View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Ulster University's Research Portal

Indoor Localisation through Object Detection on Real-Time Video Implementing a Single Wearable Camera.

Colin Shewell[1], Chris Nugent[1], Mark Donnelly[1] and Haiying Wang[1]

[1] Computer Science Research Institute and School of Computing and Mathematics, University of Ulster, Newtownabbey, United Kingdom

Abstract— **This paper presents an accurate indoor localisation approach to provide context aware support for Activities of Daily Living. This paper explores the use of contemporary wearable technology (Google Glass) to facilitate a unique first-person view of the occupants environment. Machine vision techniques are then employed to determine an occupant's location via environmental object detection within their field of view. Specifically, the video footage is streamed to a server where object recognition is performed using the Oriented Features from Accelerated Segment Test and Rotated Binary Robust Independent Elementary Features algorithm with a K-Nearest Neighbour matcher to match the saved keypoints of the objects to the scene. To validate the approach, an experimental set-up consisting of three ADL routines, each containing at least ten activities, ranging from drinking water to making a meal were considered. Ground truth was obtained from manually annotated video data and the approach was subsequently benchmarked against a common method of indoor localisation that employs dense sensor placement. The paper presents the results from these experiments, which highlight the feasibility of using off-the-shelf machine vision algorithms to determine indoor location based on data input from wearable video-based sensor technology. The results show a recall, precision, and F-measure of 0.82, 0.96, and 0.88 respectively. This method provides additional secondary benefits such as first person tracking within the environment and lack of required sensor interaction to determine occupant location.**

Keywords— **Ageing in Place, Ambient Assisted Living, Context-Aware Services, Machine Vision, Wearable Computing.**

I. Introduction

The remarkable increase in life expectancy can be viewed as one of the greatest achievements of the 20th century. As a result the oldest (aged 65 plus) in society are now regarded as the most rapidly expanding group within the population [1]. This has resulted in a surge in the increasing numbers of age related conditions, such as dementia and general cognitive decline associated with ageing. One solution to address the care provision required by these is postulated to involve technology based smart environments that have the ability to support ageing-in-place, otherwise known as Ambient Assisted Living (AAL). This solution aims to afford inhabitants the ability to remain within their own home for longer, and to maintain an acceptable level of quality of life. Thereby delaying the requirement to be re-situated within full time care facilities [1].

Over recent years 'smart' technologies for use within smart homes have gained increasing usage and acceptance, in particular, due to the widespread adoption of smart-phones, along with the introduction of wearable technology to the consumer market. This has stimulated new opportunities within the domain of pervasive computing, particularly with the advent of head-mountable wearables such as Google Glass, SmartEyeglass, and the M100. These provide a unique ability to obtain a first-person view of an inhabitant's activities and their environment. This offers opportunities within the domain of indoor localisation due to the view-point that this method provides, however, with it are an associated number of challenges such as the computational complexity that machine vision processing demands.

This paper proposes a solution to facilitate indoor localisation through the use of a single 'always on' wearable camera, which has been implemented using the Google Glass platform. Occupant location is determined using machine vision techniques that identify reference objects located within the environment which are then cross-referenced against a knowledge base that contains the objects known location. This will allow support to be given in the form of notifications/reminders in order to assist with Activities of Daily Living (ADL). This solution aims to improve context aware support through the localisation of objects within a smart environment.

The remainder of the paper is structured as follows. Section II outlines related work within the field of indoor localisation, focusing on those that use machine vision techniques. Section III discusses the methodology used, presenting an overview of the system in addition to more detailed information regarding the feature point detection and matching algorithms used. Along with a description of the routines used to carry out the experiment. Section IV presents and discusses the results that are also benchmarked against a dense sensor solution. Finally Section V provides a set of conclusions that critique these early findings and outlines the plans for future work.

II. Related Work

This Section presents a summary of the current state-of-the-art of machine vision based solutions that facilitate indoor localisation. A number of studies are reviewed, which have a focus on applying contemporary technology using machine vision techniques within the domain of AAL. The findings are promising, however, several challenges are highlighted which will need to be addressed. Dense sensor solutions are also reviewed to provide a basis for benchmarking the proposed system.

Okeyo *et al.* developed a dense sensor based solution incorporating multi-agents in order to provide services to user's within smart homes [2]. Sensors were placed on specific objects that the user would interact with which would then record the time and location associated with that sensor in order to build contextual information. While the overall results were high (1.00, 0.88, 0.88 for precision, recall, and accuracy, respectively) it still suffers from the inherent problems that exist with dense sensor based methods, such as multiple occupancy and the need for sensor interaction.

Rahal *et al.* implemented a system using anonymous dense sensor placement along with Bayesian filtering in order to determine occupant location [3]. The system was tested using a scenario of an occupants daily routine, the routine was performed by 14 subjects, one at a time. The system showed a mean localisation accuracy of 0.85, as the authors note however the system is only capable of supporting a single occupant [3].

Leotta and Mecalla [4] developed PLaTHEA (*P*eople *L*ocalization *a*nd *T*racking for *H*om*E* *A*utomation). PLaTHEA is a machine vision based system that acquires a stereo video stream from two network attached cameras in order to provide support for AAL. Two cameras are placed in each room, working in stereo, in order to ensure that as much of the room is covered and that occlusions are reduced. Foreground extraction is then performed in order to determine if occupants are present in the scene. PLaTHEA also performs identity recognition through the use of facial recognition. Facial recognition is performed using SIFT features from each face pose, which are then stored within a kd-tree data structure. At runtime a Haar classifier is applied in order to detect faces in the scene; when a face is detected SIFT features are extracted and compared to the saved features stored in the kd-tree for recognition [4]. Nevertheless, there are some limitations to the PLaTHEA system, due to the system relying on static cameras it may not be possible to ensure that the entirety of the room is viewable or that occlusions may not occur due to the opening of doors, large furniture, *etc.* In addition, an issue that was identified by the authors, were when the system was monitoring a room with a wall greater than 10 metres then it was not possible to monitor without the use of costly acquisition hardware [4]. While the issue of cost is being addressed there is also the additional cost of having to install multiple cameras within each room that support is provided within. There is also the issue of multiple occupancy, due to the use of foreground extraction to identify occupants, while this is partially mitigated through the use of facial recognition it also requires that all the occupants are known and have SIFT features saved within the system [4]. There is also the additional problem of the Haar classifier being reliant on the occupants eye's being clearly seen by the camera as this method of face detection will usually fail if the eyes are occluded [5].

Rivera-Rubio *et al.* [6] developed a system that estimated the user's location through scene recognition. The experiment was carried out using an LG Google Nexus 4 and Google Glass. A dataset was gathered of the locations by recording a video of the occupant walking through the location ten times whilst wearing a recording device (50% split between the Nexus 4 and Glass). This included a combination of day/night acquisitions and occasional strong lighting from windows. The system was tested using multiple descriptor methods (three custom designed and three standard methods) following a standard bag-of-words and kernel encoding pipeline, with HOG3D matching used as a baseline [6]. Results show errors as low as 1.6 metres over a 50 metre distance were achieved, however, for the purposes of AAL a greater level of refinement is required in order to distinguish where in a room the occupant is located and if possible what they are interacting with in order to provide relevant support.

Zhang *et al.* [7] proposed a method of indoor location using still images captured at intervals from a smart-phone worn on a lanyard. This system has the goal of assisting those with impaired vision to navigate within an indoor environment. The system relies on collecting map data of a building, that describe features/descriptors along with their 3D co-ordinates, floor plans, and other location data. Images are then captured and sent at intervals from the smart-phone to a server for processing. Images are then matched against the template map of the building in order to determine location and offer directions should the user require them. Whilst this system works well for its intended use there are limitations when applied to an AAL situation. One problem, that the authors noted, was that there were null spots, were there was not enough features to create a map image, such as when the user makes a 90° turn, for example in a hall way or entering a room [7]. One other possible issue for an AAL application is that of the intermittent image capture that may result in missing key information, such as a room transition or an interaction with an appliance, which could be vital for context.

The presented system will use a head-mounted wearable camera streaming a live video feed, this should reduce occlusions and hope to reduce missing key information that an intermittent system may produce. Along with a greater refinement in the user's location to assist in providing increasingly timely and relevant support.

III. Methodology

Our proposed approach employs off-the-shelf machine vision tools to facilitate the detection of objects. Specifically the OpenCV Oriented FAST and Rotated BRIEF (ORB) algorithm for feature detection and descriptor extraction have been used. This is paired with a Brute-Force matcher to determine when the object of focus is present in the video stream. It is hypothesised that the use of a single wearable camera to determine the inhabitant's location may facilitate inhabitant tracking within an environment. This may be used to provide enhanced contextual information based on their location. This approach has the advantage of reducing the set-up costs associated with alternative location tracking approaches, such as dense sensor placement [8]. This is achieved using machine vision techniques to identify reference objects within the patients field of view that are then cross-referenced against a knowledge base which indicates the room that the objects are located within. A high level overview of the process is shown in Figure 1, consisting of a pre-processing section where the marker templates are learned and the real-time processing section where the learned templates are matched against the real-time video feed in order to provide marker/object detection. The system was tested in the Smart Environment Research Group (SERG) smart living space which consisted of a fully sensorised kitchen and living room [9]. The environment contains a suite of sensor technology, including PIR sensors, contact sensors, and floor pressure sensors. The presented method was benchmarked against a dense binary sensor deployment consisting of 14 individual sensors.

I. Machine Vision System

As wearable devices are traditionally 'resource poor' in comparison with contemporary server hardware [10] Google Glass is responsible for capturing the video stream and delivery of reminders/notifications only. The image processing is offloaded to a server via Real Time Streaming Protocol (RTSP) for processing (Figure 1), thus decreasing the time taken for object detection and for the appropriate response to be given, along with increasing battery life on the Glass platform. To aid in the correct identification of objects unique markers where applied to the objects of interest, as shown in Figure 2a. This allows a custom identifier to be placed on each marker to distinguish between objects, as shown in Figure 2b. The unique markers are learnt during a pre-processing stage where the ORB feature points are detected and stored.

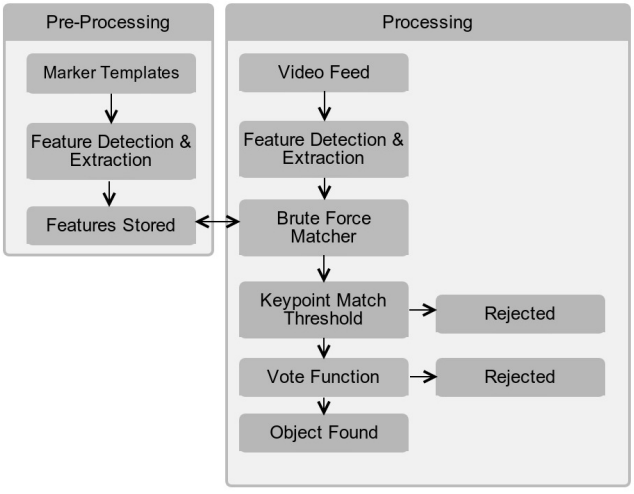
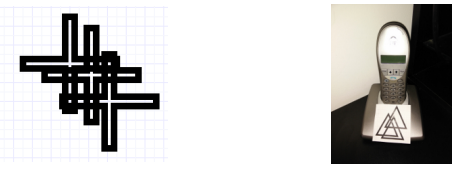


Figure 1: High level overview of machine vision system processing - consisting of a pre-processing section and a real-time processing section.

The use of markers also reduces some of the issues traditionally faced when performing object recognition, such as variations between the same objects – *i.e.* different models of appliances. Furthermore this also alleviates the problem of distinguishing between multiple identical objects in close proximity, such as kitchen cupboards/drawers [11].



(a) Example of a unique marker. (b) Marker applied to object.

Figure 2: Image (a) is an example of the markers used, Image (b) shows how the marker is applied to an object of interest, in this case a telephone.

The experiment was carried out using the OpenCV library on an Intel Core2Quad (Q9950) 2.83GHz machine, the video was transmitted at a resolution of 640x480 by Google Glass (OMAP4430 ARMv7 CPU with 773mb RAM) at 20fps. Due to the processing limitations of Google Glass a variable lag (<3s) was introduced on the stream. The lag was due to the Glass's efforts to lower the operating temperature which it achieves by reducing the clock speed of the CPU [10]. The CPU can be set to four frequencies – 300Mhz, 600Mhz, 800MHz, and 1GHz. At high temperatures the Glass firmware

limits the CPU to 600Mhz or 300MHz in order to cool down via power reduction [12].

II. ORB Feature Points and Descriptors

The chosen method of detection and extraction of feature points and descriptors is the ORB keypoint detector/extractor which was developed by Rublee *et al.* [13]. The ORB algorithm uses FAST (Features from Accelerated Segment Test) in pyramids in order to detect stable keypoints and selects the strongest features using FAST. ORB implements a simple method of corner detection, the intensity centroid as defined by Rosin [14].

III. K-Nearest Neighbour Matching

A K-Nearest Neighbour (KNN) algorithm is used to match the feature points to determine if an object is present. A simple version of an KNN is used – a Brute-Force matcher. While a Brute-Force matcher is one of the worst performing matchers in terms of time taken to establish a match (detection time as implemented is still less than one second) it is also the best performer in terms of accurately identifying the correct matches as found in [15] which benchmarked multiple algorithms for the purposes of image matching. A formal representation of a KNN algorithm finds the K closest (similar) features to a query feature among N points in a d-dimensional feature space [16]. Within this implementation the Brute-Force matcher is used to compare feature points for matching pairs, for each feature in the object the matcher finds the closest feature in the scene by trying each one. The similarity between two pairs is represented by Norm Hamming distance. A minimum Hamming distance is set to ensure that only good matches are selected. A match is considered good when the distance is less than three times the minimum Hamming distance set.

In order to dismiss the number of False Positives (FP – where an object is determined to be present when it is not) reported by the system a two stage filter was used. For the first stage the homography was used as a model for correct matches ('Keypoint Match Threshold' in Figure 1). The number of inliers that contributed to the homography were determined and compared against a threshold value, if the number of inliers match or exceed this value then it is passed onto the second stage. The second stage is a Vote Function where any further FP that have passed through the first stage are removed, this is performed by a vote count on what the object detected is thought to be, once the count has reached a pre-determined threshold value the object is determined to be present.

Table 1: Full list of activities that were performed during the three routines.

Full Activity List	
1.1	Prepare/drink water
1.2	Prepare/drink tea
1.3	Prepare/drink hot chocolate
1.4	Prepare/drink milk
2	Make/receive phone call
3.1	Prepare/eat cold meal
3.2	Prepare/eat hot meal
4	Watch TV
5	Wash dishes

IV. Contact Sensors

Dense sensor placement have also been used as a benchmark in order to provide a comparison with the machine vision system. This consists of TyneTec binary contact sensors that were placed on the same objects that also have a unique machine vision marker placed on them. There was a total of 14 TyneTec sensors which uploaded events to a MySQL database for retrieval. All components of the system were time synced with a MySQL server in order to ensure that the events were synchronised.

V. Experiment Routines

A range of activities were carried out that were representative of daily routines, with the goal of recognising the component locations that consist each activity. If prepare/drink water is taken as an example activity, then the component locations would be the kitchen door, the cup cupboard, the tap, and then finally the kitchen door again. Three routines were created, the first containing ten activities and the remaining two containing eleven activities. These ranged from simple activities such as drinking a glass of water to more complex activities, such as preparing hot food. The activities are presented in Table 1, with the full routines presented in Table 2.

These routines were performed under the same lighting conditions in order to minimise any potential discrepancy between identical activities in differing routines. In order to ensure the accuracy of the machine vision and binary sensor location systems, the ground truth was obtained from a time stamped video. The inhabitant's location reported from the location systems where then compared to the ground truth from the video.

Table 2: Breakdown of activities that took place in each routine.

Routine 1 (R1)	Routine 2 (R2)	Routine 3 (R3)
1.3	1.4	1.3
1.1	3.1	1.1
3.2	1.1	2
5	2	3.2
4	1.1	1.1
1.1	1.2	4
4	4	1.2
3.1	3.2	4
5	5	3.1
1.1	4	5
N/A	1.1	1.4

Table 3: Results of Recall, Precision, and F-Measure for the machine vision based system.

Routine	Total Events	Recall	Precision	F-Measure
R1	58	0.74	0.98	0.84
R2	56	0.88	0.94	0.91
R3	61	0.84	0.96	0.89
Total	175	0.82	0.96	0.88

IV. Results and Discussion

This Section describes the results of the machine vision localisation system, along with details on the results from the dense sensor system when compared with the ground truth from the annotated video data. Due to the high number of True Negatives (TN), over twenty thousand, from the machine vision system a skewed dataset was produced. Due to this accuracy was determined by measuring recall, precision, and F-Measure. These will be focused on to avoid the high number of TN giving an incorrect weighting to the results.

The results from the machine vision system are presented in Tables 3 and 4, and the results from the binary contact sensors are presented in Tables 5 and 6. As shown in Table 4 there is a total of 32 False Negatives (FN – where an object was present but not detected), the majority of these (11) were due to corruption within the video frame during transmission, the rest of the FN's where due to varying reasons, such as missing frames.

Table 4: Breakdown of machine vision sensor classification outcomes including TP, FN, and FP.

Routine	Total Events	TP	FN	FP
R1	58	43	15	1
R2	56	49	7	3
R3	61	51	10	2
Total	175	143	32	6

Table 5: Results of Recall, Precision, and F-Measure for the dense sensor based system.

Routine	Total Events	Recall	Precision	F-Measure
R1	58	1.00	1.00	1.00
R2	56	0.93	1.00	0.96
R3	61	0.90	1.00	0.95
Total	175	0.94	1.00	0.97

Table 6: Breakdown of dense sensor classification outcomes including TP, FN, and FP.

Routine	Total Events	TP	FN	FP
R1	58	58	0	0
R2	56	52	4	0
R3	61	55	6	0
Total	175	165	10	0

Table 7 presents a breakdown of the missed events by the machine vision system along with an attempted explanation as to why the events were missed. It should also be noted that eight sensor events were missed due to a battery failure part way through the experiment; there were three such events in R2 and five events in R3 that were missed.

While the binary contact sensors provided more accurate results this does not fully demonstrate the additional advantages the machine vision system provides over dense sensor placement. One of the key advantages this method offers is that interaction with an object is not required in order to determine the occupant's location within the environment which can offer a more timely location update compared to dense sensor placement. In the experiment the occupant's location was reported before they had interacted with the object thus offering a more timely update. Also if the occupant became confused or decided not to use the object their location would still be captured. This would have otherwise been lost in a traditional sensor based smart environment. Another potential

Table 7: A breakdown of FN machine vision events.

Cause	FN
Corrupt frame	16
Other	8
Unknown	8
Total	32

advantage is that of multiple occupancy, as each occupant will use a wearable device it would be possible to locate each occupant within the environment and to infer their activity from their own first person view. Nevertheless, this is working under the assumption that only the occupants of the environment will require support, as any visitors will not have a wearable device. If any sensor activity is detected without a corresponding machine vision event then it would be assumed that the visitors have activated a sensor and thus that event should be ignored.

V. Conclusion

A method of indoor localisation has been presented utilising a wearable camera to determine location based upon objects viewed within a scene. This was compared with a common method of indoor localisation (dense sensor placement) employing annotated video data as the ground truth. Thus supporting the hypothesis that the use of a single wearable camera allows inhabitant tracking within an environment with the goal of determining location. While the machine vision results were not as accurate as the dense sensor placement, they demonstrated that the proposed method is viable and offers other secondary advantages that are unique to this method, such as the first person view and lack of required interaction. However, there are some limitations to using such as static approach to storing the objects location within a knowledge base, such as objects being moved or certain objects that may not have a static location, for example personal devices. Future work will involve determining activity based on the objects located within the field of view, through the use of a rule-based system in order to provide support for those activities through the use of a multi-agent system with each agent governing an activity in order to provide specific support for said activity. The long term aspiration of this system is to assist those in cognitive decline with their ADL, such as in the event the occupant has became confused with a task part way through, for example making a meal; assistance could then be provided to allow the continuation of the task.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Kobayashi L. C., Wardle J., Wagner C.. Internet use, social engagement and health literacy decline during ageing in a longitudinal cohort of older English adults *Journal of Epidemiology & Community Health.* 2014;69:278–283.
2. Okeyo George, Chen Liming, Wang Hui. An Agent-mediated Ontology-based Approach for Composite Activity Recognition in Smart Homes *Journal of Universal Computer Science.* 2013;19:2577–2597.
3. Rahal Youcef, Pigot Hélène, Mabilleau Philippe. Location estimation in a smart home: System implementation and evaluation using experimental data *International Journal of Telemedicine and Applications.* 2008;2008:9.
4. Leotta Francesco, Mecella Massimo. PLaTHEA: a marker-less people localization and tracking system for home automation *Software - Practice and Experience.* 2014;39:661–699.
5. Viola Paul, Jones Michael J. Robust Real-Time Face Detection *International Journal of Computer Vision.* 2004;57:137–154.
6. Rivera-rubio Jose, Alexiou Ioannis, Bharath Anil, Secoli Riccardo, Dickens Luke, Lupu Emil C. Associating locations from wearable cameras in *British Machine Vision Conference*:1–13 2014.
7. Zhang Dong, Lee Dah Jye, Taylor Brandon. Seeing Eye Phone: A smart phone-based indoor localization and guidance system for the visually impaired *Machine Vision and Applications.* 2014;25:811–822.
8. Hightower Jeffrey, Borriello Gaetano. Location Systems for Ubiquitous Computing *Computer.* 2001;34:57–66.
9. Nugent C.D., Mulvenna M.D., Hong X., Devlin S.. Experiences in the development of a Smart Lab *International Journal of Biomedical Engineering and Technology.* 2009;2:319.
10. Ha Kiryong, Chen Zhuo, Hu Wenlu, Richter Wolfgang, Pillai Padmanabhan, Satyanarayanan Mahadev. Towards wearable cognitive assistance in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*:68–81ACM 2014.
11. Fiala Mark. Designing highly reliable fiducial markers *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2010;32:1317–1324.
12. LiKamWa Robert, Wang Zhen, Carroll Aaron, Lin Felix Xiaozhu, Zhong Lin. Draining our glass in *Proceedings of 5th Asia-Pacific Workshop on Systems*:1–7ACM 2014.
13. Rublee Ethan, Rabaud Vincent, Konolige Kurt, Bradski Gary. ORB: an efcient alternative to SIFT or SURF in *International Conference on Computer Vision*(Barcelona):2564–2571IEEE 2011.
14. Rosin Paul. Measuring Corner Properties *Computer Vision and Image Understanding.* 1999;73:291–307.
15. Cheng Jian, Leng Cong, Wu Jiaxiang, Cui Hainan, Lu Hanqing. Fast and Accurate Image Matching with Cascade Hashing for 3D Reconstruction in *Computer Vision and Pattern Recognition*(Columbus, OH):1–8IEEE Comput. Soc 2014.
16. Verma Deepika, Kakkar Namita, Mehan Neha. Comparison of Brute-Force and K-D Tree Algorithm *International Journal of Advanced Research in Computer and Communication Engineering.* 2014;3.
17. Kosaka Toru, Ohtsubo Yoshiaki, Mehuro Hiroshi. Distance-measuring appartus for camera 1991.