Gómez, M. J., García, F., Martín, D., de la Escalera, A. & Armingol, J. M. (2015). Intelligent surveillance of indoor environments based on computer vision and 3D point cloud fusion. Expert Systems with Applications, 42(21), 8156-8171.

DOI: 10.1016/j.eswa.2015.06.026

Intelligent surveillance of indoor environments based on computer vision and 3D point cloud fusion

María José Gómez, Fernando García , David Martín, Arturo de la Escalera, José María Armingol

Intelligent Systems Lab, University Carlos III de Madrid, Leganés, Spain

Abstract

A real-time detection algorithm for intelligent surveillance is presented. The system, based on 3D change detection with respect to a complex scene model, allows intruder monitoring and detection of added and missing objects, under different illumination conditions.

The proposed system has two independent stages. First, a mapping application provides an accurate 3D wide model of the scene, using a view registration approach. This registration is based on computer vision and 3D point cloud. Fusion of visual features with 3D descriptors is used in order to identify corresponding points in two consecutive views. The matching of these two views is first estimated by a pre-alignment stage, based on the tilt movement of the sensor, later they are accurately aligned by an Iterative Closest Point variant (Levenberg–Marquardt ICP), which performance has been improved by a previous filter based on geometrical assumptions.

The second stage provides accurate intruder and object detection by means of a 3D change detection approach, based on Octree volumetric representation, followed by a clusters analysis. The whole scene is continuously scanned, and every captured is compared with the corresponding part of the wide model thanks to the previous analysis of the sensor movement parameters. With this purpose a tilt-axis calibration method has been developed.

Tests performed show the reliable performance of the system under real conditions and the improvements provided by each stage independently. Moreover, the main goal of this application has been enhanced, for reliable intruder detection by the tilting of the sensors using its built-in motor to increase the size of the monitored area.

Keywords: Intelligent Surveillance System, Sensor fusion, Computer vision, 3D point cloud, Intruder detection

1. Introduction

In the latest years, the presence of surveillance systems, in outdoor or indoor environments and in public or private places, has been continuously increasing. Surveillance systems are a powerful tool to identify not authorised individuals and activities, therefore the surveillance systems demand is rising both in quality and quantity, in order to guarantee the citizens' security.

The traditional surveillance systems (first generation) basically consist of video recording and the subsequent analysis of the events, or the monitoring over a live video carried out by human operators (CCTV, Closed-Circuit Television). The necessity of human operators in these systems presents disadvantages, like high manpower costs, operators' limited capacity to monitor a

certain number of screens, or their lack of attention after sev-eral work hours. The necessity to manage, in a more effective and profitable way, all the information captured by the sensors has boosted the development of automatic surveillance systems provided with a certain degree of artificial intelligence. Intelligent Surveillance Systems (ISS) (Huihuan, Xinyu, & Yangsheng, 2011) monitor a specific environment in real time, providing an automatic interpretation of the scene and predict-ing the individuals' actions and interactions, basing on the data acquired by sensors. These systems can supplement (second generation) (Fookes et al., 2010) or even replace (Adam, Rivlin, Shimshoni, & Reinitz, 2008) (third generation) traditional systems.

Intelligent surveillance is based in three main stages: detection (Lim, Tang, & Chan, 2014; Luo & Xia, 2014), tracking (Oliveira, Noguez, Costa, Barbosa, & Prado, 2013), and behaviour analysis (Albusac, Vallejo, Castro-Schez, Glez-Morcillo, & Jiménez, 2014). The first stage deals with the low level detection of the interesting elements in the captured images. Therefore, the quality of the whole system relies on the performance of this stage. It is in the stage where the work presented on this paper focus.

One of the most used detection methods is the background subtraction, which is based on the intensity differences between background and foreground, where usually are located the interesting elements (Davis & Sharma, 2004). Although statistical techniques can be used to update changes in the background and to adapt these algorithms to the illumination conditions (Huwer & Niemann, 2000), the performance of these methods are still limited by the necessity of a hardly variable illumination in the scene. Other important group of detection algorithms are the movement-based ones, which compare consecutives captures of the scene, using 2D or 3D data. One of the most well-known algorithms is Optical Flow (Galvin, McCane, Novins, Mason, & y Mills, 1998; Horn & Schunck, 1980), which works with 2D data.

By means of a movement-based analysis, it is possible to obtain instantaneous information of the changes in a room, but this does not provide a final knowledge of the absolute changes, like objects which have been removed or added. In such cases, when the super-vised scene is a room without people transiting and the objective of monitoring it is to guarantee this situation, it is required the comparison between the real and ideal situation in the scene. The ideal situation is known and static, therefore, an effective approach is to model this in a first stage, and subsequently comparing the model with the real situation in a real-time surveillance process.

The changes detection with respect to an ideal model from 2D images (Radke, Andra, Al-Kofahi, & Roysam, 2005) is a deeply studied field with a number of applications in several disciplines, including surveillance, medical diagnosis (Kuthirummal, Bansal, Sawhney, and Eledath (2011)), buildings evaluation (Irani & Anandan, 1998), or even climate change supervision (Lu, Mausel, Brondízio, & Moran, 2004). Currently, algorithms are been designed to detect changes with respect to a 3D model (Fischer & Beyerer, 2012). Among these methods voxel-based ones (Gill, Keller, Anderson, & Luke, 2011) highlight since they allow to obtain three-dimensional knowledge of both the real scene and its model, making easier the comparison between them. Intelligent Visual Surveillance, IVS (Valera & Velastin, 2005), which utilises 2D images, presents strong constraints due to the presence of shadows and occlusions in the video sequences although they are enhanced by the use of SVM (Support Vector Machine) models (Kim, Lee, & Kim, 2014). In contrast to this, the development of 3D data acquisition devices based on 3D measurement technologies, like Time of Flight (TOF), triangulation, phase difference, Conoscopic Holography, or stereoscopic systems, allow the detection of multiple objects independently of their movement or illumination conditions.

Other alternative to improve the detection is the sensor fusion (García et al., 2013). The incorporation of 3D laser scanners or stereoscopic systems in surveillance systems leads to a high increase of their costs. Nowadays, the use of 2.5D sensors in surveillance applications is a continuously increasing phenomenon (Lee & Chung, 2006; Park, Lee, & Chung, 2006). This type of devices, such as Prime Sense's ones, measures the depth (distance between objects and sensor) with Light Coding Technology. A near-infrared light pattern is emitted by the device and then distorted by the object where it touches. A CMOS sensor catches the modified pattern and calculates the depth data with triangulation algorithms. Microsoft Kinect is a commercial device that uses this technology and allows the fusion of depth and colour data. So far, the main research in Kinect sensor applications has been the real-time 3D reconstruction of dynamic scenes (Izadi et al., 2011; Keller et al., 2013). However for the surveillance task, the objective is not to integrate the changes in the model, but detect them. Kinect sensor has boosted development of low-cost indoor surveillance systems, with applications such as human recognition (Gill et al., 2011; Savage, Clarke, & Li, 2013) and model (Tang, Luo, Tjahjadi, & Gao, 2014), movement (Song & Zhang, 2013) and behaviour analysis (Popa, Koc, Rothkrantz, Shan, & Wiggers, 2012), mapping (Hu, Hu, Wang, Gong, & Duan, 2013), and mobile robots' control (Santos et al., 2012).

3D data are not only useful for the surveillance task, but also for the previous modelling of the scene. The complete 3D model of a wide area can be created by integrating different captures of 2D or 3D data. The 3D reconstruction methods from 2D images are utilised generally to model wide outdoor zones (Müller, Smolic, Drose, Voigt, & Wiegand, 2005), where the 3D sensors technology not always works properly. One approach consists of multiple cameras placed around the area (Mishra, Ni, Winkler, & Kassim, 2007). Other solution it is the integration of several images captured by aerial vehicles using Structure from Motion (SFM) techniques. Stereoscopic systems can also be used to model the outdoor scenes, (Harville, 2004).

The state-of-the-art 3D sensors provide in each of their captures a set of 3D points that render the geometry of a part of the scene. The point clouds registration methods integrate 3D data obtained from different points of view in only one model, defining all the points regards to the same coordinate system of reference. For that purpose, semi-automatic methods (Krishnan, Saripalli, Nissen, & Arrowsmith, 2012) and even commercial programs, such as Rapidform XOR, have been designed. These methods need the user's collaboration to identify pairs of corresponding points (points that represent the same real point in two different point clouds), and then the transformation needed to make them concur in the space is automatically estimated. The automation of the search of pairs of corresponding points is one of the main research lines, where algorithms robust to noise (Myronenko & Song, 2010) and increasingly faster are sought, such as the ones based on Coarse Binary Cubes method (CBC) (Martinez, Reina, Morales, & Mandow, 2013).

In this paper it is proposes a low cost surveillance system with capacity to detect not only the intruders in a room, but also the extracted or added objects in real time. This system is based on a novel approach that detects instantaneous changes respect to a 3D model of the room independently of the illumination conditions.

This system only needs one commercial device, Microsoft Kinect, without the necessity of being moved around the scene, as it is done in Newcombe et al. (2011), to perform the surveillance task, since the size of the monitored room has significantly been increased with the using of the tilt motor built into the device. In addition, the work distance has been duplicated with respect

to others approaches (Izadi et al., 2011) that use this type of sensor, since the designed method is robust enough to deal successfully with the lack of precision in larger distances.

A scanning method based on a deep study of the built-in tilt movement of the device allows the comparison of each capture with the correct part of a larger 3D model. The analysis of the scene is carried out by a change detector and a cluster-based stage, used to catch only the real intruders and modified objects in the scene. The algorithm was based on octree structures, instead of the classical voxel grid method (Niessner, Zollhöfer, Izadi, & Stamminger, 2013). The information retrieved, can be subsequently used for advanced surveillance system in order to provide a higher level identification.

In addition, a novel modelling algorithm has been designed, dividing the registration process in two steps, the first is based on proprioceptive data from the device movement. The second uses a well-known algorithm, ICP (Fitzgibbon, 2003) improved by a previous filter, which maintains the common areas of two consecutive captures. The complex task of choosing pairs of corresponding points from two different captures of the scene has been solved using the fusion of geometric and colour data, allowing the registration of not only well-detailed surfaces (Newcombe et al., 2011), but also challenging flat surfaces. In that way, the pro-posed registration method works properly with inaccurate 3D data, allowing the use of a commercial inexpensive device from a distant position, instead of using it in close range (near mode) or requiring expensive 3D sensors, such as TOF devices (Keller et al., 2013) or structured-light rangefinders to model small objects (Niessner et al., 2013). An efficient management of 3D data is achieved by structuring the acquired point clouds in integral images, which give knowledge about the position of every points respect their neighbours.

In order to build a non-supervised system, the position of the sensor was defined to remain fixed, thus the mapping of the room is does not require a complex offline process, or manual movement of the sensor. Instead, it is done by a completely automatic stage.

Tests have been performed over to this detection system for different sensor locations and varying lighting conditions with successful results.

The paper is divided as follows, Section 2 presents a general overview of the approach. Section 3 presents the mapping algorithm. Scene analysis method is explained in Section 4. Section 5 provides the results of the tests performed to the algorithms, and finally in Section 6 conclusion and future works are detailed.

2. Overview of the intelligent surveillance of indoor environments method

The proposed indoor surveillance system analyses 3D data acquired from different orientations in a room, using Microsoft Kinect for Windows device. To carry out this task three modules have been designed for acquisition, mapping, and analysis (Fig. 1).

The analysis module performs the surveillance process by detecting strange elements (e.g. intruders, new and missed objects) in a monitored room. It acquires 3D point cloud datasets, in real time, and compares them with a 3D model of the scene, identifying the unexpected shapes. These point cloud datasets are based only on geometric information, since colour information is irrelevant for this task, this way, surveillance can be performed in absence of light.

The 3D scene model is created in an automatic process, prior to surveillance, thanks to the mapping module, which, unlike analysis module, needs illumination in the room in order to create the model. The mapping module does the registration of several colour point clouds taken from different sensor orientations, providing a wide 3D model or map, in a previous stage to the surveillance process.

The data acquisition module acquires colour and depth images and controls the tilt motor built-into the Kinect device. In order to obtain depth data, Kinect sensor uses an infrared light pattern, this does not work properly with the presence of sunlight, and consequently the system will be used at indoor scenes. However, the algorithm is suitable for other sensing devices, which provide colour and 3D data, and may be used in outdoor scenarios. The 3D point cloud data are calculated in two steps (Fig. 2):

- Matching of colour values with their corresponding depth val-ues based on the known calibration (provided by Kinect device) between colour and depth images, which are captured in con-secutive instants.
- 3D coordinates estimation from depth data using equations (1, 2, 3) based on the Pin-Hole model.

$$z = d \quad (1)$$

$$x = z\,K\,(u - w/2) \quad (2)$$

$$y = z\,K\,(h/2 - v) \quad (3)$$

where x; y; z are 3D coordinates of a point in the mobile coordinates system placed in the depth sensor (red in Fig. 3), d is the depth value of the pixel corresponding to the query point, u and v are the row and the column of this pixel in the depth image, and w and h are its total number of columns and rows, respectively. K is a constant that relates the focal distant value, f, and the size of the square pixels, s, as follow (4).

$$K = s/f \quad (4)$$

3. Complete mapping of the scene

The map must represent the entire region that the sensor is cap-able to monitor. The proposed mapping method is iterative; in each iteration a colour point cloud dataset is acquired from a dif-ferent sensor tilt angle and is registered in a global framework, until create the completed 3D model or map.

The proposed algorithm has two branches working in parallel (Fig. 3): on the one hand, a view acquired from a certain tilt angle (αi) is processed to be added to the map while, on the other hand, the motor is moved to the next acquisition angle ($\alpha i+1$).

Every point, Pi =(xi; yi; zi), belonging to a dataset acquired from a αi tilt angle, is referenced to a local coordinates system with the origin placed in the depth sensor. The position and orientation of this system change with every tilt angle value. Therefore, it is needed a fixed global coordinates system, where all the acquired points are referenced. The system chosen as the global one is the local coordinate system when tilt angle is zero.

The registration is the process responsible for estimating and applying a matrix transformation, Tiw, over each point Pi, in order to calculate the value of its coordinates in reference to the global system, Pw=(xw; yw; zw) (5).

Pw = Pi Tiw (5)

Due to the tilting movement of the sensor it can be assumed that global coordinates Pw=(xw; yw; zw) are the result of the point Pi rotation around the tilt motor axis, parallel to x axis with coordinates C(yc ; zc), a ai tilt angle. This rotation around a centre C, which does not corresponds with the origin of coordinates provided by the sensor (Fig. 4), can be rendered by the transformation matrix Tiw (6). However, C is unknown, and consequently Tiw cannot be completely calculated, so registration process is divided in two steps, pre-alignment and fine-alignment. Tiw has a rotation part only dependent on the tilt angle that is known, provided by the device, with one degree of precision. Therefore a transformation matrix Rii (9), which contains only this rotation part can be calculated and applied for every point (7). This process is called pre-alignment. Due to the low precision of the angle value, the result is an approximate rotation of the points, which require a fine adjustment of the rotated angle and the application of the translation part of Tiw. However this translation cannot becalculatedd, since it depends also on the unknown position C, but it can be estimated using an ICP-based alignment. This alignment provides the transformation matrix, Tiw0 , whose application (8) results in the points translation and a fine adjustment of their rotation to get an accurate registration. Therefore this second stage is called fine-alignment.

$$T_w^i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\alpha_i) & sin(\alpha_i) & y_c - y_c\cos(\alpha_i) - z_c\sin(\alpha_i) \\ 0 & -sin(\alpha_i) & cos(\alpha_i) & z_c + y_c\sin(\alpha_i) - z_c\cos(\alpha_i) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

$$P_{i'} = P_i \cdot R_{i'}^i \quad (7)$$

$$P_w = P_{i'} \cdot t_w^{i'} \quad (8)$$

Every acquired point cloud dataset is structured in an integral image with 640 480 points. An integral image is a point cloud organised in a structure where the data is split in rows and columns, so that the position of each point in the structure has relationship with its position in the scene from a certain point view. In this way the relationship between adjacent points is known.

Before the registration process, all those points must pass through a filtering stage, in order to improve the registration and its computational time. In this stage, first a voxel grid filter down-samples the point cloud dataset; second, a depth-based filter removes distant points due to their lack of accuracy; and finally, a statistical analysis trims from the dataset the points considered as outliers, mainly caused by measurement errors due to presence of strong illumination sources or reflective surfaces. The down-sampling is performed by a voxel grid approach, dividing the space in cubic voxels of 0.016 m and keeping one point in each voxel. Inaccurate points are eliminated with a depth threshold which maintains points nearer than 8 m. Moreover the outliers are removed by means of a statically-based method that calculates for every point, its mean distance to its 50 nearest neighbours. Later, the global mean and typical deviation of these values are obtained. If a point presents a distance from the mean value higher than three times the typical deviation value, this point is removed, such as depicted in Figs. 5 and 6.

Once a capture is filtered, the problem of registering is solved in the next step thanks to the algorithm shown in the diagram below (Fig. 7).

The pre-alignment is accomplished by rotating (7) every point,Pi (xi; yi; zi), over the x axis of the local coordinates system, from which the point cloud dataset has been acquired, using the rotation matrix Rii (9). The rotation angle has the value of the sensor tilt angle during acquisition, αi.

$$R_{i'}^{i} = R_X(\alpha_i) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\alpha_i) & sin(\alpha_i) & 0 \\ 0 & -sin(\alpha_i) & cos(\alpha_i) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

The next step is a more accurate and complicated alignment using the translation matrix (8). The matrix estimation method is based on the search of the correspondences between two given input point clouds. Points common in two different captures (both of them represent the same point, Q, in the scene) have similar values for their coordinates in the global system. So a point cloud registered in the global system is used as reference (specifically the cloud registered in the previous iteration of mapping). In this way, after the alignment, the coordinates of points, Qi, of a cloud captured from any tilt angle must have the value of the coordinates of its corresponding points, Qw, in the reference cloud. When these correspondences are perfectly known it is possible to find the trans-formation, tiw, that minimizes the distances between corresponding points.

To reduce the complexity of correspondences search, a subset is selected in the acquired and reference point cloud datasets. In both of them, the points selected are the key points, i.e. the most representative ones in the scene, independently of the capturing point of view. Key points present a special property in comparison to its neighbour points. Visual features can be detected more easily and faster than 3D features. In addition, the flat geometry of the scene makes difficult the search of 3D significant features, therefore, once the stability and reliability of the 2D features have been proved in contrast to that of the 3D ones, visual information is analysed

in the colour image to get the key points. Subsequently, the 3D coordinates of all the key points extracted are searched in the point cloud dataset. The key points search algorithm used is based on the FAST (Features from Accelerated Segment Test) algorithm (Tuytelaars & Mikolajczyk, 2007), which looks for pixels representing the object borders. If the difference between the intensity value of a pixel and that of its neighbour pixels exceeds a threshold value T, the query pixel is considered as a key point. An increment in the T value implicates a more strict condition for a point to be considered a key point, therefore the number N of extracted key points decreases.

In our extraction of key points there is an integration of two types of data: colour data from the scene image, where the inter-esting pixels are selected, and tri-dimensional data from the point cloud, from which the coordinates corresponding to each key pixel are extracted. Moreover, the matching stage is eased if each key point is identified by a certain 3D feature descriptor. This descriptor must have similar values for corresponding points, even if each one of them belongs to a different scene capture. The descriptors used are the normal vectors and the curvature change for each key point in the surface they belong to. Although the descriptor value is calculated only for key points, this estimation needs to take account their neighbour points. This task is speeded up thanks to the fact that the point cloud acquired has an integral image structure.

After the above stages, there are two point cloud subsets representing the acquired and the reference view. Each one of them is formed only by key points and they are represented by its three spatial coordinates and its descriptors values. In the next stage, a first matching is done, it is correspondences estimation; second, bad correspondences are rejected; and finally, the transformation matrix that minimizes the distance between the remained couples of corresponding points is calculated. These three steps are executed in a fast and efficient variant of the ICP (iterative closest point) algorithm, which is called LM-ICP (Levenberg–Marquardt ICP) (Fitzgibbon, 2003) until get a successful alignment of the acquired point cloud subset (source) with the reference point cloud subset (target).

In order to improve the ICP algorithm performance and its results, has been designed a filter that extracts the common region rendered in both datasets, where it is possible to find correspondences between their points (Fig. 8).

The upper and lower limits of the field of view are defined by tilted planes depending on the sensor tilt angle, whose projection in the plane YZ are straight lines.

The points selected by this filter are in the intersection of the sensor field of view in the acquisition of both, source and target point cloud. These points are the input to ICP algorithm (Fig. 9), which finally returns the transformation matrix needed to align the source point cloud with target point cloud. This resultant point cloud, source point cloud aligned, will be the target point cloud in the next mapping iteration.

The transformation matrix obtained is applied over the input point cloud dataset of the alignment process, it means that all acquired points must be transformed, not only the key points selected to search correspondences. With this matrix multiplication, the coordinates in the global coordinates system are calculated for every point, and the registration process is completed.

To get the 3D scene model all the points registered in the global coordinates system are saved in a single point cloud dataset, in a final step of each mapping iteration, called concatenation. In

each iteration, the model size increases due to the addition of new registered points, until the entire 3D scene model is obtained.

4. Intruder detection process based on scene analysis

The analysis module captures the current appearance of a room and compares it with the scene rendered in the 3D map of this room. Each captured view is represented as a 3D point cloud, like the 3D map of the room; therefore, any intruder, any moving object, or even any new or missed static element in the scene, is detected in real time by the detected changes between both point clouds. The Kinect tilt angle is modified to acquire different room areas (all of them rendered in the map) to increase the monitored region (Fig. 10).

The 3D map or model of the scene is created by the mapping module, in a previous step to the surveillance process, out of the analysis cycle. In both processes, mapping and surveillance, the Kinect device is placed in the same position.

The analysis of each acquired view is based on the comparison of its geometry and that of the map, it means, 3D coordinates values of the points of both clouds, here colour data is not needed. Hence surveillance is possible in absence of light in the monitored room. In a pre-processing step, executed previously to the surveil-lance, the 3D map is down-sampled and its colour information is removed. The colour information is consequence of the mapping process, which uses colour data in some of their stages, however at this stage this information is not longer necessary. The point cloud acquired in each surveillance iteration is filtered in a pre-processing step, where measurement errors and noise are removed, and the cloud is also down-sampled to decrease its number of points in order to allow real time processing. The down-sampling of the map and the captures of the scene are performed by a 0.03 m-voxel grid approach (like in the mapping process).

During the surveillance, point clouds are acquired from different sensor tilt angles, as occurred during the mapping process. Accordingly, the 3D coordinates of each point cloud are referenced to a mobile coordinates system placed in the depth sensor, whose position and orientation change in function of the tilt angle. For each iteration, the points of the captured cloud must be referenced to the fixed coordinate system of the map to allow the comparison of both point clouds (the 3D map and each acquired cloud) in a common system.

The registration algorithm designed in mapping module, which is based on corresponding points search between two point clouds, is not applicable in this case because of the following reasons:

- The map has a significantly larger size than each acquired view; this complicates the search of correspondences between them, since the algorithm does not know the map zone where it must search.
- The acquired views can present changes compared to the scene map, which make impossible to find the expected correspondences.
- This registration method requires a high computation time which does not allow working in real time.

Therefore, each acquired point cloud must be registered with the 3D map using a different method. As already mentioned, the 3D coordinates of each point in the fixed global system, PW (xw; yw; zw), are the result of rotating the point Pi(xi; yi ; zi), formed by its coordinates in the mobile local system, around the motor axis, parallel to x axis and placed in C(yc ; zc), a ai tilt angle. So it is possible to apply a transformation matrix, Tiw (6), that describes this rotation, like Eq. (5). The values of the elements of this matrix depend on the position of the rotation motor axis, C(yc ; zc), which is constant, and the tilt angle value in the moment of acquisition, ai, which is variable but known.

The coordinates of the motor axis in the global system, C(yc ; zc), are unknown, so an estimation method has been designed to calcu-late them. The following equations (10) and (11) allow to obtain these coordinates from the position (in reference to the global system) of any local coordinates system origin Oi(Oxi ; Oyi; Ozi) , corresponding to a certain tilt angle ai (see first diagram in Fig. 4).

$$y_c = \frac{Oy_i}{2} + \left(\frac{Oz_i \cdot sin\alpha_i}{2(1 - cos\alpha_i)} \right) \qquad (10)$$

$$z_c = Oz_i - \frac{y_c \cdot sin\alpha_i}{1 - cos\alpha_i} \qquad (11)$$

So first it is necessary to estimate the value of the origin coordinates Oi. The registration method designed in the mapping module is used with that purpose. The final alignment step of this method returns the translation vector tiw0 , which must be applied to register a point cloud acquired from a ai tilt angle in the global coordinates system (see third diagram in Fig. 4). Therefore, this vector tiw0 describes the position of the origin, Oi, corresponding to the local coordinates sys-tem in a certain tilt angle ai, in reference to the global system (12).

Oi(Oxi; Oyi; Ozi) = twi (tx; ty; tz) (12)

Since this translation vector is obtained with a not fully accurate registration method, the registration process is repeated for different tilt angle values over a special surface with a checker-board pattern (Fig. 11). From each translation vectors obtained, the values of the rotation axis coordinates C(cy; cz) are calculated with Eqs. (10) and (11), and finally a statistical analysis is carried out over all this values to select the most accurate ones. The mean of the selected values for the coordinates y and z of the rotation axis are chosen as its real estimated position (Fig. 12).

Once the rotation axis position is known the value all the elements of the transformation matrix, Tiw, can be expressed as functions of the tilt angle from which the point cloud to align was acquired, so the applying of this transformation matrix makes possible to register the points of a cloud captured from whatever tilt angle, in the global coordinates system.

When an acquired view is registered in the same coordinates system than the map (global system), it is possible the comparison between both point clouds in the change detection step. The changes in the shape of the view, in comparison to the map, are identified as the presence of an intruder, or any other modified element in the scene.

However, two clouds representing the same area can present different distributions of their points, it means, they may vary in size, resolution, density and point ordering, so it is not possible a direct comparison of the coordinates of their points. A volumetric representation of the point

clouds is necessary to get knowledge about the space rendered in them. This way, change detection for each view consists in identifying of areas which are rendered in the view but not in the map, and vice versa.

The volumetric representation of a point cloud is obtained by structuring it in an octree. The resolution parameter describes the length of the smaller voxels, in which the space is divided. The space bounded by the smallest voxels containing points of a cloud is the spatial representation of this cloud.

The shape differences between two point clouds structured in octrees are identified as the areas where there are voxels from only one of them. Therefore, the spatial changes in a scene are shown by the points which are stored at voxels of the octree structure of an acquired view which do not exist in the octree structure of the map, and vice versa.

Resolution parameter carries weight in two features of the change detection, sensibility and execution time (Fig. 13), since both of them increase with a decreasing size of voxels. The chosen value for voxel length is ten centimetres, which allows quite accurate people detection and respecting time constraints.

Ideally only the spatial changes caused by a strange element or an intruder in the scene should be detected. However, because of the noise in the acquisition and the alignment errors (caused by the poor precision of the tilt angle measurement, provided by the device), some points in the acquired clouds are detected improperly as intruders. These last points have a high dispersed spatial distribution. However, the points rendering an intruder are very close to each other forming a large size group (Fig. 14).

To analyse this properties in the detected changes, a segmentation process is carried out over the point cloud, P, which contains all the points, either well or bad detected. The result is the creation of several groups of points called clusters, Cj.

A point pi belongs to a certain cluster if the distance, dcpjk from this point to whatever point of this cluster, cjk, is smaller than the threshold distance, dth, (13). All the points representing the same object or person are grouped in the same cluster using a proper threshold distance. The points caused by errors or noise remain isolated, or forms small clusters.

The size (number of points) of each cluster, S(Cj), is compared with a constant value, Nth, which describes the minimum sized required to consider the cluster as an interesting point cloud, Q l, (14). Each one of these interesting point clouds, Q l, renderer an intruder or strange element detected in the scene.

$$\exists c_{jk} \in C_j | d_{p_i}^{c_{jk}} < d_{th} \rightarrow p_i \in C_j \qquad (13)$$

$$S(C_j) > N_{th} \rightarrow C_j := Q_i \qquad (14)$$

5. Results

This Section presents the results obtained after tests performed to the presented surveillance system as well as to several of its internal stages.

The registration algorithm designed in the mapping module is not only used in the map creation, but also in the estimation method of the centre of sensor rotation, which is a data needed for the registration of each captured view during surveillance. Therefore, the good performance of this registration algorithm determines the quality of the map, and also the efficiency of the analysis module in an indirect way. This registration is a complex process formed by several stages. Here the main stages, which determine the behaviour of the whole algorithm, are analysed using the comparison between the employed methods and others alternative solutions, and the checking of its results.

Finally, the surveillance system functionality is tested in differ-ent real scenes and conditions. In those tests, a serial of parameters are evaluated. The values obtained allow analysing the reliability of both, the map and of the intruders' detection in the scene.

5.1. Results of the main registration stages

The stages analysed in this section, i.e. ''extraction of key points'', ''descriptors calculation'' and ''ICP algorithm application'', belong to the fine alignment process (Fig. 5) on which the final precision of the registration depends.

5.1.1. Extraction of key points

There are two types of methods to extract key points depending on the property evaluated over every point of the whole set.

5.1.1.1. Methods to extract key points based on geometric features evaluated over three-dimensional data. In these methods, a point is a key point or not in function of its position in the space with regard to its neighbour points. The algorithm tested is NARF (Normal Aligned Radial Feature) (Steder, Rusu, Konolige, & Burgard, 2010), which is based on the search of points placed in the borders of uniform surfaces. This method extracts key points from a range image, with a certain angular resolution, R. A range image is a 2D image, where pixel values correspond to the distance to points in a scene from a specific point (depth sensor). The NARF algorithm has been tested with six different values of R. The Fig. 15 shows the results obtained in every case. As the R value increases, the number of extracted key points N decreases. Due to the low precision in the coordinate values at distant points, it is difficult to find uniform surfaces, which are needed for the proper performance of NARF algorithm, at longer distances from the sensor.

5.1.1.2. Methods to extract key points based on visual features evaluated over an image, it means, over bi-dimensional data. In these methods, a pixel from an image is interesting or not in function of its intensity. The method tested is FAST, whose operation is explained in the Section 5, due to it is one of the fastest extractors of key points. This algorithm has been tested with six different values of the threshold parameter, T, like showed in Fig. 16.

5.1.1.3. Comparison of results and election of method. The extraction of key points from 2D data is faster that from 3D data, in the case of the tested algorithms, FAST and NARF, respectively. This can be checked in the following diagram (Fig. 17) which represents the execution time spent by each method in function of the number of key points extracted.

On the other hand, a reliable method must extract representative points in the scene, independently of the resolution and precision of the point cloud. The method based on 2D data meets this requirement more efficiently than the other method for a same number of key points found (Fig. 18).

Because of the speed of the algorithm FAST and the reliability of its results, this is the chosen method to extract key points, and the value set for the threshold parameter is ten, which allow extracting the enough amounts of key points.

5.1.2. Features descriptors

The election of the type of descriptor utilised must be a compromise between the following premises:

The descriptor chosen must be the one that matches the best the necessities of the designed registration algorithm. This descriptor must be able to give a representative value to each key point and its computation for a large number of points must be executed inside the time restrictions.

The descriptor chosen must be the one which implementation is the best adapted to the integral images used in the registration algorithm.

The above premises are considered over following types of descriptors, as listed in Table 1:

1. Surface normal vectors and change of curvature in a point. This is the most basic descriptor of all the analysed types and its com-putation is really fast if it works over integral images. Therefore this descriptor is perfectly adaptable to the data used by the designed registration algorithm, and its computation time is appropriate to the mapping function. The calculation of the sur-face normal vectors is not possible in a point whose neighbour-hood in the integral images presents a lot of empty positions, even so, the number of descriptors calculated is enough.

2. PFH (Point Feature Histograms) descriptors. The surface normal vectors in the neighbour points of every key point are needed to calculate PFH descriptors in those points. As we mentioned above, the empty positions on the integral image produce a decrease in the number of normal vectors calculated, at the same time the lack of normal vectors drastically reduces the number of points where the calculation of PFH descriptors is viable, so they do not reach the needed number. In addition, these descriptors require a previous calculation of surface nor-mal vectors, and this fact produces an accumulation of errors and increases the computation time.

3. NARF (Normal Aligned Radial Feature) descriptor. The calcula-tion of NARF descriptor requires the previous creation of a range image, and it runs better if the key points are selected with a NARF extractor. In the previous stage this descriptor was already discarded because of a serial of disadvantages.

The descriptors applied are the surface normal vectors and change of curvature because these allow calculating their values in a large amount of points, without errors accumulation and in a short time.

5.1.3. Results of the ICP algorithm

The ICP algorithm returns the score parameter and the transfor-mation matrix needed to align a point cloud captured from a local coordinates system with other one referenced in the global coordi-nates system. In order to evaluate the quality of this matrix the fol-lowing requirements are checked:

- The score value must be lower than 0.05 m. The score value is the mean distance between the corresponding points in the two aligned clouds. If this distance is smaller than five centimetres, the alignment is accurate enough.
- The rotation matrix must approach to the identity matrix, the absolute difference between the rotation matrix and the iden-tity matrix must be lower than 0.015 for each component. The rotation between the coordinates in the local system of the points of a cloud and its coordinates in the global system is already compensated approximately in the pre-alignment pro-cess, so the transformation matrix to align a cloud must not pre-sent hardly any rotation component.
- The translation in the X axis must approach to zero, lower than 0.01 m. The sensor rotates around the X axis so the coordinates in this axis must not suffer any transformation.

In all the tests performed these requirements were met. An example of the alignment executed by the ICP algorithm is shown in the following Fig. 19.

Table 1
Evaluation of different feature descriptors.

Feature descriptor	Surface normal vectors	PFH	NARF
Computational time	$t < 2$ s	$t > 60$ s	$t > 60$ s
Type of source data	Integral image	Cloud of normal vectors	Range image

5.2. Tests results of the proposed surveillance approach

Functionality of the surveillance system designed is tested at several positions and orientations of the Kinect device base.

- Device position. The surveillance system is tested for two height values (2 m, 0.8 m) of the surface where the device is placed.
- Device orientation. The surveillance system is tested for three values (10L, 0L, 10L) of the tilt of the device support surface. Then the device placed over this tilted surface can move its tilt motor to provide different orientations for the depth and colour sensors.

In order to evaluate the quality of the system in these different situations, both the precision of the 3D model created and the later scene analysis are studied through the following parameters:

- Mean distance between corresponding points, d, in metres. In the alignment step of the mapping module, the mean value of the distances between every couple of corresponding points from two aligned point clouds is obtained. To get a representative value for the whole map, the mean of values obtained in every view registration is calculated.
- Registration time, tr , in seconds. This is the mean of time spent to run every iteration of the registration cycle, where an acquired point cloud is registered and added to the scene map.
- Map creation time, T, in seconds. This is the total time required to create the whole scene map, including the time spent in the sensor setting, which depends on the user. So this time value could be quite variable, but it is possible to offer an approximated value.
- Number of acquired views, Nv . This is the number of partial views that compose the whole scene map. This number depends on the inclination of the surface where the device is placed, since this inclination determines the scan amplitude. The sensor does its widest scan when the support surface is horizontal.
- Surveillance cycle time, t, in seconds. This is the mean of the time spent to detect strange elements in each scene capture, it means, the time taken to update surveillance data.
- Quantity of false positives, fp. A false positive is considered when something which is not a strange element is detected by the surveillance system in the analysis of a view. This is caused by problems in the registration of the views. The quantity of false positives is defined as the number of false positives detected in the analysis of 500 scene views.

The above parameters values, measured in the different tests are compiled in the following table (see Table 2)

The surveillance system has been tested with the Kinect device at four emplacements from which any intruder can be detected (Fig. 20). From a high position and a positive tilt angle the ceiling of the room is captured and from a low position and a negative tilt angle the floor is observed, so in neither of these two cases the regions are of interest. In the four studied cases the parameters values obtained are really similar. However the best results appear when the device base is high and has a negative tilt angle and when it is low and has a positive tilt angle. In these two extreme cases the map created is more accurate (d value is lower) and therefore

there are no false positives, but the surveillance cycle time gets worse, since to obtain the optimal surveillance data the analysis time required is higher. The average value of this time implies that the refresh rate to show the surveillance data is approximately 15 frames per second. The map creation time is quite low considering that this is a start-up process previous to surveillance.

Table 2
Test results.

		Height (m)	
		2 m	0.8 m
	−10°	d = 0.033 m $\boldsymbol{t_r}$ = 1.36 s $\boldsymbol{N_v}$ = 12 T = 98 s t = 0.068 s fp = 0	Not relevant data
Tilt angle (°)	0°	d = 0.041 m $\boldsymbol{t_r}$ = 1.84 s $\boldsymbol{N_v}$ = 14 T = 91 s t = 0.063 s fp = 25	d = 0.026 m $\boldsymbol{t_r}$ = 1.63 s $\boldsymbol{N_v}$ = 14 T = 77 s t = 0.061 s fp = 5
	10°	Not relevant data	d = 0.024 m $\boldsymbol{t_r}$ = 1.52 s N_v = 12 T = 90 s t = 0.063 s fp = 0

From the tests performed, the following specifications of the surveillance system are defined:

- Detection sensibility. This property determines the dimension that an element must present to be detected by the system. Firstly, the element must appear inside of the field of view of the sensor, completely or partially. Moreover, the size of the part contained in the view field must meet the two following specifications:
  1. The length must be higher than 0:10 3 m.
  2. The surface captured by the sensor must be larger than 0.05 m2 approximately.

The system is capable of detect objects with a smaller size but this cannot be guaranteed. A person has the enough size to be always detected, as demonstrated by the tests performed.

- Illumination in the scene. The illumination conditions required in the mapping are not the same than in the analysis process. During the map creation, the scene must be illuminated. In the tests, this process has been carried out in presence of both natural and artificial illumination, even with reflective surfaces, which are not convenient for the good performance of the depth sensor. The analysis of the scene, that is to say the surveillance process, is tested with and without illumination. In these two cases, the intruders' detection is properly performed as long as the illumination is not too intense.

The best results are obtained when there is no illumination in the room. Strong condition illumination problems are related to the sensing device, designed for indoor applications. However the algorithm presented can be easily adapted to further sensing devices, able to adapt to different illumination conditions.

- Both position and orientation of the Kinect device in mapping are the same than in analysis process. This fact makes easy to use the application, since the mapping can be done first and then automatically the analysis stars iteratively.

6. Conclusions

The surveillance method proposed in this work is based on a novel approach that detects instantaneous changes by means of the comparison between the real aspect of the scene and the situation represented in its 3D model or map. This fact allows identifying not only the intruders, but also the new and missing elements in the room, independently of their movement and the illumination conditions, where the nocturnal surveillance applications can be especially benefited.

A low cost surveillance system has been designed, which uses a commercial device, Microsoft Kinect. The system does not require a troublesome set up, thanks to the fact that it is based on a single device placed in a fixed position, instead of being moved around the room in order to create the map. In addition, the same device generates the 3D model of the room in a first stage of the surveillance task, without the necessity of a complex off-line processing or human monitoring. All these features make the proposed system a powerful surveillance tool even in domestic environments.

In order to increase the wideness of the monitored scene, the tilt motor built into the device has been used to modify the orientation of the sensor, scanning the whole wide room during the surveillance task. Moreover, the distance from the sensor, in which it is possible to detect intruders, has been duplicated with respect to the commonly used value of Kinect sensor. This device measures accurately the depth where objects are located until 4 m, but the designed method is robust enough to deal successfully with the lack of precision in larger distances to analyse wider areas.

To create the 3D model of the complete scene, a high accuracy algorithm has been designed for the registration of room views acquired from different orientations of the sensor. We use a novel method which divides the registration process in two main stages. The first stage is a pre-alignment process based on a proprioceptive data from the sensor tilt movement, i.e. the value of the tilt angle from which a scene view has been acquired. The second stage is a fine alignment process based on the exteroceptive data captured from the scene, which feed the Levenberg–Marquardt ICP algorithm. One of the main steps of any registration algorithm is the matching of pairs of corresponding points belonging to two consecutive captures of the scene. This complex task has been solved using the fusion of geometric and colour data, allowing the registration with light illumination changes between captures, and even the registration of challenging flat surfaces without significant geometric characteristics as well as, complexly-shaped surfaces. Moreover, the ICP algorithm has been improved by a previous filter that maintains the areas of two consecutive captures where there is certainty of find corresponding points between them. In that way ICP works properly with inaccurate 3D data, unlike others that require expensive 3D

sensors. The explained two stages, that is the registration algorithm, are capable to obtain a very accurate 3D model faithful to the real scene.

The surveillance process is viable in all the regions represented by the 3D model of the scene. The capture of the scene, performed for each specific tilt angle, are registered in the proper position of the larger model to allow the comparison between them, thanks to a registration algorithm. This algorithm proved to be fast, allowing real time intruder detection. To meet this requirement, the algorithm is based only in the proprioceptive data of the motor movement, i.e. the tilt angle and the position of the rotation axis with regards to the depth sensor. In order to know this position a calibration method has been developed. This registration method minimizes the time needed to refresh the information of surveillance, where the elements detected in the scene are shown with a frequency between 15 and 20 frames per second. In addition, this fast scanning method reduces the computational load of visual algorithms, making the system versatile and robust. The analysis of each part of the scene is carried out by a change detector and a cluster-based module to catch only real intruders and modified objects, thanks to octree structures.

Moreover, an efficient management of 3D data is achieved by means of integral images structures, this means the projection of 3D points in a matrix following the 2.5 dimensional nature of the sensor. The structuring of colour point clouds as integral images allows, not only reducing the computational time of the algorithms, but also the fusion of the colour and 3D information. This fusion gives robustness and versatility to the mapping process and consequently to the surveillance system. The system is robust against entry data that are not optimal, therefore it is possible to obtain a 3D model of any room independently of its geometric complexity. In addition, both the mapping and detection processes, work properly even with the presence of reflecting surfaces, natural or quite strong artificial illumination sources. Finally, the data fusion makes this proposed system a versatile surveillance system, which is capable to monitor different rooms from various locations and orientations of the sensor device.

Acknowledgements

This work was supported by the Spanish Government through the CICYT projects (TRA2013-48314-C3-1-R) and (TRA2011-29454-C03-02).

References

Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30, 555–560.

Albusac, J., Vallejo, D., Castro-Schez, J. J., Glez-Morcillo, C., & Jiménez, L. (2014). Dynamic weighted aggregation for normality analysis in intelligent surveillance systems. Expert Systems with Applications, 41, 2008–2022.

Davis, J. W., & Sharma, V. (2004). Robust background-subtraction for person detection in thermal imagery. In IEEE international workshop on object tracking and classification beyond the visible spectrum (pp. 1–8).

Fischer, Y., & Beyerer, J. (2012). A top-down-view on intelligent surveillance systems. In The seventh international conference on systems, ICONS (pp. 43–48).

Fitzgibbon, A. (2003). Robust registration of 2D and 3D points sets. Image and Vision Computing, 21(13–14), 1145–1153.

Fookes, C., Denman, S., Lakemond, R., Ryan, D., Sridharan, S., & Piccardi, M. (2010). Semi-supervised intelligent surveillance system for secure environments. IEEE International Symposium on Industrial Electronics (ISIE), 2815–2820.

Galvin, B., McCane, B., Novins, K., Mason, D., & y Mills, S. (1998). Recovering motion fields: An evaluation of eight optical flow algorithms. In British machine vision conference.

García, F., Jiménez, F., Anaya, J. J., Armingol, J. M., Naranjo, J. E., & y de la Escalera, A. (2013). Distributed pedestrian detection alerts based on data fusion with accurate localization. Sensors, 13(9), 11687–11708.

Gill, T., Keller, J. M., Anderson, D. T., & Luke, R. H. (2011). A system for change detection and human recognition in voxel space using the Microsoft kinect sensor. In Applied imagery pattern recognition workshop (AIPR) (pp. 1–8). IEEE.

Harville, M. (2004). Stereo person tracking with adaptive plan-view statistical templates. Tech. rep., Hewlett-Packard.

Horn, B. K. P., & Schunck, B. G. (1980). Determining optical flow. AI Memo 572, Massachusetts Institute of Technology.

Hu, J., Hu, R., Wang, Z., Gong, Y., & Duan, M. (2013). Kinect depth map based enhancement for low light surveillance image. In 20th IEEE international conference on image processing (ICIP) (pp. 1090–1094).

Huihuan, Q., Xinyu, W., & Yangsheng, Xu. (2011). Intelligent surveillance systems, intelligent systems, control and automation: Science and engineering. Springer.

Huwer, S., & Niemann, H. (2000). Adaptive change detection for real-time surveillance applications. In IEEE International Workshop on Visual Surveillance (Dublin, Ireland) (pp. 37–45).

Irani, M., & Anandan, P. (1998). A unified approach to moving object detection in 2D and 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(6), 577–589.

Izadi, S., Newcombe, R. A., Kim, D., Hilliges, O., Molyneaux, D., Hodges, S., Kohli, P., Shotton, J., Davison, A. J., & Fitzgibbon, A. (2011). KinectFusion: Real-time dynamic 3D surface reconstruction and interaction. In SIGGRAPH ‘11, ACM SIGGRAPH 2011 Talks, No.23.

Keller, M., Lefloch, D., Lambers, M., Izadi, S., Weyrich, T., & Kolb, A. (2013). Real-time 3D reconstruction in dynamic scenes using point-based fusion. In Proc. of joint 3DIM/3DPVT conference (3DV).

Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41, 1690–1700.

Krishnan, A. K., Saripalli, S., Nissen, E., & Arrowsmith, R. (2012). 3D change detection using low cost aerial imagery. In IEEE international symposium on safety, security, and rescue robotics (SSRR) (pp. 1–6).

Kuthirummal, S., Bansal, M., Sawhney, H., & Eledath, J. (2011). 3D alignment and change detection from uncalibrated eye images. In First IEEE international conference on healthcare informatics, imaging and systems biology (HISB) (pp. 299–306).

Lee, S., & Chung, W. K. (2006). Rotating IR sensor system for 2.5D sensing. In 2006 IEEE/RSJ international conference on intelligent robots and systems (pp. 814–819).

Lim, M. K., Tang, S., & Chan, C. S. (2014). ISurveillance: Intelligent framework for multiple events detection in surveillance videos. Expert Systems with Applications, 41, 4704–4715.

Lu, D., Mausel, P., Brondízio, E., & Moran, E. (2004). Change detection techniques. International Journal of Remote Sensing, 25, 2365–2401.

Luo, B., & Xia, J. (2014). A novel intrusion detection system based on feature generation with visualization strategy. Expert Systems with Applications, 41, 4139–4147.

Martinez, J. L., Reina, A. J., Morales, J., & Mandow, A. (2013). Using multicore processors to parallelize 3D point cloud registration with the coarse binary cubes method. IEEE International Conference on Mechatronics (ICM), 335–340.

Mishra, A. K., Ni, B., Winkler, S., & Kassim, A. (2007). 3D surveillance system using multiple cameras. In SPIE proceedings, 3D Sensing: Videometrics (Vol. 6491).

Müller, K., Smolic, A., Drose, M., Voigt, P., & Wiegand, T. (2005). 3-D reconstruction of a dynamic environment with a fully calibrated background for traffic scenes. IEEE Transactions on Circuits and Systems for Video Technology, 15(4), 538–549.

Myronenko, A., & Song, X. (2010). Point set registration: Coherent point drift. IEEE Transaction on Pattern Analysis and Machine Intelligence, 32(12), 2262–2275.

Newcombe, R. A., Davison, A. J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., & Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In Proc. IEEE Int. Symp. Mixed and Augmented Reality (ISMAR) (pp. 127–136).

Niessner, M., Zollhöfer, M., Izadi, S., & Stamminger, M. (2013). Real-time 3D reconstruction at scale using voxel hashing. ACM Transactions on Graphics (TOG), 32(6), 169.

Oliveira, R. R., Noguez, F. C., Costa, C. A., Barbosa, J. L., & Prado, M. P. (2013). SWTRACK: An intelligent model for cargo tracking based on off-the-shelf mobile devices. Expert Systems with Applications, 40, 2023–2031.

Park, H., Lee, S., & Chung, W. K. (2006). Obstacle detection and feature extraction using 2.5D range sensor. In SICE-ICASE, 2006. International Joint Conference (pp. 2000–2004).

Popa, M., Koc, A. K., Rothkrantz, L. J. M., Shan, C., & Wiggers, P. (2012). Kinect sensing of shopping related actions, constructing ambient intelligence. Communications in Computer and Information Science, 277, 91–100.

Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: A systematic survey. IEEE Transactions on Image Processing, 14, 294–307.

Santos, D., Sciotti, D. F., Gomes, M., Oliva, D., Wolf, D. F., & Santos, F. (2012). Mobile robots navigation in indoor environments using kinect sensor. In Second Brazilian conference on critical embedded systems.

Savage, R., Clarke, N., & Li, F. (2013). Multimodal biometric surveillance using a kinect sensor. In Proceedings of the 12th annual security conference, Las Vegas, USA.

Song, X., & Zhang, Q. (2013). Kinect-based intelligent surveillance, motion capture and 3D object recognition. CORE Project of Microsoft Research, USA.

Steder, B., Rusu, R. B., Konolige, K., & Burgard, W. (2010). NARF: 3D range image features for object recognition. In IEEE/RSJ int. conf. on intelligent robots and systems (IROS), workshop on defining and solving realistic perception problems in personal robotics (pp. 1–2).

Tang, J., Luo, J., Tjahjadi, T., & Gao, Y. Y. (2014). 2.5D multi-view gait recognition based on point cloud registration. Sensors, 14(4), 6124–6143.

Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. Foundation and Trends in Computer Graphics and Vision, 3, 177–280.

Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: A review vision. Image and Signal Processing, IEE Proceedings, 152, 192–204.

Figures

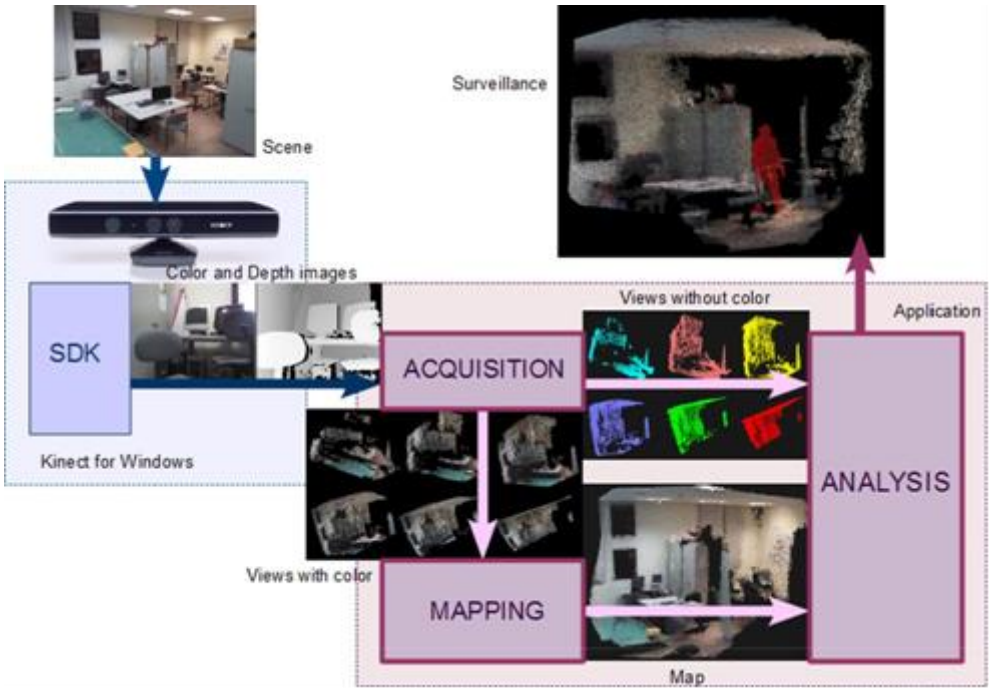

Fig. 1. Architecture and data flow between surveillance system modules with each other and the sensor.

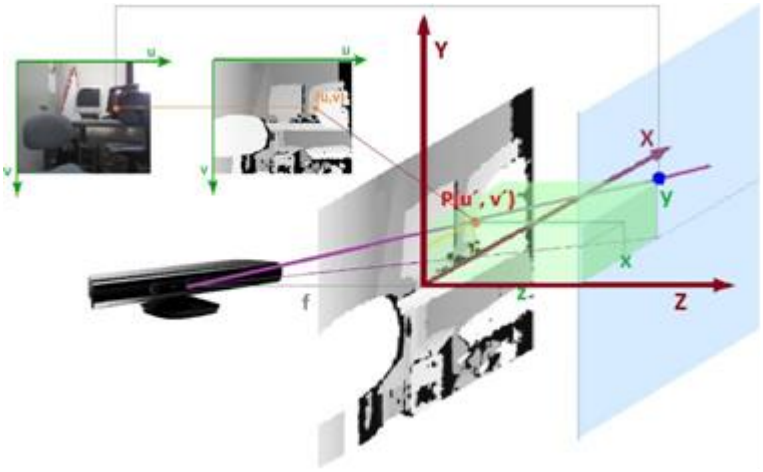

Fig. 2. 3D coordinates estimation from depth data and calibration with colour data for a point.

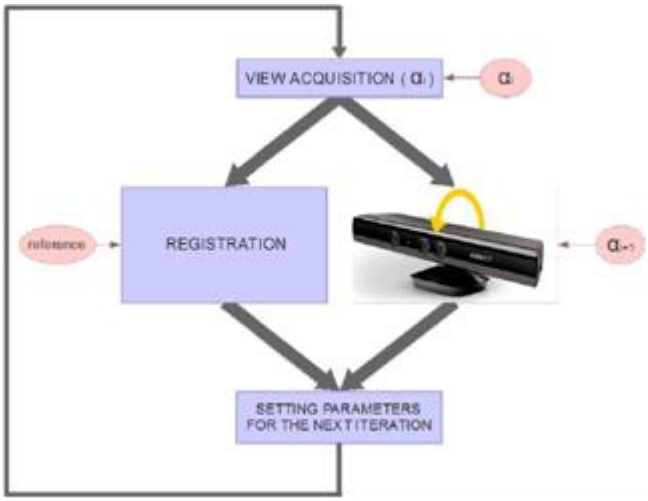

Fig. 3. Mapping method.

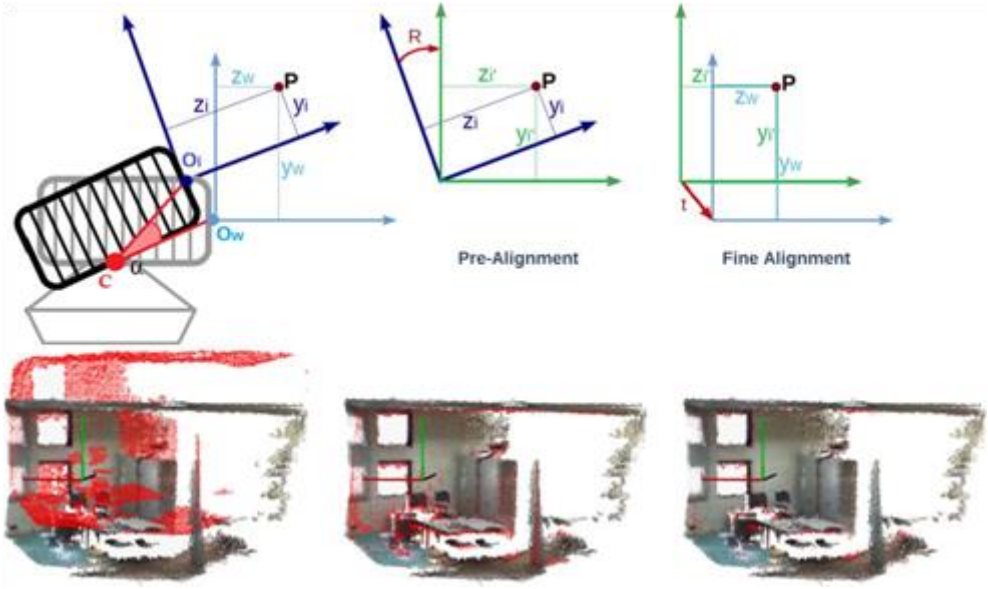

Fig. 4. Global coordinates estimation in two steps: pre alignment and fine alignment (top), and the differences (red) between a point cloud and the previous one during these phases (bottom). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

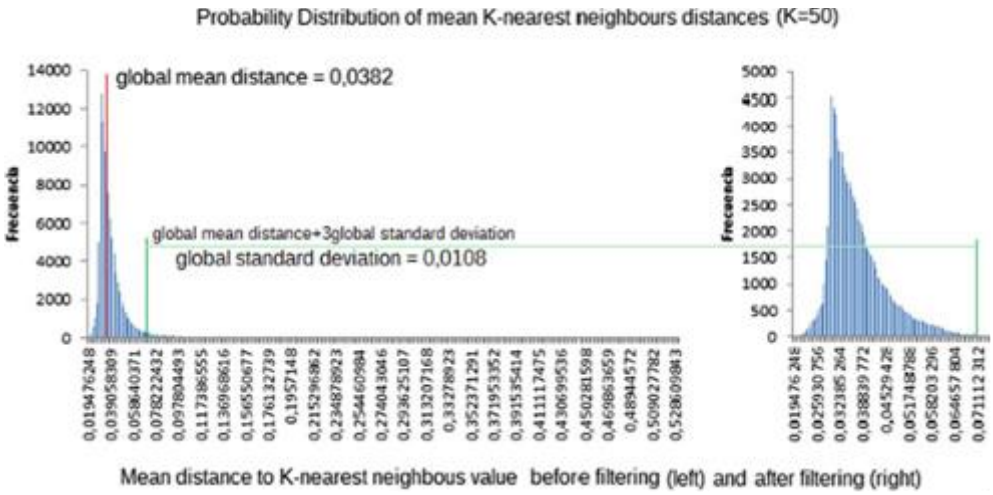

Fig. 5. Probabilistic distribution of the mean distance of each point to its 50 nearer neighbours

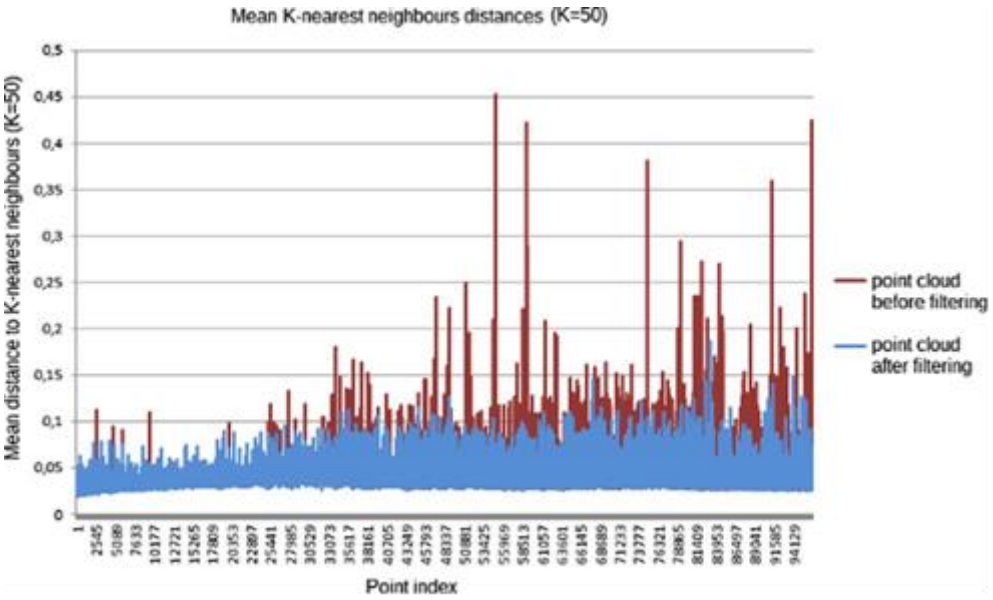

Fig. 6. Mean distance of every point to its 50 nearer neighbours with outliers values (red) and without them (blue). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

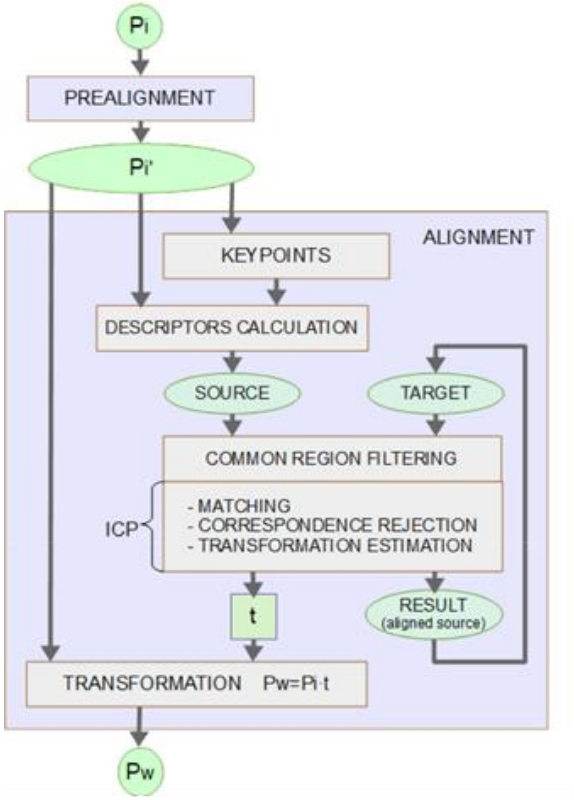

Fig. 7. Registration algorithm.

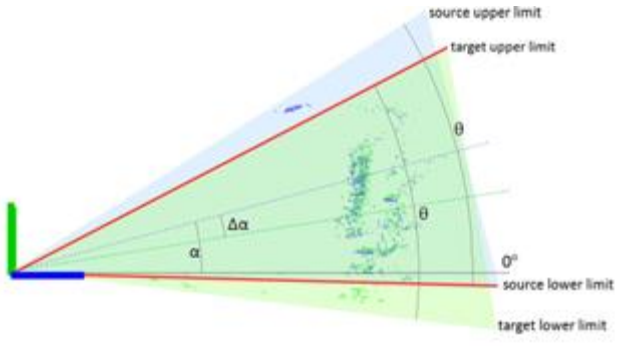

Fig. 8. Limits of the field of view for two consecutive captures.

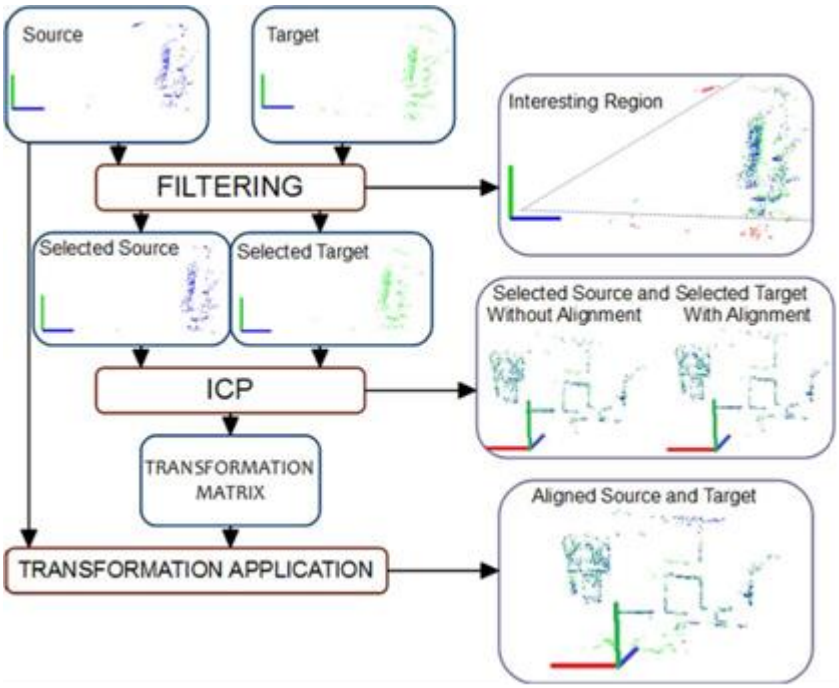

Fig. 9. Common region filtering and ICP.

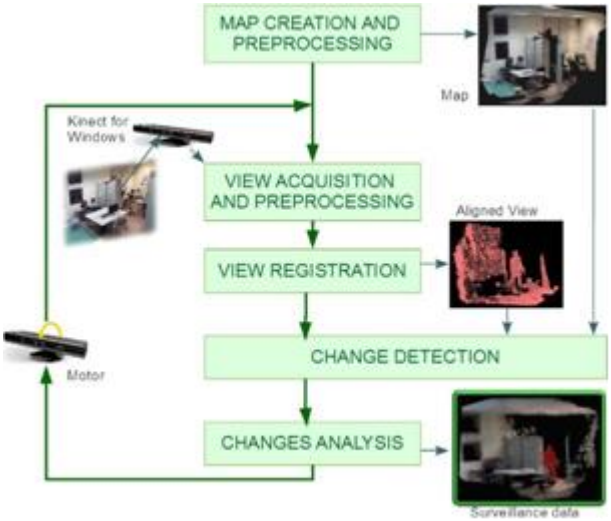

Fig. 10. Analysis module algorithm.

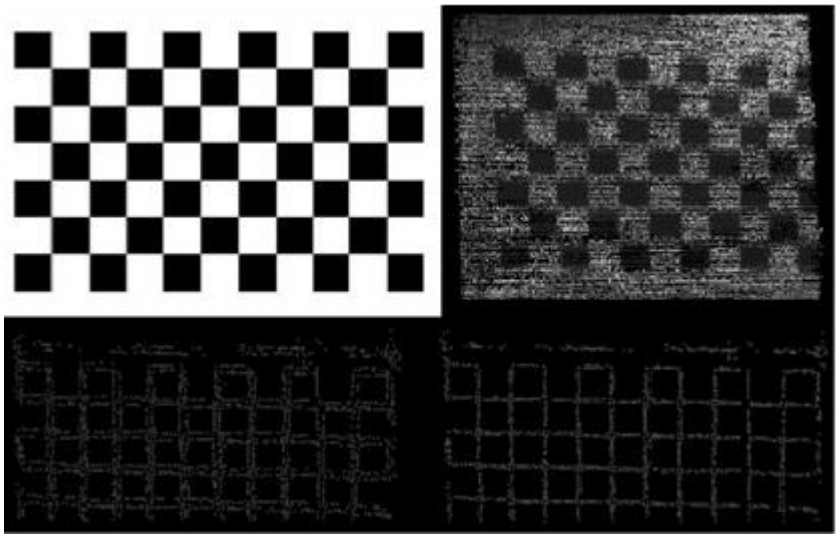

Fig. 11. Test surface for estimation of the rotation centre (top left), representative point cloud (top right) and non-aligned (bottom left) and aligned (bottom right) key points of two consecutive views.

	z	y
	-1,7252	0,0601
	-1,9773	0,06828
	-1,8502	0,1928
	-2,0381	-0,07231
	-2,0832	0,0866
	-1,7379	0,1889
	-1,0103	0,1044
	-2,3356	0,0815
	-2,1985	-0,1537
	-1,431	-0,0499
	-1,5646	-0,109
mean	-1,81381	0,036152
standard deviation	0,379351	0,116197

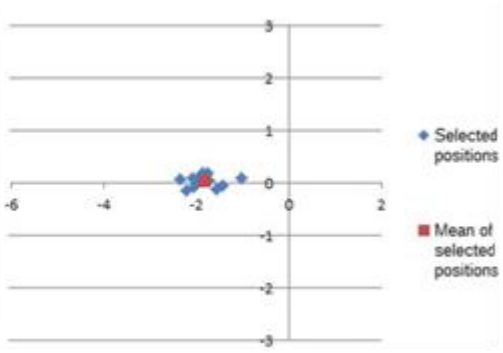

Fig. 12. Estimated values for the rotation axis coordinates (cm).

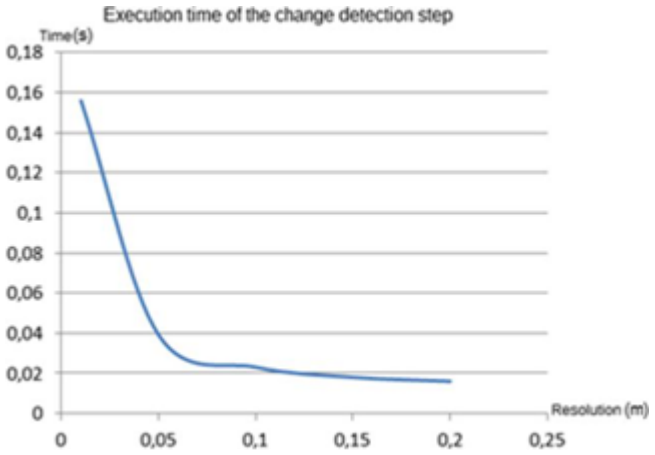

Fig. 13. Execution time for change detection depending on resolution.

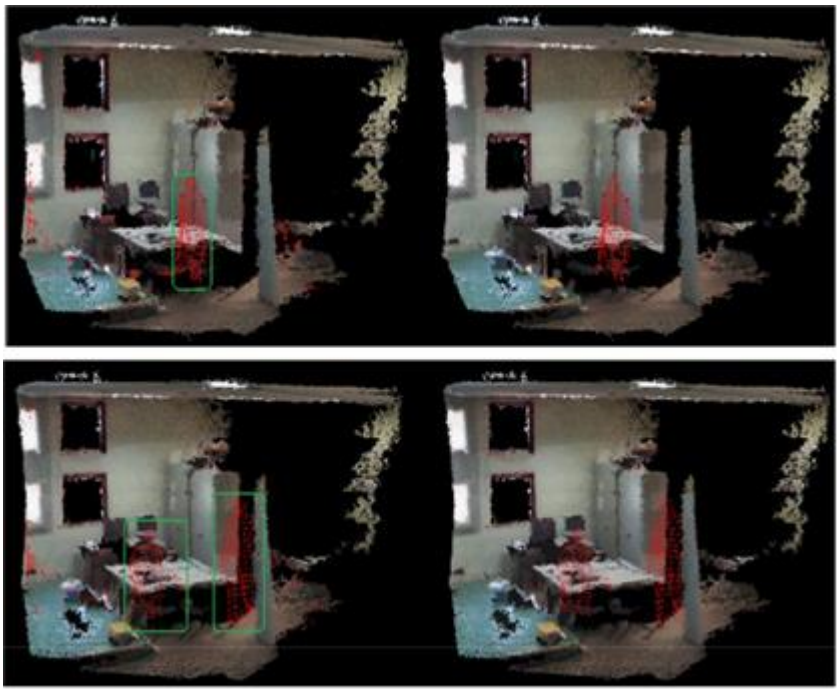

Fig. 14. Changes detected in a scene view (red) and identifying of those corresponding to an intruder (green) at left, and remained clusters after the analysis at right. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

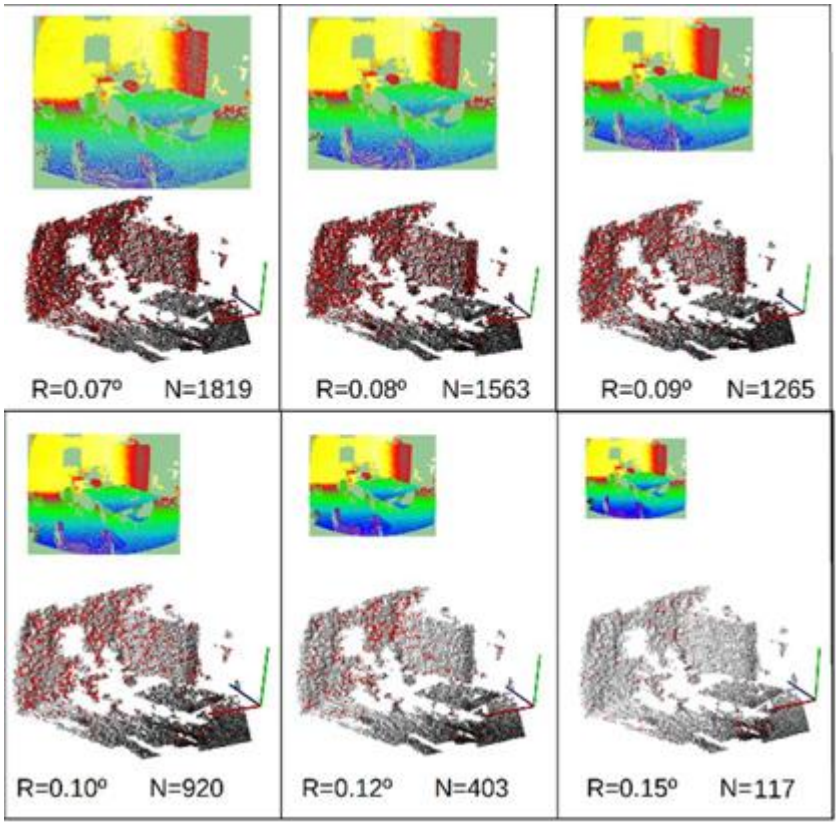

Fig. 15. Range images (top) and NARF key points (red) in the source point cloud (bottom) for different angular resolutions, R. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

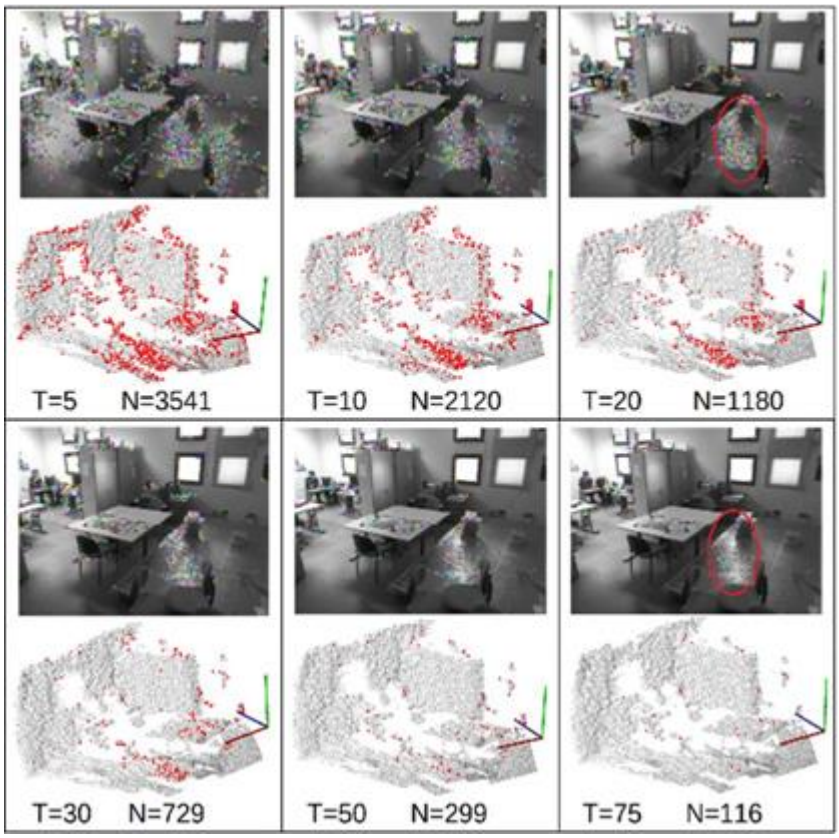

Fig. 16. Interesting pixels (coloured) in the grey-scale image (top) and their corresponding key points (red) in the source point cloud (bottom) for different threshold values, T. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

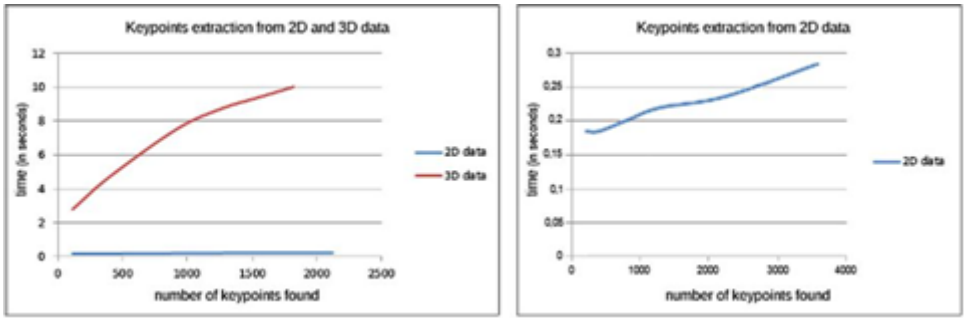

Fig. 17. Extraction time in function of the number of key points found in 3D data (red) and in 2D data (blue), this one is expanded in the right diagram. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

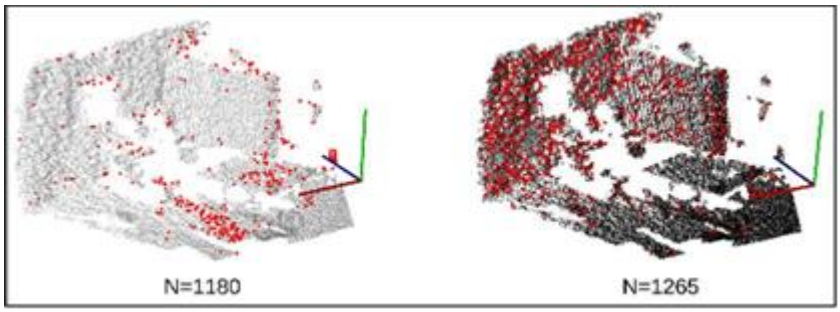

Fig. 18. Key points (red) extracted from 2D data (left) and from 3D data (right). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

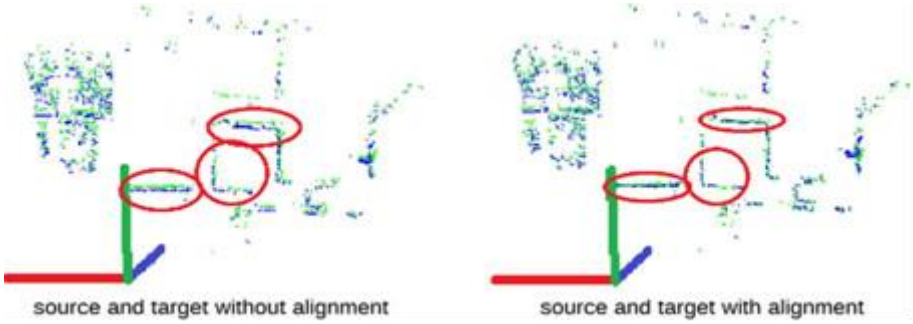

Fig. 19. Source point cloud (green) and target point cloud (blue) before (left) and after (right) the alignment. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

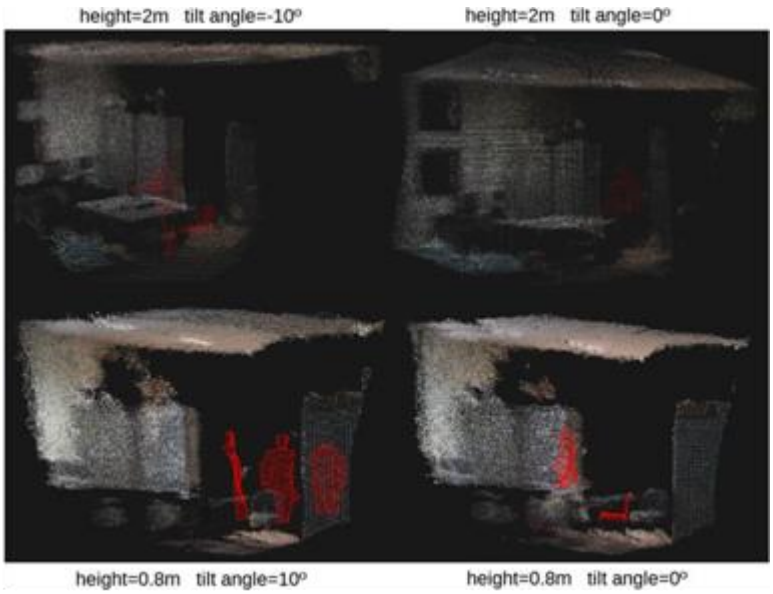

Fig. 20. Surveillance data shown to the user in the different tests.