The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Structure Learning for Activity Recognition in Robot Assisted Intelligent Environments

Douglas G. McIlwraith, Julien Pansiot, James Ballantyne, Salman Valibeik, Ahmed Elsaify and Guang-Zhong Yang

Abstract—**This paper presents a novel structure learning algorithm for the creation of distributed Bayesian networks over static and mobile Vision Sensor Network (VSN) nodes. These compose an assistive, intelligent environment for activity recognition. We provide results demonstrating a higher level of accuracy in the recognition of fine motor tasks when the environment is augmented with a mobile robot and show the ability of our learning algorithm to reduce VSN communication compared to a naïve, greedy structure learning technique.**

I. Introduction

DUE to recent advances in medical care and the adoption of increasingly healthy lifestyles, we are witnessing a demographic shift towards an increasingly aged population [1]. Consequently, considerable interest has been directed toward research into supportive environments which enable the elderly and infirm to live in their own home for longer. Tentative research has focused on the determination of activities and behavior – since significant changes in either can indicate the onset of certain diseases such as Alzheimer's or dementia. This could be through networks of privacy respectful ambient cameras [2, 3] or wearable sensors [4]. However, robot assisted environments are in a unique position to provide solutions for elder monitoring – since self navigating robots can provide high quality data on subject pose regardless of location within the environment. Furthermore, we may also see robots that provide a form of companionship and aid the elderly in achieving daily tasks [5].

Where environments are to contain multiple ambient sensors, installation may be performed by a visiting carer or those living within the domicile. Consequently, it is unreasonable to expect these to be located at optimal locations for the determination of individual activities. Furthermore, since each dwelling is unique, their relative positioning can not be assumed prior to installation thus there is a strong requirement for such networks to be self configuring. To this end, we provide a structure learning algorithm for Bayesian networks which is considerate of both inference and communication cost within ambient Vision Sensor Networks (VSNs). Using Pearl's message propagation algorithm, activity inference can be implemented in a distributed manner over the VSNs, without the requirement for a centralized data repository. Where assistive robots are present, our algorithm can seamlessly incorporate such data to augment recognition accuracy. We demonstrate the efficacy of this algorithm in a home healthcare scenario for fine motor tasks occurring at several locations within the environment.

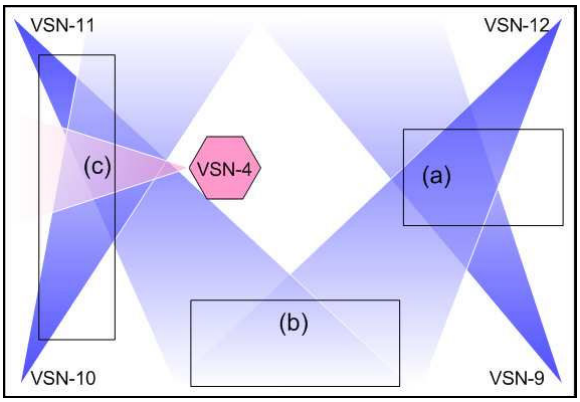

Fig. 1. Camera configuration within the sensing environment. Areas denoted by (a), (b) and (c) are seated *stations* within the experiment, whilst VSN-9 through VSN-12 are fixed, wall-mounted nodes. VSN-4 is mounted on a mobile robotic platform.

II. Related Research

For detecting Activities of Daily Living (ADLs), omni-directional cameras [6] have previously been employed to capture behavioral patterns in a household environment. For example, a system operating at multiple resolutions has been defined, with a wide angle camera directing the pan, tilt and zoom of other cameras [7]. In previous work [3], we have discussed activity recognition within the home and provided results demonstrating how the fusion of ear worn and ambient sensors can increase the accuracy of activity recognition. With this approach, certain fine motor activities could not be readily distinguished, for example, *reading* and *eating*. Robot assisted intelligent spaces [8, 9] may provide a suitable solution, allowing detailed pose information to be garnered using mobile agents, regardless of subject position within the environment. The use of ambient sensors to plan

Manuscript received August 1st, 2009. D. G. McIlwraith and A. Elsaify are with the Royal Society/Wolfson MIC Laboratory, Dept. of Computing, Imperial College London, UK (020-7594-8337; e-mail: dm05@doc.ic.ac.uk).

J. Pansiot, S. Valibeik and Guang-Zhong Yang are with the Royal Society/Wolfson MIC Laboratory, Dept. of Computing and the Institute of Biomedical Engineering, Imperial College London, UK.

J. Ballantyne is with the Division of Surgery, Oncology, Reproductive Biology and Anaesthetics, Imperial College London, UK.

robot trajectories around subject movement [10] and human following [8] have both been addressed in this context, amongst other navigation approaches [11]. Structure learning for robot localization is addressed in [12], however all features are locally obtained by the robot. Structure learning under constraints is not new [13] but to our knowledge, limited research has considered the physical location and utility of individual features to perform real time communication and inference in a robot assisted intelligent environment.

III. Intelligent Environment

The intelligent environment proposed within this work consists of four VSN nodes mounted at each corner of our home health monitoring laboratory, Fig. 1. This laboratory has been designed to simulate a room within a typical home and includes a dining table (a), study area (b) and sofa (c). Cameras are arranged to provide coverage for particular areas with some overlap, however they are not optimal for all regions. For example, it can be seen that VSN-9 provides no coverage of the sofa at station (c). In addition, a VSN node is also mounted upon a mobile robot programmed to follow subjects within the environment.

A. Video Sensor Nodes

The VSN nodes comprise an Omnivision OV9655 1.3 megapixel camera, a 500MHz Analog Devices Blackfin BF537 Processor, 256MB SDRAM, 32MB SPI Flash, and a Lantronix Matchport WLAN 802.11g/b Wi-Fi board for wireless communication. For each individual pixel captured by the device a statistical model is built and maintained to account for slow changes in light conditions as well as object displacement. For this purpose, a mixture model of three Gaussian distributions is employed, as suggested by Lee [14]. Once segmentation has been performed, further erosion and dilation filters are applied in order to remove inherent high-frequency noise, see Fig. 2. The center and axis-aligned

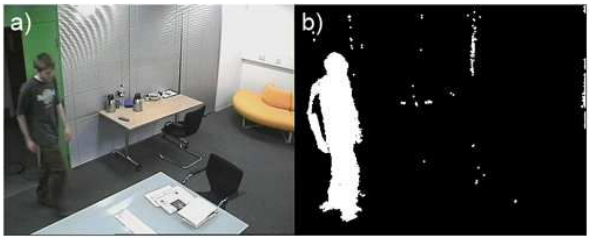

Fig. 2. Camera View from VSN-12 before (a) and after (b) background segmentation.

bounding boxes (AABB) of the subject are then computed, followed by the Oriented Bounding Boxes (OBB), based upon Principal Component Analysis (PCA).

In order to discount the mobile robot when generating features from the wall mounted VSN nodes, a luminous skirting was applied to the robot's exterior, Fig. 3. The colour of this skirting is known in advance by the environmental sensors, allowing the robot to be discounted from the background model during both learning and testing phases. In order to obtain ground truth data, video is collected from all nodes before off-node processing is performed. On-node versions of our algorithms have been implemented [15] but are not used due to data segmentation practicalities.

B. Mobile Robot

The Peoplebot Robot [16], Fig. 3, is equipped with an autonomous navigation system using a time-of-flight camera [17]. The system locks onto the person by building a shape descriptor and tracks using an Interacting Multiple Model Filter (IMMF) [18]. The robot is programmed to keep the person centered in the time-of-flight range map and maintain a distance of 1.5m. Path-planning uses a kinematic dynamic window approach to build a safe and accurate path to the person. In order to ensure the quality of features obtained from the robot, background models are learnt only once the robot is static. In addition to the features generated by fixed VSN nodes, optical flow is also generated using the method proposed by Horn and Schunck [19].

IV. Structure Learning Algorithm

We propose a greedy structure learning algorithm based upon the Bayesian Information Criterion (BIC) [20, 21, 22], which maximizes the information gain of a structure, whilst minimizing the communication required in obtaining remote features. During learning, the algorithm first performs a *preprocessing step,* before *structure learning*. During the final phase, the accuracy of inference is evaluated, which can be used to determine the processor responsible for maintenance of activity recognition at a particular station.

A. Preprocessing

For each data set and for each class, the set of cameras and associated feature sets are identified through analysis of person presence within the field of vision. Processing units within these cameras then become candidates for the management of these activities – known as the *candidate set*, and the union of their associated features is the maximal feature set that can be used to infer activities at this location.

B. Structure Learning

An individual VSN node can build a Bayesian network which is used to infer the activities for which it is a candidate. In order to reduce the size of conditional probability tables that are stored, the number of causal ancestors for a given node is initially limited to 2 and these must be obtained locally. At each iteration of the algorithm the BIC gain, Equation 2, is evaluated at each potential

branch point and with each unused feature. The structure which exhibits the greatest gain is then chosen.

$$BIC(G,\Theta) = \log P(D \mid G,\Theta) - \frac{\log N}{2}\dim(G) \qquad (1)$$

$$\Delta BIC(f) = BIC(\{f,G\},\Theta^{ML}_{\{f,G\}}) - BIC(G,\Theta^{ML}_{G}) \qquad (2)$$

Equation 1 [20, 21, 22] presents the BIC score. D denotes the set of data points and N the magnitude of this set. G denotes the graph structure under investigation whilst, Θ^{ML}_{G} denotes the maximum likelihood parameters for graph G with data D. To maximize this equation, G must model D without over-fitting. This is achieved through the second term penalty factor, which rewards low complexity structures through the dimensionality function **dim**. Equation 2 illustrates the information gain achieved when feature f is included within network G, although the structure of the composed graph is not presented.

Fig. 3. The Peoplebot Robot. Note the luminous skirting for background segmentation as well as the time of flight camera, used for navigation, mounted at the top of the robot.

Where $\max(\Delta BIC)$ drops below a fraction, α, of the gain achieved during the first iteration, there is a high probability that additional local features will not contribute to the accuracy of the classifier regardless of their location in G. Consequently, during the next iteration, all features from the entire candidate set (which includes the remaining local features) are considered and may join at any branch within the network. Constraints are then reintroduced, such that causal links may only be formed between local features and the maximum branching factor of nodes is returned to 2. This process iterates until either a maximum number of features within the graph is reached or features are exhausted. The purpose of this algorithm is to localize computation and minimize communication requirements between sensor nodes. This is achieved by forcing individual cameras to provide sub-decisions in the inference process. We note the efficacy of this approach to home networks where communication may be transient. Should communication be lost, the probability of remote data items can be marginalized out, effectively removing the requirement for communication. This algorithm is summarized in Figure 4.

TABLE I
COMMUNICATION CHARACTERISTICS OF STRUCTURES LEARNT

Method	Average Number of Links	Average Transmission Size (Relative)
SL (NR)	0.89	12.78
U-SL (NR)	0.89	22.22
SL (R)	0.78	8.15
U-SL (R)	0.56	7.78

Average number of communication links to remote sensors with the relative average transmission sizes. Note that the transmission size will vary dependent upon the instantiation status of the sending variable. SL denotes the structure learning algorithm introduced within this paper, whilst U-SL denotes *unrestricted* structure learning – where remote and local features are considered equally. NR illustrates that the robot is n*ot present* and R, that the robot is *present.*

C. Robot Feature Fusion

In addition to the wall mounted VSN nodes, the robot illustrated in Fig. 3 is free to roam within the environment. For the purpose of this experiment we assume that the robot never loses visual contact with the subject under surveillance. Consequently, the correlation of data between mobile and static VSN nodes is not a concern. If the experimental environment was not constrained to a single subject, sensor correlation techniques such as those discussed in [2] could be utilized.

In order to account for the difference in relative position between robot and subject, several position invariant features have been used. Furthermore, the robot navigation scheme discussed in Section III-B ensures that the robot always stops at a given distance from the target. Where the robot is present at a given station, the VSN node mounted on the robot (VSN-4) is treated in the same manner as static, wall mounted nodes, aside from the additional generation of optical flow features.

V. EXPERIMENT DESIGN

Our experiment was carried out using 3 subjects within a simulated home environment consisting of 4 static and 1

```
Parameters

Alpha; Activities; Branching_constraints;
Graph_size_limit

Pseudocode

if(first iteration)
     Prev_BIC = select local feature returning
       maximal BIC when direct causal ancestor of
       Activities
     Curr_BIC = choose further local ancestor to
       maximise BIC of graph subject to
       Branching_constraints
     Init_Delta_BIC = Curr_BIC - Prev_BIC
endif


while(features AND size(graph)<Graph_size_limit)
     Prev_BIC = Curr_BIC
     Curr_BIC = Choose causal ancestor subject to
       Branching_constraints
     Delta_BIC = Curr_BIC - Prev_BIC
       if(Delta_BIC < Alpha*Init_Delta_BIC)
         relax Branching_constraints for 1
         iteration of while loop
       endif
endwhile
```

Fig. 4. Structure learning algorithm employed. Note that once a feature has been selected for inclusion in the structure, it cannot be used again. For details regarding *branching constraints*, refer to Section IV-B.

mobile VSN node, as per Fig. 1. Each user entered the room via the door in the bottom right hand corner and proceeded to perform a series of 5 activities at each of the 3 different locations; s*itting still, eating, writing, reading newspaper* and *reading book.* The robot tracked and followed each subject – and each subject waited for the robot to maneuver into place before beginning their activities.

VI. Results

After *preprocessing* for subject movement it was found that VSN-9 and VSN-12 provided data on station (a), VSN-11 and VSN-12 on station (b) and VSN-11 and VSN-10 on station (c), as per Fig. 1. Where the robot was present, each candidate set was augmented by the robot mounted camera, VSN-4.

In order validate our approach we provide results from several experiments. In all cases, training of the resulting networks is performed using Expectance Maximization (EM) and inference using message passing where all observable features have been instantiated [23]. Training and testing is done on a per subject basis using sub-sampled data with a 2:1 training to test ratio. Firstly, we show activity recognition results using our structure learning scheme with individual VSN nodes only – thus no external communication can be performed as the candidate feature set is restricted to local features. Secondly, we compare our structure learning algorithm at each location using static VSN node candidate sets, against a completely unrestricted scheme – where at each iteration, any feature can be selected provided the generating VSN node is static. Finally, we repeat this experiment with the environment augmented by a mobile robot. In each of these experiments α=0.8. In order to obtain representative results for each station, each algorithm is applied and results averaged over all VSN nodes within the candidate set, before being further averaged over all subjects.

A. Activity Recognition using Local Features

In Fig. 5 the average accuracy of activity recognition using local features only at individual VSN nodes is presented. We see a high variance in accuracy, ranging from approximately 62% at VSN-11, location (c) to approximately 86% using only the robot's VSN node at location (b). Clearly, the use of a mobile camera for home monitoring is advantageous as, within this experiment, the highest accuracy recognition at each location is seen using the robot alone (VSN-4).

B. Structure Learning Validation without Robot

Fig. 6 presents the average percentage accuracy at each of the 3 locations, averaged over 3 subjects, 5 activities and all VSN nodes in the location *candidate set,* which does not include the robot. The introduced technique provides a level of overall accuracy that is within 2.5 percentage points of unrestricted learning whilst reducing average communication requirement by 42%, See Table I and Fig. 6.

C. Structure Learning Validation with Robot

Finally, we demonstrate the application of our algorithm when the robot is present within the environment. Using our structure learning algorithm the communication requirement between sensors has dropped by 63% of that required by the unrestricted learning algorithm without the robot, and by 36% compared to the experiment using our algorithm without the robot present, Table I. This can be partly explained by the efficacy of the robot in performing activity recognition – since often only a single data summary is needed from the robot in order to increase overall accuracy of inference at a given camera location. Consequently, we note a negligible difference in communication requirements comparing the proposed structure learning against unrestricted learning when the robot is used. Fig. 7 illustrates the average accuracy in these two cases. Because of the robot's ability to generate discriminatory features for fine motor tasks, unrestricted learning favors the robot for most branch selections, minimizing communications by default. Overall, using our algorithm with the robot yields a slight increase in accuracy (2.5 percentage points) and detailed analysis shows a reassignment of class accuracy, for example, where VSN-10 is responsible for *sitting still* at location (c), the structure produced by our algorithm performs better by 20 percentage points. Nevertheless, where our algorithm is used with mobile VSNs it will always learn structures *faster* due to a reduction in branch possibilities.

TABLE II
COMPARISON OF CLASS ACCURACIES

Method	Sitting	Book	News Paper	Writing	Eating
SL (RO)	81%	77%	77%	84%	92%
SL (NR)	85%	75%	76%	69%	72%
SL (R)	88%	86%	83%	84%	86%

For each method, results are obtained by averaging over all subjects and VSN nodes within the *candidate set*, where applicable. *SL (RO)* provides results when the structure learning algorithm was applied to features from the robot only (RO), whilst *SL (NR)* details algorithm application without the robot present. *SL (R)* presents recognition rates when both fixed and mobile VSN nodes were used in structure learning.

D. Class Accuracy

In order to demonstrate the utility of this system for activity recognition, we provide the average class accuracy obtained during several of the aforementioned experiments, Table II. Firstly, *SL Robot Only (RO)* presents class specific results when only features from the robot were used. Secondly, with only wall mounted VSN nodes used, results were as found under *SL No Robot (NR).* Finally, when our algorithm was used to learn structures over both fixed and mobile VSN nodes, we provide the results as *SL Robot (R).* All results are averaged over all subjects and, where applicable, all VSN nodes within the associated candidate set. We see that, for *sitting*, *book reading* and *newspaper reading,* the robot alone and the fixed camera system are roughly comparable, however the robot outperforms the fixed system during *writing* and *eating*. This could be due to the use of optical flow from the robot mounted VSN node – which is ideally placed to pick up fine motor motion, such as that of the hands. We note that through fusion of fixed and mobile cameras, the system performs better in every activity than if fixed cameras were used alone, with an average increase of approximately 10 percentage points per activity.

VII. DISCUSSION

In previous sections we have shown the utility of our algorithm to reduce the communication required within VSNs during inference in a home healthcare environment, whilst maintaining activity recognition accuracy – specifically when the robot it not present (Table I, Figure 6). We have also seen that intelligent environments can benefit from the use of mobile autonomous agents with regard to activity recognition, although respectable accuracy can be achieved through fixed devices alone (Table II). This leads us to believe that a hybrid system, comprised of multiple Bayesian Networks that are dependent upon robot location and availability, may be the most suitable solution for in-home activity monitoring.

Unfortunately, due to the intensive nature of the structure learning algorithm it would not be possible to determine and compare structures rooted at different VSN nodes in real time – this would need to occur offline and during an initial training phase. We also note that, due to the nature of our algorithm in restricting causal ancestors and since all feature values were instantiated in our experiments, many features within the structure would have remained unconsidered during inference. Interestingly, due to the power of message propagation to operate with incomplete data this raises two further research issues. Firstly, it is possible to effectively reduce the computational requirements on a given node by producing only subsets of the aforementioned features during each clock cycle. Secondly, sub-trees rooted at remote VSN nodes could self-test the quality of their features though partial instantiation of local features, observing the probability of those not instantiated and comparing against locally sensed values.

VIII. CONCLUSION

In this paper we present a VSN based system which achieves high accuracy activity recognition through the inclusion of a mobile agent. We also demonstrate a structure learning algorithm that minimizes communication during distributed inference when that agent is not present. Such an approach is vital to maximize the battery life of ubiquitous vision sensors for home health monitoring. In the future, we intend to extend this framework to support wearable sensor data – providing additional discriminatory power to robot assisted intelligent environments.

REFERENCES

[1] J. Riley. Rising Life Expectancy: A Global History. Cambridge University Press, 2001

[2] D.G. McIlwraith, J. Pansiot, S. Thiemjarus, B.P.L. Lo and G.Z. Yang, Real-Time Correlation of Ambient and Wearable Sensors, Proceedings of the 5th International Workshop on Wearable and Implantable Body Sensor Networks (BSN08).

[3] J. Pansiot, D. Stoyanov, D. McIlwraith, B.P.L. Lo and Yang G.Z., Ambient and Wearable Sensor Fusion for Activity Recognition in Healthcare Monitoring Systems, Proceedings of 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN07).

[4] B. Lo, L. Atallah., O. Aziz, M. ElHew, A. Darzi and G.Z. Yang, Real-Time Pervasive Monitoring for Postoperative Care, in Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN07), pp. 122 -127.

[5] M. Pollack, S. Engberg, J.T. Matthews, S. Thrun, L. Brown, D. Colbry, C. Orosz, B. Peintner, S. Ramakrishnan, J. Dunbar-Jacob, C. McCarthy, M. Montemerlo, J. Pineau, and N. Roy, Pearl: A Mobile Robotic Assistant for the Elderly, Workshop on Automation as Caregiver: the Role of Intelligent Technology in Elder Care (AAAI), August, 2002.

[6] J. Seki, A. Kiso and S. Tadakuma, Omni-Directional Vision Sensor Based Behavior Monitoring System Using Bayesian Network, Proceedings of SICE Annual Conference, 2007.

[7] R. Bodor, R. Morlok and N. Papanikolopoulos, Dual-Camera System for Multi-Level Activity Recognition, Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.

[8] K. Morioka, J-H Lee, H. Hashimoto, Human-Following Mobile Robot in a Distributed Intelligent Sensor Network, IEEE Transactions on Industrial Electronics, 51 (1), 2004.

[9] J-H Lee and H. Hashimoto, Intelligent Space -- Concept and Contents, Advanced Robotics 16 (1), pp 265-280, 2002

[10] G. Appenzeller, J. Lee, and H. Hashimoto, Building Topological Maps by Looking at People: An example of cooperation between

intelligent spaces and robots, Proceedings of the 1997 International Conference on Intelligent Robots and Systems, 1997.
[11] F. Capezio, F. Mastrogiovanni, A. Sgorbissa and R. Zaccaria, Mobile Robots and Intelligent Environments, Proceedings of AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, pp 781-788, 2007.
[12] H. Zhou and S. Sakane, Learning Bayesian Network Structure from Environment and Sensor Planning for Mobile Robot Localization, Multisensor Fusion and Integration for Intelligent Systems, MFI2003. Proceedings of IEEE International Conference on, 2003.
[13] Henry Schneiderman, Learning a Restricted Bayesian Network for Object Detection, IEEE Conference on Computer Vision and Pattern Recognition, June, 2004.
[14] Dar-Shyang Lee, Effective Gaussian Mixture Learning for Video Background Subtraction, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (5), pp. 827-832, May, 2005.
[15] Julien Pansiot, Ahmed Elsaify, Benny Lo, and Guang-Zhong Yang. RACKET: Real-time Autonomous Computation of Kinematic Elements in Tennis. In to appear in the Fifth IEEE Workshop on Embedded Computer Vision (ECVW), Kyoto, Japan, 2009.
[16] MobileRobots, ActivMedia Robots, http://www.mobilerobots.com
[17] Mesa Imaging, http://www.mesa-imaging.ch
[18] J. Ballantyne, S. Valibeik, A. Darzi A and G.Z. Yang, Assisting Elderly Patients through a Mobile Robot - First Steps towards a Safe Robotic Platform, 4th International Congress of Minimally Invasive Robotic Association (MIRA 2009)
[19] B.K.P. Horn and B.G. Schunck, (1981) Determining Optical Flow. Artificial Intelligence, 17, pp 185-203.
[20] G. Schwarz, Estimating the Dimension of a Model, Annals of Statistics 6(2), pp 461-464, 1978.
[21] P. Leray and Olivier François. Bayesian Network Structural Learning and Incomplete Data. Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, pp 33-40, Espoo, Finland, June 2005.
[22] Kevin Murphy. Bayes Net Toolkit. http://bnt.sourceforge.net
[23] J. Pearl, Probabilistic Reasoning In Intelligent Systems, Networks Of Plausible Inference, Elsevier Science & Technology

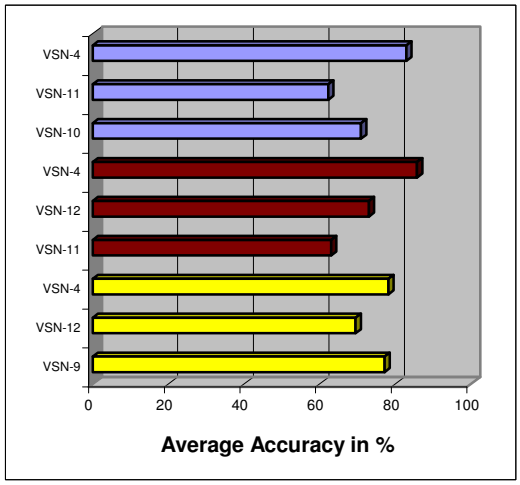

Fig. 5. Percentage accuracy over all 3 subjects and 5 activities using local features at individual VSNs. Location (a) is given in yellow (bottom), Location (b) in burgundy (middle) while location (c) is in blue (top).

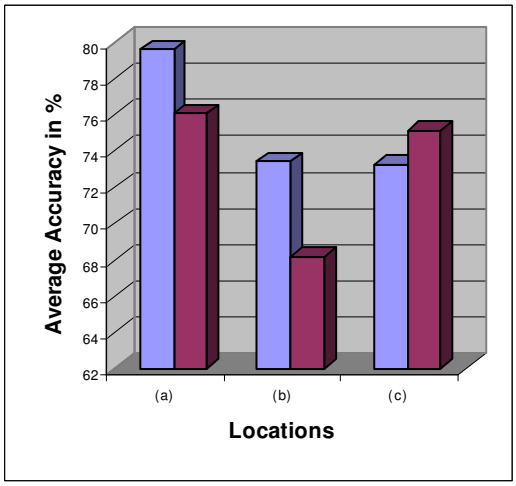

Fig. 6. Average percentage accuracy at each of the 3 locations, averaged over 3 subjects, 5 activities and all VSNs in the locations *candidate set*. The robot was not used. Blue shows the average activity recognition accuracy where the structure learning algorithm of Figure 4 is used (left). In burgundy, cameras adopted the unrestricted learning method, where features may be selected regardless of location (right). The introduced technique provides a level of overall accuracy which is comparable (within 2.5 percentage points) to unrestricted learning whilst reducing average communication requirement by 42%.

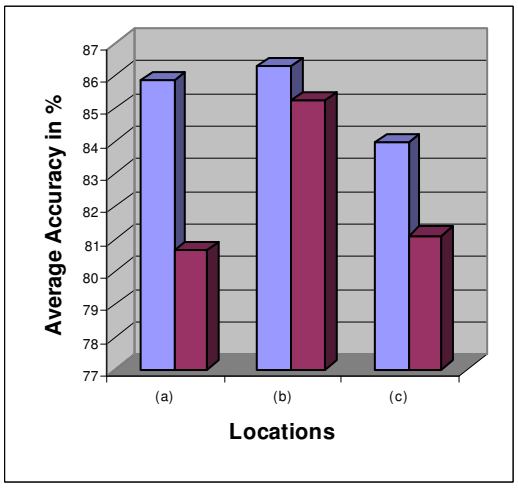

Fig. 7. Average percentage accuracy at each of the 3 locations, averaged over 3 subjects, 5 activities and all VSNs in the locations *candidate set* – which now includes the robot. Blue shows the average activity recognition where our structure learning algorithm is employed (left), burgundy shows the results using the unrestricted algorithm (right). Overall there is a 9.9 percentage point increase in accuracy when compared to a system where the robot is not used (either with or without the introduced structure learning algorithm). Furthermore, do to the efficacy of the robot in activity recognition; the overall requirement for communication between sensor nodes has dropped by 63% of that required by unrestricted learning without the robot.