-occlusions, noise in the segmentation process, etc. Therefore, 3D methods try to solve these disadvantages. The first works were focused on recovering the 3D articulated posed over time. These methods take advantage of the large amount of a priori knowledge about kinematics and shape properties of the human body to make the problem tractable (1, 5, 10, 15, 17, 24, 47, 54, 56). However, to develop a real-time system that employs this sort of models in classical action recognition problems is still an open problem due to the high computational requirements. As a consequence, methods that do not require explicit models have received high interest in last years. In these approaches, the aim is to work directly with 3D information and perform the rest of the tasks in a four dimensional spatio-temporal framework.

Multi-view approaches for action recognition can be divided in two groups depending on how they treat 3D information: those that fuse cues from different views at different stages (such as segmentation, learning and classification) and those that obtain an explicit 3D representation first, and work with it in subsequent processes. We focus in the latter since a preliminary 3D representation simplifies the subsequent processes. In order to obtain such 3D representation from several monocular cameras, one of the most used techniques is Shape-from-Silhouette (SfS). This is a well-known approach to reconstruct the 3D structure of objects using a set of silhouettes obtained from different views. Roberts (45) was the first to introduce the machine perception of 3D models, while Baumgart (6), a decade after, introduced new concepts about the

3D geometric modelling. But it was not until 1991 that Laurentini (34) defined the concept of the Visual Hull (VH) as the closest 3D convex solid equivalent to the real object that explains the silhouettes extracted. The VH is the geometric intersection of all visual cones explaining the projection of a silhouette in its corresponding camera image. Consequently, as the number of views increases, so does the precision of the reconstructed object (35), although it only happens when the image silhouettes are totally consistent (i.e., there are no errors in the segmentation process).

The first contribution of this Thesis is a new descriptor for three-dimensional action recognition (20). The proposed descriptor exhibits invariance to scale, translation, rotation and reflection, by projecting the visual hull into a much smaller subspace that maximize the discrimination of actions. The descriptor allows to reduce the amount of information in an order of magnitude while improving the recognition rates when compared to state of the art descriptors.

Nonetheless, a major problem of standard SfS methods is that they are strongly linked to the principle of silhouette consistency, i.e., the set of silhouettes employed must explain precisely the real object. A single inconsistency in one of the silhouettes could distort the reconstructed VH regarding the expected one, i.e., a hole in a silhouette is automatically propagated to the three-dimensional reconstruction. As a consequence, the likelihood of obtaining false negatives in the reconstructed volumes increases with the number of views. Nevertheless, total consistency hardly ever happens in real-life scenarios due to several factors such as inaccuracies in camera calibration, foreground extraction errors (21, 26, 28), and occlusions. Therefore, SfS methods have been usually confined to problems under controlled conditions (22, 48).

In recent years, several works have addressed the inconsistency problem for SfS in different ways (13, 22, 32, 33, 48). These approaches aim to exploit the information redundancy wisely in order to overcome the inconsistency problem using probabilistic approaches. Despite the improvements achieved, they have missed the use of an important piece of information which can be exploited in order to improve reconstruction: the positional relationship of the cameras. Information about the relative camera locations can be used to detect inconsistencies. For instance, if a camera indicates that a voxel is part of the object and another camera close to the former indicates the opposite, then there is an inconsistency that might be solved by a third camera.

The second contribution of this Thesis is a novel technique (19) to improve the SfS technique by employing information about the camera relationships using the Dempster-Shafer (DS) theory of evidence (2, 7, 11, 16).

Most SfS algorithms are designed using a grid of voxels for representing and handle 3D information. These "voxels-sets" are arranged homogeneously in the working area so that accessing them is trivial. Octrees, on the other hand, are tree-based structures in which the first node, usually called "root", is recursively divided into eight cubes of the same size until a desired resolution is reached. They were first used as a method for representing volumes in an efficient way. Then, several works were proposed to optimise their creation, storage and manipulation (14, 23, 27, 29, 55). In Fig.1.2 are depicted the divisions in both representations.

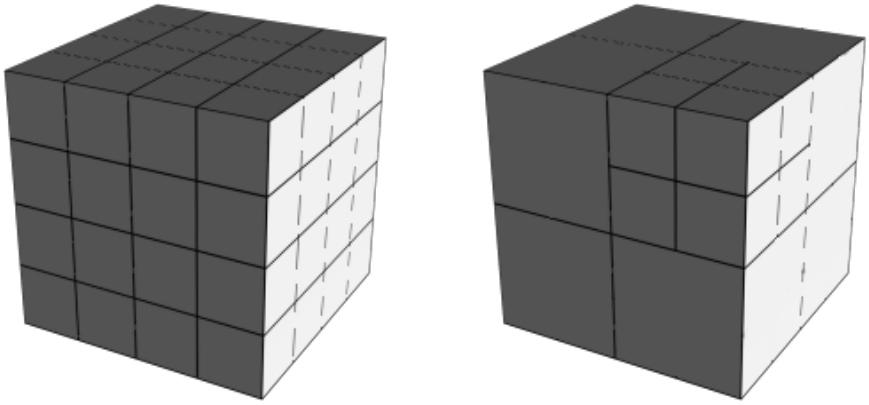

Figure 1.2: In voxel grids every voxel have the same size while in Octrees the sizes depend on the octree level.

Octrees have some advantages over voxel set approaches. First, the inherent multi-scale structure used in the representation of the volume. Second, only those nodes that represent part of an existing object in the scene need to be refined. Therefore, a better performance is obtained when compared with voxel set methods. However, the octree-based approach has also drawbacks. First, it is a more complex and dynamic structure. Second, the size of a voxel in the octree is function of the octree's level.

SfS methods require a projection test to determine whether a voxel is occupied. In a voxel set based approach, voxels always have the same volume, and there is a relatively small ratio of projected surface area to total volume in the working space, making the

design of the projection test simple. However it does not happen with the octree-based approach.

The third contribution of this Thesis is a novel approach for reconstructing 3D volumes with Octrees employing a similar SfS approach than the presented in the second contribution (18). In this work we have developed a method for handling the issue with voxels of different size.

1.1 Contributions

This Thesis offers three main contribution which have lead to three papers accepted in international journals indexed in the JCR. Following, we provide an overview of the main contributions along with their references.

Action recognition using volume integrals We propose the volume integral (VI) as a new descriptor for three-dimensional action recognition. The descriptor transforms the actor's volumetric information into a two-dimensional representation by projecting the voxel data to a set of planes that maximize the discrimination of actions. Our descriptor reduces significantly the amount of data of the 3D representations yet preserving the most relevant information. As a consequence, the action recognition process is greatly speeded up while achieving very high success rates. The method proposed is therefore especially appropriate for applications in which limitations of computing power and space are significant aspects to consider, such as real-time applications or mobile devices.

Luis Díaz-Más, Rafael Muñoz Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Three-dimensional action recognition using volume integrals. *Pattern Analysis and Applications*, pages 1–10, September 2011

SfS using DS theory A major problem of standard SfS methods is that they are strongly linked to the principle of silhouette consistency. A single inconsistency in one of the silhouettes could distort the reconstructed VH regarding the expected one. Therefore, SfS methods have been usually confined to problems under controlled conditions. Some researchers have addressed the inconsistency problem for SfS in different ways. They have exploited the information redundancy in

order to overcome the inconsistency problem using probabilistic approaches. Despite the improvements achieved, they have missed the use of an important piece of information which can be exploited in order to improve reconstruction: the positional relationship of the cameras. Our contribution solves the SfS problem by employing information about the camera relationships using the DS theory of evidence. The method, named SfSDS, provides 3D reconstructions robust to silhouette inconsistencies.

Luis Díaz-Más, Rafael Muñoz Salinas, F.J. Madrid-Cuevas, and Rafael Medina-Carnicer. Shape from silhouette using Dempster–Shafer theory. *Pattern Recognition*, **43**(6):2119–2131, June 2010

SfS using DS theory with Octrees Octree-based SfS approaches have several advantages, such as computational efficiency and storage, compared with the voxel set ones. However, the octree-based approach has a main drawback: the size of a voxel in the octree is a variable that is a function of the octree's level. All of the methods for SfS apply a projection test to determine whether or not a voxel is occupied. In a voxel set-based approach, the voxels always have the same volume, and there is a relatively small ratio of projected surface area to total volume in the working space, making the design of the projection test simple. However, it does not hold with the octree-based approaches. Our proposal extends the SfSDS method previously developed to octree-based representations.

Luis Díaz-Más, Francisco José Madrid-Cuevas, Rafael Muñoz Salinas, Angel Carmona-Poyato, and Rafael Medina-Carnicer. An octree-based method for shape from inconsistent silhouettes. *Pattern Recognition*, **45**:3245–3255, March 2012.

Part II

Publications

In this chapter are attached a copy of the main contributions published by the author of the thesis during its research stage. Only are presented those works in which the author of the thesis appears as first author.

2

First contribution: Three-dimensional action recognition using volume integrals

Pattern Anal Applic
DOI 10.1007/s10044-011-0239-5

SHORT PAPER

Three-dimensional action recognition using volume integrals

Luis Díaz-Más · Rafael Muñoz-Salinas · F. J. Madrid-Cuevas · R. Medina-Carnicer

Received: 11 May 2010 / Accepted: 13 September 2011

Abstract This work proposes the volume integral (VI) as a new descriptor for three-dimensional action recognition. The descriptor transforms the actor's volumetric information into a two-dimensional representation by projecting the voxel data to a set of planes that maximize the discrimination of actions. Our descriptor significantly reduces the amount of data of the three-dimensional representations yet preserves the most important information. As a consequence, the action recognition process is greatly speeded up while achieving very high success rates. The method proposed is therefore especially appropriate for applications in which limitations of computing power and space are significant aspects to consider, such as real-time applications or mobile devices. Additionally, the descriptor is sensitive to reflected actions, i.e., same actions performed with different limbs can be differentiated. This paper tests the VI using several Dimensionality Reduction techniques (namely PCA, 2D-PCA, LDA) and different Machine Learning approaches (namely Clustering, SVM and HMM) so as to determine the best combination of these for the action recognition task. Experiments conducted on the public IXMAS dataset show that the VI compares favorably with state-of-the-art descriptors both in terms of classification rates and computing times.

Keywords Action recognition · View invariance · Multi-camera · Motion descriptor

L. Díaz-Más (✉) · R. Muñoz-Salinas · F. J. Madrid-Cuevas · R. Medina-Carnicer
Department of Computing and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain
e-mail: i22dimal@uco.es

1 Introduction

Action recognition has become an important research topic in several fields, such as video analysis, video surveillance, and human–computer interaction, thus attracting a great deal of attention in recent years [1, 2]. The problem has been addressed using different approaches, both in the 2D and 3D domains, with either single monocular cameras [3–5], single stereo cameras [6–8], or multiple monocular cameras [9–12]. However, developing robust methods of recognition independent of the view employed remains an unsolved problem [13].

Action recognition approaches can be roughly divided into two main categories: holistic and sequential. Holistic approaches consider the whole action as a shape varying in time and space, which can be represented by a single-feature vector. Then, inferences can be drawn by different methods such as Clustering, Neural Networks, or SVM. In contrast, sequential approaches consider the action as a series of individual poses, so that a different feature vector is extracted for each frame. Then, inferences can be drawn by models such as HMM or Conditional Random Fields, which capture the intrinsic temporal structure of the action as a sequence of state transitions.

While most research efforts have been focused on recognition from arbitrary single view-points, when multiple views are available during the recognition stage, three-dimensional information can be exploited to improve the success of recognition. The works of Weinland et al. [9] and Peng et al. [14], constitute the two most notable examples of three-dimensional viewpoint-invariant action recognition in the holistic and sequential approaches, respectively. In the first work, the authors propose the MHV as a method for compressing the voxel's occupancy of a whole action into a single voxelset. Then, the spectrum

Published online: 30 September 2011

of the Fourier transform is obtained, which is rotation invariant, and the result is reduced using PCA. Finally, recognition is performed using a Clustering method. In the second work, the voxel data of each frame is reduced using Multilinear Analysis and Hidden Markov Models (HMM) for inference.

In both cases, manipulating the large amount of information of three-dimensional voxel representations poses a major problem. As a consequence, standard Dimensionality Reduction techniques must be applied prior to the use of Machine Learning approaches. However, a direct application of Dimensionality Reduction techniques to voxel data results in sub-optimal strategies, both in terms of computational efficiency and recognition performance. The reason is that they require the use of large matrices that are difficult to manipulate and determine accurately, given the small number of training examples frequently available. Therefore, a more compact representation of voxel data would be preferable.

This work proposes the use of the volume integral (VI), an intermediate representation of voxel data that can be used before applying standard Dimensionality Reduction techniques in both holistic and sequential approaches. The VI reduces voxel data with a minimum loss of information by projecting it onto a set of planes that maximize the discrimination of the actions. Our descriptor allows a reduction of data by one and three orders of magnitude, respectively, compared with Weinland's and Qian's approaches. Consequently, Dimensionality Reduction techniques can be more efficiently applied without affecting classification performance.

Another remarkable property of the VI is that, unlike Weinland's descriptor, it is sensitive to reflection. In other words, it is able to differentiate gestures performed with different limbs. We consider this as an important advantage of our method since it doubles the number of possible gestures that can be represented. For instance, waving the right hand might trigger an action different from the one triggered by waving the left hand.

Our descriptor has been evaluated in the public dataset Inria Xmas Motion Acquisition Sequences (IXMAS) to compare it with Weinland's and Qian's works, which employ the same dataset. The VI has been applied using several Dimensionality Reduction techniques and different Machine Learning approaches (both sequential and holistic) to determine the best strategy for the action recognition task. The results show that the proposed descriptor improves on the previous work in terms of both computing times and classification rates.

The remainder of this paper is structured as follows: The rest of this section is devoted to related work in the field of action recognition. Section 2 defines the basis of three-dimensional action representation approaches, including the definition of our proposal. Section 3 presents the experiments carried out, while Sect. 4 draws conclusions.

1.1 Related work and proposed contribution

View-invariant action recognition approaches can be broadly divided into two main categories: model-based and view-based. The former rely on a parametric model of the person (normally comprised of a set of connected joints in a tree structure), which is matched against the input observations [15–17]. The latter, however, rely on a set of observations from which a set of features are learned and employed for recognition. Our work is related to the latter category, which has two main sub-categories: holistic and sequential.

Our approach can be seen as the 3D extension of the classical *silhouette projection histograms*, a very popular method for two-dimensional action recognition proposed by Haritaoglu et al. [18]. They showed that silhouette projection histograms are a very simple yet effective representation mechanism [19–21]. Their main disadvantage is that the results depend strongly on the view. Thus, a great part of the effort on view-based invariant gesture recognition has been focused on recognizing actions observed from single arbitrary viewpoints, using models acquired from either single or multiple views. One of the earliest holistic approaches is the Motion-History Image (MHI), proposed by Bobick and Davis [3]. In their work, temporal templates are created by analyzing movement information, and gestures are identified by measuring distances to templates previously stored from multiple views. Later, Yilmaz and Shah propose the use of spatio-temporal volumes (STVs), which are automatically generated for actions observed from any viewing direction using a graph-theoretic approach. STVs are later extended to multiple-views by Pingkun et al. [12]. They propose the use of 4D action feature model (4D-AFM) for recognizing actions from arbitrary viewpoints. The 4D-AFM is created by concatenating in time a sequence of reconstructed visual hulls (VHs) and then computing the differential geometric properties of the STVs. Also, Souvenir and Parrigan [22] propose a manifold learning-based framework which is tested on MHI and on the $\mathcal{R}$ Transform Surface Motion Descriptor.

When a sequential approach is selected, HMM are most frequently employed. Lv and Nevatia [23] propose an example-based action recognition method by the use of *Action Nets*. The action nets are automatically generated graph models that contain the 2D representation of one view of 3D key poses, in which links represent transitions between key poses. Inspired by that work, Ji and Liu [24] propose a simpler approach, in which actions are first clustered in the 3D space and then projected onto virtual

cameras. On the other hand, Weinland et al. [25] propose the use of a wrapper approach to determine the key exemplars; Chakraborty et al. [5] propose a different approach, in which individual classifiers are trained for specific body parts seen from specific viewpoints. Then, actions are recognized by a component-wise HMM that takes as input the classifier results.

In any case, recognition from a single monocular view suffers from ambiguities produced by projecting on two-dimensional images, e.g., it is difficult to determine if a person is pointing towards the camera or away from it. Thus, some authors have proposed the use of stereo vision. Shin et al. [6] were the first to extend MHIs to 3D using stereo input sequences for generating Motion History Models (MHMs). Later, Muñoz-Salinas et al. [8] presented stereo-based extensions to several monocular approaches, showing that using stereo information leads to better results in all cases. Recently, Roh et al. [7] proposed an extension of MHI to three-dimensional space called volume motion template (VMT). It is obtained by projecting stereo information to a 2D orthogonal plane that maximizes the discrimination of the action.

Despite the great attention that has been paid to the recognition of actions from a single viewpoint, there are applications in which multiple views are available during the recognition phase. These approaches rely on voxel data, a grid-based division of the 3D space into cubes of the same size that can be either occupied or empty. The set of voxels occupied by an actor forms the so-called visual hull (VH), which is a rich description of the actor's shape. Although 3D approaches can yield more descriptive models and thus achieve higher success rates, few view-based approaches have been proposed to exploit this fact.

The work of Weinland et al. [9] is one of them. The authors propose the motion history volume (MHV), a natural 3D extension of MHI, which is a representation of the sequence of VHs that form an action. The MHV is processed using the Fourier transform in order to obtain its spectrum, which is invariant under rotation. Finally, the spectrum is reduced using principal component analysis (PCA), and inference is performed using Clustering. Nonetheless, their approach has two main drawbacks. First, the number of features that result from the Fourier spectrum extraction is very high. After taking advantage of spectrum symmetry, the number of features to be reduced with (PCA) is $(n_\theta \times n_z \times n_r)/2$, where n_θ, n_z and n_r are the numbers of subdivisions on each axis. As a consequence, the Dimensionality Reduction process is slow, and in some cases, the number of observations is insufficient to obtain reliable covariance matrices. Second, a side effect of using the Fourier spectrum is that the resulting descriptor is invariant under reflection. Thus, gestures performed with different limbs produce similar descriptors, which might be a problem in certain applications.

In contrast, the work of Peng et al. [14] presents a sequential approach based on HMM. It consists of reducing voxel data by pose tensor decomposition using the High Order Singular Value Decomposition method, which is a generalization of the singular value decomposition method for matrices of arbitrary dimensions. Although their method improves on the results obtained by Weinland's, it suffers from several weaknesses. First, it requires a discretization of all possible body orientations so as to create tensors in each direction. Second, and partially as a consequence of the former, the pose tensor requires a large amount of memory. It requires $n_x \times n_y \times n_z \times r \times n$ features, where n_x, n_y and n_z denote the number of subdivisions in each axis; the parameter r represents the number of possible body orientations; and n is the number of key poses (these are employed to calculate the core tensor). These values are set to $n_x = n_y = n_z = 32, r = 16$ and $n = 25$ in their work. We consider that the large amount of information required by this method makes it inappropriate for many applications and that the use of lighter methods would be preferable. Our comparison to this method will be based on the results reported by the authors.

2 Action representation

This section explains the basis of the methods employed for action representation. The first two sections explain the techniques employed for representing 3D actions. First, Sect. 2.1 explains the concept of VH and how it is created using the Shape-from-Silhouette (SfS) algorithm. Whereas sequential approaches use the VH generated at each frame for recognition, holistic approaches use the MHVs, which compresses a set of frames into a single voxelset (Sect. 2.2).

Once a suitable 3D representation of the actions is obtained, the information is transformed so as to reduce the amount of data, and to achieve rotation invariance. Section 2.3 explains Weinland's approach while Sect. 2.4 describes our proposal, the VI.

2.1 Visual hull extraction

Let us assume that the monitored area can be divided into cubes of the same volume (called voxels) denoted by

$$V = \{v^i | i = (x^i, y^i, z^i)\}, \tag{1}$$

where $i = (x^i, y^i, z^i)$ represents the voxel's center in Cartesian coordinates and $v^i \in \{1, 0\}$ depending on whether the voxel is occupied or empty.

The VH of an actor consists of all the occupied voxels that belong to him, and it can be calculated using the well-known SfS algorithm [26], although more recent approaches can also be employed [27, 28]. For that purpose, we require a set of cameras surrounding the actor and placed at known locations determined by calibration. In addition, a background subtraction method is required to obtain the actor's silhouettes.

The traditional SfS method examines the projections of the voxels in the foreground images in order to determine whether they belong to the shapes in question. This is achieved by means of a projection test. Each voxel is projected in all the foreground images, and if its projection lies completely inside a silhouette in all the foreground images, then it is considered occupied. However, if the voxel projects to a background region in any of the images, it is considered empty. Finally, if the voxel projects partially onto a foreground region it is considered to belong to a border, and a decision must be made. In the end, the result is a Boolean decision {0,1} indicating whether the region of the space represented by the voxel is empty or occupied.

Figure 1a shows the VHs obtained for three instants of time in one of the *kick* sequences of the IXMAS dataset.

2.2 Motion history volumes

The motion history volumes compress a set of VH into a single probabilistic representation of the voxels' occupancy. Let us define v_t^i as the value of voxel v^i at time t. Then the MHV function is defined by

$$\mathcal{V}_\tau(i,t) = \begin{cases} \tau & \text{if } v_t^i = 1 \\ \max(0, \mathcal{V}_\tau(i,t-1) - 1) & \text{otherwise} \end{cases}, \quad (2)$$

where τ is the duration, in number of frames, of the action represented. At instant t, $\mathcal{V}_\tau(i,t)$ has values close to τ if the corresponding voxel has been recently observed as occupied. Values of $\mathcal{V}_\tau(i,t)$ near 0 indicate that the voxel has not been occupied recently, and the value 0 indicates that it has not been observed in occupation of any of the sequence frames.

In order to perform a reliable comparison between the MHVs, they must be normalized. Assuming that the start and end time of the action is known, the MHV is normalized as

$$\mathcal{V}(i) = \frac{\mathcal{V}_\tau(i,\tau)}{\tau}. \quad (3)$$

Invariance to translation and scale are achieved by calculating the means μ_x, μ_y, μ_z and variances $\sigma_x, \sigma_y, \sigma_y$ of the non-empty voxels in the MHV and then shifting and scaling it so that $\mu_x = \mu_y = \mu_z = 0$ and $\sigma_x = \sigma_y = \sigma_z = 1$.

Figure 1c shows the MHV of the action shown in Fig. 1a. Red colors represent values near 1 while blue correspond to values close to 0. Then, more recently occupied voxels correspond to redder areas.

The main advantage of the MHV is that it compresses a whole sequence into a single voxelset. However, it has two main drawbacks. First, it requires knowledge of the start and end of the sequence. Second, the speed at which the action is performed affects the MHV obtained. On the other hand, sequential approaches are in theory less sensitive to that problem because they are capable of modeling the intrinsic dynamics of the action as state transitions.

2.3 Weinland's descriptor

Although the VHs and MHV are invariant under scaling and translation, they are not invariant under rotation. The descriptor proposed by Weinland et al. [9] achieves rotation invariance by using the Fourier transform (FT), taking into account the Fourier *shift theorem*. This states that a

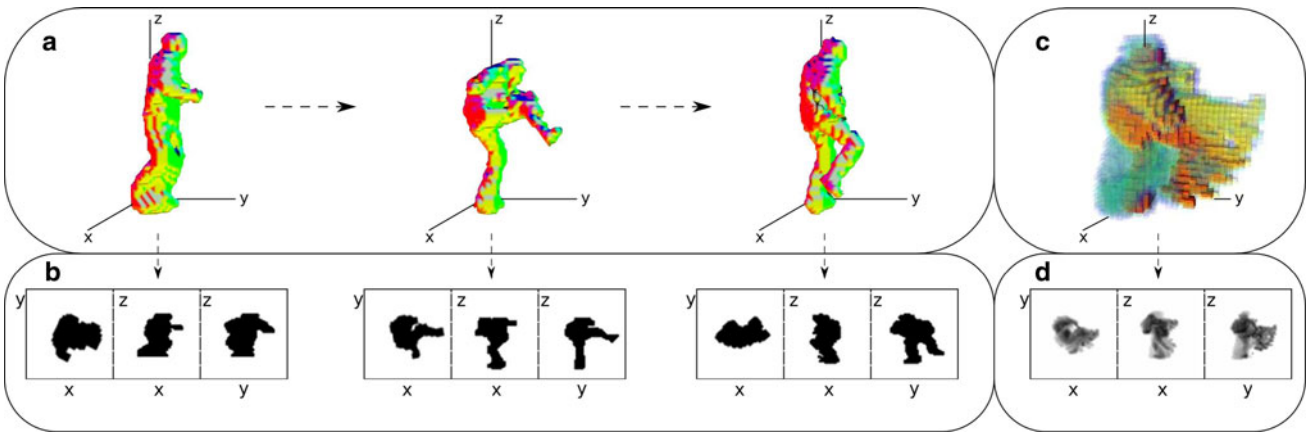

Fig. 1 **a** Set of visual hulls of the *kick* action. **b** Volume integrals of the actions' visual hulls. **c** MHV of the action. **d** Volume integral of the MHV

function $f_0(x)$ and its translated counterpart $f_t(x) = f_0(x - x_0)$ differ only by a phase modulation in the Fourier space:

$$F_t(k) = F_0(k)\mathrm{e}^{-j2\pi k x_0}.$$

In this descriptor, voxels are expressed in a cylindrical coordinate system (see Fig. 2) defined as follows:

$$\mathcal{C}(r, \theta, z) = \mathcal{V}(r\cos(\theta), r\sin(\theta), z)$$

Thus, rotations around the vertical z-axis result in cyclical translation shifts:

$$\mathcal{V}(x\cos(\theta_0) + y\sin(\theta_0), -x\sin(\theta_0) + y\cos(\theta_0), z) \rightarrow \mathcal{C}(r, \theta + \theta_0, z).$$

Hence, rotation invariance along θ for each pair of values (r, z) is achieved using the Fourier spectrum values $|\mathsf{C}(r, k_\theta, z)|$ of the 1D Fourier transform

$$\mathsf{C}(r, k_\theta, z) = \int_{-\pi}^{\pi} \mathcal{C}(r, \theta, z)\mathrm{e}^{-j2\pi k_\theta \theta}\mathrm{d}\theta. \quad (4)$$

The descriptor is then defined as the vector

$$\mathbf{W} = \{\mathbf{w}^{z,r} | z = 1, \ldots, n_z, r = 1, \ldots, n_r\}, \quad (5)$$

where

$$\mathbf{w}^{z,r} = \{|\mathsf{C}(r, k_\theta^0, z)|, \ldots, |\mathsf{C}(r, k_\theta^{n_\theta}, z)|\}, \quad (6)$$

The parameters n_z, n_r and n_θ represent the number of divisions employed in r-, θ- and z-axes, respectively.

Due to the trivial ambiguity of 1D-Fourier magnitudes with respect to the reversal of the signal, motions that are symmetric with respect to the z-axis (i.e., identical actions performed with different limbs) result in the same motion descriptors. Invariance under reflection can be considered either as a loss of information or as a useful property, depending on the final application. In any case, considering the symmetry of the Fourier spectrum, the total number of features of the descriptor is $(n_\theta \times n_z \times n_r)/2$, which is very large.

Although in his original work Weinland applied the descriptor only to reduce MHVs, note that the descriptor can also be employed to reduce individual VHs. Actually, this work demonstrates (in the experimental section) that Weinland's descriptor obtains better results when it is employed in combination with a sequential approach (HMM) than when the MHV is employed.

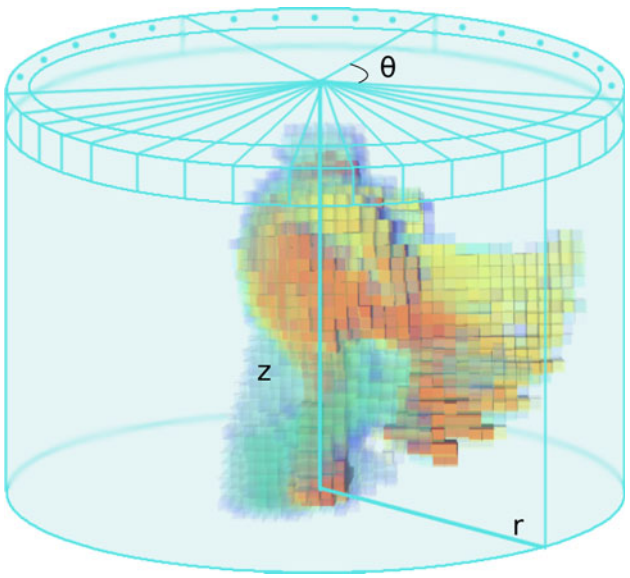

Fig. 2 Representation of Weinland's descriptor. Fourier transforms over θ are computed for (r, z) pairs forming the feature vector

2.4 Volume integral

As can be observed, the amount of information obtained in a voxel-based representation is very high. Hence, it is desirable to reduce it before applying Machine Learning techniques. To that end, we propose the VI, which is an alternative representation of a voxelset that can be applied both to VHs and MHVs. The idea is to project the three-dimensional information to a set of planes that maximize the action's discriminability.

For the calculus of the VI, it is assumed that there is a voxelset invariant under translation and scaling. This is achieved using the transformation procedure explained above for the MHV (i.e., calculating the means and variances of the voxelset and then shifting and rescaling it so that $\mu_x = \mu_y = \mu_z = 0$ and $\sigma_x = \sigma_y = \sigma_z = 1$). Afterwards, rotation invariance is obtained by finding the angle that aligns two identical voxelsets. The actor's main direction is then obtained and the voxelset rotated accordingly. Assuming that people stand while performing actions, the angle search can be restricted to the x–z plane. Thus, the actor's main direction is defined as the vector from the center of the volume to the farthest occupied voxel.

The resulting voxelset is then integrated over the three axes in pairs, so that the 3D information is represented by three 2D images with a minimum loss of information. Let us define the integrals of the voxelset as

$$\mathcal{P}_z(x, y) = \frac{1}{\sum_{i=1}^{n_z} \delta(\mathcal{V}(x, y, z^i))} \sum_{i=1}^{n_z} \mathcal{V}(x, y, z^i), \quad (7)$$

$$\mathcal{P}_y(x, z) = \frac{1}{\sum_{i=1}^{n_y} \delta(\mathcal{V}(x, y^i, z))} \sum_{i=1}^{n_y} \mathcal{V}(x, y^i, z), \quad (8)$$

and

$$\mathcal{P}_x(y, z) = \frac{1}{\sum_{i=1}^{n_x} \delta(\mathcal{V}(x^i, y, z))} \sum_{i=1}^{n_x} \mathcal{V}(x^i, y, z). \quad (9)$$

The function $\mathcal{V}(x, y, z)$ is the value of the voxel with center (x, y, z). When calculating the VI from a VH, $\mathcal{V}(x, y, z)$

represents the voxel's state, i.e., $\mathcal{V}(x,y,z) \in \{0,1\}$, indicating whether the voxel is empty or occupied. However, when the VI represents a MHV, $\mathcal{V}(x,y,z) \in [0,1]$ is the continuous value given by Eq. 3. The function δ is 0 if the voxel's occupancy is 0 and 1 otherwise. Therefore, the integrals are normalized by dividing by the number of non-empty voxels. If all the voxels are empty for a specific pixel of the plane then its final value is 0.

Note that the integral is normalized by dividing by the number of non-empty voxels instead of by the total number of subdivision (n_x, n_y or n_z). We have experimentally found that this approach yields better results, especially when dealing with MHV. The reason is that for regions such as limbs, the sum of the occupancy is very low, and dividing by the non-empty voxels better preserves the information for these parts.

Figure 1c, d shows the VI corresponding to the voxel-sets of Fig. 1a and b. It can be observed that our normalization approach produces binary silhouettes when applied to VHs, but it preserves the most essential spatio-temporal properties when applied to MHVs.

Finally, a unique descriptor is formed by concatenating all the features in $\mathcal{P}_y(x,z), \mathcal{P}_z(x,y)$ and $\mathcal{P}_x(y,z)$. Figure 3 outlines the main steps of the proposed method. Note that the two first steps of the algorithm are explained in Sect. 2.2

When compared with Weinland's descriptor, the VI achieves a substantial reduction of the data because the final feature vector contains only $n_x \times n_y + n_x \times n_x + n_y \times n_z$ elements. For instance, for a voxel set of resolution $64 \times 64 \times 64$ (as employed in our experiments), Weinland's descriptor produces 131,072 features, while ours produces 12,288, i.e., approximately ten times fewer features. Compared with Qian's descriptor, our descriptor produces one hundred times fewer features. As we show in the experimental section, these are sufficient to provide better results than Weinland's descriptor. In addition, the VI produces different descriptors for reflected actions. Hence, similar actions performed with different limbs can be distinguished.

Volume Integral creation
INPUT: Volume set $\mathcal{V}$
OUTPUT: Volume integral

1. **Shift** voxels in $\mathcal{V}$ so that the resulting volume has $\mu_x = \mu_y = \mu_z = 0$ and $\sigma_x = \sigma_y = \sigma_z = 1$.
2. **Rotate** voxels in $\mathcal{V}$ by finding the main's direction as the vector from the center of the volume to the farthest occupied voxel.
3. **Project** voxels in the three planes xy, xz and yz so as to obtain the volume integrals as:
$\mathcal{P}_z(x,y) = \frac{1}{\sum_{i=1}^{n_z} \delta(\mathcal{V}(x,y,z^i))} \sum_{i=1}^{n_z} \mathcal{V}(x,y,z^i)$
$\mathcal{P}_y(x,z) = \frac{1}{\sum_{i=1}^{n_y} \delta(\mathcal{V}(x,y^i,z))} \sum_{i=1}^{n_y} \mathcal{V}(x,y^i,z)$
$\mathcal{P}_x(y,z) = \frac{1}{\sum_{i=1}^{n_x} \delta(\mathcal{V}(x^i,y,z))} \sum_{i=1}^{n_x} \mathcal{V}(x^i,y,z)$
4. **Concatenate** $\mathcal{P}_z(x,y)$,$\mathcal{P}_y(x,z)$ and $\mathcal{P}_x(y,z)$ creating a single 2D feature matrix representing the volume integral.

Fig. 3 Algorithm that transform a 3D volume into a volume integral

3 Experimental results

This section presents the experiments conducted to evaluate the proposed descriptor. For testing purposes, the public IXMAS dataset has been employed. It is a well-known 3D dataset in the field of action recognition that contains 13 actions, each one performed three times by 12 different actors. The actors freely change their orientations in each acquisition so that rotation invariance can be analyzed.

The goal of our experiment is fourfold. First, we seek to evaluate the performance of the VIs using different combinations of Dimensionality Reduction techniques and Machine Learning approaches in order to find the most appropriate one. Second, we compare the discriminability power of our descriptor to the one proposed by Weinland [9]. For that purpose, the descriptors are evaluated using the original dataset and the results are presented in Sect. 3.1. Third, the tests evaluate the discriminatory power of the descriptors when reflected actions are considered, i.e., identical actions performed with different limbs. For that purpose, a total of four new actions are added to the original dataset by applying a reflection operator to some of the original actions. The results obtained are shown in Sect. 3.2. Finally, in Sect. 3.3, we also aim to evaluate the improvement in the computational time achieved by our descriptor when compared with Weinland's.

We have employed the same evaluation methodology employed in Ref. [9] so as to obtain comparable results. This means that: (a) we have used 10 of the 12 actors in the dataset; (b) a leave-one-out cross-validation is applied, i.e., one actor is chosen for testing and the rest are used for training; (c) actions 11 (point) and 13 (throw over head) have not been included, because they are performed very differently by the actors, and (d) the space is discretized into 64 subdivisions for each axis. In addition, the VHs and ground-truth provided with the dataset have been employed.

3.1 Comparative evaluation in the IXMAS dataset

The first test compares our descriptor to Weinland's in the original IXMAS dataset using different combinations of Dimensionality Reduction and Machine Learning approaches. The results of the tests are shown in Table 1.

Table 1 Classification results in the original IXMAS dataset

Mach. Learn	Dim. Red	Weinland	Volume integrals
Clust.	PCA	93.33	90.90
Clust.	2D-PCA	–	85.15
Clust.	PCA+LDA	92.42	90.90
SVM-RBF	PCA	80.90	90.00
SVM-RBF	2D-PCA	–	77.27
SVM-RBF	PCA+LDA	89.39	86.69
SVM-LIN	PCA	89.09	88.87
SVM-LIN	2D-PCA	–	86.66
SVM-LIN	PCA+LDA	86.69	88.48
HMM	PCA	71.21	76.06
HMM	2D-PCA	–	87.57
HMM	PCA+LDA	93.64	98.45

The first column indicates the Machine Learning approach employed, and the second column indicates the Dimensionality Reduction technique applied. In the first column, the *SVM-RBF* stands for Support Vector Machines (SVM) using radial basis function, while *SVM-LIN* indicates that a linear kernel is being employed. The third and fourth columns show the results obtained for the different descriptors. For the *SVM-RBF*, *SVM-LIN* and HMM methods, the results presented correspond to the best results after testing with different values of the methods' parameters. As previously explained, the Clustering, *SVM-RBF* and *SVM-LIN* methods are applied using the MHVs. However, the HMM tests employ the individual VHs generated at each frame.

We note several remarks regarding the Dimensionality Reduction techniques. First, the complete two-dimensional principal component analysis (2D-PCA) technique has been applied to our descriptor but not to Weinland's descriptor, since the technique requires the data to be a two-dimensional matrix. Different compression ratios have been tested, and the values in the Table correspond to those that provided the best results. Second, to determine the number of PCA components, we have employed the minimum-error formulation suggested by Bishop [29]. For the VI, we retain, approximately, the first 123 principal components out of the 12,288 features given by the descriptor. For Weinland's descriptor, we employed approximately 233 components out of the 131,072 features. Finally, for the linear discriminant analysis (LDA) method, we retained only the ten most relevant features, i.e., the number of classes minus one.

The results obtained show that Weinland's descriptor obtains better results than ours when the Clustering and SVM-LIN methods are applied. However, the VI obtains better results when using the SVM-RBF and HMM methods. In general, the results show that the combination of HMM+ (PCA+LDA) obtains the best results for both descriptors. In particular, our method obtains, in the best case, a 98.45% success rate. This result is obtained using HMM with 12 states and no skips between states.

In order to compare both methods more formally, we have used a *Paired-Samples* test using the data in Table 1 (excluding the 2D-PCA results because they are not available in both methods). For more information on statistical analysis see [30, 31]. The first three rows of Table 2 show the performance statistics of both methods, while the last row shows the statistics calculated by the Paired-Samples test. As indicated in the table, the significance of the null hypothesis is 0.290 (following a two-tailed *T-Distribution* with 7 degrees of freedom (DOF)). Using a p value > 0.05, the test indicates that there is not enough evidence to reject the null hypothesis. In other words, there is insufficient evidence to suggest that one method is better than the other in terms of performance in the dataset employed.

Finally, we must indicate that the work of Qian et al. [14], reported a success rate of 94.59%. To allow replication of our experiments, we provide the complete set of VIs corresponding to the VHs and MHVs of the IXMAS actions, along with references to the libraries we have used.

Note to reviewers: The final location of this data has not yet been determined since it might be located in the journal servers. However, the dataset is temporarily available at http://www.uco.es/users/i22dimal/files/data-vi.tar.bz2 for reviewing purposes.

Table 2 Statistics of the distributions and results of the paired-samples test

Method	Mean	Std. deviation	Std. error			
Weinland	0.8708	0.0765	0.0270			
VI	0.8879	0.0621	0.0219			
	Mean	Std. deviation	Std. mean error	t	DOF	Sig. (2-tailed)
Weinland-VI	−0.0171	0.0422	0.0149	−1.145	7	0.290

See text for discussion

3.2 Comparative evaluation using reflected actions

The goal of this test is to evaluate the ability of our method to deal with reflected actions. For that purpose, a new dataset has been created by adding four new actions to the original IXMAS dataset. The new actions are created by applying a reflection operator to the actions check watch, wave hand,kick, and punch. An example can be seen in Fig. 4, which shows the original VH of one of the sequences of the punch action, along with its reflected counterpart. It can be observed that the action is identical except for the fact that it seems to be performed with the opposite arm.

The results using the two descriptors are shown in Table 3. The parameters of the Machine Learning and Dimensionality Reduction techniques have been obtained as in the previous test.

As previously explained, the descriptor proposed by Weinland is not capable of distinguishing between reflected action because of the symmetry of the Fourier spectrum. This explains the bad results obtained in this test.

Fig. 4 Reflection of action "punch"

Table 3 Classification results in the IXMAS dataset using reflected actions

Mach. Learn	Dim. Red	Weinland	Volume integrals
Clust.	PCA	70.00	84.88
Clust.	2D-PCA	–	80.66
Clust.	PCA+LDA	67.99	85.55
SVM-RBF	PCA	68.44	83.55
SVM-RBF	2D-PCA	–	70.66
SVM-RBF	PCA+LDA	70.22	84.88
SVM-LIN	PCA	64.22	78.44
SVM-LIN	2D-PCA	–	70.80
SVM-LIN	PCA+LDA	59.99	81.77
HMM	PCA	52.2	72.50
HMM	2D-PCA	–	79.20
HMM	PCA+LDA	70.90	91.90

In contrast, the proposed VI achieves better classification results using all combinations of Machine Learning and Dimensionality Reduction techniques. As in the previous case, the best results are obtained for a combination of HMM and (PCA+LDA). The HMM had eight states and no skips between states.

The confusion matrices of each descriptor for the best case are shown in Figs. 5 and 6. The actions identified with the suffix *R* correspond to those reflected. It can be clearly observed that most of the errors of Weinland's descriptor correspond to confusion between an action and its reflected counterpart.

As in the previous case, we have employed the Paired-Samples test to compare the results. Table 4 shows the

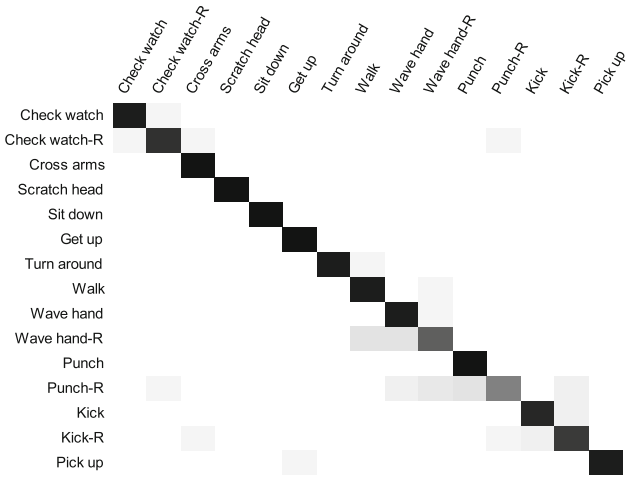

Fig. 5 Confusion matrix for the VI descriptor in the test with reflected actions. The results correspond to the combination of (PCA+LDA) and HMM

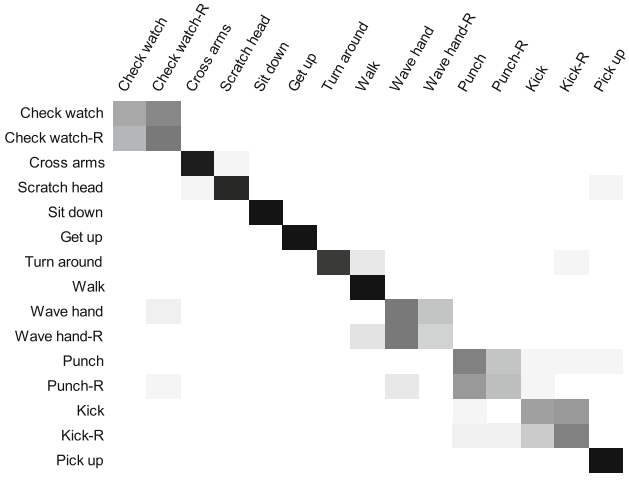

Fig. 6 Confusion matrix for Weinland's descriptor in the test with reflected actions. The results correspond to the combination of (PCA+LDA) and HMM

Table 4 Statistics of the distributions and results of the paired-samples test using the database of reflected actions

Method	Mean	Std. deviation	Std. error			
Weinland	0.6549	0.0649	0.0229			
VI	0.8293	0.0567	0.0200			
	Mean	Std. deviation	Std. mean error	t	DOF	Sig. (2-tailed)
Weinland-VI	−0.1743	0.0315	0.0111	−15.622	7	0.00

See text for discussion

Table 5 Times employed in computing and reducing data using PCA

Descriptor	Desc. Comp. (s)	PCA-Comp. (s)	PCA-Reduct. (s)
VI	4.6^{-3}	3.7	1.04^{-2}
Weinland	1.1^{-2}	37.5	1.14^{-1}

performance statistics of both methods, along with the statistics calculated by the test. In this case, the null hypothesis is rejected using a p value > 0.05, so we can affirm that the proposed method is significantly better than Weinland's in distinguishing actions performed with different limbs.

3.3 Evaluation of the computational cost

The experiments conducted so far show that the proposed descriptor outperforms Weinland's method in several situations while using far fewer features. The results also suggest that a combination of PCA+LDA provides the best performance. LDA computation is not a time-consuming process, since the number of features obtained after PCA is small. However, calculating the PCA eigenvectors and eigenvalues, as well as projecting the patterns to its main components, is computationally demanding processes.

This section aims to analyze the benefit of using the proposed VI descriptor in terms of computing time. To that end, we have measured the times needed to calculate the most relevant phases of the action recognition process for the two descriptors; the results are presented in Table 5. The first column indicates the descriptor employed, and the second column indicates the time required to calculate it from a single voxelset. The third column indicates the average time required to calculate the PCA eigenvectors and eigenvalues for the MHVs of all the train patterns in the IXMAS database (i.e., a total of 297 patterns). Finally, the last column indicates the amount of time required to reduce a single voxelset to its main components. The tests have been performed on a Intel Quad Core 2.66 Ghz with 4 Gb of RAM, running Linux, and the routines employed for computing PCA and the Fourier transform are those provided by the OpenCv library [32].

Note that the calculation of the VI requires half the time required by Weinland's descriptor. In addition, the computing times required for the PCA process are reduced by one order of magnitude. We propose that the reduction in the computational cost makes the proposed descriptor more suitable for a wide variety of applications in which time is a crucial aspect.

4 Conclusions

This paper proposes a new descriptor, the VI, for three-dimensional action recognition. Our descriptor (which is invariant under rotation, scaling and translation) is created by integrating voxel data over a set of planes that maximize the discriminability of actions. The main advantage of our descriptor is that it significantly reduces the amount of data of three-dimensional representations yet preserves the most important information. As a consequence, the action recognition process is greatly sped up, but very high success rates are maintained. In addition, the proposed descriptor can distinguish between reflected actions, which was not possible with some of the previous approaches.

The VI is tested in the IXMAS dataset and compared with the descriptor proposed by Weinland et al. [9]. The experiments evaluate the performance of the descriptors using several Dimensionality Reduction and Machine Learning techniques, so as to determine the most appropriate combination. The results show that: (a) the proposed descriptor obtains similar performance to Weinland's descripor in the IXMAS data set, and better performance on mirrored action; (b) the combination of PCA+LDA and HMM obtains the best classification results; and (c) the proposed descriptor reduces the amount of data to be manipulated by one order of magnitude compared with Weinland's approach. As a consequence, the time employed to reduce the dimensionality of our feature vector is also reduced by an order of magnitude.

Acknowledgments This work was developed with the support of the Research Project "TIN2010-18119" financed by Science and Technology Ministry of Spain.

References

1. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28:976–990
2. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circuits Syst Video Technol 18(11):1473–1488
3. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23:257–267
4. Ikizler N, Duygulu P (2009) Histogram of oriented rectangles: a new pose descriptor for human action recognition. Image Vis Comput 27(10):1515–1526
5. Chakraborty B, Rudovic O, Gonzalez J (2008) View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In: 2008 8th IEEE international conference on automatic face & gesture recognition. IEEE, pp 1–6
6. Shin H-K, Lee S-W, Lee S-W (2005) Real-time gesture recognition using 3D motion history model. In: Proceedings of ICIC (1), pp 888–898
7. Roh M-C, Shin H-K, Lee S-W (2010) View-independent human action recognition with volume motion template on single stereo camera. Pattern Recognit Lett 31(7)
8. Muñoz-Salinas R, Medina-Carnicer R, Madrid-Cuevas FJ, Carmona-Poyato A (2008) Depth silhouettes for gesture recognition. Pattern Recognit Lett 29:319–329
9. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104(2):249–257
10. Yang Y, Hao A, Zhao Q (2008) View-invariant action recognition using interest points. In: International multimedia conference
11. Cherla S, Kulkarni K, Kale A, Ramasubramanian V (2008) Towards fast, view-invariant human action recognition. In: 2008 IEEE Computer Society conference on computer vision and pattern recognition workshops. IEEE, pp 1–8
12. Pingkun Y, Khan SM, Shah M (2008) Learning 4D action feature models for arbitrary view action recognition. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–7
13. Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. IEEE Trans Syst Man Cybernet C (Appl Rev) 40(1):13–24
14. Peng B, Qian G, Rajko S (2009) View-invariant full-body gesture recognition via multilinear analysis of voxel data. ICDSC
15. Brubaker MA, Fleet DJ, Hertzmann A (2009) Physics-based person tracking using the anthropomorphic walker. Int J Comput Vis 87(1–2):140–155
16. Corazza S, Mündermann L, Gambaretto E, Ferrigno G, Andriacchi TP (2009) Markerless motion capture through visual hull, articulated ICP and subject specific model generation. Int J Comput Vis 87(1–2):156–169
17. Li R, Tian T-P, Sclaroff S, Yang M-H (2009) 3D human motion tracking with a coordinated mixture of factor analyzers. Int J Comput Vis 87(1–2):170–190
18. Haritaoglu I, Harwood D, Davis LS (2000) W4: real-time surveillance of people and their activities. IEEE Trans Pattern Anal Mach Intell 22:809–830
19. Haritaoglu I, Cutler R, Harwood D, Davis LS (1999) Backpack: detection of people carrying objects using silhouettes. Comput Vis Image Underst 81:102–107
20. Cucchiara R, Grana C, Prati A, Vezzani R (2005) Probabilistic posture classification for human-behavior analysis. IEEE Trans Syst Man Cybernet A: Syst Humans 35(1):42–54
21. Juang C-F, Chang C-M (2007) Human body posture classification by a neural fuzzy network and home care system application. IEEE Trans Syst Man Cybernet A: Syst Humans 37(6):984–994
22. Souvenir R, Parrigan K (2009) Viewpoint manifolds for action recognition. EURASIP J Image Video Process 2009:1–13
23. Lv F, Nevatia R (2007) Single view human action recognition using key pose matching and viterbi path searching. In: IEEE conference on computer vision and pattern recognition, pp 1–8
24. Ji X, Liu H (2009) View-invariant human action recognition using exemplar-based hidden Markov models. Lect Notes Comput Sci 5928:78–89
25. Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3D exemplars. In: 2007 IEEE 11th international conference on computer vision. IEEE, pp 1–7
26. Laurentini A (1991) The visual hull: a new tool for contour-based image understanding. In: Proceedings of seventh Scandinavian conference on image processing, pp 993–1002
27. Díaz-Más L, Muñoz-Salinas R, Madrid-Cuevas FJ, Medina-Carnicer R (2010) Shape from silhouette using dempster-shafer theory. Pattern Recognit 43(6):2119–2131
28. Landabaso JL, Pardàs M, Ramon Casas J (2008) Shape from inconsistent silhouette. Comput Vis Image Underst 112:210–224
29. Bishop CM (2007) Pattern recognition and machine learning (information science and statistics), 1st edn, 2006. Springer. corr. 2nd printing edition, October 2007
30. Sheskin DJ (2007) Handbook of parametric and nonparametric statistical procedures, 4th edn. Chapman & Hall/CRC
31. Devore JL, (2008) Probability and statistics for engineering and the sciences, 7th edn. Thomson Brooks/Cole
32. Intel. OpenCV: Open source Computer Vision library. http://www.intel.com/research/mrl/opencv.

3

Second contribution: Shape from silhouette using Dempster–Shafer theory

Pattern Recognition 43 (2010) 2119–2131

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Shape from silhouette using Dempster–Shafer theory

L. Díaz-Más *, R. Muñoz-Salinas, F.J. Madrid-Cuevas, R. Medina-Carnicer

Department of Computing and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain

ARTICLE INFO

Article history:
Received 10 July 2009
Received in revised form
26 October 2009
Accepted 5 January 2010

Keywords:
Dempster–Shafer
Shape-from-silhouette
Multi-camera
Visual hull

ABSTRACT

This work proposes a novel shape from silhouette (SfS) algorithm using the Dempster–Shafer (DS) theory for dealing with inconsistent silhouettes. Standard SfS methods makes assumptions about consistency in the silhouettes employed. However, total consistency hardly ever happens in realistic scenarios because of inaccuracies in the background subtraction or occlusions, thus leading to poor reconstruction outside of controlled environments.

Our method classify voxels using the DS theory instead of the traditional intersection of all visual cones. Sensors reliability is modelled taking into account the positional relationships between camera pairs and voxels. This information is employed to determine the degree in which a voxel *belongs to* a foreground object. Finally, evidences collected from all sensors are fused to choose the best hypothesis that determines the voxel state.

Experiments performed with synthetic and real data show that our proposal outperforms the traditional SfS method and other techniques specifically designed to deal with inconsistencies. In addition, our method includes a parameter for adjusting the precision of the reconstructions so that it could be adapted to the application requirements.

1. Introduction

Shape from silhouettes (SfS) is a well known approach to reconstruct the three-dimensional structure of objects using a set of silhouettes obtained from different views. Baumgart [2] was the first in introducing some concepts about the 3D geometric modelling, but it was not until 1991 that Laurentini [26] defined the concept of the Visual Hull (VH) as the largest 3D solid equivalent to the real object that explains the silhouettes extracted. The VH is obtained as the geometric intersection of all visual cones explaining the projection of a silhouette in its corresponding camera image. Consequently, as the number of views increase, so does the precision of the reconstructed object [27].

A major problem of standard SfS methods is that they are strongly linked to the principle of *silhouette consistency*, i.e., the set of silhouettes employed must explain precisely the real object. A single inconsistency in one of the silhouettes makes the VH reconstructed no longer equivalent to the real object. However, total consistency hardly ever happens in real-life scenarios due to several factors such as inaccuracies in camera calibration, foreground extraction errors [15,20,22], occlusions, etc. Therefore, SfS methods have been usually confined to problems under controlled conditions [8,17,40,41,47].

In recent years, a number of works have addressed the SfS problem in different ways [7,17,24,25,40]. These approaches aim to exploit the information redundancy wisely in order to overcome the inconsistency problem using probabilistic approaches. Despite the improvements achieved they have missed the use of an important piece of information which can be exploited in order to improve reconstruction: the positional relationship of the cameras. Information about the relative camera locations can be used to detect inconsistencies as we show in this work. For instance, if a camera indicates that a voxel is part of the object and another camera close to the former indicates the opposite, then there is an inconsistency that might be solved by a third camera.

This paper proposes a novel approach to solve the SfS problem by employing information about the camera relationships using the Dempster–Shafer (DS) theory of evidence [1,3,5,11,18,36]. In our approach, camera pairs are employed as sensors instead of individual cameras. Each sensor employs a confidence model which takes into account the relative position of their cameras in relation to the voxels analyzed. Besides, sensors provide degrees of evidence about the occupancy of voxels which are fused using the Dempster's rule of combination in order to provide a final value representing the probability of a voxel to be occupied. The experiments conducted (both in synthetic and real environments) show that our method outperforms traditional and new SfS approaches (namely Ref. [25]), specially in the presence of strong noise.

The remainder of this paper is structured as follows. The rest of this section provides an overview of the most relevant works

* Corresponding author. Tel.: +34 625680220; fax: +34 957 211035.
E-mail addresses: piponazo@gmail.com, i22dimal@uco.es (L. Díaz-Más), rmsalinas@uco.es (R. Muñoz-Salinas), ma1macuf@uco.es (F.J. Madrid-Cuevas), rmedina@uco.es (R. Medina-Carnicer).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.01.001

related to ours along with its main contributions. In Section 2, the basis of the DS theory of evidence are explained. The standard SfS method along with its limitations are shown in Section 3 while Section 4 details our proposal. Finally, Section 5 explains the results obtained and some final conclusions are drawn in Section 6.

1.1. Related works

Many SfS algorithms have been proposed for creating volumetric models since the initial definition of VH by Laurentini [26]. While some methods employ a set of cameras distributed around the area of work, others employ a single camera and place the object on a turntable [8,41,47]. Either way, foreground extraction plays an essential role in the extraction of the VH. Traditionally, the two classical approaches are voxel sets and octrees. In the first approach, the entire area of interest is discretized in a three-dimensional grid of voxels of the same volume. Then, voxels are projected in all the images to check whether they belong to foreground objects. The second approach is the octree structure which is based on a tree of voxels. Octree-based methods [6,21], starts with a cube that cover all the working area which is recursively subdivided into eight voxels until they reach an homogeneous content (shape or background) or a maximum resolution has been reached. In this work, we are interested in voxel sets.

Standard SfS methods are strongly linked to the principle of *silhouettes consistency*. However, full consistency hardly even happens in realistic scenarios mainly because of segmentation errors. As a result, the models reconstructed might contain holes or added spots in the background area. Morphological operators can be applied in a cleanup phase, but only to mitigate the problem. Recently, some approaches have addressed the problem of silhouette inconsistencies proposing different algorithms that minimizes the propagation of 2D miss-detections to the 3D models.

A first approach which minimizes the probability of voxel miss-classification is shown in [7]. In their work, Cheung et al. propose the *sparse pixel occupancy test* (SPOT) algorithm, which determines the minimum number of foreground pixels lying inside each voxel projection as a constraint to pass the projection test. This projection test must be passed in all the views to classify the voxel as part of the shape as in traditional SfS methods. Although this approach constitutes improvements over the original SfS algorithm, they do not employ the concept of inconsistent silhouettes so that 2D miss-classifications are propagated to the 3D reconstruction.

In [40], Snow et al. do not work with traditional 2D silhouettes. Instead, they replace the silhouettes hard constraint with a soft constraint that minimizes an energy function. They also suggest a relaxation of the visual cones intersection constraint. Instead of requiring the intersection of a number of visual cones equal to the total number of cameras, they suggested the intersection of $C-X$ cones, where X is the number of acceptable false alarms among the set of C cameras. Using this approach, a single miss does not break the consistency constraint of silhouettes. The main disadvantage of their approach is that the VHs reconstructed are larger than the original ones, specially when several objects are present in the scene.

Later, in [17] the SfS problem is restated proposing a new framework for multi-view silhouette cue fusion. For that purpose, they use a space occupancy grid as a probabilistic 3D representation of scene contents. The idea is to consider each camera pixel as a statistical occupancy sensor and to use them jointly to infer where, and how likely, matter is present in the scene. As in previous approaches, inconsistencies in silhouettes are not taken into account thus propagating 2D miss-classifications.

Finally, [25] introduces the concepts of inconsistent volume (IV) and unbiased Hull (UH) to propose a novel reconstruction scheme taking into account inconsistencies in realistic scenarios. In an initial step, the VH is calculated with the standard SfS method. Then, a decision on the voxels not forming part of the VH is taken in a second step by minimizing the error probability on each voxel independently and forming the IV. In the last step the UH is calculated taking into account the number of occlusions and inconsistencies. The main drawback of their algorithm is that it is based in the minimization of the 3D miss-classification probability using both prior probabilities of voxels forming part of the background or shape and 2D miss-classifications probabilities. However, determining these prior probabilities in realistic scenes is a rather difficult problem.

In spite of that, the work in Ref. [25], constitutes a leap towards the use of SfS methods in tasks such as people tracking [16,32,30,23]. As previously indicated, the need of silhouette consistency imposes a hard constraint to the applicability of SfS methods in uncontrolled environments where illumination changes and clutter causes great errors in the silhouettes extracted. Thus, it is required to improve the robustness of these methods in order to successfully apply them in this real-life scenarios.

This work proposes a new SfS method specially designed to work in complex scenarios by means of exploiting both the information redundancy and the relative camera positions. In Ref. [35], Pribanić studied the influence of the camera setup in the accuracy of the 3D reconstructions. He demonstrated that the higher reconstruction accuracy takes place for cameras forming an angle of 90° with the object. As the angle formed by the camera variates from this, there is a reduction in the reconstruction accuracy. Nevertheless, none of the previous SfS approaches have explicitly considered positional information in their formulations.

1.2. Proposed contribution

The contribution of this paper is three-fold. Firstly, we propose an algorithm that use information about the relative positions between cameras and voxels in the reconstruction problem. Our approach employs Pribanić's principles in order to deal with silhouettes inconsistencies. Secondly, the algorithm is based on the DS theory to classify voxels instead of the classical intersection of visual cones. It allows us to define sensor uncertainty models which help to reduce the propagation of 2D miss-classifications to the 3D model. Finally, the proposed model has a parameter that allows us to specify the degree of precision of the reconstructed scene.

The problem is formulated in terms of the DS theory of evidence, which is a generalisation of the Bayes theory of subjective probability for the mathematical representation of uncertainty. The algorithm models the voxel classification problem using a sensor fusion in the DS sense. In contrast to other approaches, our sensors are camera pairs instead of individual cameras. So, we compute a degree of confidence for each sensor, which depends on the angle between their cameras and the voxel analyzed. Sensors provide degree of evidences about the occupancy of voxels which are fused using the Dempster's rule of combination in order to provide a final value representing their probability of occupancy. The fusion process takes sensor confidence into account. Therefore, most of the 2D miss-classifications produced in silhouettes are not propagated to the 3D reconstruction when the evidence of correct detections is greater than the evidence of 2D miss-classifications. In addition,

our algorithm introduces a parameter that allow us to choose between an accurate reconstruction, similar to the one of original SfS algorithm, and a more permissive reconstruction in which the bulk of objects is recovered at the expense of a lower precision.

The proposed method is specially appropriated for shape reconstruction in uncontrolled environments where there is not much control over lighting conditions [32]. The experiments conducted (both in synthetic and real environments) show that our method outperforms traditional and new SfS approaches (namely Ref. [25]), specially in the presence of strong noise.

2. Dempster–Shafer theory of evidence

The DS theory, which is also known as the evidence theory, is a generalisation of the Bayes theory of subjective probability. It includes several models of reasoning under uncertainty such as the Smets' Transferable Belief Model (TBM) [37]. It has been applied to several disciplines such as people tracking [32], fraud detection [33], classification [14], risk analysis [10], clustering [13,28], image processing [4,3,19,34], autonomous robot mapping [46], human-computer interaction [44], land mine detection [29] and diagnosis [45], amongst others.

The DS approach employs degrees of evidence that are a weaker version of probabilities. The management of uncertainty in the DS theory is especially attractive because of its simplicity and because it does not require specifying priors or conditionals that might be unfeasible to obtain in certain problems. In the DS domain, it is possible to set a degree of ignorance to an event instead of being forced to supply prior probabilities adding to unity.

Let us consider a variable ω taking values in the frame of discernment Ω and let us denote to the set of all its possible subsets by 2^{Ω} (also called power set). A basic belief assignment (bba)

$$m : 2^{\Omega} \rightarrow [0, 1],$$

is a function that assign masses of belief to the subsets A of the power set, verifying:

$$\sum_{A \in \Omega} m(A) = 1. \tag{1}$$

While the evidence assigned to an event in the Bayesian approach must be a probability distribution function, the mass $m(A)$ of a power set element can be a subjective function expressing how much evidence supports the fact A. Furthermore, complete ignorance about the problem can be represented by $m(\Omega) = 1$.

The original Shafer's model imposes the condition $m(\emptyset) = 0$ in addition to that expressed in Eq. (1), i.e., the empty subset should not have mass of belief. However, Smets' TBM model relaxes that condition so that $m(\emptyset) > 0$ stands for the possibility of incompleteness and conflict (see Ref. [37]). In the first case, $m(\emptyset)$ is interpreted as the belief that something out of Ω happens, i.e. accepting the *open-world assumption*. In the second case, the mass of the empty set can be seen as a measure of conflict arising when merging information from sources pointing towards different directions.

Nonetheless, a renormalization can transform a Smets' bba m into a Demspter's bba m^* as

$$m^*(\emptyset) = 0, \quad m^*(A) = \frac{m(A)}{1 - m(\emptyset)} \quad \text{if} \quad A \neq \emptyset. \tag{2}$$

One of the most attractive features of DS theory is the set of methods available to fuse information from several sources. Let us consider two bbas m_1 and m_2 representing distinct pieces of evidences, the standard way of combining them is using the conjunctive sum operation [38] defined as

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega. \tag{3}$$

The Dempster's rule of combination can be derived from Eq. (3) by imposing normality (i.e., $m(\emptyset) = 0$) as

$$(m_1 \otimes m_2)(A) = \frac{1}{1 - K} \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega, \ A \neq \emptyset, \tag{4}$$

with

$$K = (m_1 \otimes m_2)(\emptyset), \tag{5}$$

thus spreading the conflict among the elements of the power set. This approach would certainly discard useful information about the problem. For instance, large values of conflict might indicate an inappropriate formulation of the sensors or even that the solution to the problem lays out of the power set defined. So, disregarding the conflict should not be done in problems with high conflict.

The above rules assume that the sources manage independent pieces of information. However, if information is correlated, the cautious rule should be employed [12].

In some applications it is necessary to make a decision and choose the most reliable single hypothesis ω. To do so, Smets [39] proposed the use of the pignistic transformation that is defined for a normal bba as

$$BetP(\omega) = \sum_{A \subseteq \Omega, \omega \in A} \frac{m(A)}{|A|}, \tag{6}$$

where $|A|$ denotes the cardinality of A.

3. Standard voxel-based SfS

Since our method is a new approach for reconstructing shapes from silhouettes, it is important to explain first the basic concepts of standard SfS methods and to expose the reasons why they are unable to reconstruct scenes from inconsistent silhouettes.

We assume a three-dimensional work area that is divided into cubes of the same volume called voxels. Let us denote by

$$\mathbf{V} = \{v^i = \{x, y, z\} | i = 1, \ldots, n\},$$

the voxel set, where n represents the total number of voxels and $v^i = \{x, y, z\}$ the center of the i-th voxel.

Let us assume that there is a set of cameras placed at known locations (extracted using calibration) and that we have a background subtraction method that obtains the silhouettes of the foreground objects. We denote these foreground images as

$$\mathbf{F} = \{\mathcal{F}_c | c = 1, \ldots, C\},$$

where C is the number of cameras. A pixel $p \in \mathcal{F}_c$ is *true* if it is classified as belonging to the foreground and *false* otherwise.

SfS methods examine voxel projections in the foreground images in order to determine whether they belong to the shape of objects or not. This is achieved by means of a projection test. Each voxel is projected in all the foreground images and if its projection lays completely into a silhouette in all the foreground images, then it is considered *occupied*. However, if the voxel projects in a background region in any of the images it is considered $\neg$ *occupied*. Finally, if the voxel projects partially in a foreground region it is considered to belong to a border and a decision must be made.

Projection tests play an essential role in SfS algorithms. The most simple one consists in projecting only the center of the voxel. More complex approaches consist in testing either all the pixels or a subset of pixels within the polygon formed by the voxel projection. Either way, the result is a boolean decision indicating whether the voxel is occupied or not.

An overview of the SfS algorithm is shown in Algorithm 1. At the first, all voxels are assumed to be occupied. Then, voxels projections are examined in all the images using the projection test. If the projection test indicates that the voxel does not belong to the silhouette of any foreground image, then it is considered as not occupied independently of its projection on the rest of images. The result of the algorithm is the set of VHs of the objects in the scene. In case of consistent silhouettes (i.e., these resulting from error-free background subtraction techniques), the reconstruction of VHs is correctly performed. However, consistent silhouettes rarely occur in realistic scenarios thus leading to several types of miss-classifications.

Algorithm 1. Classical SfS algorithm

Require: Foreground images: **F**
Require: Projection Test Function: $PT(v^i, \mathcal{F}_c)$
for all $v^i \in \mathbf{V}$ **do**
  $v^i \leftarrow occupied$
  for all $\mathcal{F}_c \in \mathbf{F}$ **do**
    if $PT(v^i, \mathcal{F}_c)$ is false **then**
      $v^i \leftarrow \neg occuppied$
      examine another voxel
    end if
  end for
end for

3.1. Type of miss-classifications

Two types of errors can be found in the process of background subtraction which lead to inconsistent silhouettes, namely false positives (*FP*) and false negatives (*FN*).

As previously indicated, a voxel must pass the projection test in all views in order to be considered as belonging to an object. Therefore, a single FN in any view causes a miss-classification in the three-dimensional reconstruction. As a result, FN constitute the main problem for SfS methods. As an example see Figs. 1a and b. Fig. 1a shows a scene in which four cameras observe a pair of objects. However, an error in the background subtraction method causes a FN in the second camera which makes impossible to reconstruct the second object.

In contrast, a FP in a single silhouette does not propagate the error to the 3D space, unless the visual cone that is erroneously created intersects simultaneously with $C-1$ visual cones (probably produced by consistent detections). If the intersection happens, then a region of the scene is wrongly reconstructed. In Fig. 1c we have shown both cases. Cam4 produces a FP that intersects with the remainder $C-1$ visual cones and therefore, the miss-classification is propagated to the reconstruction of the scene (Fig. 1d). On the other hand, cam2 produces a FP that intersects with less than $C-1$ so that the error is not propagated. Note that in Fig. 1d we differentiate between FPs due to inconsistencies (in yellow) and those produced by the very definition of the VH (in red).

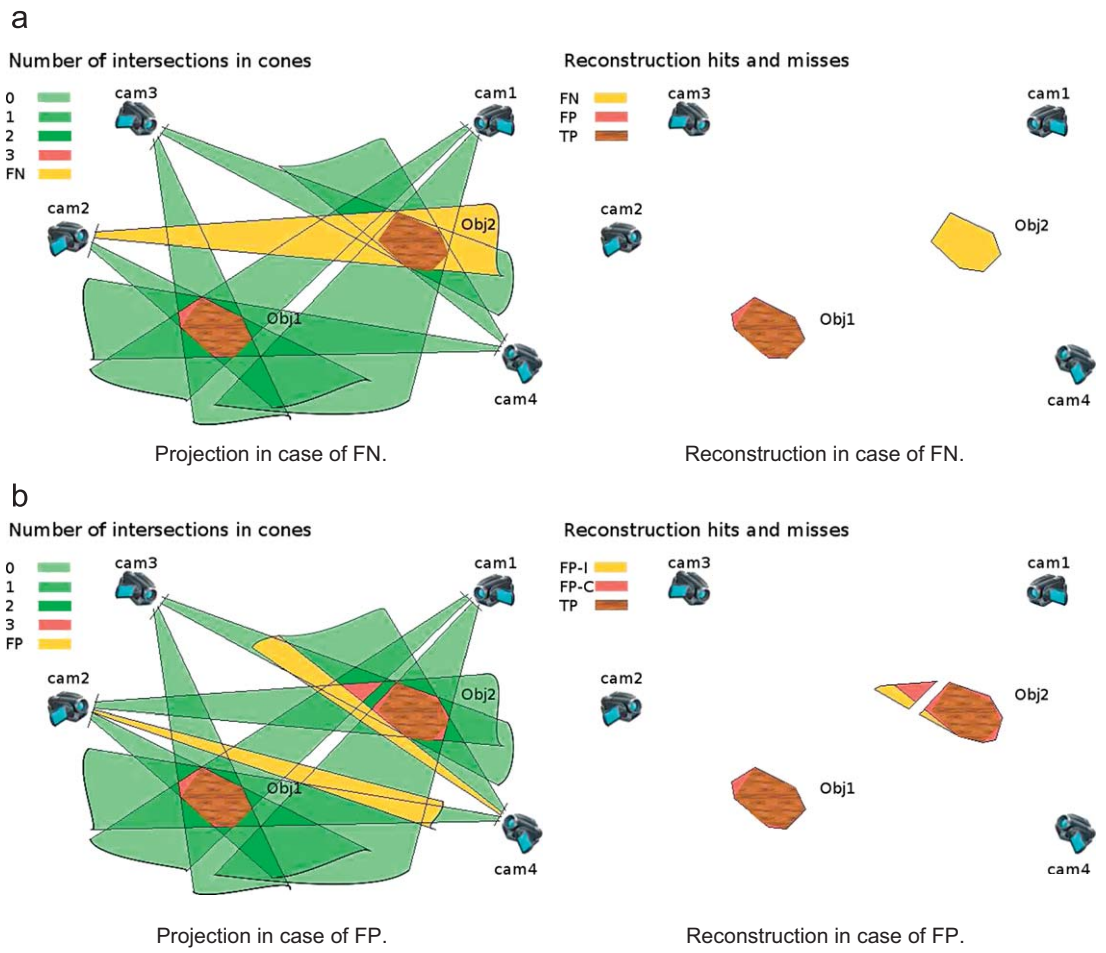

Fig. 1. Typical cases of miss-classification in SfS algorithms. Yellow areas represent miss-classifications: (a) scene where cam2 experiences a FN, (b) reconstruction of the scene in (a), (c) scene where cam2 and cam4 produce FPs and (d) reconstruction of the scene in (c). See text for further details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

There is a final observation on classical SfS methods. The extended idea "as the number of cones increases, the object is reconstructed with higher precision" [27] was established in the context consistent silhouettes. However, this hardly ever holds true in realistic scenarios where lightning can only be roughly controlled. As a consequence, the higher the number of images employed, (assuming a low but non null rate of randomly distributed 2D miss-classifications), the lower is the probability of reconstructing the real shape of the object.

In conclusion, standard SfS algorithms work fine when the assumption about consistent silhouettes is satisfied, but do not deal properly with habitual miss-classifications produced in the foreground extraction process.

4. Shape from silhouette using DS theory

The goal of the shape from silhouette using Dempster–Shafer theory (SfSDS) proposed in this work is to achieve the classification of voxels as *occupied* or ¬ *occupied*, like in any SfS problem, but considering the presence of inconsistent silhouettes.

The main difference between the proposed method and other SfS approaches is that ours does not classify voxels by intersecting visual cones or using other strategies derived from this intersection (see Section 1.1). Instead, our approach classify voxels fusing information from camera pairs using the DS theory which has proven to be a powerful tool for managing uncertainty and lack of knowledge. For each voxel, each sensor is asked two questions. First, which is the degree of confidence of the sensor in the calculus of a voxel occupancy? This is answered taking into account the relative position of the sensor's cameras and the voxel. And second, to what extent is the voxel occupied according to the camera pair? The answers given by sensors to each voxel are employed to assign the evidences to the facts described as in Section 4.1. In a final stage, once all the sensors have provided a degree of evidence about the voxels, evidences are fused in order to classify voxels as *occupied* or ¬ *occupied*. Below, we provide a detailed explanation of the algorithm details.

4.1. SfS problem formulation using DS theory

We shall denote the facts to be evaluated about each voxel as

$$\mathcal{X}=\{occupied,\neg occupied\},$$

so that the power set of our problem is

$$\mathbb{P}(\mathcal{X})=\{\emptyset,\{occupied\},\{\neg occupied\},\Omega\}.$$

For each sensor, a bba must be defined for the elements of $\mathbb{P}(\mathcal{X})$. By

$$M_s^i=\{m_s^i(occupied),m_s^i(\neg occupied),m_s^i(\Omega)\},$$

we shall denote the bba provided by the s-th sensor about the i-th voxel with regards to the subsets in the power set. The mass $m_s^i(occupied)$ represents the degree of evidence assigned by the s-th sensor to the fact that the voxel v^i *belongs to* the shape of any of the objects in the scene. On the other hand, $m_s^i(\neg occupied)$ represents the evidence that this voxel belong to the background of the scene. Finally, $m_s^i(\Omega)$ represents the degree of evidence of the sensor itself, in other words, its own unreliability.

On the basis of the power set just defined, the parameter

$$\mathbb{M}^i=\{M_s^i|s=1\ldots S\},$$

represents all the bbas *provided by the S sensors to the voxel* v^i. Please notice that since we are not assuming any particular camera configuration, all voxels might not project in all cameras. Therefore, if a voxel does not project on a camera, sensors including this camera are considered completely unreliable, thus setting all the mass in the Ω set.

Let $\mathcal{M}^i$ be the bba resulting from fusing the evidences in $\mathbb{M}^i$. We have employed in this work the Dempster's combination rule (Eq. (4)), thus assuming independence between the cameras employed. This is in our opinion the best choice given the fact that correlated cameras are assigned with a low confidence in our model as we shall show later. However, this combination rule discard conflict so it should be used only when the amount conflict is low. We will show later (in the experimental section) that this assumption holds true in our problem.

Finally, voxels are classified using the pignistic probability (Eq. (6)) of the events in $\mathcal{M}^i$. To decide whether a voxel is occupied of not, we use:

$$e^*=\underset{e\in\mathcal{X}}{\operatorname{argmax}}(BetP(e)). \tag{7}$$

4.2. Degrees of evidence calculation

This section explains the basic belief assignment proposed for the masses M_s^i of each sensor about each voxel.

The mass $m_s^i(\Omega)$ represents the degree in which the s-th sensor cannot provide a solution to the problem. This can be seen as the uncertainty or the inability of the sensor to decide between the possible states of the voxels to be evaluated (*occupied* and ¬ *occupied*).

We define the unreliability of a sensor in determining the occupancy of a voxel by a generic function $f(\alpha_s^i)$:

$$m_s^i(\Omega)=f(\alpha_s^i), \tag{8}$$

where α_s^i represents the angle between the segments formed by the voxel v^i and the cameras of the sensor s (see Fig. 2a). The function $f(\alpha_s^i)$ can be defined in different ways as far as they satisfy the constraint given by Pribanić [35], i.e., perpendicular cameras should be considered to be more reliable than parallel ones. It means that $m_s^i(\Omega)$ must tend to 0 as $\alpha_s^i\to\pm\frac{\pi}{2}$, and it must tends to 1 otherwise. In this work, we have tested the four different possibilities for Eq. (8) shown in Fig. 2 b. The main difference in these functions is the smoothness of the transition from low to high values. The results of the proposed method using these functions are reported in the experimental section (Fig. 3).

The angle α_s^i is calculated using trigonometry given that the position of the cameras are known (let us denote them by P_{c1}^s and P_{c2}^s):

$$\alpha_s^i=\arccos\left(\frac{\overrightarrow{P_{c1}^s v^i}\cdot\overrightarrow{P_{c2}^s v^i}}{|\overrightarrow{P_{c1}^s v^i}|\cdot|\overrightarrow{P_{c2}^s v^i}|}\right). \tag{9}$$

The masses $m_s^i(occupied)$ and $m_s^i(\neg occupied)$ are calculated once it has been obtained the uncertainty of the sensor so that the total sum of the three masses is equal to one. Besides, the amount of mass assigned to the *occupied* and ¬*occupied* elements is the remaining of $m_s^i(\Omega)$ to obtain one, i.e.,

$$m_s^i(occupied)+m_s^i(\neg occupied)=1-m_s^i(\Omega).$$

The mass $m_s^i(occupied)$ is calculated first. The overall idea is that if at least one of the cameras of the sensor obtains a very low occupancy degree for the voxel's projection, then there is faint chance that the voxel is occupied. Nevertheless, if both cameras observe a high occupancy degree in the projections, the voxel is likely to be occupied. Therefore, the first thing we have to define is the projection test employed. Let occ_c^i represents the degree of occupancy of the voxel projection v^i in the camera c. It is

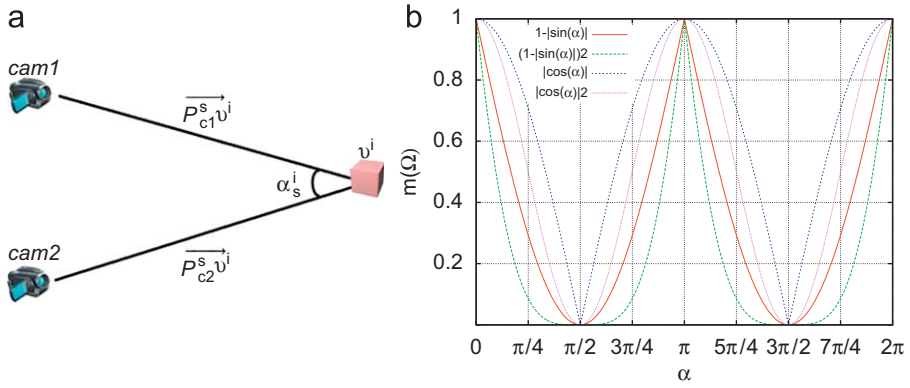

Fig. 2. (a) Angle between the segments formed from a voxel to each camera and (b) functions employed to calculate $m_s^i(\Omega)$.

calculated as

$$occ_c^i = \frac{1}{|P(\mathcal{F}_c, v^i)|} \sum_{p \in P(\mathcal{F}_c, v^i)} \delta(p), \tag{10}$$

where $p \in P(\mathcal{F}_c, v^i)$ represents the pixels of the $\mathcal{F}_c$ image where the voxel v^i projects and δ is defined as

$$\delta(p) = \begin{cases} 1 & \text{if } p = true \\ 0 & \text{otherwise} \end{cases}$$

The sum in Eq. (10) represents the total number of pixels in the projection that belong to the foreground. So, occ_c^i obtains values close to 1 when most of the pixels inside the voxel projection are foreground pixels and values close to 0 in the opposite case. Please notice that this projection test differs from the traditional ones which returns a boolean decision (see Section 3). Instead of taking a decision about the voxel occupancy at this stage, our methods delays the decision until information from all the cameras is fused.

Given the results of the projection tests for the cameras of a sensor, we defined the occupancy degree of the sensor s about voxel v^i as generic function:

$$Occ_s^i = g(occ_{c1}^i, occ_{c2}^i), \tag{11}$$

where the parameters occ_{c1}^i and occ_{c2}^i represent the occupancy degree of the two cameras of the sensor s as given by Eq. (10).

Different functions can be employed for Eq. (11) as long as $Occ_s^i \in [0, 1]$. The idea is that Occ_s^i must take values near 1 if both cameras provide a high occupancy degree. However, the value must tend to zero if the occupancy observed by the cameras is low. In this work, we have tested two different approaches for this function, namely:

$$g_a(occ_{c1}^i, occ_{c2}^i) = \left(\frac{occ_{c1}^i + occ_{c2}^i}{2}\right)^n, \tag{12}$$

and

$$g_b(occ_{c1}^i, occ_{c2}^i) = (occ_{c1}^i \cdot occ_{c2}^i)^n. \tag{13}$$

The first one can be seen as a general way of averaging the occupancy values of the cameras with a free parameter $n \in (0, \infty)$. For $n = 1$, the function represents the arithmetic mean. When $n > 1$ the function forces both cameras to provide high occupancy values in order to obtain high value. However, as n tends to 0, the measure becomes more "permissive", in the sense that inconsistent voxels can be assigned with a high value of Occ_s^i. Indeed, very low values of the parameter n might result in incoherent results since they become independent of the data itself.

The second approach (Eq. (13)) is a more restrictive fusion method which also contains a free parameter $n \in (0, \infty)$ modulating the degree of consensus required for obtaining high values. For $n = 0.5$, it represents the geometric mean. As in the previous case, values of $n < 1$ makes Eq. (13) more permissive and $n > 1$ makes it more strict. The two approaches are evaluated later in Section 5 and their results compared.

Finally, we shall define $m_s^i(occupied)$:

$$m_s^i(occupied) = (1 - m_s^i(\Omega)) \cdot (Occ_s^i). \tag{14}$$

As can be noticed, $m_s^i(occupied)$ has high values when both the reliability of the sensor and the occupancy degree indicated by the cameras are high. As can be notices, our approach differs substantially from classical SfS methods that classify voxels as not occupied if they do not pass the projection test in any of the cameras.

Finally, we define

$$m_s^i(\neg occupied) = 1 - (m_s^i(\Omega) + m_s^i(occupied)), \tag{15}$$

so that the sum of masses is equal to one.

5. Experimental results

This section presents the experimentation carried out. First, we show the behavior of our algorithm in the reconstruction of different kinds of objects for the different $m_s^i(\Omega)$ and Occ_s^i functions previously proposed. Then, the results of our algorithm are compared with these of the SfS and the SfIS [25] methods in two different experiments. In the first one, we analyze the reconstruction performance of the three algorithms from a set of synthetic images. In the second experiment, images obtained in our laboratory are employed to show the reconstruction results in a realistic tracking scenario.

The performance of the algorithms is evaluated using three verification measures commonly employed in the information retrieval field. In the first place, the *recall* measure, also known as detection rate, indicates the percentage of detected true positives in the reconstructed scene in relation to the total number of true positives in the ground truth (GT):

$$recall = \frac{TP}{TP + FN}, \tag{16}$$

where TP is the number of true positives, FN is the number of false negatives and the denominator refers to the number of occupied voxels in the GT. Nonetheless, the *recall* measure is not enough to compare different methods so that it is generally used in conjunction with the *precision* measure, which indicates the percentage of TP in relation to the total number of voxels detected by the method:

$$precision = \frac{TP}{TP + FP}. \tag{17}$$

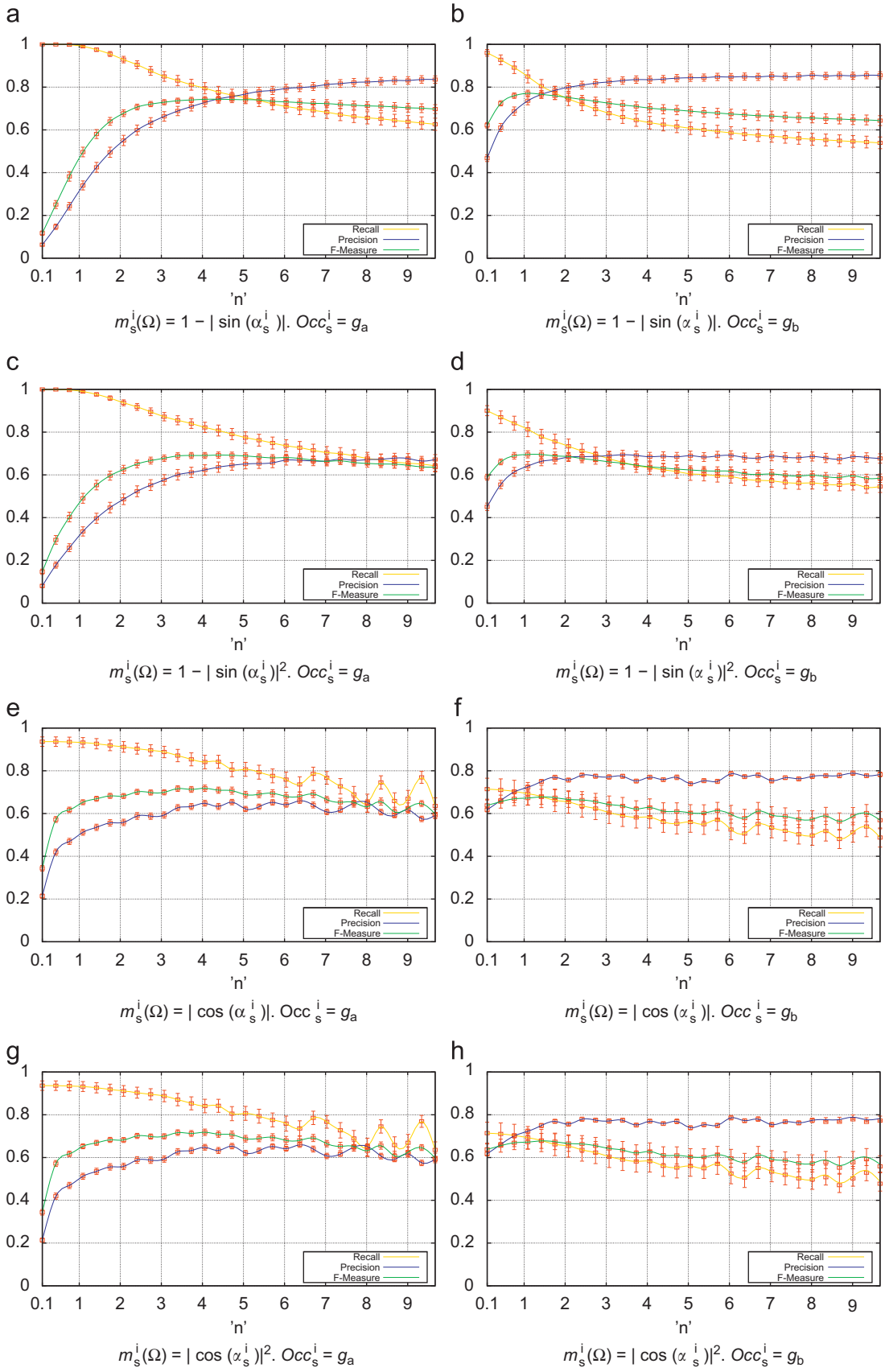

Fig. 3. Verification measures of SfSDS for several values of n. Columns show results obtained employing different functions for Occ_s^i, while in rows different functions for $m_s^i(\Omega)$ have been used: (a) $m_s^i(\Omega)=1-|\sin(\alpha_s^i)|$. $Occ_s^i=g_a$, (b) $m_s^i(\Omega)=1-|\sin(\alpha_s^i)|$. $Occ_s^i=g_b$, (c) $m_s^i(\Omega)=1-|\sin(\alpha_s^i)|^2$. $Occ_s^i=g_a$, (d) $m_s^i(\Omega)=1-|\sin(\alpha_s^i)|^2$. $Occ_s^i=g_b$, (e) $m_s^i(\Omega)=|\cos(\alpha_s^i)|$. $Occ_s^i=g_a$, (f) $m_s^i(\Omega)=|\cos(\alpha_s^i)|$. $Occ_s^i=g_b$, (g) $m_s^i(\Omega)=|\cos(\alpha_s^i)|^2$. $Occ_s^i=g_a$ and (h) $m_s^i(\Omega)=|\cos(\alpha_s^i)|^2$. $Occ_s^i=g_b$.

Finally, we have considered the *f-measure* measure, which represents the weighted harmonic mean of *precision* and *recall*:

$$f\text{-}measure = \frac{2 \times recall \times precision}{recall + precision}. \quad (18)$$

5.1. Evaluation of the SfSDS algorithm

This section aims to analyze our algorithm's performance for the porposed functions of $m_s^i(\Omega)$ and Occ_s^i (Eqs. (8) and (11)).

The scene employed for this experiment is shown in Fig. 4a. It contains a virtual avatar seen from eight different points of view. We have extracted the GT of the scene using the SfS method (Algorithm 1) on the synthetic images without noise (i.e., no inconsistencies). Under these circumstances, the SfS method extract perfectly the convex hull of objects. Then, we have added random spot noise (both background and foreground spots) of different shapes (namely circles and rectangles) which have been placed in the image using an uniform distribution. The size of the spots, in pixels, is also randomly selected using a uniform distribution in the range [5,50]. The number of spots of each type varies between 1 and 15 in each image. The images have a size of 320×240 pixels. An example of noisy images employed for these tests are shown in Fig. 4b. We have created a total of one hundred noisy images that have been used as input to our algorithm.

In order to test the influence of $m_s^i(\Omega)$ in the algorithm's performance, we have employed the four functions shown in Fig. 2b. For each function $m_s^i(\Omega)$, the avatar has been reconstructed using the two Occ_s^i functions proposed (Eqs. (12) and (13)) with different values of the parameter n. The average result of these experiments are shown in Figs. 3(a–h) with 95% confidence intervals.

The figures show a very similar tendency in all cases: while the best values of *precision* are obtained for high values of n, high values of *recall* are achieved for low values of n. In fact, this is the expected result because when n tends to zero all voxels tends to be considered as occupied. As n increases, so does the *precision* at the expenses of reducing *recall*. With regards of the function $m_s^i(\Omega)$, we can observe that the results obtained for $m_s^i(\Omega) = 1 - |\sin(\alpha_s^i)|$ (Figs. 3(a and b)) are better than the others, specially in the *f-measure* which averages the other two measures.

With regards the Occ_s^i function, we can observe in both figures that Eqs. (12) and (13) provide comparable results. Initially, the *f-measure* increases with n up to a certain point in which the measure decreases. This is because *recall* decreases faster than *precision* increases. Although this tendency is observed in both cases, the deterioration of the *f-measure* is faster for Eq. (13) than for Eq. (12). In fact, the latter allows to obtain a more precise control over the results since the variations of the measures take place slower. In other words, Eq. (12) obtains very similar results to Eq. (13) but allowing a finer control over the precision–recall trade-off. Therefore, we consider it is a better approach. Thus, we are employing $m_s^i(\Omega) = 1 - |\sin(\alpha_s^i)|$ and Eq. (12) for the rest of the experiments.

In order to better show the effect of n in the reconstruction obtained, we have performed another pair of tests. We have created a pair of synthetic scenes: one with a cube and another one with a sphere (see Fig. 5a). We have chosen these shapes in order to analyze the behavior of the SfSDS algorithm in both planar and curved surfaces. For space reasons, Fig. 5a shows only four of the eight images employed. In these tests, noise has not been added to the images. The algorithm has been tested on these scenes using $n = \{1, 4, 8\}$ and compared to the results of the SfS algorithm as GT. The results can be seen in Table 1.

The projections of the reconstructed shapes are shown in Figs. 5b–d. The second row (Fig. 5b) shows the reconstructions performed by our algorithm with $n = 1$. Note that several FPs are introduced in both shapes, smoothing the planar surfaces of the box. In the third row (Fig. 5c) it is shown that the reconstruction with $n = 4$ fits better the real objects, specially the sphere.

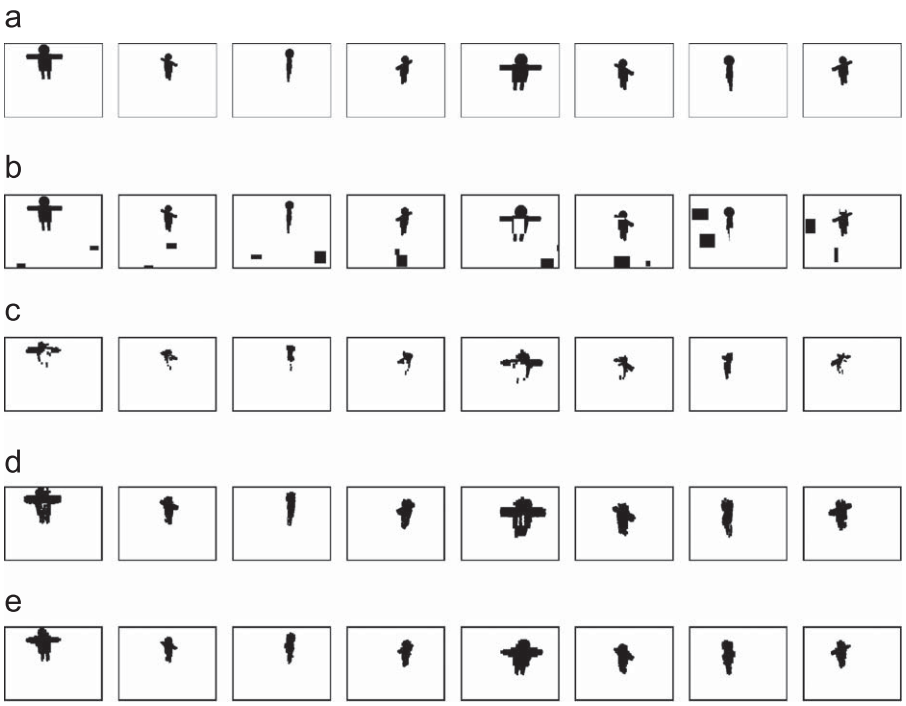

Fig. 4. The first row show the silhouettes employed in one of the experiments carried out. The second row shows the resulting silhouettes after adding noise. The rest of rows show the reconstruction result of the methods analyzed. See text for further discussion: (a) original silhouettes, (b) silhouettes after adding spot noise, (c) reconstruction with the method SfS, (d) reconstruction with the method SfIS and (e) Reconstruction with the method SfSDS ($n = 4$).

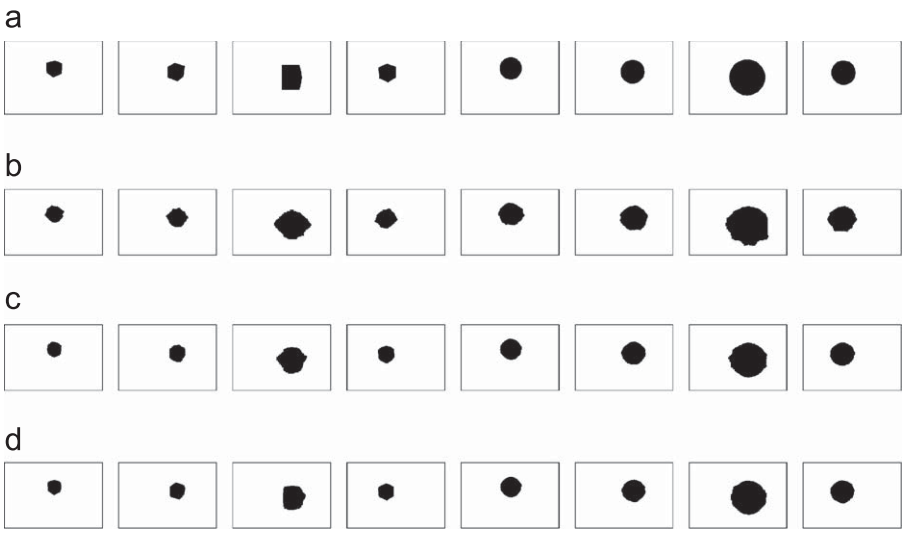

Fig. 5. SfSDS reconstructions with different values of n. As n increases, so does the *precision* of the reconstruction: (a) original silhouettes, (b) SfSDS reconstruction with $n=1$, (c) SfSDS reconstruction with $n=4$ and (d) SfSDS reconstruction with $n=8$.

Table 1
Results of SfSDS for the box and sphere experiments.

	Box			Sphere		
	$n=1$	$n=4$	$n=8$	$n=1$	$n=4$	$n=8$
Recall	1.000	0.983	0.948	1.000	0.996	0.977
Precision	0.547	0.758	0.850	0.701	0.922	0.964
f-measure	0.707	0.856	0.897	0.824	0.957	0.970

However, in the box shape there are still some FPs smoothing its surface. In the bottom row (Fig. 5d) it is shown the reconstruction for $n=8$. For the sphere, the reconstruction is very precise, with very few FPs. For the box, the reconstruction is more precise than using $n=4$, but there are still a few FPs that smooth some of the box corners.

In conclusion, the experiments conducted *have* shown that the proposed method is able to reconstruct the objects with a parametrizable degree of *precision* which is controlled by the parameter n. While high values of n increase the *precision*, low values of it enlarge the shape borders thus increasing *recall*.

5.1.1. Conflict analysis

As previously indicated, the fusion approach employed in this work (Eq. (4)) assumes that the level of conflict is low enough to be discarded by sharing the conflict among the elements of the power set (normalization). This can be seen as assuming a close-world problem, i.e., the solution to the problem lays within the power set defined so that other solutions out of the initial propositions are not considered.

In order to examine whether this assumption holds true in our problem, we have performed an analysis of the degree of conflict in our model. Fig. 6 shows the probability distribution of the conflict in our problem for the one hundred noisy images previously employed for testing. The test has been performed using values of n ranging from 1 to 9 so that the figure represents the probability distribution for all the tests performed. The horizontal axis represents the values of $m(\emptyset)$ set when using the conjunctive sum rule (Eq. (3)). The vertical axis represents the probability using a logarithm scale so as to better visualize the results.

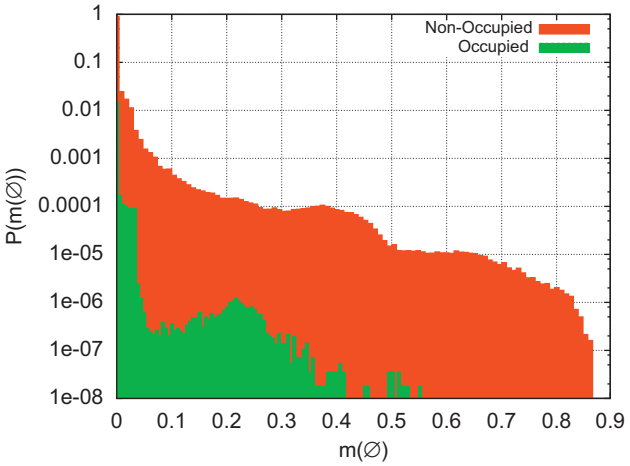

Fig. 6. Probability distribution of the conflict with our method.

The distribution shows that 90% of the voxels have no conflict at all and that the voxels with higher conflict are the non-occupied. However, 99% of the voxels in both cases have a conflict smaller than 0.05. Thus, we can consider this level of conflict low enough to safely ignore it.

5.2. Comparison with other approaches

This section aims to compare the results of our algorithm with previously proposed approaches, namely the original SfS algorithm (see Algorithm 1) and the SfIS [25] algorithm, in noisy conditions. The first method has been chosen because it is the original algorithm on which others are based on. On the other hand, the SfIS is a recently proposed approach that has been proven to overcome the problems of the original method by taking into account inconsistencies in the silhouettes. In both SfS and SfSDS the projection test used is the one explained in (Eq. (10)). For the SfS method, a voxel is considered as $\neg$*occupied* if $occ_c^i < 0.5$ in any camera. Nevertheless, for the SfIS method we have used the *sampled pixels projection test* as defined in [25].

The test have been performed using the noisy images employed in the previous section so that the results can be

compared. Please note that the GT is obtained using the original SfS method of the scene without noise and that all methods employed a voxel size of 6 cm. Table 2 presents the average results obtained with a 95% confidence interval. The table also includes some of the results previously shown in Fig. 3 in order to ease the comparison task.

It can be observed that our method compares very favorably to the SfIS method in all the measures. In relation to the SfS method, ours performs better in *recall* and the *f-measure*. However, the original SfS method gives the higher *precision* at the expense of a low *recall*.

Table 2
Average results with 95% confidence intervals.

	Recall	*Precision*	*f-measure*
SfS	0.366 ± 0.060	0.931 ± 0.036	0.460 ± 0.062
SfIS	0.641 ± 0.045	0.489 ± 0.009	0.539 ± 0.021
SfSDS ($n=1$)	0.983 ± 0.007	0.266 ± 0.018	0.410 ± 0.021
SfSDS ($n=4$)	0.781 ± 0.043	0.683 ± 0.022	0.693 ± 0.022
SfSDS ($n=8$)	0.657 ± 0.049	0.808 ± 0.019	0.680 ± 0.035

Table 3
Results of methods in the experiment shown in Fig. 4.

	SfS	SfIS	SfSDS		
			$n=1$	$n=4$	$n=8$
Recall	0.308	0.676	1.000	0.918	0.792
Precision	1.000	0.500	0.342	0.750	0.913
f-measure	0.471	0.575	0.510	0.846	0.848

Fig. 4 shows the reconstruction results of one of the many test images employed. The first row (Fig. 4a) depicts the silhouettes of the original synthetic scene without spot noise. The second row (Fig. 4b) shows the corresponding set of silhouettes after adding noise. The remainder rows in Fig. 4 show the projection of the models reconstructed by each method. The numerical results of this particular test are presented in Table 3.

The third row (Fig. 4c), corresponds to the reconstruction of the using the SfS method. Please note that most of the errors of this algorithm correspond to FN. Therefore, the *recall* value is very low but the *precision* very high. The fourth row (Fig. 4d) corresponds to the reconstruction of the experiments using the SfIS method. This method is able to recover the shape better than SfS, but at the expense of introducing FPs thus obtaining a lower *precision*. Moreover, it is appreciable that some parts of the body are not totally recovered, e.g., the torso. Finally, the bottom row (Fig. 4e) depicts the reconstruction of our approach for $n=4$. With SfSDS most of the shape have been reconstructed but introducing some FPs. However, the results obtained show that not only does our method outperforms the other two in terms of the *f-measure*, but also that it obtains a high *recall*.

5.3. Experiments in a real scenario

The goal of this section is to perform a qualitative comparison of the three algorithms in a complex reconstruction task. The scenario chosen is our lab, a room of approximately 5×6 m that is

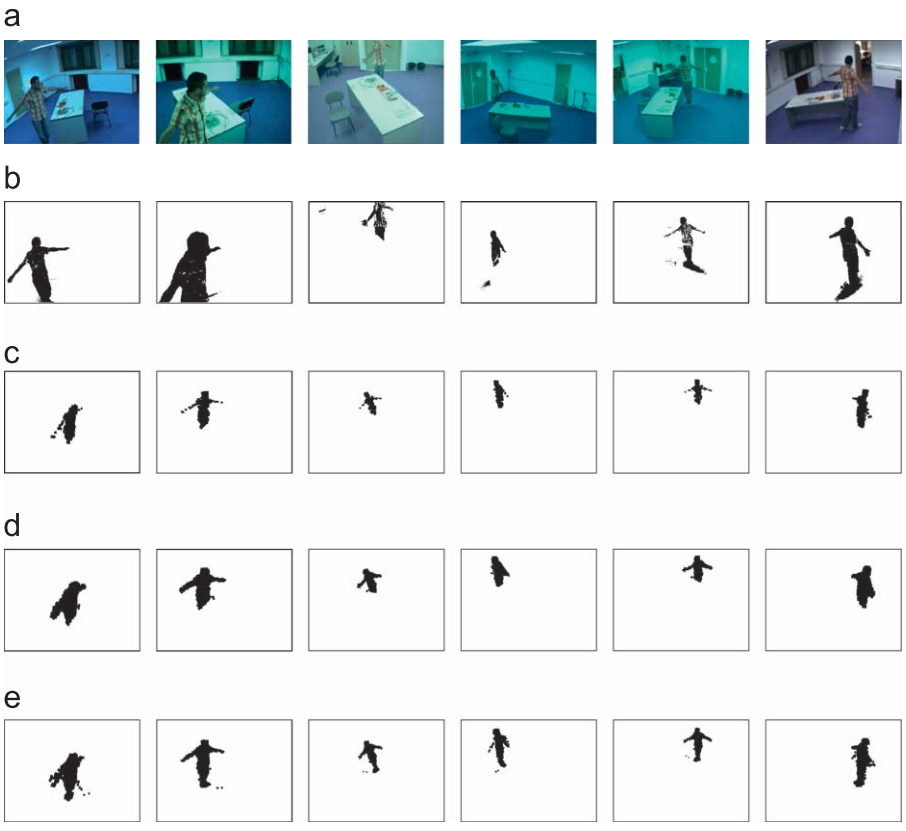

Fig. 7. The first row show the images captured in the lab. The second row show the results of extracting foreground. The rest of the rows represent the results of the reconstruction performed by the methods evaluated. See text for further discussion: (a) original images from the video sequence in the six camera, (b) silhouette images extracted from the original ones, (c) reconstruction with the SfS method, (d) reconstruction with the SfIS method and (e) reconstruction with the SfSDS method.

equipped with a total of six synchronised firewire cameras placed at a height of 2.3 m in slanting positions. Cameras record with a resolution of 640×480 pixels and we have employed the background subtraction technique proposed by Horprasert [20] to create the foreground images.

As previously indicated, our main goal is to create a robust algorithm that can be employed for people tracking purposes. The people tracking problem in a realistic scenario poses a major challenge to reconstruction algorithms for several reasons. Firstly, this type of scenario imposes a reduced control over the lighting conditions so that the quality of the segmentation technique is low. Moreover, it might be expected to find shadows generated by the people being tracked. Secondly, cluttered environments are fraught with occlusions which are often referred as "systematic false negatives".

The voxel set employed covers and area of $4.2 \times 2.1 \times 4$ m with a total of $71 \times 35 \times 68$ voxels using an edge size of 6 cm. We have chosen this voxel size in order to achieve fast 3D reconstruction since speed is an important aspect for tracking if the method is to be applied to real-time problems. The time required for reconstruction the scene is approximately 70 ms on a Quadcore @ 2.4 GHz. An optimization to our implementation consists in the use of the integral image [42] in order to analyze the occupancy of voxel projections. To do so, voxels projections are approximated by the bounding boxes of the projected corner points.

Fig. 7a shows one of the frames captured in our lab. It shows a person moving around a table placed in the middle of the room. It can be seen that the table occludes the person's legs in some views. Also, the person is not fully seen in all views. Fig. 7b shows the foreground images extracted. It is well worth noting that the foreground images contains many miss-classifications. Not only can we find false negatives produced by either light conditions and occlusions, but also false positives produced by shadows are present in the scene. The reconstruction results of the SfS, SfIS and SfSDS (using $n = 4$) methods are shown in Figs. 7c–e, respectively. Please note that the reconstructions depicted in the figure correspond to points of view farther from the scene than the original cameras. As a result, the images showing the reconstruction do not present the silhouettes in the same position than in the foreground images.

The results obtained clearly demonstrate that the SfS method gives a very precise reconstruction, but it is unable to properly recover person's body shape, e.g. the legs are missed. The results of the SfIS have a higher *recall* but still miss the person's legs. In addition, it can be observed that the reconstruction of the arms and body is complete but adding some FPs. Finally, although our method also introduces some FPs, it is able to properly reconstruct the whole person's body including his legs.

In Fig. 8, we show another complex scene employed to evaluate the algorithms. In this case, the environment shown is the same but there are two people and the light conditions are worse than in the previous scene. Because the foreground images extracted have so many miss-classifications, the reconstruction becomes very difficult. For this scenario, we have required to set $n = 1.5$ in order to obtain appropriated reconstruction results.

While Fig. 8b shows the foreground images extracted, Figs. 8c–e show the reconstruction results of the SfS, SfIS and SfSDS methods respectively. The SfS method gives a very poor reconstruction, since the silhouettes have a lot of miss-classifications. The results of the SfIS and SfSDS are much better than these

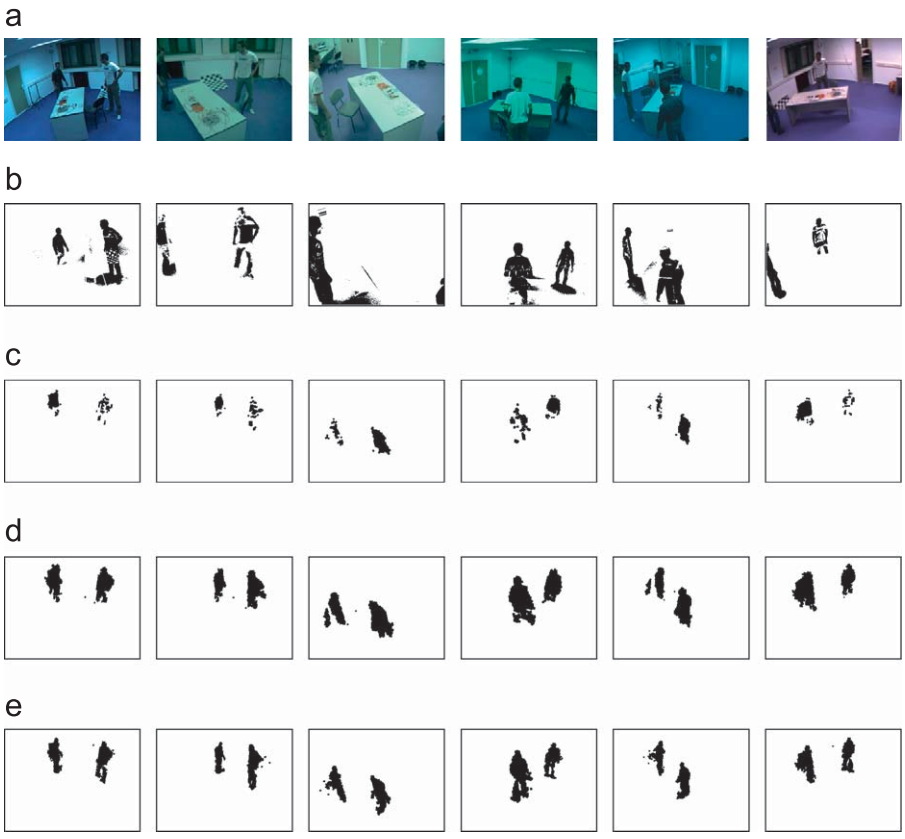

Fig. 8. Original images, silhouettes and reconstructions for a specific frame in a video sequence with two persons: (a) original images from the video sequence in the six camera, (b) Silhouette images extracted from the original ones, (c) reconstruction with the SfS method, (d) reconstruction with the SfIS method and (e) reconstruction with the SfSDS method.

of the SfS method. However, in our method the amount of FP introduced seems to be lower than in SfIS. In addition, it can be observed that the reconstruction of the legs is complete in our case.

We would like to stress that the capability of our method to tune the trade-off between *precision* and *recall* makes it suitable for different tasks. For instance, in tracking problems the main goal is to determine the location of the person. To do so, a high *precision* in the reconstruction is not necessary. On the contrary, it is more appropriate having a reconstruction of the person robust to occlusions (even if it is rough) than having a precise reconstruction that misses the person in case of occlusion. Nevertheless, other problems might take advantage of more precise reconstructions such as gesture recognition [9,31,43]. In any case, it is required a tunning process to determine the most appropriated value of n for the particular problem employed. A method for automatically determining the most appropriate value of n would be a topic of future works.

A limitation of our approach is that it is slower than the other SfS algorithms. The problem is that the number of sensors increases combinatorially with the number of cameras. A possible solution to that problem in applications in which time is crucial would be to prune out the set of possible sensors by considering only these with $m(\omega)$ below a certain threshold. Therefore, sensors formed by nearly parallel cameras (thus providing little information) would not be computed.

6. Conclusions and future works

In this paper, we have proposed a novel SfS method that deals with inconsistent silhouettes. The contribution of this paper is three-fold. Firstly, we propose an algorithm that use information about the relative positions between cameras and voxels. Secondly, the algorithm is based on the Dempster–Shafer theory to classify voxels instead of the classical intersection of visual cones. Finally, the proposed model has an useful parameter that allows to specify the trade-off between *precision* and *recall* in the reconstruction. Our approach is particularly attractive due to its simplicity and because it does not require specifying priors not conditionals that might be difficult to obtain in complex scenarios. The proposed method has been compared to the standard SfS and the SfIS proposed in Ref. [25], and the results show that it obtains better reconstructions, specially under noisy conditions.

Finally, we would like to point out two possible future works. First, we consider the possibility of replacing the binary output of the background subtraction method by a soft measure that employs the distance of the pixels to the background model. This would allow to manage uncertainty from the early stages of the visual processing. Second, the creation of an automatic tuning method for estimating the best n for a particular application.

Acknowledgment

This work has been financed by the Spanish National Company for Radioactive Waste Management.

References

[1] A. Aregui, T. Denoeux, Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities, International Journal of Approximate Reasoning 49 (2008) 575–594.
[2] B.G. Baumgart, Geometric modeling for computer vision, Ph.D. thesis, CS Department, Stanford University, 1974. AIM-249, STAN-CS-74-463.
[3] I. Bloch, Defining belief functions using mathematical morphology—application to image fusion under imprecision, International Journal of Approximate Reasoning 48 (2008) 437–465.
[4] A.-O. Boudraa, A. Bentabet, F. Salzenstein, Dempster–Shafer's basic probability assignment based on fuzzy membership functions, Electronic Letters on Computer Vision and Image Analysis 4 (2004) 1–10.
[5] F. Caro, B. Ristic, E. Duflos, P. Vanheeghe, Least committed basic belief density induced by a multivariate Gaussian: formulation with applications, International Journal of Approximate Reasoning 48 (2008) 419–436.
[6] H.H. Chen, T.S. Huang, A survey of construction and manipulation of octrees, Computer Vision, Graphics and Image Processing 43 (1988) 409–431.
[7] K.M. Cheung, T. Kanade, J.-Y. Bouguet, M. Holler, A real time system for robust 3d voxel reconstruction of human motions, Proceedings of Computer Vision and Pattern Recognition, vol. 2, IEEE Computer Society, 2000, pp. 714–720.
[8] R. Cipolla, M. Yamamoto, Stereoscopic tracking of bodies in motion, Image and Vision Computing 8 (1) (1990) 85–90.
[9] C. Colombo, A.D. Bimbo, A. Valli, Visual capture and understanding of hand pointing actions in a 3-D environment, IEEE Transactions on Systems, Man and Cybernetics—Part B 33 (2003) 677–686.
[10] S. Démotier, W. Schön, T. Denoeux, Risk assessment based on weak information using belief functions: a case study in water treatment, IEEE Transactions on Systems, Man and Cybernetics C 36 (2006) 382–396.
[11] T. Denoeux, Constructing belief functions from sample data using multinomial confidence regions, International Journal of Approximate Reasoning 42 (2006) 228–252.
[12] T. Denoeux, Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence, Artificial Intelligence 172 (2008) 234–264.
[13] T. Denoeux, M. Masson, Evclus: evidential clustering of proximity data, IEEE Transactions on Systems, Man and Cybernetics B 34 (2004) 95–109.
[14] T. Denoeux, P. Smets, Classification using belief functions: the relationship between the case-based and model-based approaches, IEEE Transactions on Systems, Man and Cybernetics B 36 (2006) 1395–1406.
[15] A.M. Elgammal, D. Harwood, L.S. Davis, Non-parametric model for background subtraction, in: Lecture Notes in Computer Science, vol. 1843, 2000, pp. 751–767.
[16] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 267–282.
[17] J.S. Franco, E. Boyer, Fusion of multi-view silhouette cues using a space occupancy grid, in: 10th IEEE International Conference on Computer Vision (ICCV 2005), 2005, pp. 1747–1753.
[18] M. Ha-Duong, Hierarchical fusion of expert opinions in the transferable belief model, application to climate sensitivity, International Journal of Approximate Reasoning 49 (2008) 555–574.
[19] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, Facial expression classification: an approach based on the fusion of facial deformations using the transferable belief model, International Journal of Approximate Reasoning 46 (2007) 542–567.
[20] T. Horprasert, D. Harwood, L.S. Davis, A statistical approach for real-time robust background subtraction and shadow detection, in: Seventh IEEE International Conference on Computer Vision, Frame Rate Workshop (ICCV '99), 1999, pp. 1–19.
[21] C.L. Jackins, S.L. Tanimoto, Oct-trees and their use in representing three-dimensional objects, Computer Graphics and Image Processing 14 (3) (1980) 249–270.
[22] M. Karaman, L. Goldmann, Da Yu, T. Sikora, Comparison of static background segmentation methods, in: Visual Communications and Image Processing 2005, vol. 5960, 2005, pp. 2140–2151.
[23] S.M. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: Lecture Notes in Computer Science, vol. 3954, 2006, pp. 133–146.
[24] J.L. Landabaso, M. Pardàs, Foreground regions extraction and characterization towards real-time object tracking, in: Proceedings of Multimodal Interaction and Related Machine Learning Algorithms, Lecture Notes in Computer Science, Springer, 2005.
[25] J.L. Landabaso, M. Pardàs, J. Ramon Casas, Shape from inconsistent silhouette, Computer Vision and Image Understanding 112 (2008) 210–224.
[26] A. Laurentini, The visual hull: a new tool for contour-based image understanding, in: Proceedings of Seventh Scandinavian Conference on Image Processing, 1991, pp. 993–1002.
[27] A. Laurentini, The visual hull concept for silhouette-based image understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 150–162.
[28] M.-H. Masson, T. Denoeux, Ecm: an evidential version of the fuzzy c-means algorithm, Pattern Recognition 41 (2008) 1384–1397.
[29] N. Milisavljevic, I. Bloch, S. Broek, M. Acheroy, Improving mine recognition through processing and Dempster–Shafer fusion of ground-penetrating radar data, Pattern Recognition 36 (2003) 1233–1250.
[30] R. Muñoz-Salinas, A Bayesian plan-view map based approach for multiple-person detection and tracking, Pattern Recognition 41 (2008) 3665–3676.
[31] R. Muñz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, Depth silhouettes for gesture recognition, Pattern Recognition Letters 29 (2008) 319–329.
[32] R. Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, Multi-camera people tracking using evidential filters, International Journal of Approximate Reasoning 50 (5) (2009) 732–749.

[33] S. Panigrahi, A. Kundu, S. Sural, A.K. Majumdar, Use of Dempster–Shafer theory and bayesian inferencing for fraud detection in mobile communication networks, in: Lecture Notes in Computer Science, 2007, pp. 446–460.
[34] W. Pieczynski, Multisensor triplet Markov chains and theory of evidence, International Journal of Approximate Reasoning 45 (2007) 1–16.
[35] T. Pribanić, M. Cifrek, S. Tonković, The choice of camera setup in 3d motion reconstruction systems, in: Proceedings of the 22th Annual EMBS International Conference, 2000, pp. 163–165.
[36] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.
[37] P. Smets, The combination of evidence in the transferable belief model, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 447–458.
[38] P. Smets, Belief functions: the disjunctive rule of combination and the generalized bayesian theorem, International Journal of Approximate Reasoning 9 (1993) 1–35.
[39] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–243.
[40] D. Snow, P. Viola, R. Zabih, Exact voxel occupancy with graph cuts, in: Proceedings of Computer Vision and Pattern Recognition, IEEE Computer Society, 2000, pp. 345–353.
[41] S. Sullivan, J. Ponce, Automatic model construction and pose estimation from photographs using triangular splines, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (10) (1998) 1091–1096.
[42] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 511–518.
[43] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2006) 249–257.
[44] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, Sensor fusion using Dempster–Shafer theory, in: IEEE Instrumentation and Measurement Technology Conference, 2002, pp. 21–23.
[45] J. Yen, Gertis: a Dempster–Shafer approach to diagnosing hierarchical hypotheses, Communications of the ACM archive, 1989, pp. 573–585.
[46] Z. Yi, H.Y. Khing, C.C. Seng, Z.X. Wei, Multi-ultrasonic sensor fusion for autonomous mobile robots, in: SPIE Proceedings Series: Architectures, Algorithms and Applications IV, 2000, pp. 314–321.
[47] J.Y. Zheng, Acquiring 3-d models from a sequence of contours, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 163–178.

About the Author—LUIS DÍAZ MÁS received the Bachelor degree in Computer Science in 2008 from University of Córdoba, Spain. He is currently studying for the Ph.D. degree with a research grant. Since 2007 he has been collaborating in the research group of Applications of Artificial Vision in the University of Córdoba. His research is focused on 3-D modelling and action recognition.

About the Author—RAFAEL MUÑOZ SALINAS was born in Córdoba, Spain, in 1979. He received both his M.S. degree in Computer Science in 2005 and his Ph.D. in Computer Science in 2006, both from the University of Granada, Spain. He is currently Assistant Professor at the University of Córdoba. His research interests include Autonomous Robots, Soft Computing, Stereo Vision and Tracking.

About the Author—FRANCISCO JOSE MADRID-CUEVAS received the Bachelor degree in Computer Science from Málaga University (Spain) and the Ph.D. degree from Polytechnic University of Madrid (Spain), in 1995 and 2003 respectively. Since 1996 he has been working with the Department of Computing and Numerical Analysis of Córdoba University, currently he is associated professor. His research is focused mainly on image segmentation, 2-D object recognition and evaluation of computer vision algorithms.

About the Author—RAFAEL MEDINA-CARNICER received the Ph.D. degree in Computer Science from the Polytechnic University of Madrid (Spain) in 1992. Since 1993 he has been a lecturer of Computer Vision in Cordoba University (Spain) and he has been participant of four research projects related with computer vision systems. His research is focused on edge detection, evaluation of computer vision algorithms and pattern recognition. He is a reviewer for Pattern Recognition Letters.

4

Third contribution: An octree-based method for shape from inconsistent silhouettes

Pattern Recognition 45 (2012) 3245–3255

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

An octree-based method for shape from inconsistent silhouettes

L. Díaz-Más *, F.J. Madrid-Cuevas, R. Muñoz-Salinas, A. Carmona-Poyato, R. Medina-Carnicer

Dpto. de Informática y Análisis Numérico, Universidad de Córdoba, Campus de Rabanales, s/n, 14071 Córdoba, Spain

ARTICLE INFO

Article history:
Received 24 March 2011
Received in revised form
15 March 2012
Accepted 17 March 2012
Available online 28 March 2012

Keywords:
Shape from Silhouette
Octrees
Evidence theory

ABSTRACT

Shape-from-Silhouette (SfS) is the widely known problem of obtaining the 3D structure of an object from its silhouettes. Two main approaches can be employed: those based on voxel sets, which perform an exhaustive search of the working space, and those based on octrees, which perform a top-down analysis that speeds up the computation. The main problem of both approaches is the need for perfect silhouettes to obtain accurate results. Perfect background subtraction hardly ever happens in realistic scenarios, so these techniques are restricted to controlled environments where the consistency hypothesis can be assumed. Recently, some approaches (all of them based on voxel sets) have been proposed to solve the problem of inconsistency. Their main drawback is the high computational cost required to perform an exhaustive analysis of the working space. This paper proposes a novel approach to solve SfS with inconsistent silhouettes from an octree based perspective. The inconsistencies are dealt by means of the Dempster–Shafer (DS) theory and we employ a Butterworth function for adapting threshold values in each resolution level of the octree. The results obtained show that our proposal provides higher reconstruction quality than the standard octree based methods in realistic environments. When compared to voxel set approaches that manage inconsistency, our method obtains similar results with a reduction in the computing time of an order of magnitude.

1. Introduction

Shape-from-Silhouette (SfS) is a well-known approach to reconstruct the 3D structure of objects using a set of silhouettes obtained from different views. Baumgart [1] was the first to introduce concepts about the 3D geometric modelling, but it was not until 1991 that Laurentini [2] defined the concept of the Visual Hull (VH) as the closest 3D solid equivalent to the real object that explains the silhouettes extracted. The VH is the geometric intersection of all visual cones explaining the projection of a silhouette in its corresponding camera image. Consequently, as the number of views increases, so does the precision of the reconstructed object [3].

Two classical approaches are used to analyse and represent 3D information: voxel sets and octrees. In the first approach, the entire area of interest is a discrete 3D grid of voxels of the same volume. Then the voxels are projected into all the images to check whether they belong to foreground objects. The second approach is the octree structure, which is based on a tree of voxels. Octree based methods [4,5] start with a cube that covers all of the working area; the working area is recursively subdivided into eight voxels until a homogeneous content (shape or background) is reached, or until a maximum resolution has been obtained.

A major problem of standard SfS methods is that they are strongly linked to the principle of *silhouette consistency*, i.e., the set of silhouettes employed must explain precisely the real object. A single inconsistency in one of the silhouettes could distort the reconstructed VH regarding the expected one. Nevertheless, total consistency hardly ever happens in real-life scenarios due to several factors such as inaccuracies in camera calibration, foreground extraction errors [6–8], and occlusions. Therefore, SfS methods have been usually confined to problems under controlled conditions [9–13].

In recent years, a number of investigations have addressed the inconsistency problem for SfS in different ways [10,11,14–17]. These approaches aim to exploit information redundancy wisely to overcome the inconsistency problem. However, all of these works use the voxel set-based approach.

From our point of view, the octree-based approach has several advantages compared with the voxel set approach, among which we can highlight the following two:

- The inherent multi-scale structure used in the representation of the volume.
- A better performance is obtained when the working space is analysed compared with voxel set-based methods.

However, the octree-based approach has a major inconvenience regarding the voxel set: the size of a voxel in the octree is a variable that is a function of the octree's tree level. All of the methods for SfS apply a projection test to determine whether or

* Corresponding author. Tel.: +34 957 211035; fax: +34 957 218630.
E-mail addresses: i22dimal@uco.es, piponazo@gmail.com (L. Díaz-Más).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.03.012

not a voxel is occupied. In a voxel set-based approach, the voxels always have the same volume, and there is a relatively small ratio of projected surface area to total volume in the working space, making the design of the projection test simple. However this does not happen the same way with the octree-based approach, where the variable size of the voxel causes two main problems that must be resolved:

- First, in the high levels of the octree, the projected area of a voxel will be large in comparison to the size of the silhouette. This can cause only a small portion of the projected area to be occupied, which implies that the projection test should be permissive in these levels to prevent losing portions of the volume from the first levels of the octree.
- Second, as the size of the voxel varies, a static criterion cannot be fixed to determine the occupation of a voxel, but rather should be a dynamic criterion as a function of the level of the octree.

To design a projection test in spite of these inconvenient issues is not a trivial problem, as will be seen later in this paper.

In our previous work [17], we advanced the solution of the problem of inconsistency in the silhouettes by introducing data provided by the relative position of the cameras with respect to the object in the re-construction process, and by using the theoretical working frame of the theory of the evidence of Dempster–Shafer (DS) [18] to fuse the data provided by the different sensors. This approach has been proven using the voxel set-based alternative and the experimental results obtained were superior to those obtained from alternative approaches.

In this work, we propose a novel approach to solve the problem of SfS by combining the advantages of using an octree to analyse the working space with the robustness of the theory of the evidence of DS to address the inconsistency of silhouettes. The final objective is to apply our proposal to problems of volumetric reconstruction in real scenarios to carry out detection, tracking and analysis of people's activity.

The remainder of this paper is structured as follows. The rest of Section 1 provides an overview of the most relevant work related to ours, along with its main contributions. Section 2 details our proposal. Section 3 explains the results we obtained, and some final conclusions are drawn in Section 4.

1.1. Related works

1.1.1. Standard octree-based SfS

Many SfS algorithms have been proposed for creating volumetric models using octrees since its initial definition in [19]. The octrees were first used as a method of representation of a volume in an efficient way and several works were proposed to optimise their creation, storage and manipulation [20,21].

Later on, the problem of the shape-from-X using an octree was approached [22]. Several techniques were proposed, and SfS was among them. Initially, to carry out SfS, "restricted orthographic projections" were used [23,24]. Then several authors argued that three views were not enough to obtain a good reconstruction [25,26]. It was proposed to increase the number of views, classifying them into two groups: edge views (view direction parallel to the plane of two axes) and corner views (isometric view where the view direction is along the line joining a corner and the centre of a voxel).

Finally, several authors proposed methods to carry out SfS using multiple arbitrary views with a perspective projection. Two of the first works following this line of reasoning [27,28] present almost identical proposals, where an octree is generated for each silhouette and then the octrees are combined with a logical "and" operation to generate the final octree. Szeliski [29] argues that to generate an octree for each view to merge those later is not very efficient because it becomes necessary to analyse more voxels. He proposes a method similar in essence to the one proposed by Potmesil [28], but only one octree is generated, merging the views at the voxel level. In this paper, we call the proposal of Szeliski the "standard octree based SfS".

The standard algorithm constructs the octree in a hierarchical coarse-to-fine fashion. At a given octree resolution, the inner loop of the algorithm incrementally constructs the 3D volume by applying a projection test to the octree voxels against a sequence of silhouettes. Those voxels whose occupancy is uncertain are subdivided and a new iteration starts at the next resolution level. This process is repeated until a final resolution is reached. Fig. 1 shows an example of an octree, and Fig. 2 shows the described standard algorithm.

To implement the projection test on a set of silhouettes, Szeliski uses a coarse projection test, where the projection of the voxel is converted into the bounding square and an one-sided version of the chess-board distance transform is used to distinguish the state of occupation of the voxel among the possibilities: *occupied* (black), ¬*occupied* (white) and *unknown* (gray). Given a voxel, the projection test is applied in sequence to all of the silhouettes, and the results are merged using Table 1.

1.1.2. Addressing the inconsistency problem

None of the analysed octree-based works approaches the problem of the inconsistency in the silhouettes. All use controlled systems where the extraction of the background is a trivial process and for that reason they can assume that the extracted silhouettes are consistent.

However, full consistency hardly even happens in realistic scenarios mainly because of segmentation errors. As a result, the

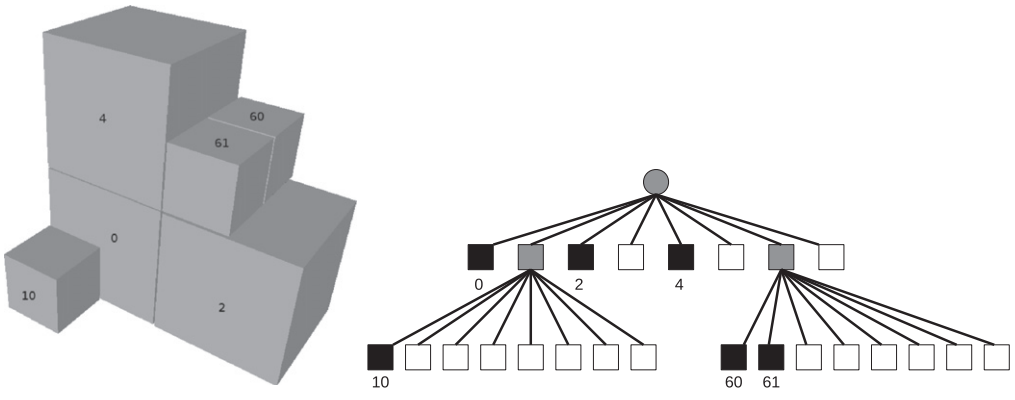

Fig. 1. An example of an octree.

Table 1
Method used by the standard octree-based SfS algorithm to combine the result of the projection test of a voxel with its current occupation state.

Test result	Current state		
	Black	Gray	White
occupied	Black	Gray	White
unknown	Gray	Gray	White
¬occupied	White	White	White

models reconstructed might contain holes or added spots in the background area. Morphological operators can be applied in a cleanup phase, but only to partially mitigate the problem. Recently, some approaches (all based on the voxel set approach to SfS) have addressed the problem of silhouette inconsistencies, proposing different algorithms that minimise the propagation of 2D mis-detections to the 3D models.

In a previous work, Diaz et al. [17], a complete discussion is given on how the problem of the inconsistency of the silhouettes is approached by methods based on voxel sets.

In this work, Diaz et al. propose a solution to the problem of the inconsistency in the silhouettes by exploiting both the information redundancy and the relative camera positions. In Ref. [30], Pribanić studied the influence of the camera setup on the accuracy of the 3D reconstructions. He demonstrated that a higher reconstruction accuracy takes place for cameras forming an angle of 90° with the object. As the angle formed by the camera deviates from 90°, there is a reduction in the reconstruction accuracy. Nevertheless, none of the previous SfS approaches have explicitly considered positional information in their formulations.

However the application of the voxel set-based proposal of Díaz et al. cannot be made in a direct way in an octree-based approach, as an important adaptation of the projection test has to be carried out to solve the problems discussed in Section 1.

Summarising:

- All of the octree-based SfS methods work with the hypothesis of consistent silhouettes. As a result, these proposals are not adequate for practical applications with actual scenarios that have uncontrolled illumination and a complex background. Under these natural environmental conditions one cannot assume the restriction of consistency in the extracted silhouettes.
- It is important to remark that in conditions of low noise and small number of cameras, classical approaches obtain in many cases better results. However, we think that these demanding conditions only could be achieved in environments very controlled and our algorithm has been developed for working in more flexible conditions.
- The inconsistency problem has been approached only by methods based on voxel sets. Although very good results have been obtained by Díaz et al., the use of voxel sets in large work spaces does not result in good performance in computation time due to the exhaustive analysis that is carried out. The application of the octree-based approach would allow one to obtain a better performance because it carries out a more intelligent analysis of the working space; however, the octree-based approach is not a trivial adaptation of the method proposed by Díaz et al. to solve the problem of inconsistency.

2. Proposed method

We have used for our proposal the octree-based SfS algorithm shown in Fig. 2. The main difference with the standard method of

- **Definitions:**
  - Let R be the resolution of the octree.
  - Let $\mathbb{F} \equiv \{\mathcal{F}_i\}, 1 \leq i \leq V$ be the set of silhouette images from the V views.
  - Let $c : \{x, y, z, d, l\}$ be an octree's voxel at depth $d \in \{0, \ldots, R-1\}$ with the label $l \in \{occupied, \neg occupied, unknown\}$ and octree's integer coordinates $x, y, z \in \{0, \ldots, 2^d - 1\}$. From now on we use the notation c^d_{xyz} to notate such a octree's voxel.
  - Let $\mathbb{C}$ be the set of all possible voxels c defined as above for a given number R of resolution levels.
  - Let $\mathbb{O} \subseteq \mathbb{C}$ be an octree.
  - Let $\mathrm{PT}(c, \mathbb{F}) \rightarrow \{occupied, \neg occupied, unknown\}$ be the projection test of a voxel $c \in \mathbb{C}$ on the silhouette set $\mathbb{F}$. Upon reaching the maximum level of resolution R, the only possible states will be "*occupied*" or "*¬occupied*".
  - Let $\mathrm{SP}(c^d_{xyz}) \rightarrow \mathbb{C} : \{c^{d+1}_{2x+i,2y+j,2z+k}\}, i, j, k \in \{0, 1\}$ the set of eight next resolution level voxels obtained by division of the voxel $c : \{x, y, z, d, l\}$.
  - Let $\mathbb{R} \subseteq \mathbb{C}$ and $\mathbb{R}' \subseteq \mathbb{C}$ be the current and the next iteration set of analysed voxels respectively.
- **Inputs:**
  - The octree's root voxel c^0_{000}.
  - The set of silhouette images $\mathbb{F}$.
- **Outputs:**
  - The octree $\mathbb{O} \subseteq \mathbb{C}$.

```
𝕆 ← {∅}
ℝ ← {c^0_000}
while ℝ ≠ {∅} do
  ℝ' ← {∅}
  for-all c ∈ ℝ do
      case-of PT(c, 𝔽) in
          case occupied:
              c ← occupied
          case ¬occupied:
              c ← ¬occupied
          case unknown:
              c ← unknown
              ℝ' ← ℝ' ∪ SP(c)
      end-case
      𝕆 ← 𝕆 ∪ {c}
  end-for
  ℝ ← ℝ'
end-while
return 𝕆
```

Fig. 2. Standard octree-based SfS algorithm.

Szeliski [29] is in the projection test that we have designed to provide robustness regarding inconsistent silhouettes.

In the standard method the hypothesis of consistency of the silhouettes is assumed and a projection test is used to make an intersection (logical "and" operation) of the visual cones restricted to projections of the analysed voxel on each of the silhouettes.

However, our projection test does not classify a voxel by intersecting visual cones or by using other strategies derived from this intersection (see Section. 1.1). Instead, our approach classifies voxels by fusing information from camera pairs (we define a logical sensor for each possible pair of cameras), using the DS theory which has proven to be a powerful tool for managing uncertainty and lack of knowledge.

With more detail, given a voxel, each logical sensor is asked two questions. First, to what extent is the voxel occupied according to the camera pair? The answers given by logical sensors are employed to assign evidence to the facts *occupied* and ¬*occupied*, as is described in Section 2.1. Second, what is the degree of confidence of the logical sensor in the calculus of voxel occupancy? This question is answered by taking into account the relative positions of the cameras that form the logical sensors regarding the voxel. In a final stage, once all the logical sensors have provided a degree of evidence about the voxel, these evidence levels are fused to classify the voxel's state as *occupied* or ¬*occupied*. Also it could be possible that none of these states can be determined for the voxel. This event represents the *unknown* state. Below, we provide a detailed explanation of the algorithm.

2.1. Projection test formulation using DS theory

Next we provide notation and constructs to represent the facts to be evaluated about a given voxel c. (For details on the Dempster–Shafer theory and notation, see [17].)

$$\mathcal{X}^c = \{occupied, \neg occupied\},$$

so that the power set of our problem is

$$\mathbb{P}(\mathcal{X}^c) = \{\emptyset, \{occupied\}, \{\neg occupied\}, \Omega\}.$$

For each logical sensor s, a basic belief assignment (bba) must be defined for the elements of $\mathbb{P}(\mathcal{X}^c)$. By

$$M_s^c = \{m_s^c(occupied), m_s^c(\neg occupied), m_s^c(\Omega)\} \tag{1}$$

we shall denote the bba provided by the s-th logical sensor about the voxel c with regard to the subsets in the power set. The mass $m_s^c(occupied)$ represents the degree of evidence assigned by the s-th logical sensor to the fact that the voxel c belongs to the shape of any of the objects in the scene. On the other hand, $m_s^c(\neg occupied)$ represents the evidence allocated to the fact that this voxel belongs to the background of the scene. Finally, $m_s^c(\Omega)$ represents the degree of evidence of the logical sensor's ignorance about the real voxel's status. How the masses are allocated to the events *occupied*, ¬*occupied* and *unknown* is explained in Sections 2.2 and 2.3.

On the basis of the power set just defined, the set

$$\mathbb{M}^c = \{M_s^c | s = 1 \dots |\mathbb{S}|\},$$

where $\mathbb{S}$ is the set of logical sensors and $|\mathbb{S}|$ is its cardinality, represents all the bba provided by the $|\mathbb{S}|$ logical sensors for the voxel c. Note that because we are not assuming any particular camera configuration, the voxel might not project into all of the cameras. Therefore, if a voxel does not project into a camera, then the logical sensors that include this camera are considered to be completely ignorant, resulting in setting all of the mass in the Ω set.

Let $\mathcal{M}^c$ be the bba resulting from fusing the evidences in $\mathbb{M}^c$. We have employed in this work Dempster's combination rule [31], thus assuming independence between the cameras employed. This is in our opinion a good election from a practical point of view given the fact that correlated cameras are assigned with a low confidence in our model, as we shall show later.

As the projection test for a voxel, we propose to use the pignistic probability transformation (BetP [32]) of the events in $\mathcal{M}^c$ to classify its state in the following way :

$$\mathrm{PT}(c,\mathbb{F}) = \begin{cases} unknown & \text{if } \mathrm{BetP}(\{occupied\}) > \mathrm{BetP}(\{\neg occupied\}) \\ & \quad \text{and } (d(c) < (R-1)), \\ occupied & \text{if } \mathrm{BetP}(\{occupied\}) > \mathrm{BetP}(\{\neg occupied\}) \\ & \quad \text{and } (d(c) = (R-1)), \\ \neg occupied & \text{otherwise,} \end{cases} \tag{2}$$

where $d(c)$ is the depth of the voxel c.

2.2. Degrees of evidence calculation

This section explains the basic belief assignment proposed for the masses M_s^c of each logical sensor about each voxel.

As mentioned above, each logical sensor $s_{ij} \in \mathbb{S}$ is formed by the two cameras i and j. It is assumed that the intrinsic and extrinsic calibration parameters of each camera are known, so that we can project an image point into the coordinates of the 3D reference space in each of the silhouettes obtained by the silhouette extraction algorithm in each camera.

Given a voxel c, a camera i and its associate silhouette image $\mathcal{F}_i \in \mathbb{F}$, we define its Occupation Area Rate (OAR) as

$$occ_i^c = \frac{1}{|P(\mathcal{F}_i, c)|} \sum_{p \in P(\mathcal{F}_i, c)} \delta_i(p), \tag{3}$$

where $P(\mathcal{F}_i, c)$ represents the set of pixels of the image $\mathcal{F}_i$ to which the voxel c projects, $|P(\mathcal{F}_i, c)|$ is the area of this projection in pixels, and δ_i is defined as

$$\delta_i(p) = \begin{cases} 1 & \text{if } p \text{ is inside of the silhouette } \mathcal{F}_i, \\ 0 & \text{otherwise.} \end{cases}$$

A fundamental difference regarding the previous voxel set-based approach [17] is the necessity of a mechanism to adapt the OAR obtained by each camera based on the level of the octree. In the octree approach, voxels have different sizes and therefore it will have very different OARs because of the huge differences in the area projected in silhouette images.

We propose to use a Butterworth function to compensate input OARs (occ_i^c) to new compensated OARs ($\overline{occ_i^c}$) before they can be used in the DS formulation. This function allows a high level of control for the interpretation of the input OARs and it will lead us to define an automatic and dynamic threshold to distinguish between *occupied* and ¬*occupied*. The expression proposed to calculate the compensated OAR is as follows:

$$\overline{occ_i^c} = \begin{cases} occ_i^c & \text{if } f_c = 1, \\ \dfrac{1}{1 + \left(\dfrac{1 - occ_i^c j}{1 - f_c}\right)^{o_c}} & \text{in other case.} \end{cases} \tag{4}$$

The cutoff frequency f_c and the order o_c of the Butterworth function for each voxel c are calculated automatically depending on the resolution level $l(c)$ and on the average OAR of all the views as follows:

$$f_c = \frac{1}{2^{l(c)}} \mathrm{avg}(\{occ_i^c | i \in \{1,..,|\mathbb{F}|\}, occ_i^c > 0\}),$$

where avg() represents the average operation and

$$o_c = 2^{l(c)+1}.$$

Note here the difference between the level and depth of voxels ($l(c) = R-1-d(c)$).

Both control variables of Butterworth function are mainly controlled by the voxel level $l(c)$. At high levels, close to the octree's root, the size of voxels is huge compared to deepest

voxels so their uncompensated OAR will be likely very small. In these first levels, we want that f_c will be small (near to 0) and the function's slope very accentuated to be more permissive. In this way, voxels with low uncompensated OAR will not be rejected in first levels of refinement. On the other hand, as we reach deeper levels we are not interested in continuing the refinement of voxels with low uncompensated OAR. So at deepest voxels the function's slope will be smaller and f_c will be nearer to the average uncompensated OAR in all the views. It will make the occupancy voxel decision not as dependent of $l(c)$ as in the former case.

Fig. 3 shows the Butterworth function for different resolution values considering two possible values of average OAR.

Given $\overline{occ_i^c}$ and $\overline{occ_j^c}$ for the voxel c in the cameras i,j, we allocate evidence mass to the *occupied* event for the sensor s_{ij} built with these cameras in the voxel c as follows:

$$m^c_{s_{ij}}(\{occupied\}) = \left(\frac{\overline{occ_i^c}+\overline{occ_j^c}}{2}\right)^n. \tag{5}$$

Eq. (5) can be seen as a general way of averaging the occupancy values of the cameras with a free parameter $n \in (0,\infty)$. For $n=1$, the function represents the arithmetic mean. When $n\to\infty$ the function forces both cameras to provide high occupancy values to obtain a high value. However, as $n\to 0$, the measurement becomes more "permissive", in the sense that inconsistent voxels can be assigned with a high value of $m^c_{s_{ij}}(\{occupied\})$. Indeed, very low values of the parameter n might result in incoherent results since they become independent of the data itself.

Once calculated the evidence mass associated with the event *occupied*, the evidence mass associated with the event $\neg occupied$ for the voxel c in the logical sensor s_{ij} is calculated in the following way:

$$m^c_{s_{ij}}(\{\neg occupied\}) = 1 - m^c_{s_{ij}}(\{occupied\}) \tag{6}$$

2.3. Calculating the reliability of a logical sensor

In Ref. [30], Pribanić studied the influence of the camera setup on the accuracy of the 3D reconstructions. He demonstrated that higher reconstruction accuracy takes place for cameras forming an angle of 90° with the object. As the angle formed by the camera deviates from 90°, there is a reduction in the reconstruction accuracy. Nevertheless, none of the previous octree based SfS approaches have explicitly considered positional information in their formulations.

As we have already mentioned in Section 2.2, we form a logical sensor $s_{ij} \in \mathbb{S}$ by combining two cameras i,j. We define the unreliability of the sensor s_{ij} in determining the occupancy of a voxel c as

$$\alpha^c_{s_{ij}} = |\cos(\beta^c_{s_{ij}})|, \tag{7}$$

where $\beta^c_{s_{ij}}$ represents the angle between the cameras with respect to the analysed voxel. We have studied two forms for calculating this angle:

- The first alternative is more precise but has a greater computational cost and measures the angle formed by the segments that bind the centres of projection of the cameras that form the logical sensors and the centre of the studied voxel (See Fig. 4a.)
- The second alternative has a low computational cost but is less precise and measures the angle formed by the two view directions associated with the two cameras that form the logical sensor, independent of the studied voxel (Fig. 4b).

Both alternatives to compute $\alpha^c_{s_{ij}}$ satisfy the constraint given by Pribanić [30], i.e., perpendicular cameras should be considered to be more reliable than parallel ones.

Once having calculated the grade of unreliability $\alpha^c_{s_{ij}}$ associated with the sensor, we propose to obtain the masses $m^c_{s_{ij}}\{occupied\}$ and $m^c_{s_{ij}}\{\neg occupied\}$ using Eqs. (4) and (6), respectively, to apply a "discounting rate" [18,31] to the bba in the following general way:

$$\begin{cases} m^\alpha(A) = (1-\alpha)m(A) & \forall A \subset \Omega, \\ m^\alpha(\Omega) = (1-\alpha)m(\Omega)+\alpha. \end{cases} \tag{8}$$

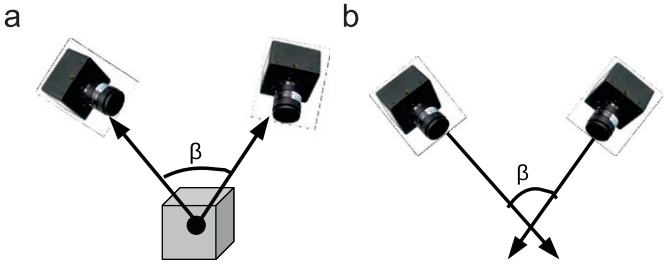

Fig. 4. (a) Angle formed by the segments from a voxel's centre to the centre of projection of each camera that forms a logical sensor. (b) Angle formed by the camera view directions.

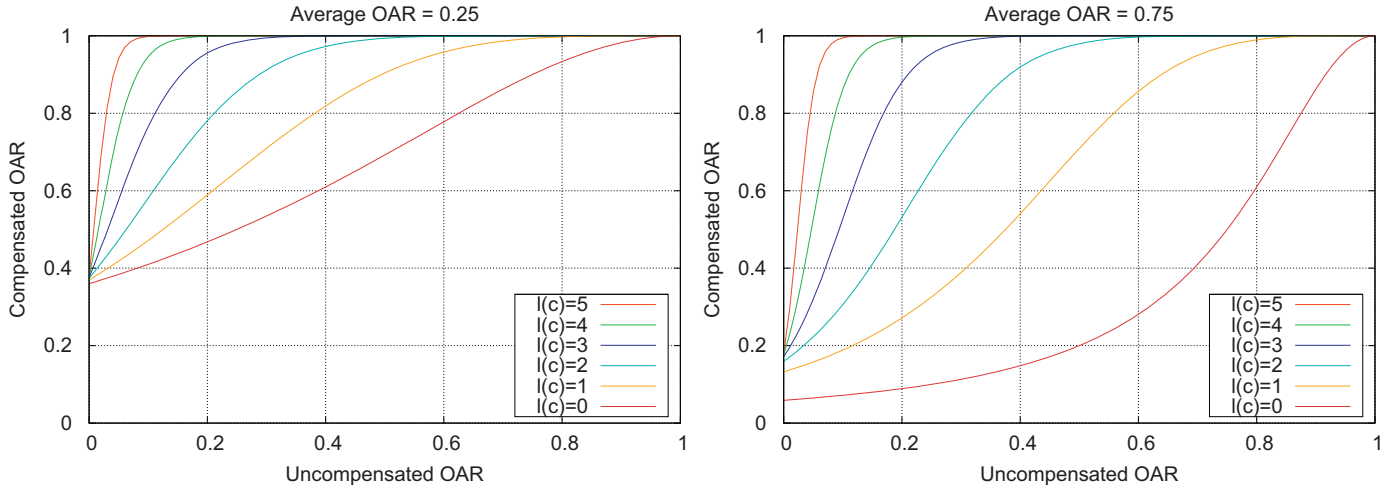

Fig. 3. Response of the Butterworth function for different resolution values ($l(c)$) and fixed average OARs equal to 0.25 and 0.75 for f_c calculation.

Particularising in our problem, the reader should note that $m(\Omega)=0$ before the discount has been realised. Also note that when the cameras are parallel, the α value is near to 1 and practically the entire evidence mass is allocated to the *unknown* event (represented by the Ω set), indicating the ignorance of the logical sensor to establish the state of occupation of the voxel. Eq. (8) is used to obtain the final bba $M^c_{s_{ij}}$ (Eq. (1)) associated with a logical sensor s_{ij}.

3. Experimental results

In this section, we present the results of three experiments that we have performed to contrast our proposal with existing methods. In the first experiment, we evaluate the quality of the obtained reconstruction in an objective way. In the second experiment, we study the dependence of the quality of the reconstruction with regard to the number of used views. In the third experiment, we evaluate our proposal in a real environment of the application.

In the three experiments, we compare two versions of our proposal: the first that calculates the grade of reliability of a logical sensor as a function of the angle formed by the segments that bind the centre of projection of each camera with the centre of the studied voxel (Oct-DS-Vox) and the second that calculates this grade independently of the voxel using the angle formed by the lines of vision of the two cameras (Oct-DS-Gen). We also compare these approaches with two versions of the voxel set-based approach [17], also with the two forms of calculating the grade of reliability of a logical sensor, which we call "VS-DS-Vox" and "VS-DS-Gen". A further comparison is made with the octree-based classic method of intersection of visual cones due to Szeliski [29], which we call "Oct-AND".

3.1. Experiment 1

We have used the database of images "Dancer" [33]. This database mixes an actual scenario with an avatar carrying out different evolutions. Each scene is taken from eight different points of view and the given silhouettes fulfil the restriction of consistency. With the consistent silhouettes we use the octree-based classic method of Szeliski to obtain the reconstructions that will be used as Ground Truth (GT) in the comparisons.

We have selected 10 scenes from the database. To each selected scene, synthetic noise has been added to simulate errors in the extraction of the silhouettes. The process of generating synthetic noise is the following: to each silhouette we added randomly between 1 and 14 "spots". A "spot" is a circle with a random radius in the interval [8, 30] pixels. The proportion of false negative (FN)/false positive (FP) errors generated was 30%/70%, respectively. The spatial disposition of the noise has been random, with priority in areas that have more silhouette density. As an example, Fig. 5 shows four views of one of the selected scenes together with the versions contaminated with synthetic noise.

The quality of the volumetric reconstruction obtained by the algorithms is evaluated using three verification measures commonly employed in the information retrieval field. In the first place, the recall measure, also known as the detection rate, indicates the percentage of detected true positives in the reconstructed scene in relation to the total number of true positives in the GT:

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{9}$$

where TP is the number of true positives, FN is the number of false negatives and the denominator refers to the number of occupied voxels in the GT. Nonetheless, the recall measure is not enough to compare different methods, so it is generally used in conjunction with the precision measure, which indicates the percentage of true positive (TP) in relation to the total number of voxels detected by the method:

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}. \tag{10}$$

We have also considered the f-measure, which represents the weighted harmonic mean of precision and recall:

$$\text{f-measure}=\frac{2\times\text{recall}\times\text{precision}}{\text{recall}+\text{precision}}. \tag{11}$$

With the intention of obtaining a better estimate of the values of the measures, the experiment was repeated 100 times (with different noise spots) to obtain averages.

Three levels of maximum resolution were used for the generation of the octree/voxel set: five, six, and seven levels. Figs. 6, 7 and 8 and Table 2 show the values of the f-measures for each resolution level obtained by the different proven methods and for different values of n (see Eq. (5)).

The results obtained show that our proposal is a significant improvement on the classic octree-based technique when we have inconsistent silhouettes. We also observe that the classic proposal is penalised with regard to the number of resolution levels used, while our proposal achieves best f-measure values in a "stable" range of values for the parameter n with the three resolutions levels tested. (See Table 2.)

With regard to the approach based on voxel set, our proposal has maintained the good results obtained by the proposal based on voxel set, and even, a lightly superior quality has been obtained with the biggest proven resolution level. This result should be seen in connection with the computation times that are

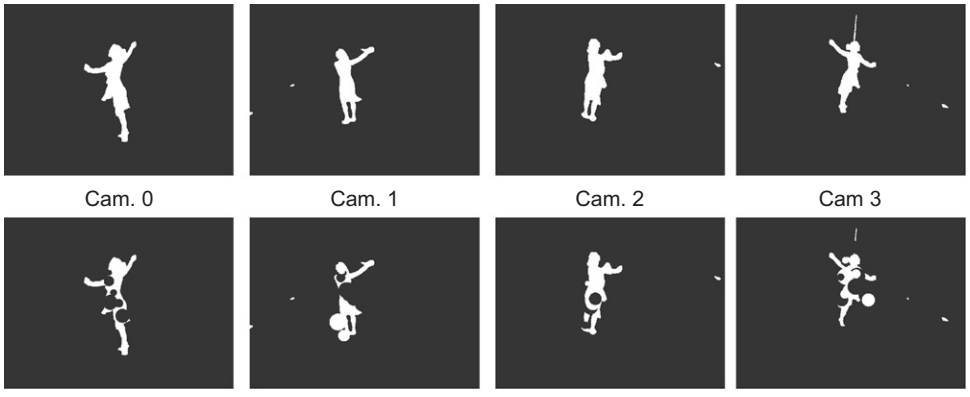

Fig. 5. Example of the first four views (top) and their respective versions contaminated with noise (below) corresponding to scene number 614 of the database "Dancer".

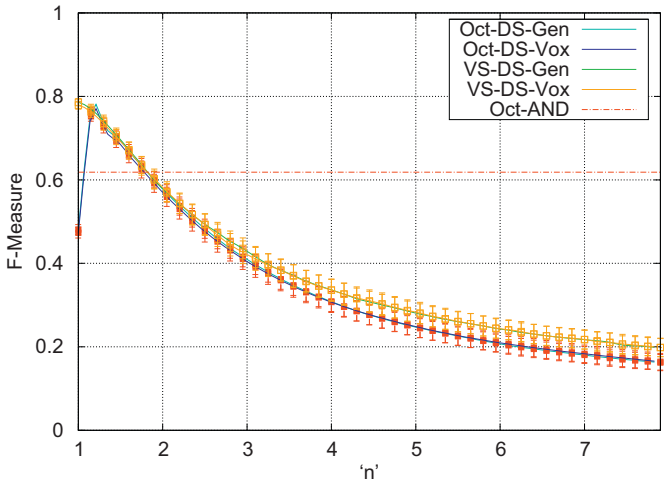

Fig. 6. Results obtained with five resolution levels. f-Measure curves are shown for the classic method (Oct-AND) and the two versions of our proposal (Oct-DS-Gen and Oct-DS-Vox) and voxel-sets proposal (VS-DS-Gen and VS-DS-Vox).

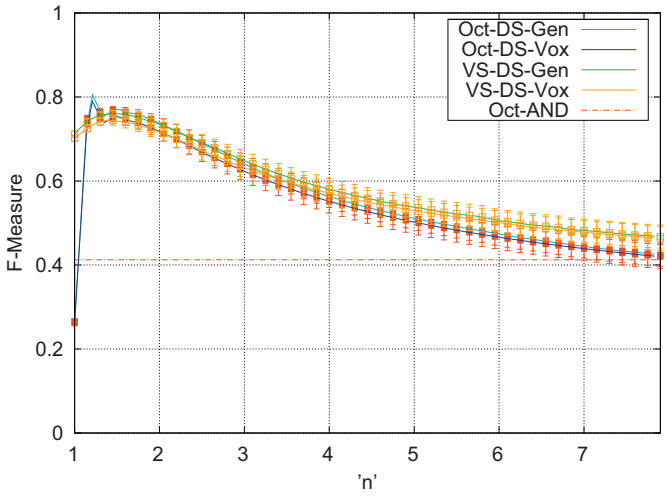

Fig. 7. Results obtained with six resolution levels. f-Measure curves are shown for the classic method (Oct-AND) and the two versions of our proposal (Oct-DS-Gen and Oct-DS-Vox) and voxel-sets proposal (VS-DS-Gen and VS-DS-Vox).

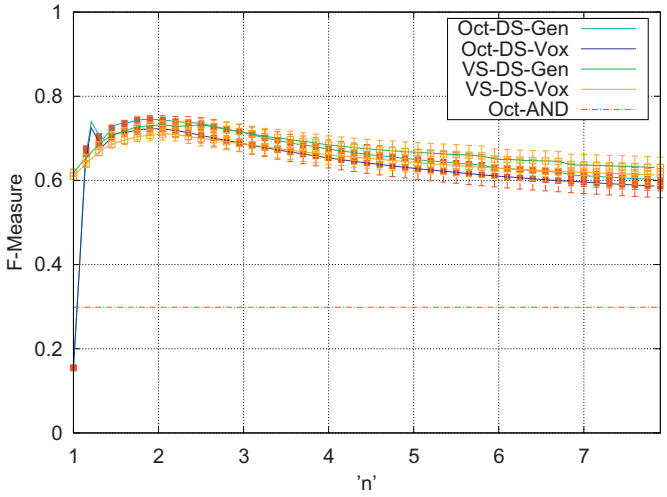

Fig. 8. Results obtained with seven resolution levels. f-Measure curves are shown for the classic method (Oct-AND) and the two versions of our proposal (Oct-DS-Gen and Oct-DS-Vox) and voxel-sets proposal (VS-DS-Gen and VS-DS-Vox).

Table 2
Values of f-measures obtained by the alternative methods, using three resolution levels. The confidence intervals are also shown with $\alpha = 0.95$. For Dempster–Shafer based methods we show the best result values of the parameter n.

Method	Resolution levels		
	5	6	7
Oct-AND	0.62 ± 0.02	0.41 ± 0.03	0.30 ± 0.03
Oct-DS-Gen	0.76 ± 0.01, $n = 1.15$	0.77 ± 0.01, $n = 1.45$	0.75 ± 0.01, $n = 1.9$
Oct-DS-Vox	0.75 ± 0.01, $n = 1.15$	0.75 ± 0.01, $n = 1.45$	0.72 ± 0.01, $n = 1.9$
VS-DS-Gen	0.78 ± 0.01, $n = 1.0$	0.76 ± 0.01, $n = 1.45$	0.73 ± 0.01, $n = 2.05$
VS-DS-Vox	0.77 ± 0.01, $n = 1.0$	0.74 ± 0.01, $n = 1.45$	0.71 ± 0.01, $n = 2.05$

Table 3
Computation times in mile-seconds on a Quadcore 2.4 GHz.

Res/method	Oct-AND	Oct-DS-Gen	Oct-DS-Vox	VS-DS-Gen	VS-DS-Vox
5	1.65	9.95	12.50	104.25	130.19
6	4.20	50.20	64.65	794.90	995.37
7	17.75	313.65	406.50	6323.10	7872.26

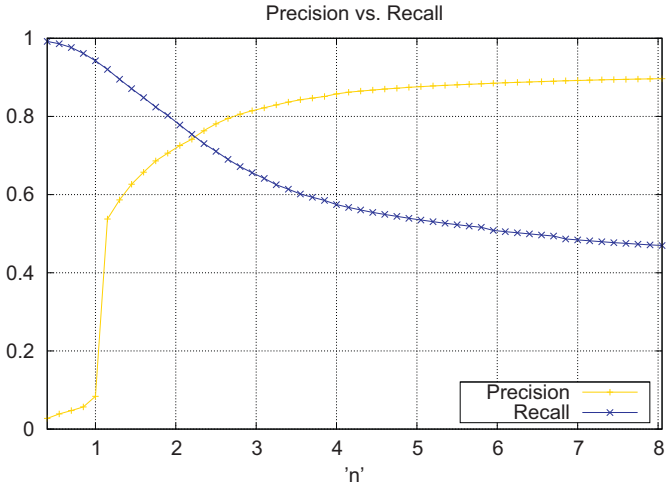

Fig. 9. Evolution of the measures precision and recall obtained by our approach when the free parameter n varies. Seven resolution levels were used.

shown in Table 3, where it is observed as our proposal it has been run 20 times faster with seven resolution levels maintaining the quality of the obtained reconstruction.

With regard to the two forms tested to obtain the unreliability parameter of a logical sensor in our proposal, the values of f-measures obtained do not indicate which of the two alternatives is best. Using the angle formed by the lines of view has a significantly smaller computational cost because this angle is independent of the analysed voxel and it can easily be pre-calculated; this alternative, called Oct-DS-Gen, will be the alternative used in the rest of experiments.

Finally, the results shown in Figs. 6–8 and Table 2 also indicate that the free parameter n of our proposal has its best values when considering the f-measure in the interval $[1, 2]$. The adjustment of this parameter allows us to vary the balance between the detection rate and the precision in the reconstruction obtained as a function of the required application. This is shown better in Fig. 9, where the measures recall and precision vary separately. In this figure, it can be observed that when $n \to 0$, the recall increases

at the expenses of worsening the precision. However when $n \gg 1$ the precision increases but the recall decreases.

3.2. Experiment 2

In a second experiment, we evaluated the influence of the number of views of the scene on the quality of the reconstruction obtained. It is commonly accepted that the quality of the reconstruction is directly proportional to the number of views of the scene [3], but this belief pertains to when the extracted silhouettes are consistent.

We selected a scene from the database "Dancer". To this scene, segmentation noise was added in a similar fashion as in the previous section.

We have eight cameras that provide us eight different views of the same scene. We now are interested in studying how the algorithm behaves when the number of available views varies from 3 to 8: $V \in \{3,4,5,6,7,8\}$. To carry out the experiment with $V=3$ views we could work with $\binom{8}{3} = 56$ possible subsets of cameras. To simplify, for each number of views, we have only used the subset of cameras that minimises the sum of the corresponding reliability coefficients (Eq. (7)) of the logical sensors obtained by combination of all the possible pairs of cameras belonging to this subset. This process is repeated for subsets of 4, 5, 6 and 7 views. For the experiment with 8 views, there is an only possible subset of cameras.

As can be observed in Fig. 10, our proposal shows better behaviour as the number of views increases, while the opposite happens with the classic method. It is also observed that our proposal is less sensitive regarding the resolution used to generate the octree. Because the results obtained by voxel set-based proposals are very similar to those obtained from our proposal, they are not shown, to make the figure more clear.

Fig. 11 explains the poor results obtained by the classic method when the number of views increases. Due to the inconsistency in the extracted silhouettes, there is a higher probability that an FN affects some portion of the volume in some of the views and, due to the projection test used being the intersection of visual cones, the result would be that a portion of the volume is not reconstructed correctly. This causes the recall to decrease significantly when the number of views used increases.

Fig. 12 shows the detection and precision curves obtained separately using different numbers of views and for three different resolutions. Consistent with that shown by Fig. 9 is observed

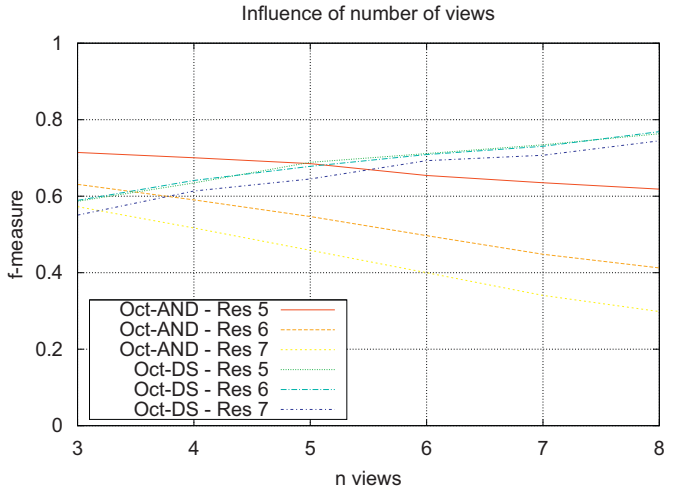

Fig. 10. Influence of the number of views on the reconstruction error. The classic method (Oct-And) is compared with our proposal (Oct-Ds-Gen), considering three levels of maximum resolution in the octree.

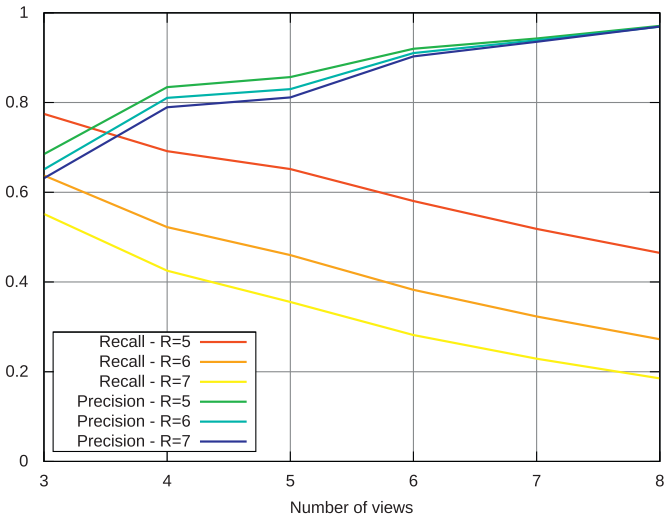

Fig. 11. Values of precision and recall from the classic method as a function of the number of views used for the reconstruction.

again that the free parameter n allows to change the balance between detection and precision. Fig. 12 also shows that for a given resolution, increase the number of views for better results.

3.3. Experiment 3

The goal of the third experiment is to carry out a qualitative comparison of the algorithms in a complex reconstruction task. The scenario chosen is our lab, a room of approximately 5×6 m that is equipped with a total of six synchronised firewire cameras placed at a height of 2.3 m in slanting positions. Cameras record with a resolution of 640×480 pixels and we have employed the background subtraction technique proposed by Horprasert [7] to extract the silhouettes.

As previously indicated, our main goal is to create a robust algorithm that can be employed for the purpose of tracking people. The people tracking problem in a realistic scenario possesses a major challenge to reconstruction algorithms for several reasons. First, this type of scenario imposes reduced control over the lighting conditions so that the quality of the segmentation technique is low. Moreover, it might be expected to find shadows generated by the people being tracked. Second, cluttered environments are fraught with occlusions which are often referred as "systematic false negatives".

The space analysed covers a volume of $4.2 \times 2.1 \times 4$ m. The time required for reconstructing the scene is approximately 50 ms on a Quadcore 2.4 GHz. The voxel projection is approximated by bounding box of the projected voxel's corner points.

Fig. 13 shows one of the frames captured in our lab. This figure shows a person moving around a chair with a box placed in the middle of the room. It can be seen that the chair occludes the person's legs in some views. Also, the person is not seen fully in all views. The silhouette images extracted are also shown. It is well worth noting that the silhouette images contain many misclassifications. Not only can we find FN produced by either lighting conditions or occlusions, but we also can find FP produced by shadows that are present in the scene. The back-projection of the reconstruction results of the standard method Oct-AND and our proposal Oct-DS-Gen, using several values for the parameter n, is shown. The results for the voxel set-based approach are not shown because they are very similar to the results obtained by our proposal and are in concordance with the results obtained in Experiment 1.

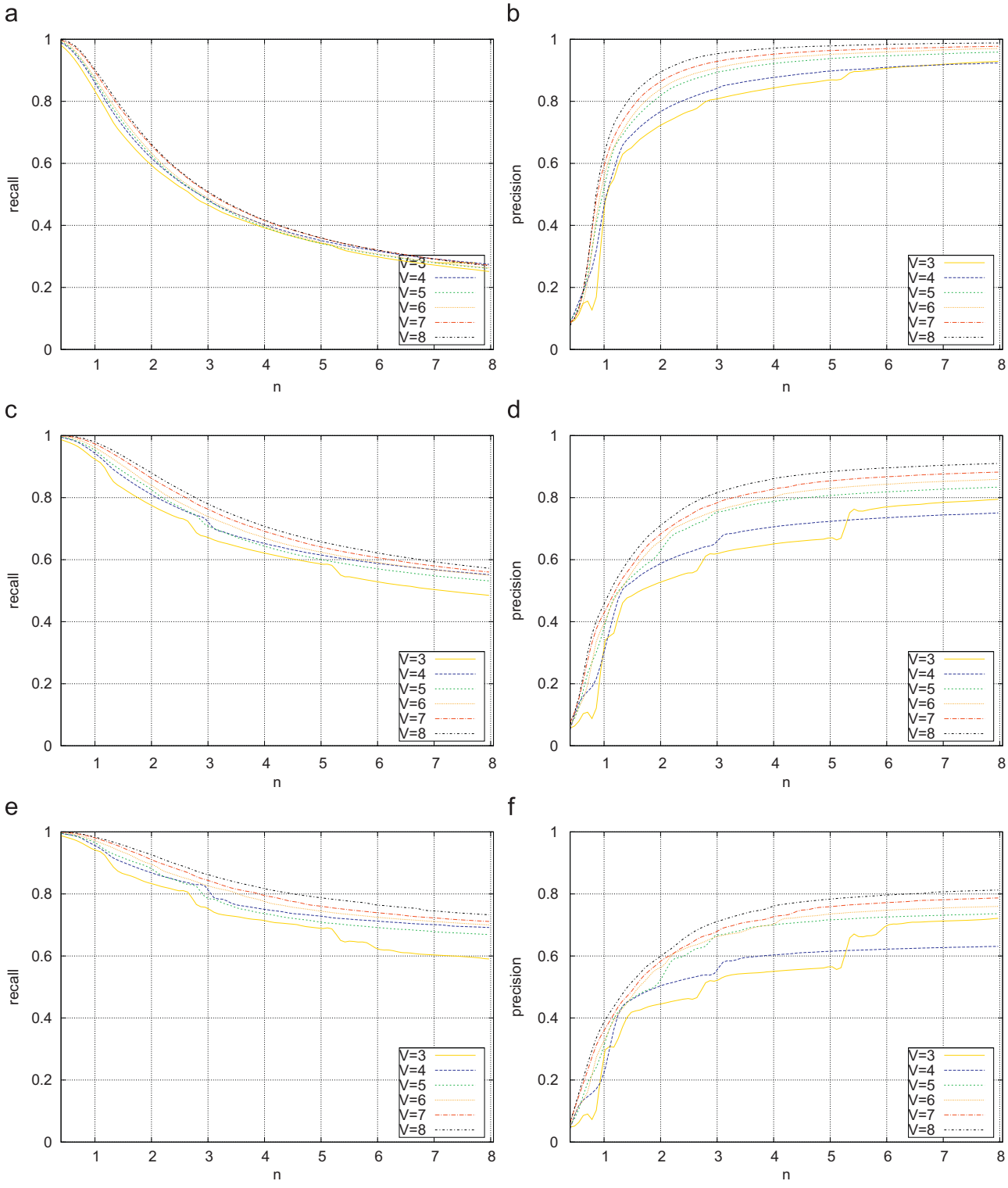

Fig. 12. Recall and precision obtained using different number of views with resolution 5 (a and b), resolution 6 (c and d), and resolution 7 (e and f).

The results obtained clearly demonstrate that the standard Oct-AND method gives a precise reconstruction, but is unable to properly recover a person's body shape, e.g., the legs are missing. Although our method introduces some FP, it is able to properly reconstruct the whole person's body, including his legs. As can be seen in Fig. 13, the precision can be adjusted by tuning the value of the parameter *n*.

4. Conclusions

In this paper, we have presented a novel technique to solve the problem SfS with an approach based on octree. In contrast to the classic octree method based on the intersection of visual cones, our proposal does not require consistency of the extracted silhouettes. To process the ambiguity resulting from the inconsistencies generated

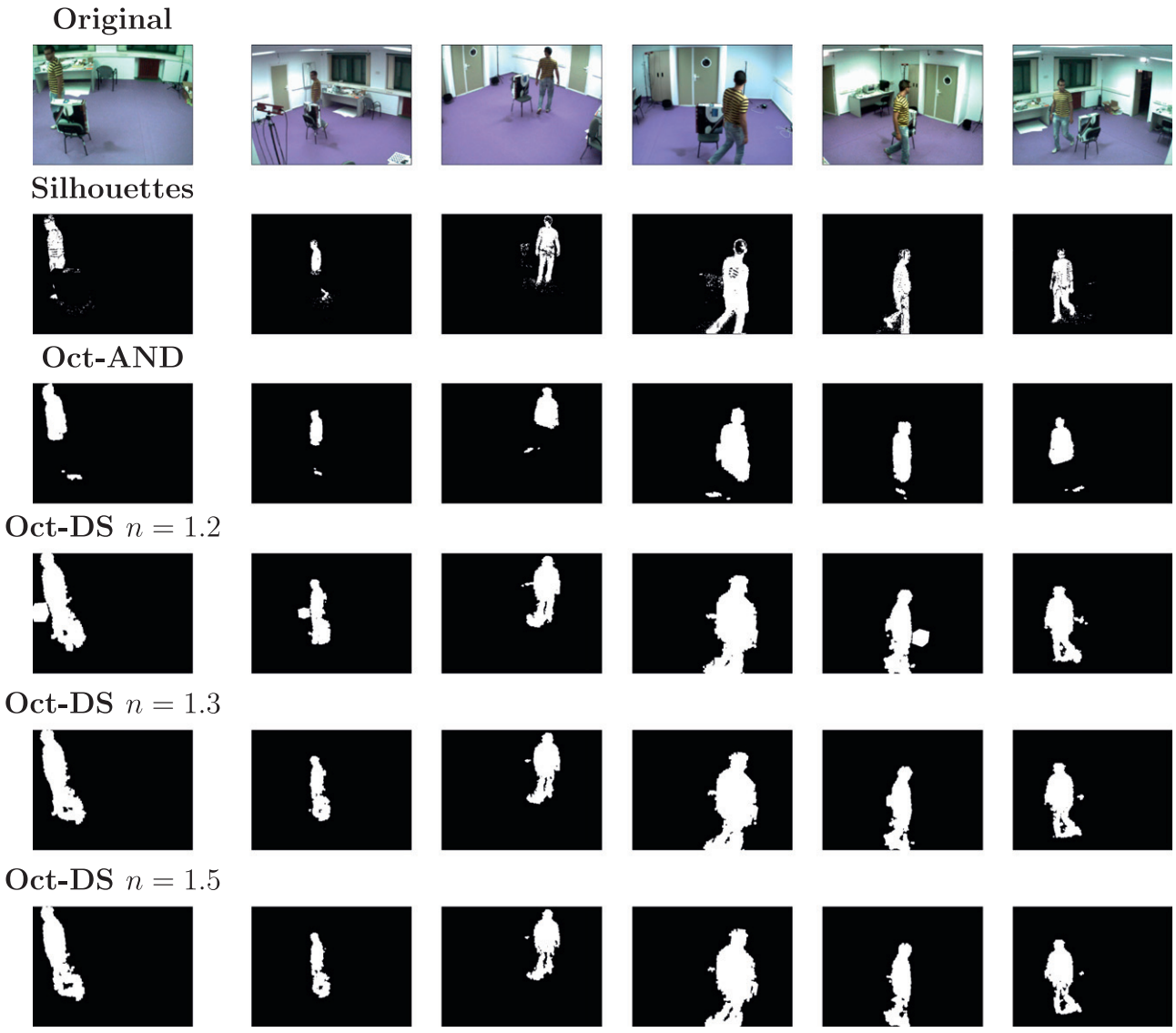

Fig. 13. Results in a real scenario. The first two rows show the six views used and the silhouettes extracted, respectively. The following rows show the back-projections on each view of the volume reconstructed by the classic method (Oct-AND) and by our proposal (Oct-DS-Gen), with values of the parameter $n=1.3$, $n=1.2$ and $n=1.5$, respectively.

in the extraction of the silhouettes, we propose a projection test based on data fusion of a group of logical sensors formed from pairs of cameras.

Recently, the problem of the inconsistency of the silhouettes has also been approached by other methods, namely SfS based on voxel sets. The main inconvenience of these methods is the high computational cost from the exhaustive analysis of the working space that these methods perform.

The proposed method was compared to the standard octree based SfS proposed by Szeliski [29], and the results show that it obtains better reconstructions under noisy conditions when the extracted silhouettes are inconsistent.

We have also compared our proposal with the method based on voxel sets proposed by Díaz et al. [17], which solves the problem of the inconsistency. The results obtained are similar considering the quality of the 3D reconstruction, but our proposal obtains a reduction in computation time of an order of magnitude.

Our approach is particularly attractive due to its simplicity and its high performance and because it does not require specifying priors or conditionals that might be difficult to obtain in complex scenarios, especially in detection tasks and in tasks that involve tracking people.

Acknowledgements

This work has been financed by the Project TIN2010-18119 of MICINN of Spain.

References

[1] B. Baumgart, Geometric Modeling for Computer Vision, Ph.D. Thesis, CS Department, Stanford University, aIM-249, STAN-CS-74-463 1974.
[2] A. Laurentini, The visual hull: a new tool for contour-based image understanding, in: Proceedings of Seventh Scandinavian Conference on Image Processing, 1991, pp. 993–1002.
[3] A. Laurentini, The visual hull concept for silhouette-based image understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 150–162.
[4] H. Chen, T. Huang, A survey of construction and manipulation of octrees, Computer Vision, Graphics and Image Processing 43 (1988) 409–431.
[5] C. Jackins, S. Tanimoto, Oct-trees and their use in representing three-dimensional objects, Computer Graphics and Image Processing 14 (3)

(1994) 249–270. http://dx.doi.org/10.1016/0146-664X(80)90055-6. URL 〈http://www.sciencedirect.com/science/article/B7GXF-4KM8SJC-4/2/3489b20a22b84079394d7a5306865324〉.
[6] A. Elgammal, D. Harwood, L. Davis, Non-parametric Model for Background Subtraction, Lecture Notes in Computer Science, vol. 1843, 2000, pp. 751–767.
[7] T. Horprasert, D. Harwood, L. Davis, A statistical approach for real-time robust background subtraction and shadow detection, in: 7th IEEE International Conference on Computer Vision, Frame Rate Workshop (ICCV '99), 1999, pp. 1–19.
[8] M. Karaman, L. Goldmann, D. Yu, T. Sikora, Comparison of static background segmentation methods, in: Visual Communications and Image Processing 2005, vol. 5960, 2005, pp. 2140–2151.
[9] R. Cipolla, M. Yamamoto, Stereoscopic tracking of bodies in motion, Image and Vision Computing 8 (1) (1990) 85–90.
[10] J.S. Franco, E. Boyer, Fusion of multi-view silhouette cues using a space occupancy grid, in: 10th IEEE International Conference on Computer Vision (ICCV 2005), 2005, pp. 1747–1753. http://dx.doi.org/10.1109/ICCV.2005.105.
[11] D. Snow, P. Viola, R. Zabih, Exact voxel occupancy with graph cuts, in: Proceedings of Computer Vision and Pattern Recognition, IEEE Computer Society, 2000, pp. 345–353.
[12] S. Sullivan, J. Ponce, Automatic model construction and pose estimation from photographs using triangular splines, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (10) (1998) 1091–1096.
[13] J. Zheng, Acquiring 3-d models from a sequence of contours, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 163–178.
[14] K. Cheung, T. Kanade, J.-Y. Bouguet, M. Holler, A real time system for robust 3d voxel reconstruction of human motions, in: Proceedings of Computer Vision and Pattern Recognition, IEEE Computer Society, vol. 2, 2000, pp. 714–720.
[15] J. Landabaso, M. Pardàs, Foreground regions extraction and characterization towards real-time object tracking, in: Proceedings of Multimodal Interaction and Related Machine Learning Algorithms, Lecture Notes in Computer Science, Springer, 2005.
[16] J. Landabaso, M. Pardás, J. Ramon Casas, Shape from inconsistent silhouette, Computer Vision and Image Understanding 112 (2008) 210–224.
[17] L. Díaz-Más, R. Muñoz-Salinas, F.J. Madrid-Cuevas, R. Medina-Carnicer, Shape from silhouette using Dempster–Shafer theory, Pattern Recognition 43 (6) (2010) 2119–2131. doi:http://dx.doi.org/10.1016/j.patcog.2010.01.001.
[18] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[19] B. Baumgart, Geometric Modeling For Computer Vision, Aim-249, stan-cs-74-463, Stanford University, 1974.
[20] C.L. Jackins, S.L. Tanimoto, Oct-trees and their use in representing three-dimensional objects, Computer Graphics and Image Processing 14 (1980) 249–270.
[21] I. Gargantini, Linear octrees for fast processing of three dimensional objects, Computer Graphics and Image Processing 20 (1982) 356–374.
[22] H.H. Chen, T.S. Huang, SURVEY a survey of construction and manipulation of octrees, Computer Vision, Graphics, and Image Processing 43 (1988) 409–431.
[23] C. Chien, J. Aggarwal, A volume/surface octree representation, in: 7th International Conference on Pattern Recognition, Montreal, Canada, 1984, pp. 817–820.
[24] W. Martin, J. Aggarwal, Volumetric description of objects from multiple views, IEEE Transactions on Pattern Analysis and Machine Intelligence 5 (2) (1983) 150–158.
[25] J. Veenstra, N. Ahuja, Efficient octree generation from silhouettes, in: Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Miami, USA, 1986, pp. 537–542.
[26] C. Chien, J. Aggarwal, Recognition and matching of 3-D objects using quadtrees/octrees, in: 3rd Workshop on Computer Vision, Bellaire, MI, 1985, pp. 94–54.
[27] T.-H. Hong, M.O. Shneier, Describing a robot's workspace using a sequence of views from a moving camera, IEEE Transactions on Pattern Analysis and Machine Intelligence 7 (6) (1985) 721–726. http://dx.doi.org/10.1109/TPAMI.1985.4767730. URL 〈http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4767730〉.
[28] M. Potmesil, Generating octree models of 3D objects from their silhouettes in a sequence of images, Computer Vision, Graphics, and Image Processing 40 (1) (1987) 1–29. URL 〈http://linkinghub.elsevier.com/retrieve/pii/0734189X87900533〉.
[29] R. Szeliski, Rapid octree construction from image sequences, Computer Vision, Graphics, and Image Processing: Image Understanding 58 (1) (1993) 23–32.
[30] T. Pribanić, M. Cifrek, S. Tonković, The choice of camera setup in 3d motion reconstruction systems, in: Proceedings of the 22th Annual EMBS International Conference, 2000, pp. 163–165.
[31] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, International Journal of Approximate Reasoning 9 (1993) 1–35.
[32] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–243.
[33] INRIA, Dancer image data base, Internet, February 2011. URL 〈http://4drepository.inrialpes.fr/public/viewgroup/1〉.

L. Díaz-Más received the Bachelor degree in Computer Science in 2008 from University of Córdoba, Spain. He is currently studying for the Ph.D. degree with a research grant. Since 2007 he has been collaborating in the research group of Applications of Artificial Vision in the University of Córdoba. His research is focused on 3-D modelling and action recognition.

F.J. Madrid-Cuevas received the Bachelor degree in Computer Science from Malaga University (Spain) and the Ph.D. degree from Polytechnic University of Madrid (Spain), in 1995 and 2003, respectively. Since 1996 he has been working with the Department of Computing and Numerical Analysis of Cordoba University, currently he is an assistant professor. His research is focused mainly on image segmentation, 2-D object recognition and evaluation of computer vision algorithms.

R. Muñoz-Salinas received the Bachelor degree in Computer Science from Granada University (Spain) and the Ph.D. degree from Granada University (Spain), in 2006. Since 2006 he has been working with the Department of Computing and Numerical Analysis of Cordoba University; currently he is an assistant professor. His research is focused mainly on mobile robotics, human-robot interaction, artificial vision and soft computing techniques applied to robotics.

A. Carmona-Poyato received his title of Agronomic Engineering and Ph.D. degree from the University of Cordoba (Spain), in 1986 and 1989, respectively. Since 1990 he has been working with the Department of Computing and Numerical Analysis of the University of Cordoba as lecturer. His research is focused on image processing and 2-D object recognition.

R. Medina-Carnicer received the Bachelor degree in Mathematics from University of Sevilla (Spain). He received the Ph.D. in Computer Science from the Polytechnic University of Madrid (Spain) in 1992. Since 1993 he has been a lecturer of Computer Vision in Cordoba University (Spain). His research is focused on edge detection, evaluation of computer vision algorithms and pattern recognition.

5

Conclusions

Three-dimensional action recognition is becoming an active topic of research. In last years, new devices have appeared enabling new ways of interacting with computers. Actions performed by people in an unrestricted scene can be performed in any direction, different speeds in the movements, and equivalent gestures could vary much between them depending on the actors. This thesis has proposed three main contributions to help incorporating action recognition in daily-life scenarios by increasing their robustness. Our work has achieved the following objectives:

- It has been demonstrated that the new descriptor proposed for representing 3D actions (the "Volume Integral") minimizes the amount of information necessary for recognizing actions without losing discriminative power (20).
- In (19), we have demonstrated that the robustness of SfS algorithms can be increased using the Dempster-Shafer theory of evidence by considering the relative position between cameras.
- Finally, it has been proven that it is possible to formulate a multi-scale projection test in SfS algorithms and therefore integrate the usage of Octree structures in them (18). Thanks to this integration SfS algorithms obtain the benefits of Octree structures.

As an additional conclusion, we want to remark that the achievements of this thesis can be used to help in the solutions of other specific problems like people tracking. For example, in (41) we have proposed a system that uses the Shape from Silhouette

using Dempster-Shafer theory (SfSDS) approach for tracking people in uncontrolled environments.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, **28**(1):44–58, January 2006.

[2] Astride Aregui and Thierry Denœux. Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, **49**(3):575–594, November 2008.

[3] C. Barron and I.A. Kakadiaris. Estimating anthropometry and pose from a single image. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, **1**, pages 669–676. IEEE Comput. Soc, 2000.

[4] C. Barron and I.A. Kakadiaris. On the improvement of anthropometry and pose estimation from a single uncalibrated image. In *Proceedings Workshop on Human Motion*, pages 53–60. IEEE Comput. Soc, 2000.

[5] Carlos Barrón and Ioannis A. Kakadiaris. Estimating Anthropometry and Pose from a Single Uncalibrated Image. *Computer Vision and Image Understanding*, **81**(3):269–284, March 2001.

[6] Bruce Guenther Baumgart. *Geometric Modelling for Computer Vision, Rep*. PhD thesis, Standford, 1974.

[7] Isabelle Bloch. Defining belief functions using mathematical morphology – Application to image fusion under imprecision. *International Journal of Approximate Reasoning*, **48**(2):437–465, June 2008.

[8] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(3):257–267, March 2001.

[9] François Brémond, Monique Thonnat, and Marcos Zúñiga. Video-understanding framework for automatic behavior recognition. *Behavior Research Methods*, **38**(3):416–426, August 2006.

[10] Gabriel J Brostow, Irfan Essa, Drew Steedly, and Vivek Kwatra. Novel skeletal representation for articulated creatures. In *In Proc. European Conf. on Computer Vision*, pages 66–78, 2004.

[11] François Caron, Branko Ristic, Emmanuel Duflos, and Philippe Vanheeghe. Least committed basic belief density induced by a multivariate Gaussian: Formulation with applications. *International Journal of Approximate Reasoning*, **48**(2):419–436, June 2008.

[12] Andrea Cavallaro. AVSS 2007. In *2007 IEEE International Conference on Advanced Video and Signal based Surveillance*, 2007.

[13] GKM Cheung and T Kanade. A real time system for robust 3D voxel reconstruction of human motions. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, pages 714 – 720, 2000.

[14] CH Chien. Volume/surface octrees for the representation of three-dimensional objects. *Computer Vision, Graphics, and Image*, **36**(1):100–113, 1986.

[15] Chi-Wei Chun, O.C. Jenkins, and M.J. Mataric. Markerless kinematic model and motion capture from volume sequences. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, **2**, pages 475–482. IEEE Comput. Soc, 2003.

[16] Thierry Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, **42**(3):228–252, August 2006.

[17] Jonathan Deutscher and Ian Reid. Articulated Body Motion Capture by Stochastic Search. *International Journal of Computer Vision*, **61**(2):185–205, February 2005.

[18] Luis Díaz-Más, Francisco José Madrid-Cuevas, Rafael Muñoz Salinas, Angel Carmona-Poyato, and Rafael Medina-Carnicer. An octree-based method for shape from inconsistent silhouettes. *Pattern Recognition*, **45**:3245–3255, March 2012.

[19] Luis Díaz-Más, Rafael Muñoz Salinas, F.J. Madrid-Cuevas, and Rafael Medina-Carnicer. Shape from silhouette using Dempster–Shafer theory. *Pattern Recognition*, **43**(6):2119–2131, June 2010.

[20] Luis Díaz-Más, Rafael Muñoz Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Three-dimensional action recognition using volume integrals. *Pattern Analysis and Applications*, pages 1–10, September 2011.

[21] Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric Model for Background Subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 751–767, London, UK, 2000. Springer-Verlag.

[22] J.-S. Franco and E. Boyer. Fusion of multiview silhouette cues using a space occupancy grid. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1747–1753 Vol. 2. IEEE, 2005.

[23] Irene Gargantini. Linear Octtrees for Fast Processing of Three-Dimensional Objects. *Computer Graphics and Image Processing*, **20**(4):365–374, 1982.

[24] DM Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, **73**(1):82–98, January 1999.

[25] Rafael C. González and Richard Eugene Woods. *Digital image processing*. Prentice Hall, third edition, 2008.

[26] Thanarat Horprasert, David Harwood, and L. S. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *Proc. IEEE ICCV*, pages 1–19, 1999.

[27] C JACKINS and S TANIMOTO. Oct-trees and their use in representing three-dimensional objects. *Computer Graphics and Image Processing*, **14**(3):249–270, November 1980.

[28] Mustafa Karaman. Comparison of static background segmentation methods. In *Proceedings of SPIE*, **5960**, pages 596069–596069–12. SPIE, 2005.

[29] Aaron Knoll. A Survey of Octree Volume Rendering Methods. 2008.

[30] N. Krahnstoever and R. Sharma. Articulated models from video. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **1**, pages 894–901. IEEE, 2004.

[31] N. Krahnstoever, M. Yeasin, and R. Sharma. Automatic acquisition and initialization of articulated models. *Machine Vision and Applications*, **14**(4):218–228, September 2003.

[32] José Landabaso and Montse Pardàs. Foreground Regions Extraction and Characterization Towards Real-Time Object Tracking. In *Machine Learning for Multimodal Interaction*, pages 241–249, 2006.

[33] Jose-Luis Landabaso, Montse Pardàs, and Josep Ramon Casas. Shape from inconsistent silhouette. *Computer Vision and Image Understanding*, **112**(2):210–224, November 2008.

[34] A. Laurentini. The visual hull: A new tool for contour-based image understanding. In *Proc. 7th Scandinavian Conf. Image Analysi*, pages 993–1002, 1991.

[35] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(2):150–162, 1994.

[36] M.K. Leung. First Sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(4):359–377, April 1995.

[37] Haojie Li, Si Wu, Shan Ba, Shouxun Lin, and Yongdong Zhang. Automatic detection and recognition of athlete actions in diving video. In *Proceedings of the 13th International conference on Multimedia Modeling - Volume Part II*, MMM'07, pages 73–82, Berlin, Heidelberg, 2006. Springer-Verlag.

[38] Ying Luo, Tzong-Der Wu, and Jenq-Neng Hwang. Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks. *Computer Vision and Image Understanding*, **92**(2-3):196–216, November 2003.

[39] Microsoft. Kinect, 2011.

[40] TB Moeslund and Adrian Hilton. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image*, **104**(2-3):90–126, November 2006.

[41] Rafael Muñoz Salinas, Enrique Yeguas-Bolivar, Luis Díaz-Más, and Rafael Medina-Carnicer. Shape from pairwise silhouettes for plan-view map generation. *Image and Vision Computing*, **In press**, February 2012.

[42] Jan Neumann, Cornelia Fermüller, and Yiannis Aloimonos. Animated heads: from 3d motion fields to action descriptions. In *DEFORM '00/AVATARS '00: Proceedings of the IFIP TC5/WG5.10 DEFORM'2000 Workshop and AVATARS'2000 Workshop on Deformable Avatars*, pages 1–11, 2001.

[43] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **2**, pages 16–22. IEEE, 2004.

[44] Ramprasad Polana and Randal C. Nelson. Detection and Recognition of Periodic, Nonrigid Motion. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, **23**(3):261 – 282, 1997.

[45] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[46] Alessandro Saffiotti, M. Broxvall, M. Gritti, K. LeBlanc, R. Lundh, J. Rashid, B.S. Seo, and Y.J. Cho. The PEIS-Ecology project: Vision and results. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2329–2335, Nice, France, September 2008. IEEE.

[47] Cristian Sminchisescu and Bill Triggs. Estimating Articulated Human Motion With Covariance Scaled Sampling. *International Journal of Robotics Research*, **22**:2003, 2003.

[48] Dan Snow and Paul Viola. Exact voxel occupancy with graph cuts. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, pages 345 – 352, 2000.

[49] Yang Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(7):814–827, July 2003.

[50] SONY. Cámara USB para EyeToy, 2003.

[51] Camillo J. Taylor. Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. *Computer Vision and Image Understanding*, **80**(3):349–363, December 2000.

[52] Graham W. Taylor, Ian Spiro, Christoph Bregler, and Rob Fergus. Learning invariance through imitation. In *CVPR 2011*, pages 2729–2736. IEEE, June 2011.

[53] N. Vaswani, A.K. Roy-Chowdhury, and R. Chellappa. "Shape Activity": a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection. *IEEE Transactions on Image Processing*, **14**(10):1603–1616, October 2005.

[54] S. Wachter and H.-H. Nagel. Tracking Persons in Monocular Image Sequences. *Computer Vision and Image Understanding*, **74**(3):174–192, June 1999.

[55] Juyang Weng and Narendra Ahuja. Octrees of objects in arbitrary motion: Representation and efficiency. *Computer Vision, Graphics, and Image Processing*, **39**(2):167–185, August 1987.

[56] M. Yamamoto and K. Yagishita. Scene constraints-aided tracking of human body. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, **1**, pages 151–156. IEEE Comput. Soc, 2000.

[57] Wei Yan and D A Forsyth. Learning the Behavior of Users in a Public Space through Video Tracking. Technical Report UCB/CSD-04-1310, EECS Department, University of California, Berkeley, 2004.

Impact report

This report show the impact factors and positions of the journal in which the works presented in the thesis have been published.

- Luis Díaz-Más, Rafael Muñoz Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Three-dimensional action recognition using volume integrals. *Pattern Analysis and Applications*, pages 1–10, September 2011
  Journal: Pattern Analysis and Applications
  Impact factor: JCR 2010: 1.097
  Position of journal: 65 of 108
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE

- Luis Díaz-Más, Rafael Muñoz Salinas, F.J. Madrid-Cuevas, and Rafael Medina-Carnicer. Shape from silhouette using Dempster–Shafer theory. *Pattern Recognition*, **43**(6):2119–2131, June 2010 Journal: Pattern Recognition
  Impact factor: JCR 2010: 2.682
  Position of journal: 15 of 108
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE

- Luis Díaz-Más, Francisco José Madrid-Cuevas, Rafael Muñoz Salinas, Angel Carmona-Poyato, and Rafael Medina-Carnicer. An octree-based method for shape from inconsistent silhouettes. *Pattern Recognition*, **45**:3245–3255, March 2012 Journal: Pattern Recognition
  Impact factor: JCR 2010: 2.682
  Position of journal: 15 of 108
  Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE