INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2017;**30**:e3223.
Published online 6 October 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/dac.3223

Comparison of fiducial marker detection and object interaction in activities of daily living utilising a wearable vision sensor

C. Shewell[1,*,†], J. Medina-Quero[2], M. Espinilla[2], C. Nugent[1], M. Donnelly[1] and H. Wang[1]

[1] *Computing and Mathematics, University of Ulster at Jordanstown, Newtownabbey, UK*
[2] *Computer Science, Universidad de Jaén, Jaén, Spain*

SUMMARY

This paper presents a comparison between algorithms (Oriented FAST and Rotated BRIEF (ORB) and Aruco) for the detection of fiducial markers placed throughout a smart environment. A series of activities of daily living (ADL) were conducted while monitoring a first-person perspective of the situation; this was achieved through the usage of the Google Glass platform. Fiducial markers were employed, as a means to assist with the detection of specific objects of interest, within the environment. Each marker was assigned unique Identification (ID) and was used to identify the object. Three activities were performed by a participant within the environment. On subsequent trials of the solution, lighting conditions were modified to assess fiducial marker detection rates on a frame-by-frame basis. This paper presents the results from this investigation, detailing performance measure for each object detected under various lighting conditions, motion blur and distance from the objects. An intelligent system was developed to specifically consider distance estimation in order to aid with the filtering out of false interactions. A linear filtering method was applied along with a fuzzy membership function to estimate the degree of user interaction, which assists in removing false positives generated by the occupant. The intelligent system returns an average precision, recall and an F-Measure of 0.99, 0.62 and 0.49, respectively. Copyright © 2016 John Wiley & Sons, Ltd.

Received 18 December 2015; Revised 22 August 2016; Accepted 1 September 2016

KEY WORDS: Aruco; fiducial; localisation; machine-vision; ORB; wearable

1. INTRODUCTION

The use and manipulation of objects is of key importance when carrying out activities of daily living (ADL) [1]. Unfortunately, those suffering from cognitive decline often find that their ability to independently carry out ADLs independently is reduced. Cognitive decline is typically attributed to a condition such as Alzheimer's disease, or due to the effects of stroke or traumatic brain injury. The symptoms include impaired memory, which can affect recognition with respect to people, objects or locations, in addition to causing a degradation in both short-term or long-term memory [2]. Smart Environments have long been postulated as a means to improve the quality of life of those suffering from cognitive decline, offering increased independence and postponing the need for full time care or institutionalisation [3].

These sensorised Smart Environments typically monitor an environment and its occupant and aim to reason on available sensor inputs towards offering some level of support to occupants. Support can range from automated temperature management through to detailed support with complet-

*Correspondence to: Colin Paul Shewell, University of Ulster at Jordanstown, Computing and Mathematics, Shore Road, Newtownabbey, Co. Antrim, Newtownabbey, UK.
†E-mail shewell-c@email.ulster.ac.uk

ing ADLs. Within a health context, Smart Environments offer a number of potential benefits that include reducing the number of accidents, providing support and intervention for specific illnesses or conditions as well as providing general assistance; previous systems include [4–6].

Given that ADLs range widely in terms of variety and complexity, varying approaches have been investigated in an effort to support automatic recognition. A common approach has been to employ dense sensor placement within an environment to determine which object(s) is being interacted with [7]. However, this method has limitations, due to the binary nature of the sensors and its ability to only determine the occupant's location if an object has been interacted with. In addition, this approach has difficulty handling a multiple occupancy scenario as it may not be possible to determine which occupant has interacted with a particular object. There are also more practical issues to contend with, such as the typical need for retrofit installation, and the requirement for ongoing maintenance costs of such systems [8].

Machine vision techniques have been postulated as one potential solution to the abovementioned challenges. These techniques offer the ability to track an occupant's activity throughout an environment. Rather than relying on embedded sensors within the environment, the sensing is carried out directly through machine vision processing of the environment. This also offers the advantage that it works on an unmodified environment; therefore, a smart environment is not needed. Additional data streams can also be augmented with the vision stream, such as accelerometer readings to assist in inferring user context.

The system proposed in the current work makes use of fiducial markers to assist in the process of detecting objects. Fiducial markers, in the context of the current work, are defined as being images placed within a physical environment which can be used in support of tracking, alignment and identification of objects or location [9]. They can both be placed either on a mobile person/object in order to determine the location/identity of that person/object or, as in the work presented, they can be placed on fixed objects in order to determine the relative location of a moving camera. Fiducial markers do not have to be purposely placed within a scene as the use of natural markers within a scene can be used to determine location. An example of this would be a scene of a kitchen that contains a cooker; the cooker itself would be able to function as a fiducial marker thus assisting in identifying the scene. Other features such as windows and other miscellaneous objects would also suffice, so long as they make up a unique set of feature points to assist in identifying the scene as being unique [10]. The use of fiducial markers also reduces some of the traditional issues reported when performing object recognition, such as the requirement to learn variants of the same objects, for example, different models of a household appliance. They also aid in alleviating the problem of distinguishing between multiple identical objects in close proximity, such as kitchen cupboards [11].

This paper proposes a novel non-invasive solution to that of occupant localisation and object interaction, offering a unique first-person view of the environment. The proposed method reduces the invasiveness normally associated with the installation and maintenance of traditional systems, for instance, dense sensor based or static camera methods, along with the costs involved with the additional financial acquisition and deployment from the aforementioned systems. Additionally, the use of fiducial markers negates the need for training to each unique environment, which the system may be deployed within, as they will all share common objects that the occupant can interact with. The issue of multiple occupancy is also addressed with each occupant wearing a vision device that offers a first person view of each occupant in turn allowing individual support to be given. However, this is assuming that it will only be the occupants whom require support.

In a real world situation, different objects have different means of interaction; some objects require direct interaction during their use, whereas some objects only require passive interaction [12]. An indirect effect of this is that the distance between the user and object will differ depending on the type and level of interaction, an example of this would be a toaster and a television. A toaster requires a direct interaction to operate, that is, putting the bread in, turning on the toaster then removing the toast. A television would require a passive degree of interaction, the occupant would be viewing the object at a distance and would not require direct interaction with the object for it to be considered 'in use'. As a result of this, an intelligent system has been developed that allows the determination of whether an occupant is interacting with an object versus if they are viewing the object due to general gaze activity. An example of this would be looking around the environment while locating an object/item or viewing objects while navigating throughout the environment. In order to determine interaction, a threshold value is set by a human expert, which determines if an interaction is taking place [12]. The distance from the object is then calculated in real time and compared with the threshold value to establish interaction. These interactions between humans and objects are highly useful to infer both the activity and temporal information. In this paper, we aim to study the detection of interactions between users and objects by means of a vision sensor.

The future growth of vision sensors, driven by devices such as Google Glass, offers benefits to those in cognitive decline. An example of these benefits is the ability to record images to boost memory recall, or as this paper will focus on, to detect interactions with the future goal of providing timely and relevant assistance to aid in ADLs. Section 2 will discuss the current state-of-the-art in indoor localisation leveraging fiducial markers, Section 3 will detail the system along with the algorithms and markers used and the filtering process to remove false positives (FP). Section 5 will present the results gathered from evaluating the system and finally Section 6 will provide concluding remarks and detail the direction in which the work will proceed.

2. RELATED WORK

This Section will present an overview of the current state-of-the-art in machine vision solutions, facilitating indoor localisation through the use of fiducial markers, with the goal of supporting applications in the domain of ambient assisted living (AAL).

Rivera-Rubio *et al.* [13] implemented a solution that estimated the occupant's location through scene recognition. The study was implemented using an LG Google Nexus 4 paired with Google Glass. A dataset of the locations was obtained by recording a video of the occupant walking through the environment 10 times while wearing the relevant device (a 50/50 split between the Nexus and Glass). The recorded scenes simulated both daytime and nighttime lighting conditions with occasional strong lighting assessed via windows within the environment. The system was tested using a range of descriptor methods; three being custom designed and three standard methods. A bag-of-words and Kernel encoding pipeline method was used along with HOG3D matching to establish a baseline. Their results demonstrated an error rate as low as 1.6 m over a 50 m distance. However, for the purposes of AAL, a greater level of refinement is required in order to distinguish the occupant's location within a single room.

Zhang *et al.* [14] proposed a method of indoor localisation using still images captured at regular intervals from a smart-phone worn via a lanyard. The goal of the approach was to assist navigation throughout a familiar environment for those with impaired vision. The system relied on collecting data of a building that describes its features and descriptors along with relevant 3D co-ordinates, floor plans and other location data. Images were captured and sent at regular intervals to a server for processing, where they were matched against the template map of the building to determine location and offer assistance if required. Some challenges faced by this system, as noted by the authors, were that there were null spots caused by a lack of features in the image to create a map. This tended to happen when the user made a 90 degree turn, for example, when entering a room. A further shortcoming, related to intermittent images, was due to their intermittent nature; as there was a period of time between images being captured where data can be lost. This could lead to interactions being missed, such as an interaction with an object; which could be vital for determining an activity.

Orrite *et al.* [15] developed a system titled 'Memory Lane' which aimed at providing a contextualised life-blog for those with special needs. It contained images and sounds, as perceived by the user, which would be chronologically ordered and automatically tagged by the system, thereby providing contextual meaning. From the occupants environment, a data-set of images were gathered from which feature points were computed using SIFT with RANSAC. During each RANSAC iteration, a candidate fundamental matrix was calculated using the eight-point algorithm [16], normalising the problem to improve robustness to noise. The system consisted of a wearable camera, which would systematically record still images as the occupant moved throughout

DOI: 10.1002/dac

the environment. These images would be matched against the data-set of images that were gathered previously, in order to determine the occupant's location. To determine the distance from the object, a match correspondence amongst features, based on scale, is used. This solution involves generating a variable circle cantered on the average position of the detected features and comparing it to the average position in the next image. When the radius increased, it was determined that the occupant had moved closer to the object. This solution has some limitations; due to the intermittent nature of the images some key information could be lost, such as room transitions or the image lacking sufficient features in order to perform a match.

Zeb *et al.* [17] developed a system that supported blind users with navigating throughout a known environment. It achieved this via the user holding a web-cam in their hand and moving through the environment. The web-cam continuously took video frames from the environment, which were then processed for relevant markers. Whenever a relevant marker was detected, the detection and identification module compared it to the stored markers in a database, returning a unique ID that associated the user's position and direction. The main drawback from this system was that it required constant interaction from the user in the form of having to manipulate a hand-held camera at all times, in order for the system to detect markers.

The proposed approach within this paper aim to address these aforementioned shortcomings via the use of a head worn camera that requires no direct interaction by the user. As the system performs marker detection on each individual frame, it addresses the problem of data being lost due to intermittent images being taken. This method increases robustness; if the marker was not identified in an image, it may be identified in the following frames. In a system that captures intermittent images; if the marker is not detected then key information may be lost.

3. APPROACH

This Section details the methodology adopted to develop the system. The design of the fiducial markers that were used to identify the objects is presented along with a detailed overview of the algorithms used in the evaluation of the system. A description of the feature point identification method along with the implemented matching process is also presented. Finally, the analysis of occupant-object interaction was carried out in order to determine whether an interaction was a true positive, or a false positive generated via the occupant's navigation through the environment.

The initial approach aimed to compare the performance of two 'off-the-shelf' algorithms for performing fiducial marker recognition when coupled with a wearable Google Glass vision sensor

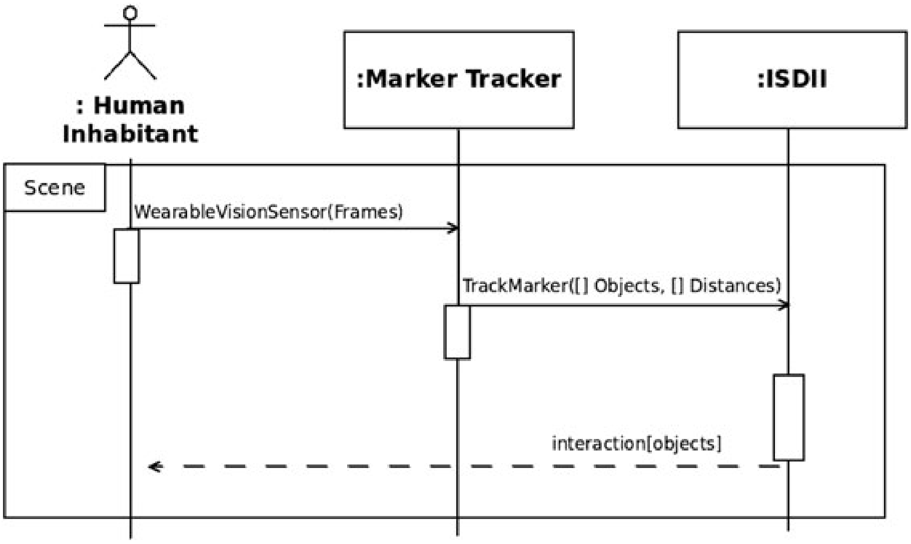

: Human Inhabitant
:Marker Tracker
:ISDII
Scene
WearableVisionSensor(Frames)
TrackMarker([] Objects, [] Distances)
interaction[objects]

Figure 1. Sequence diagram of the wearable vision sensors in ADLs.

towards accurate discrimination of occupant-object interaction. Figure 1 demonstrates the general sequence of events and presents: (i) frames that are returned from the wearable vision sensor; (ii) fiducial markers are then located within the returned frames; (iii) the degree of occupant-object interaction is established as a quantifiable metric.

Google Glass (Explorer) platform was employed to provide a first-person view of the user's environment. Google Glass facilitates the recording of high definition video (1280x720) and accept audio based commands from wearers of the device via natural spoken language commands. Pertinent information can also be presented to the wearer via a small prism display that is located in front of the eye.

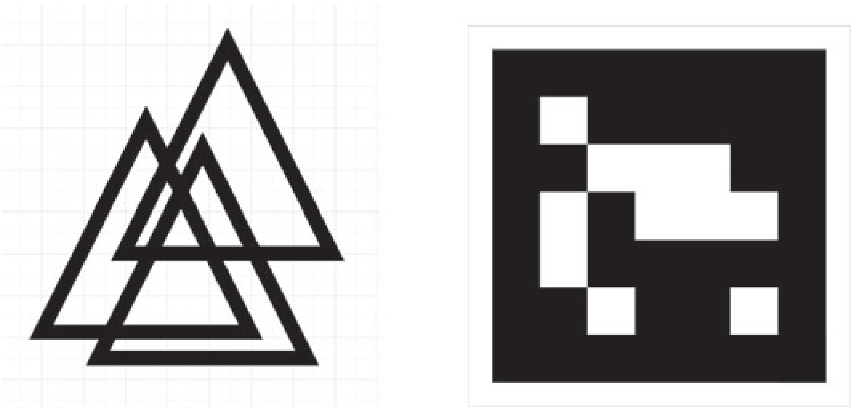
Traditionally, the impact of wearable computing devices has been partly slowed by their lack of streaming [18]. In an effort to overcome this, a Glass App was developed in our previous work that supports transmission of live video to a cloud-based server via Real Time Streaming Protocol (RTSP) [19]. This approach does however introduce a short latency between (<4 s) due to Glass' efforts to lower its temperature during high load situations, such as streaming. This is achieved by reducing the clock speed of the CPU [20].

Each fiducial marker has a custom identifier applied to it to represent the object it is associated with. The occupant's location is then estimated by means of a 3D reconstruction method that incorporates the known size of the markers, along with the calibration parameters of the vision sensor. Occupant location is of key importance when supporting ADLs; in the presented work distance is estimated to determine the degree of occupant-object interaction. Two feature point algorithms were employed to detect the markers located in the environment, using inputs from the vision sensor. In this work, we have integrated two detection algorithms, to detect the markers located in the environment, using vision sensors, they are:

3.1. ORB algorithm

The first method employs the OpenCV implementation of the ORB algorithm for both feature detection and description. This method was developed by Rublee *et al.* [21], and implements FAST in pyramids to facilitate the detection and selection of stable key-points. ORB implements the intensity centroid method of corner detection as defined by Rosin [22].

A Brute Force algorithm (k-Nearest Neighbour) has been implemented as a feature point matcher to determine if a marker is present in the frame. A formal representation of a k-Nearest Neighbour algorithm locates the k nearest features to a query feature N points in a D-dimensional space. Even though a Brute Force matcher is often one of the worst performing algorithms, in terms of time taken to resolve a match, this is counterbalanced by high levels of accuracy in identifying the correct matches. This can be found in [23], which benchmarked multiple techniques for the purposes of



(A) Example of ORB fiducial marker. (B) Example of Aruco fiducial marker.

Figure 2. A) example of ORB fiducial marker. B) example of Aruco fiducial marker.

image matching. Within this implementation for each feature in the marker, the matcher locates the closest feature in the scene by systematically trying each feature point. The similarity between feature points is represented by Norm Hamming distance. With a minimum distance set ensuring good matches are selected: a match is deemed to be good when the distance is less than three times the minimum distance set.

In order to reduce the number of FP found by the system, a key-point match threshold was used, where the number of inliers that contributed to the homography was calculated and compared against a threshold value. If the number of inliers met or exceeds the threshold, then a marker was deemed to be present. A strength of the approach is that the markers can be partly freely designed; refer to Figure 2.

3.2. Aruco algorithm

The second algorithm is Arcuo [24], developed around the concept of fiducial markers. The markers are automatically generated by Aruco by means of a marker dictionary [25] and is focused on extracting the binary code from the rectangles that make up the fiducial marker, see Figure 2. This process involves image segmentation, based on local adaptive thresholding. In order to increase robustness to varied lighting conditions, contour extraction and filtering, marker code extraction to obtain the internal binary code, and dictionary based correction once the binary code is extracted. This tracker is developed under Open Source license: the Berkeley Software Distribution. It has been deployed in several research and enterprise projects[‡].[§]

4. INTELLIGENT SYSTEM FOR DETECTING INHABITANT-OBJECTS INTERACTIONS

During testing of the vision algorithms, it was discovered that FP were being generated through general gaze activity due to the occupant looking around the environment when locating an object of interest. Further FP was generated through the occupant's navigation of the environment as various objects came into their field of view as they moved through the environment. An intelligent filter was developed with the aim to detect the degree of interaction between the occupant and the object, this is based around the observation that when the occupant is interacting with an object of interest they are in a close proximity to the object. This also aids in taking account of the differing forms of interaction that certain objects require, namely passive or active interaction, those objects that require active interaction will have a much closer distance threshold compared with those that passive objects which are interacted with from a larger distance – such as viewing TV. It is known as the intelligent system for detecting inhabitant-object interaction (ISDII). The output from the marker detection algorithms serve as the input for the ISDII system. These consist of the unique ID associated with the detected markers and the distance of the occupant to the marker. A three stage process is employed:

1. The first stage is to collect and analyse the scenes where interaction occurs between the occupant and the object.
2. Thresholds are then determined by an expert, establishing the distance at which occupant-object interaction is known to be occurring.
3. Once the threshold distances have been established, ISDII is able to identify interaction on a real-time basis.In order for ISDII to recognise if occupant-object interactions are occurring, a preliminary threshold value needs to be set by a human expert. An initial process was carried out that consisted of recording scenes where an occupant interacted with a series of objects throughout the environment and threshold distances were then set by a human expert; a sequence diagram detailing this step is presented in Figure 3. This allows ISDII to calculate, in real time, the distance between the occupant and the object and determine whether an interaction is taking place; the pseudo-code is presented in Algorithm 1.

[‡] http://www.vision4uav.com/?q=node/386
[§] http://vision4uav.eu/?q=researchline/seeAndAvoid_CE_MFandRules

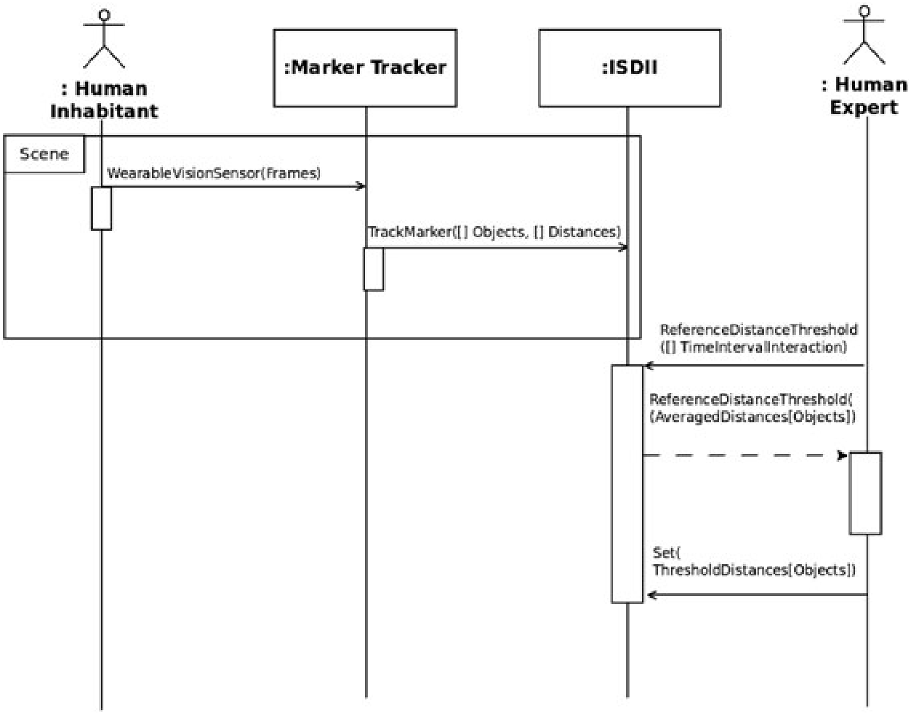

: Human Inhabitant
:Marker Tracker
:ISDII
: Human Expert
Scene
WearableVisionSensor(Frames)
TrackMarker([] Objects, [] Distances)
ReferenceDistanceThreshold ([] TimeIntervalInteraction)
ReferenceDistanceThreshold((AveragedDistances[Objects])
Set(ThresholdDistances[Objects])

Figure 3. Sequence diagram of studying scenes of user-object interactions.

When estimating object interaction in real time scenes, uncertainty is introduced due to missed marker detections in the video stream and measurement errors introduced by the algorithms. In order to manage this uncertainty a two stage filter has been developed. The first stage is to remove the high frequency noise using a low-pass filter. The exponential smoothing [26, 27], is defined in equation 1:

Algorithm 1. Estimation of reference distance thresholds to objects.

```
$distances = \emptyset$
$detections = \emptyset$
for $marker \in detectedM\ arkers$ do
  for $interval \in InteractionIntervals$ do
    if $marker.time \in interval$ then
      $distances[marker.object]+ = marker.distance$
      $detections[marker.object] + +$
    en d i f
  end for
end for
$threshold = \emptyset$
for $object \in objects$ do
  $threshold[object] = distances[marker.object]/detections[marker.object]$
end for
return $threshold$
```

$$s_0 = d_0, s_t = \omega_0 d_t + (1 - \omega_0)s_t{}_{-1}, \omega_0 \in [0, 1] \quad (1)$$

Where d_0 is the initial distance to a marker, t is the temporal index $\in [0, N)$ being N the final size of the set of distances, s_t is the filtered output, d_t the measured data – the distance from the marker, and ω_0 is the smoothing factor; this method is widely used in control applications [28, 29].

The second filter is designed to mitigate two main causes of FP – removing isolated detections, where a marker is detected due to general gaze activity. Fitting the window of interaction to the true occupant-object interaction, that is, removing the preceding time where the occupant is approaching the object and the proceeding time where the occupant is finished interacting with the object. In order to achieve this, a fuzzy membership function was developed. Fuzzy logic [30] has been successfully applied in sensor based signal processing applications [31]. In the context of fuzzy logic, the semantics of the linguistic terms are given by fuzzy sets; where the membership degree of the elements x of the base set X in the fuzzy set A, $\mu_A : X \rightarrow [0, 1]$ is defined. The smoothing distance of the markers from the first stage was evaluated by the fuzzy membership function which describes the linguistic term *'there is interaction with'*.

For each object o_i a membership function μ_{Oi} is defined which evaluates the distance between the occupant and the object s_t into a degree of occupant-object interaction between [0,1]. The membership function is parameterised by the threshold value of the object d_{oi}, and two weighted factors, ω_1 and ω_2, representing the lower and upper cut-off threshold for interaction respectively, (as presented in Figure 4).

$$\mu_{\tilde{O}}\,\mu_{\tilde{O}_i}(s_t, d_{o_i}) = \left\{ \begin{array}{ccc} 1 & if & s_t \le \omega_1 \cdot d_{o_i} \\ \dfrac{s_t - \omega_2 \cdot d_{o_i}}{\omega_1 \cdot d_{o_i} - \omega_2 \cdot d_{o_i}} & if & s_t \in [\omega_1 \cdot d_{o_i}, \omega_2 \cdot d_{o_i}] \\ 0 & if & s_t \ge w_2 \cdot d_{o_i} \end{array} \right\} \qquad (2)$$

ISDII provides a degree of interaction representing the occupant-object interaction within the environment. It should be noted that an upper threshold can be applied using $\alpha - cut$ between [0,1] above which an interaction is determined to have taken place. Pseudo-code detailing the second stage filter is presented in Algorithm 2 along with a sequence diagram presented in Figure 5.

5. RESULTS AND DISCUSSION

This Section presents the experimental use case scenarios. A series of markers were applied to objects within a smart lab. Three different scenarios were evaluated that required an occupant to enter the environment and proceed to complete pre-defined activities, while wearing a pair of Google Glass. The three activities were: making a hot drink; preparing a hot snack and washing dishes/cutlery. A sequential breakdown of the objects interacted with during the completion of each activity is presented in Table I.

Algorithm 2. Detecting Object Interaction.

```
degree = ∅
detection = ∅
for marker ∈ detectedM arkers do
    distance[marker.object] = $\omega_0$ • marker.distance + (1 − $\omega_0$) • distance[marker.object]
    degree[marker.object] = $\mu_{Oi}$ (distance[marker.object], threshold[marker.object]) ˜
end for
for object ∈ objects do
  if degree[object] > $\alpha$ then
     detection[object] = true
  en d i f
end for
return [degree, detection]
```

To facilitate the experiments, a total of 18 markers (9 unique), were placed within the environment on: kitchen door, cupboard doors, a microwave, a refrigerator, a tap and a chair. Multiple lighting conditions were simulated via the use of blinds and artificial lighting to provide a realistic context to the scenarios.

Low brightness and high motion blur situation. B) High brightness and low motion blur situation.

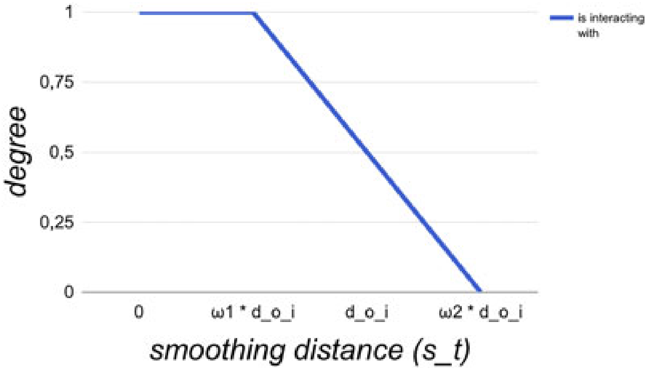


Figure 4. Membership function to obtain the degree of interaction with the object.

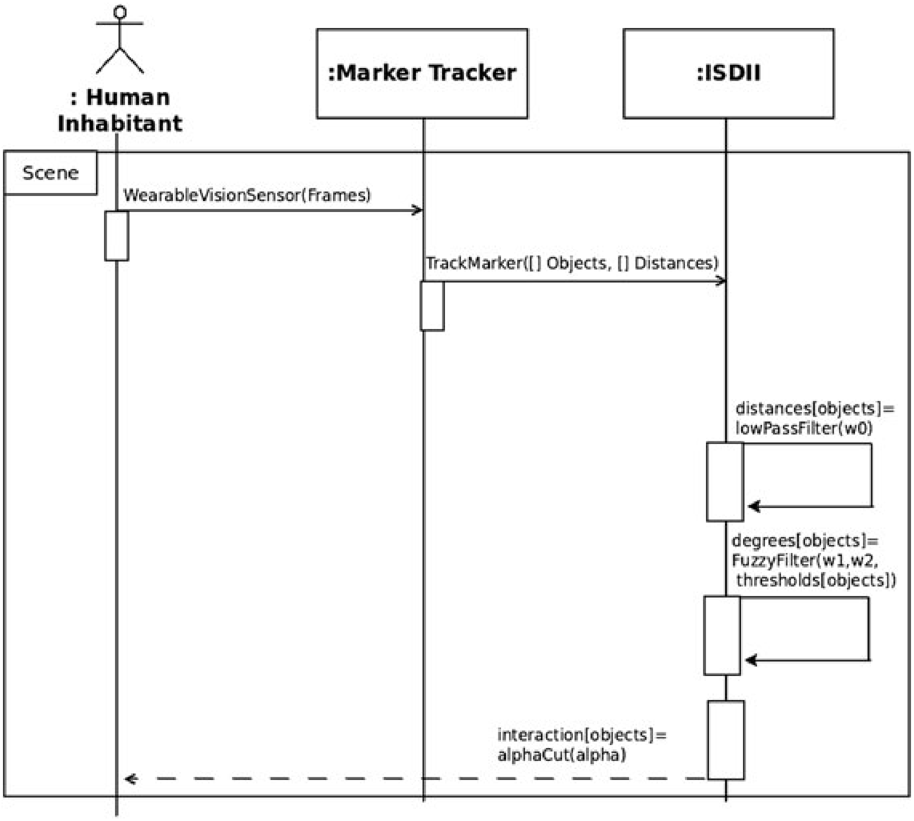


Figure 5. Sequence diagram of detecting object interaction in real-time scenes.

Videos conformed to Google Glass specification and were recorded at 24 fps in mp4 format. The video recordings can be previewed here[¶]: The quantitative findings from the three case scenes are described in Table II. Each scene is represented by the total number of frames, the duration of the scene and the percentage of frames during which an object was correctly identified (true positive rate).

5.1. Analysing algorithm performance

As can be seen from the results, both algorithms provide good performance in low blur and high brightness situations, with Aruco displaying higher accuracy in general. The strength of ORB is its

[¶] https://drive.google.com/file/d/0B_rp8F6H7iwDNFVsUGpxQ1RqeDg/view?usp=sharing

Table I. Breakdown of activities.

Hot chocolate	Hot snack	Washing dishes
Kitchen door	Kitchen door	Kitchen door
Cup cupboard	Fridge	Tap
Fridge	Plate cupboard	Cup cupboard
Microwave	Microwave	Cutlery cupboard
Tea/coffee cupboard	Cutlery cupboard	Tea/coffee cupboard
Cutlery cupboard	Microwave	Plate cupboard
Microwave	Chair	Kitchen door
Tea/coffee cupboard	Kitchen door	N/A
Kitchen door	N/A	N/A

Table II. Scenes and general statistics.

	Parameters			Detection ratio	
Scene	Total frames	duration (s)	Object frames	Aruco (%)	ORB (%)
1	2574	96	658	44.8	25.9
2	1567	52	624	44.8	22.7
3	1663	96	604	36.5	28.3

ability to accommodate low brightness conditions; this is in part due to ORB's implementation of the Harris Corner Detection algorithm, which has been shown to have strong performance in low lighting conditions [32, 33]. An example of favourable and unfavourable conditions regarding movement and brightness are presented in Figure 6. In addition to these statistics, the results from this evaluation will provide the initial threshold distance references for ISDII to be adjusted by an expert.

Tables III, IV and V detail the objects sequentially interacted with during each scene, along with the average distance that each object was detected, the number of frames and duration of frames that the occupant-object interaction took place within. Tables III and IV and V also specifies the lighting conditions during the interaction with each object, along with the calculated distance from the occupant's view point to the marker. Details of the simulated conditions are provided, specifying the amount of motion blur during the interaction and the level of ambient lighting. The detection ratio of ORB and Aruco algorithms are presented, displaying the percentage of frames where an object was detected within the duration window.

5.2. *Adjusting and evaluating ISDII thresholds*

As discussed in Section 4, an initial threshold value for objects was generated during the algorithm evaluation. These values can then be adjusted by an expert to determine at what distance an occupant is determined to be interacting with an object. Table VI details the average distance of detection as found by ISDII as well as the final threshold distance after being modified by a human expert for each object.

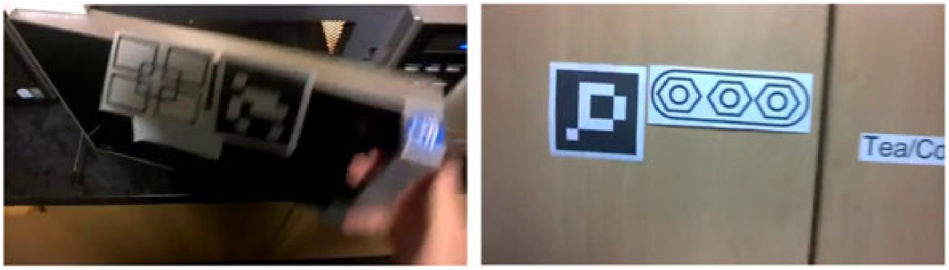


Figure 6. Frames from the wearable vision sensor showing first person view of interactions with objects. A)

Table III. Scene 1 and statistics of object interactions.

Objects	Output			Simulated conditions		Detection ratio	
Interaction order	Avg. distance (m)	Frames	Duration (s)	Brightness	motion blur	Aruco (%)	ORB (%)
Door is opened in	0.36	95	3.96	High	Normal	50.00	43.48
Cupboard-A is opened	0.36	32	1.33	High	Low	78.79	50.0
Cupboard-A is closed	0.19	34	1.42	High	Low	61.29	66.67
Refrigerator is opened	0.29	47	1.96	High	High	56.25	21.28
Refrigerator is closed	0.24	44	1.83	High	High	62.22	45.45
Microwave is opened	0.47	54	2.25	Low	High	3.64	0.00
Microwave is closed	0.37	50	2.08	Low	High	17.65	6.00
Cupboard-B is opened	0.22	38	1.58	Normal	Low	61.54	5.26
Cupboard-B is closed	0.30	49	2.04	Normal	Low	68.00	2.04
Cupboard-C is opened	0.29	29	1.21	Low	High	0.0	31.02
Cupboard-C is closed	0.26	26	1.08	Low	High	0.0	11.54
Microwave is opened	0.44	42	1.75	Low	Normal	13.95	4.76
Microwave is closed	0.37	24	1.00	Low	Normal	24.00	5.88
Cupboard-D is opened	0.31	29	1.21	High	Low	80.00	10.34
Cupboard-D is closed	0.19	35	1.45	High	Low	69.44	2.86
Door is opened out	0.20	125	5.21	Normal	High	44.44	23.33

Table IV. Scene 2 and statistics of object interactions.

Objects	Output			Simulated conditions		Detection ratio	
Interaction order	Avg. distance (m)	Frames	Duration (s)	Brightness	motion blur	Aruco (%)	ORB (%)
Door is opened in	0.35	95	3.96	High	Normal	52.08	24.21
Turn tap on	0.32	101	4.28	Low	Low	39.22	3.79
Cupboard-C is opened	0.21	41	1.71	High	Low	4.76	14.63
Cupboard-C is closed	0 24	24	1.00	High	Low	0.0	14.63
Cupboard-A is opened	0.23	32	1.33	High	Low	78.79	81.08
Cupboard-A is closed	0.18	34	1.42	High	Low	80.00	87.50
Cupboard-B is opened	0.32	54	2.25	Normal	Low	70.91	5.55
Cupboard-B is closed	0.20	35	1.46	Normal	Low	66.67	2.85
Cupboard-D is opened	0.25	45	1.88	Normal	Low	43.48	48.88
Cupboard-D is closed	0.22	35	1.46	Normal	Low	58.33	60.00
Door is opened out	0.32	111	1.46	Normal	High	58.33	7.82

Table V. Scene 3 and statistics of object interactions.

Objects	Output			Simulated conditions		Detection ratio	
Interaction order	Avg. distance (m)	Frames	Duration (s)	Brightness	motion blur	Aruco (%)	ORB (%)
Door is opened in	0.28	58	2.42	High	Low	72.80	22.41
Refrigerator is opened	0.30	48	2.00	High	Low	79.59	45.83
Refrigerator is closed	0.19	29	1.21	High	Low	60.00	44.82
Cupboard-D is opened	0.24	29	1.21	Normal	Low	13.33	87.50
Cupboard-D is closed	0.18	27	1.13	Normal	Low	60.71	88.88
Microwave is opened	0.25	35	1.46	Low	Normal	22.22	11.42
Microwave is closed	0.23	50	2.08	Low	Normal	5.88	0.00
Cupboard-C is opened	0	30	1.25	High	Low	0	46.15
Cupboard-C is closed	0	26	1.08	High	Low	0	50.00
Chair interaction	0.35	159	6.63	Normal	Normal	40.00	13.20
Door is opened out	0.25	111	4.63	Normal	High	22.32	21.52

Table VI. Threshold distances to objects.

Object	Average distance	Final threshold distance
Chair	0.350	0.350
Cupboard-A	0.240	0.235
Cupboard-B	0.260	0.250
Cupboard-C	0.240	0.250
Cupboard-D	0.230	0.235
Door	0.296	0.300
Microwave	0.355	0.355
Refrigerator	0.255	0.255
Tap	0.320	0.320

The precision and recall have been evaluated from the ISDII output against the time window defined by an expert. An interaction has been determined when the interaction degree exceeds $\alpha - cut = 0.95$. The evaluation has included the full range of options for estimating the $\omega_0 \in [0, 1]$, $\omega_1 \in [0, 5]$, $\omega_2 \in [0, 5]$, $\omega_1 < \omega_2$ with a step offset of 0.05. Table VII presents the best precision results from the three scenes in function of ω_0, ω_1, ω_2 and Table VIII displaying the best results for recall. (Table IX).

While the precision results obtained by ISDII determine if an interaction is a true positive are promising, it relies on a high accuracy of detections from the marker detection algorithm in order to return a high recall. The lack of detections in the results from Section 5 results in a low recall which cannot be improved through the filtering and estimation process. The Averaged Ratio Detection (ARD) from the detection algorithm in each scene must match the distance threshold value to be able to analyse the recall obtained by ISDII. This improves the ratio of marker detection due to the exponential smoothing filter. The averaged parameters have been set to allow a comparison of ISDII interaction estimations to expert-defined interaction estimations. The results are displayed in Figure 7, which presents the human expert defined degree of interaction along with an overlay of the ISDII defined interaction.

Table VII. Best precision from scenes in function of ω_0, ω_1, ω_2.

Scene	Precision	ω_0, ω_1, ω_2
1	1.00	[0.95, 0.00, 0.05]
2	0.98	[0.95, 0.00, 0.80]
3	1.00	[0.95, 0.00, 0.60]

Table VIII. Best recall from scenes in function of ω_0, ω_1, ω_2.

Scene	Recall	ARD	*Recall/ARD*	ω_0, ω_1, ω_2
1	0.45	0.43	1.05	[0.95, 0.20, 4.90]
2	0.45	0.47	0.95	[0.95, 0.00, 2.40]
3	0.37	0.34	1.09	[0.95, 0.00, 3.10]

Table IX. Best $F_{\beta} = 1._5$ from scenes in function of ω_0, ω_1, ω_2.

Scene	$F_{\beta = 15}$	ω_0, ω_1, ω_2
1	0.51	[0.95, 0.20, 2.20]
2	0.52	[0.95,0.00,2.45]
3	0.43	[0.95,0.00,1.65]
Average	0.49	[0.95, 0.00, 2.10]

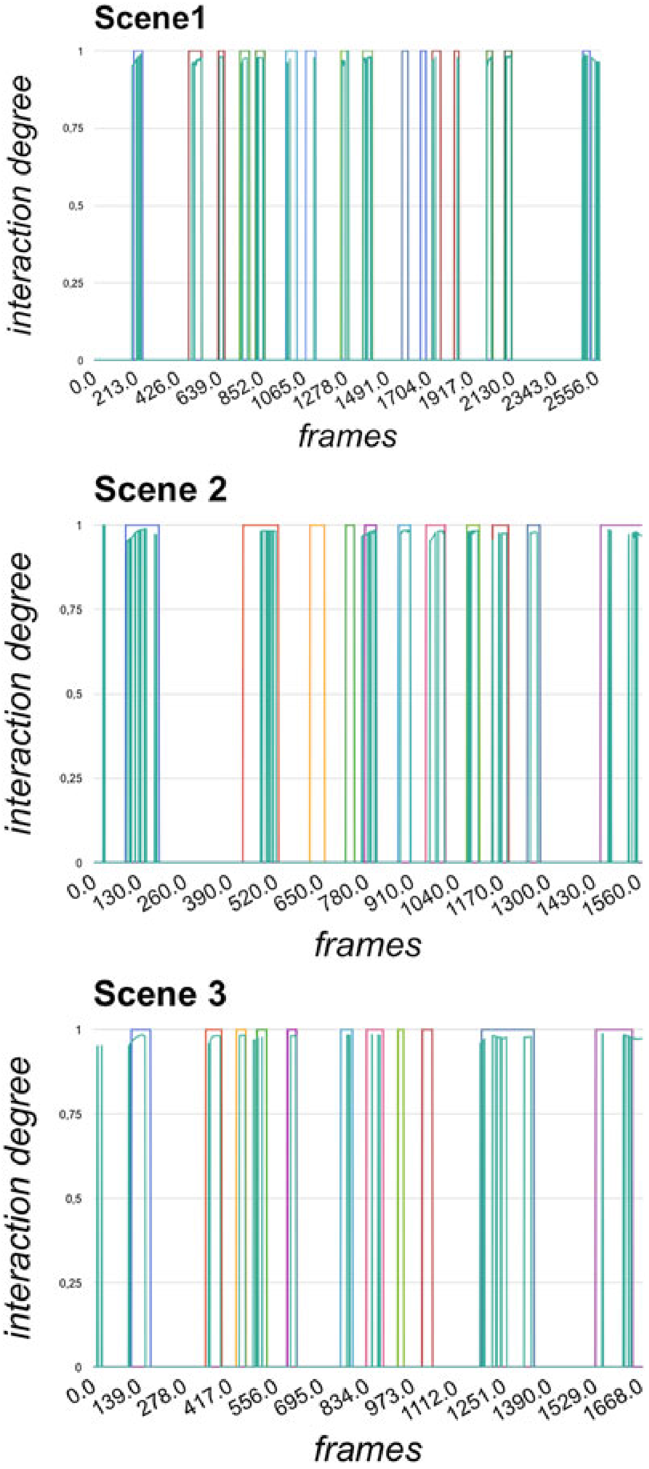


Figure 7. Comparison of ISDII vs human-defined interactions: a) the human-defined interaction is shown by the solid columns and b) the blue line represents the estimation degree as determined by ISDII.

Adjusting the threshold of object interaction offers improved performance when the detection algorithm provides a high rate of detection, as the lack of detections shown in some scenes results in a loss of occupant-object interactions reported from ISDII. The final values of ω_0, ω_1 and ω_2 provide the best averaged parameters in all scenes, and results in a low computational overhead method of determining object interaction, as well as a method of isolating FP.

6. CONCLUSIONS AND FUTURE WORK

The proposed method offers many advantages/innovations over existing methods to determine object interaction within the domain of AAL. One of the methods biggest strengths is the ease to which it is able to be deployed within differing environments, the use of fiducial markers with associated ID's negates the need for specific training to each environment. This is due to the markers being associated with common static items that are commonly found within home environments, with the ID of the object being tied to the marker rather than any features of the object itself. Secondly, the use of a moving camera couple with static objects reduces the issues traditionally seen with a static camera solution such as the limited field of view, which may require the installation of multiple cameras within an environment. Occlusions that may be created through environmental objects, such as doors and large items of furniture, or occlusions generated by the user themselves, such as hands/head/torso occluding objects that they are interacting with [34]. This coupled with being a superior solution for object interaction due to the added advantages a head-mounted camera provides. Firstly, occlusions of the manipulated object tend to be lessened as the object being interacted with is usually the centre of attention for the user [34]. As the object is the centre of the users attention, the object is usually in the centre of the image and in focus, providing a high quality image for processing [34]. Because of the high levels of noise that are typically present in egocentric videos many FP are unavoidable [35]. It can be difficult to identify the correct object as it is possible that multiple objects can be within the occupants' field of view. This is due to some areas of the environment being densely populated with relevant objects, such as the kitchen.

As can be seen in Section 5, a detailed comparison has been carried out on the ORB and Aruco algorithms. The results show that the Aruco algorithm is generally more accurate, with the ORB algorithm providing better performance in extreme light conditions. Based on the information from marker trackers, we have proposed an Intelligent System for Detecting Inhabitant-objects Interaction. It determines if the interaction is a true positive by using two filters: a low-pass filter and a fuzzy filter. A study has been carried out to determine the performance of ISDII, showing an improved precision by removing FP. However, it is highly sensitive to missed detections from the detection algorithm which can result in a deteriorated recall result.

The proposed solution offers a non-intrusive method of detecting occupant object interaction and localisation. The use of a single head-worn camera provides a unique first person view of the environment and their activities, offering additional opportunities within the domain. This solution also minimises the cost in terms of hardware, implementation and maintenance costs associated with alternative solutions, for example, dense sensor placement or static camera approaches. Future work will focus on translating the results to the next generation of wearable vision devices, such as Google Glass 2.0, and the inclusion of the analysis of ISDII commercial markers and tracker developed by companies.**

ACKNOWLEDGMENTS

This contribution has been supported by research projects: UJA2014/06/14, TIN-2012-31263, and the Santander Mobility Scholarship.

REFERENCES

1. Gibson K. Tools, language and intelligence: evolutionary implications. *Man* 1991; 255–264.
2. Giebel CM, Sutcliffe C, Stolt M, Karlsson S, Renom-Guiteras A, Soto M, Verbeek H, Zabalegui A, Challis D. Deterioration of basic activities of daily living and their impact on quality of life across different cognitive stages of dementia: a European study. *International Psychogeriatrics* 2014; **26**(8): 1283–1293.
3. Hong X, Nugent C, Mulvenna M, McClean S, Scotney B, Devlin S. Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing* 2009; **5**(3): 236–252.
4. Sernani P, Claudi A, Palazzo L, Dolcini G, Dragoni A. In *Home Care Expert Systems for Ambient Assisted Living: A Multi-Agent Approach*, The Challenge of Ageing Society: Technological Roles and Opportunities for Artificial Intelligence , 2013.Retrieved from http://ceur-ws.org/Vol-1122/paper1.pdf

** https://developer.vuforia.com/

5. Yuan B, Herbert J. Context-aware hybrid reasoning framework for pervasive healthcare. *Personal and Ubiquitous Computing* 2013; **18**(4): 865–881. DOI:10.1007/s00779-013-0696-5.
6. Yuan J, Tan KK, Lee TH, Choon G, Koh H. Power-efficient interrupt-driven algorithms for fall detection and classification of activities of daily living. *Sensors* 2015; **15**(3): 1377–1387.
7. Chen, L., & Khalil, I. (2011). Activity Recognition: Approaches, Practices and Trends. In *Activity Recognition in Pervasive Intelligent Environments* (Vol. **4**, pp. 131). Doi: 10.2991/978-94-91216-05-3_1
8. Roy, N., Roy, A., & Das, S. K. (2006). Context-aware resource management in multi-inhabitant smart homes: a Nash H -learning based approach. In Fourth Annual IEEE International Conference on Pervasive Computing and Communications (pp. 148–158). Washington. Retrieved from http://www.percom.org/2006/doucments/roymarkweiser.pdf
9. Owen CB, Xiao FXF, Middlin P. What is the best fiducial? *The First IEEE International Workshop Agumented Reality Toolkit* 2002; **15**(11): 3317. DOI:10.1109/ART.2002.1107021.
10. Ma, M., Jain, L. C., & Anderson, P. (2014). *Virtual, Augmented Reality and Serious Games for Healthcare 1 (First)*. Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-54816-1
11. Fiala M. Designing highly reliable fiducial markers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010; **32**(7). DOI:10.1109/TPAMI.2009.146.
12. Pirsiavash H, Ramanan D. IEEE conference on computer vision and pattern recognition. 2012; 2847–2854.
13. Rivera-Rubio J, Alexiou I, Bharath AA. Appearance-based indoor localization: a comparison of patch descriptor performance. *Pattern Recognition Letters* 2015; **66**: 109–117. DOI:10.1016/j.patrec.2015.03.003.
14. Zhang D, Lee DJ, Taylor B. Seeing eye phone: a smart phone-based indoor localization and guidance system for the visually impaired. *Machine Vision and Applications* 2014; **25**(3): 811–822. DOI:10.1007/s00138-013-0575-0.
15. Orrite C, Soler J, Rodrguez M, Herrero E, Casas R. Image-based location recognition and scenario modelling. *International Conference on Computer Vision Theory and Applications* 2015; 216–221. DOI:10.5220/0005352702160221.
16. Hartley, R., & Zisserman, A. (2004). Multiple view geometry in computer vision. *Optics and Lasers in Engineering* (Second, Vol. **37**). Cambridge University Press. doi: 10.1016/S0143-8166(01)00145-2
17. Zeb A, Ullah S, Rabbi I. Indoor vision-based auditory assistance for blind people in semi controlled environments. *Image Processing Theory, Tools and Applications* 2014; 16.
18. Ackerman, E. (2013). Google gets in your face [2013 tech to watch]. *IEEE Spectrum*, **50**(1), 26–29.
19. C. Shewell, C. Nugent, M. Donnelly, & H. Wang. (2016). Indoor localisation through object detection on real-time video implementing a single wearable camera. In Mediterranean Conference on Medical and Biological Engineering and Computing. (Accepted)
20. LiKamWa, R., Wang, Z., Carroll, A., Lin, F. X., & Zhong, L. (2014). Draining our glass. In Proceedings of 5th Asia-Pacific Workshop on Systems (pp. 17). ACM. doi: 10.1145/2637166.2637230
21. Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: an efcient alternative to SIFT or SURF. In International Conference on Computer Vision (pp. 25642571). Barcelona: IEEE.
22. Rosin P. Measuring corner properties. *Computer Vision and Image Understanding* 1999; **73**(2): 291–307. DOI:10.1006/cviu.1998.0719.
23. Cheng, J., Leng, C., Wu, J., Cui, H., & Lu, H. (2014). Fast and accurate image matching with cascade hashing for 3D reconstruction. In *Computer Vision and Pattern Recognition* (pp. 18). Columbus, OH: IEEE Comput. Soc. doi: 10.1109/CVPR.2014.8
24. Garrido-Jurado S, Muoz-Salinas R, Madrid-Cuevas FJ, Marn-Jimnez MJ. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 2014; **47**(6): 2280–2292.
25. Garrido-Jurado, S., Muoz-Salinas, R., Madrid-Cuevas, F. J., & Medina-Carnicer, R. (2015). *Generation of fiducial marker dictionaries using Mixed Integer Linear Programming*. Pattern Recognition.
26. Brown RG. In *Exponential Smoothing for Predicting Demand*, (Little AD ed.).: Cambridge, Mass, 1956.
27. Holt CC. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting* 2004; **20**(1): 5–10.
28. Masry E. Alpha-stable signals and adaptive filtering. *IEEE Transactions on Signal Processing* 2000; **48**(11): 3011–3016.
29. Brookner E. In *Tracking and Kalman filtering made easy*, Wiley: New York, 1998.
30. Zadeh LA. Fuzzy sets. *Information and Control* 1965; **8**(3): 338–353.
31. Mendel JM. Uncertainty, fuzzy logic, and signal processing. *Signal Processing* 2000; **80**(6): 913–933.
32. Gil A, Mozos OM, Ballesta M, Reinoso O. A comparative evaluation of interest point detectors and local descriptors for visual SLAM. *Machine Vision and Applications* 2010; **21**(6): 905–920. DOI:10.1007/s00138-009-0195-x.
33. Pibyl B, Chalmers A, Zemk P. Feature point detection under extreme lighting conditions. *Conference on Computer Graphics* 2012; 156–163.
34. Nguyen, THC., Nebel, JC., & Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: a review, sensors (Switzerland), vol. **16**, no. 1.
35. Xiong, B., Kim, G., & Sigal, L. (2015). Storyline representation of egocentric videos with an applications to story-based search, in IEEE International Conference on Computer Vision, pp. 4525–4533.