

Using Human-Object Interaction Signatures to Find and Label Chairs, Floors

Patrick Peursum Svetha Venkatesh Geoff West Hung H. Bui¹

Dept of Computing, Curtin University of Technology GPO Box U1987, Perth, Western Australia

Telephone: +61 8 9266 7680, Fax: +61 8 9266 2819

{peursum, svetha, geoff, buihh}@cs.curtin.edu.au

Version #3

Date of Submission: 2nd August 2004

1 Introduction

A fundamental component in making smart homes truly live up to their ‘smart’ moniker is the ability to recognise objects in an indoor scene and detect when and how a human interacts with them. Without object recognition, smart homes cannot make full use of video cameras since the vision system cannot provide context to the human activities it is monitoring, consequently under-utilising a rich and versatile source of information. However, traditional shape-based object recognition tends to fail when presented with typical scenes that a smart home would need to cope with — that is, wide-angle views of indoor scenes containing a variety of objects cluttered together. Partial occlusions, unconstrained orientations, irregular shapes, relocation of objects by humans and lack of detail for distant objects are all factors that make it difficult to recognise an object by its shape. Unfortunately, these same factors are also defining characteristics of household environments. This means that very little object recognition research has proven robust enough to be deployed in smart home testbeds such as Microsoft’s EasyLiving project [11] and the AwareHome at the Georgia Institute of Technology [6].

In order to gain access to the benefits of context, many researchers are forced to manually label objects or areas of interest. This can facilitate the intelligent control of devices and simplify the task of human behaviour monitoring. For example, recent work by Kimberle Koile *et al* allows a user to personalise the behaviour of a smart home’s devices by maintaining a map of manually-named context areas and a list of preferences. The user can then define rules such as having a nearby projector switch on when a person sits at a meeting table. Another example is to use object recognition to provide context for recognising human actions — knowing the position of the phone simplifies the task of recognising when a person is making a call [1]. Similarly, the act of page-flipping is more probable when the object being used has been recognised as a book [12].

To address the issues facing object recognition in a smart home, we explore an action-*centred* approach to learning and classifying functional objects in an indoor laboratory monitored by stationary cameras. The premise of this approach is that it is much easier to interpret human motion since it is constrained by the structure of the human body, as opposed to recognising arbitrary objects. Moreover, humans tend to interact differently with objects that differ in their functionality, so it should be possible to identify an object by analysing the motions that a human performs in order to manipulate that object. We call these motions the human-object *interaction signature*.

¹Currently at AI Center, SRI International, email bui@sri.com.au

The advantage of interaction signatures is that they can be used to perform object recognition without considering the object's physical structure, thus bypassing many of the difficulties inherent in shape-based recognition. Although this means that objects which humans never interact with cannot be labelled (for instance, walls and ceilings), such objects are generally less relevant than manipulated objects. Another advantage is the fact that the occupants in a smart home will frequently and repeatedly interact with household objects, and this can be used to build up evidence for object locations and labels. Object labels are strengthened or weakened as interaction signature evidence is accumulated over time, and this has the added benefit of being able to adapt to the scene as it changes.

To demonstrate the potential of our approach, we use the bounding box statistics of a moving human to recognise the interaction signatures that represent the interaction with chairs and floors (ie: sitting and walking). Although our features are very coarse, we are still able to label the objects by employing standard, well-known action recognition algorithms. We also show that the system can adapt its labelling after a human relocates a chair to another position in the scene. Note that the investigation is restricted to labelling chairs and floors since recognising the interaction signatures for other objects (such as cups or telephones) requires the detection of motions that are too subtle for our coarse features.

2 Related Work

Most current approaches to object recognition rely on classifying an object by comparing a shape-based model of the object against a database of known objects [4]. However, there are several serious drawbacks to this approach. Firstly, the possible variation in shapes and orientation for any particular class of objects is often very large. Louise Stark and Kevin Bowyer proposed to address this issue by function-based object recognition, where object models are classified based on their functional components [15]. For example a chair could be defined as any object that has a flat, stable sitting surface. Unfortunately, this still suffers from the basic problem of trying to extract the object's 3D model from its 2D image. Moreover, actually finding and segmenting the object's 2D image out of a wide-angle view is a difficult task. Recent work by Brandon Sanders *et al* [14] in using video to segment objects is one approach to the latter problem, where objects that are occasionally moved by humans (dubbed *quasi-static objects*) can be segmented accurately by using background subtraction and temporal evidence. In contrast to our work, Sanders is focussed on segmenting an object from the image without being concerned with *what* the object is, whereas we wish to observe the human's actions in order to infer both the location and label of the object.

Other researchers have begun to use human activity to reason about the contents of a scene. Some applications of this are in finding the paths in outdoors scenes, either for the purposes of detecting unusual behaviour or for determining the extent of the pathways and obstacles that exist within the scene [7, 10]. Similarly, Kimberle Koile *et al* [8] mapped areas of interest they call 'activity zones' by accumulating evidence of human activity in the scene. From this evidence, the most heavily used areas are designated as activity zones. However, they were limited to manually providing descriptive labels for each zone. As an attempt to use action recognition to assist in labelling objects, Darnell Moore *et al* [12] tracked the movement of a human's hands interacting with an object to refine an initial classification of the object done by traditional shape-based object recognition. They worked with top-down, close-up views of office desks monitored by a camera. Whilst successful (and incorporated into the AwareHome project) the method has limited deployment opportunities in a smart home due to the necessity for uniplanar scenes (such as the flat surface of a desk), its reliance on initial shape-based object classification and the need for very close-up views. These factors constrain its potential deployment to areas of the home that are both fixed and experience significant, cohesive activity within a small area, such as dining tables or kitchen sinks.

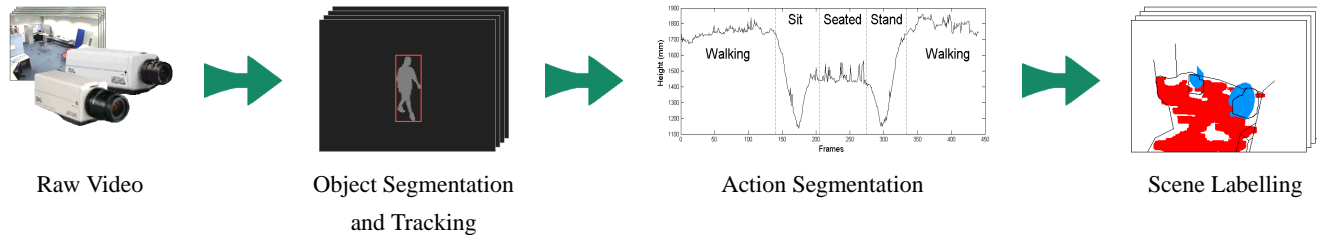


Figure 1. Major steps in interaction signature scene labelling.

3 Tracking, Action Segmentation and Labelling

3.1 Foreground Object Segmentation and Tracking

We use background subtraction to segment objects of interest from the video stream, using a mixture model of Gaussian distributions to model the background [16]. This background model was chosen since it can robustly adapt to changes in the background definition over time, which is essential for our research. Foreground objects (ie: people) are segmented out from the background, outlined by a bounding box and tracked using a Kalman filter. Background segmentation can often fail to perfectly segment the person, but since we only measure coarse features such as height and width, only major segmentation failures will significantly affect the measurements. More details on our tracker and test environment can be found in our previous work [13].

Each view is pre-calibrated to the world coordinate system via a set of landmark points and correspondences are found between different views of a person by their proximity in the world coordinate system (assuming that the person is standing on the ground plane). Additionally, we detect partial occlusions of people by comparing the world heights and positions of the person in all views. If the person’s lower portion is occluded in one view it will report a smaller height than the other views, indicating occlusion by an object in that view.

3.2 Object Relocation

We detect the addition or removal of scene objects by relying on the fact that household objects generally do not move without intervention by humans — what Sanders calls quasi-static objects [14].

The addition of a new object will result in a new foreground blob suddenly appearing in the scene. A quasi-static object will not move on its own, so the system can infer that the new blob relates to a new, unknown object at that location. The system eliminates any existing labels in the area to reflect the fact that a new object is there and adds the blob to the statistical background so that it no longer appears in the foreground.

Removal of objects presents an additional complexity — when an object is removed, it leaves behind a ‘ghost’ where it once was located. This ghost occurs because when the object is removed, the scene behind the object is uncovered and no longer matches the background (which has been learned with the object at that position). This causes a foreground blob to appear, even though no physical object is there (hence the term ‘ghost’). Fortunately, the colour of the ghost will generally match its surroundings since there is now no object in the way. We use this to distinguish between *removed* objects (since the blob colour matches its surroundings) versus *introduced* objects (since the blob does *not* match its surroundings). However, we do not make the connection that the removal of an object and its addition to another area indicates a transfer of the same object — currently, they are considered two different objects.

3.3 Action Segmentation

We trained four Hidden Markov Models (HMMs) with features extracted from the video, one HMM for each interaction signature. HMMs were used due to their proven aptitude in modelling human motion [2, 3]. The four modelled actions are:

- Walking
- Sitting down into a chair
- Person seated in a chair
- Standing up from a chair

Training data consisted of six examples with four views per example (24 sequences in total) of a person walking into the room, sitting down into a chair, standing back up and leaving the room. For each sequence, the chair was positioned at different orientations and positions within the room. Each sequence was manually segmented into the four constituent actions, which we then used to train the HMMs. Training features are:

- Real-world height (in mm).
- Change in height between this frame and the previous frame, expressed as a proportion of the total height to minimise dependency on the object's height.
- Change in width, also expressed as a proportion of the total width.
- Ground speed of the object (absolute velocity, in mm/frame).

These HMMs then form the basis for automatically segmenting test video sequences into blocks of actions that each relate to one particular interaction signature. Automated segmentation is preferable manual segmentation since we wish to demonstrate that our proposed evidence-based labelling is robust to noisy action segmentation. We used a simple sliding window approach [2] where segmentation occurs by only considering the frames that fall within a fixed-size moving window. For our actions, a single window size of 30 frames was found to provide the best results. The features from frames within this window are then used to calculate the log-likelihood for each HMM, with the most likely HMM selected as the action label for the block of frames. The selected action is then estimated to have begun halfway in the window, based on the reasoning that an HMM will become dominant over the previous action's HMM when at least half the frames of the window relate to the new action. The window is then moved one frame forward and the entire process is repeated.

However, sliding windows tend to produce short bursts of incorrect action labelling since an incorrectly-classifying HMM can temporarily become more probable than the correctly-classifying HMM. This can be due to background subtraction failures, occlusions or other random factors. To solve this, we introduced a heuristic confidence test on the HMM log-likelihoods which mandates that the most likely HMM must significantly outperform the next most likely model. To decide this, the ratio between the highest and second highest HMM log-likelihoods is calculated, with 'significant' difference being defined by an arbitrary threshold (currently 0.75). If no significant HMM is found, the last significant action is re-instated.

Each view performs action segmentation independently of all other views. To improve action segmentation further, each view then casts an equally-weighted vote as to the action being performed — the current model is re-instated if a deadlock occurs. The elected action is then used by all views to perform scene labelling, again independently of each other. Although it would be possible to fuse the features from each camera into a single corresponded set, the features we measure are not fine-grained enough to benefit from such a fusion. Because of this, the voting method will generally provide better segmentation accuracy since voting is essentially a form of bagging (where a more reliable classification is made by combining the results of several classifiers).

3.4 Scene Object Labelling

We label objects in the scene by taking each frame of an action block and updating the view based on the action being conducted and the position of the person. This is done by maintaining a weight for each label (chair or floor) for every pixel in a view's background image. The weights lie within the range 0 to 1, and are initialised to 0. When a pixel is updated, all weights are updated via the following exponential-forgetting function:

$$w_{t+1}^L(x,y) = w_t^L(x,y) \cdot (1 - \eta) + (\eta \cdot \kappa),$$
$$\kappa = \begin{cases} 1 & \text{if L = detected object,} \\ 0 & \text{otherwise} \end{cases}$$

- w is the weight of the L^{th} label (chair or floor) at time t and pixel (x,y) .
- η is the learning rate for learning labels, and is generally very small (less than 0.05) to avoid building up weights too quickly.
- κ is the update value that controls which label will be strengthened.

This function ensures that older evidence becomes less and less important as new evidence is observed. Also, conflicts in evidence (such as a single pixel having similar evidence for both chair and floor labels) are quickly resolved since rival labels are decayed when new observations occur that supports one of the labels over the others.

In terms of the concept of interaction signatures, a label for a particular object label should occur whenever that object's associated action is observed. Thus chairs are labelled whenever the sitting, seated or standing up actions are detected. The fitted ellipse of the seated person is used as the labelling area in preference to the bounding box since it is a closer match to the person's silhouette, and by implication closer to the area of the chair. Floor space is labelled when the walking action occurs, with the heuristic that only the lowest 5% of the fitted ellipse is labelled as floor space since this area generally corresponds to the feet of the person.

3.5 Higher Level Constraints on Labelling

In addition to the basic labelling-from-action, we can also affect labels by detecting higher level interaction signatures that are not specific to any particular type of object, including partial occlusions of the person and transference of objects about the scene.

Partial occlusions of a person who walks on the far side of a chair are used to refine chair labels and affect the future learning of chair labels in the area. When the chair occludes the person's legs, we can infer that the chair does not extend into the unoccluded area, so we remove all chair labels and slow the rate of future labelling within this unoccluded area. If the occluding object is not labelled (such as a table or an unlabelled chair) there are no labels to refine, but the system now has evidence that there is *some* object there, and will slow future labelling in the unoccluded area. Slowing the rate of labelling ensures that incorrect chair labels are not quickly reinstated. The rate of retardation is heuristically defined as linear with respect to the number of times occlusion is observed in the region — more instances mean a slower re-learning rate. The learning rate is not reduced to zero since we still wish to recover from mistakes in defining the area of partial occlusion. All these responses to partial occlusion are applicable to almost any object even though we only have shown it for chairs, but note that it cannot be applied to floors since the floor cannot occlude the person.

Another type of high-level interaction signature is detecting when a person relocates an object in the scene. When this occurs, we must destroy the labels that are no longer valid. If the system detects that a chair has been taken away from an area, we remove all chair labels in the area to reflect that the chair no longer exists in that location. Similarly, if the system detects a chair has been placed in an area, we remove all non-chair (ie: floor) labels for that area since the chair now occludes the floor and chair labels must take priority.

4 Results and Analysis

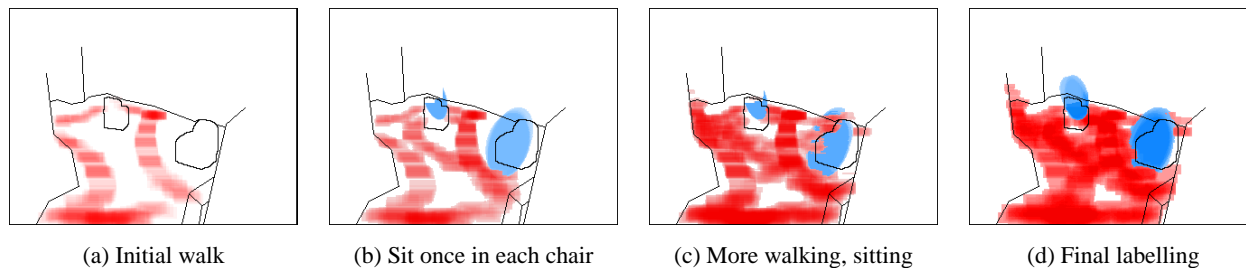


Figure 2. Sequence of images showing the progression of floor (red) and chair (blue) labelling for one run. Outlines indicate chair, floor and obstacle boundaries. Intensity indicates weight of the label, and strengthens as more evidence is accumulated.

4.1 Data

The system performs four major operations in sequence to produce a labelled image of a given scene (see Figure 1). Video is captured at 25 frames per second from four ceiling-mounted cameras monitoring the scene (a laboratory). Captured video is saved to disk in MPEG-4 format, which is then processed offline for object segmentation and tracking to produce the raw data needed for action segmentation and scene labelling. This is then separately processed to produce a labelled image of the scene from all four camera views.

Labelling accuracy was tested using three video sequences, each around one to two minutes in duration and comprising of four views taken of a person alternately moving about and sitting down in the target chairs. The chairs remained in fixed positions throughout the experiments. Action segmentation and scene labelling was performed on each of the sequences to produce three sets of four labelled images (one image for each view of the scene — see Figure 3).

A second set of experiments was conducted to evaluate the effectiveness of the system in dealing with a person moving a chair about the scene. For this, three video sequences were taken where a person moved about the scene, repeatedly sitting in and relocating a chair.

4.2 Action Segmentation

A ground-truth for the starting frame for each action instance was estimated by manually determining the start and end times of each action. The uncertainty for the ground truth is roughly ± 5 frames, though this is a subjective judgement. Table 1 shows the difference (in frames) between the ground-truth and the automatic action segmentation, indicating how noisy the segmentation is and whether the subsequent labelling process has a reasonable chance of succeeding.

The ‘walk’ and ‘sit’ actions are segmented quite accurately (mean error of -2.94 and 5.15 respectively), especially given that the ground-truth uncertainty is ± 5 frames. Also significant is that the ‘sit’ action is generally segmented slightly later than the actual ‘sit’ action, and conversely the ‘walk’ action is segmented slightly earlier. This means that the beginning and end of chair interactions are conservatively estimated.

The most concerning aspect of the data is that only two ‘seated’ instances were actually detected (out of 19), and the system generally detects the beginning of the ‘stand’ actions far too early. In fact, the two failures are related — the motion profile of the last third of the ‘sit’ action looks strikingly similar to the first third of the ‘stand’ action. (see Figure 1: ‘Action Segmentation’ for an example). This results in the ‘stand’ action becoming prematurely dominant at the time when the ‘seated’ action actually begins. The mistake is not corrected over the next few frames since our log-likelihood confidence

Model	Num. Instances	Found Instances	Mean Error (frames)	Error Var. (frames)
Walk	19	19	-2.94	24.56
Sit	19	19	5.15	48.31
Seated	19	2	0	8
Stand	19	19	-50.24	726.32

Table 1. Error means and variances for action segmentation.

threshold prevents the ‘seated’ model from replacing the ‘stand’ model. Fortunately, our evidence-accumulation framework means that the loss of the ‘seated’ action label merely results in less evidence for the chair labelling, which is easily offset by observing more instances of a person sitting in the chair.

Segmentation accuracy is severely limited by the coarseness of the measurements of the human actor (bounding box statistics and speed) — the failure in finding the ‘seated’ action is a direct consequence of these simple features. Selecting better features would improve segmentation accuracy and allow us to classify more interaction signatures, such as drinking from a cup. Pose estimation could be used to obtain better features, where techniques include skeletonisation [5] or model-based methods [9]. Also, more sophisticated segmentation techniques could be used to further improve the results, such as using the HMM’s Viterbi state sequence [3]. That said, the segmentation results are adequate for our purposes.

4.3 Scene Labelling Accuracy

Chair labelling was evaluated by comparing the area labelled as ‘chair’ against the true extent of the chairs in each of the views (where a chair’s extent also includes the space between the chair legs). Table 2 shows these results.

	Classified as (pixels)			Recall
	Chair	Floor	Other	
Chair	18,127	3,068	5,083	69%
Floor	11,503	168,320	72,435	67%
Other	7,313	7,686	628,065	98%
Precision	49%	94%	89%	

(a)

	Classified as (pixels)		Recall
	Chair	Floor	
Chair	18,127	3,068	85%
Floor	11,503	168,320	94%
Precision	61%	98%	

(b)

Table 2. Confusion matrices for chair labelling. ‘Other’ relates to unlabelled pixels where no significant action occurred. Table 2b shows the confusion matrix if ‘Other’ is not taken into account

The system achieved a recall rate of 69% for chair labelling, representing the correctly-labelled percentage of the total chair area across all views. Thus it is evident that chair labelling manages to locate chairs fairly successfully, with nearly 7 out of 10 chair pixels found. Inaccuracies are mostly due to the fact that the labels are produced from the seated person, who is almost always offset slightly from the chair itself since they sit *on* the chair rather than *within* it.

To measure how closely chair labelling was able to fit within chair boundaries, it is necessary to refer to precision. Even though the precision value of 49% seems quite low (indicating that about half the chair labels were outside of chairs), it is not unexpected since the seated person’s extent is nearly always larger than the chair itself. For example, the person’s head and shoulders are almost always higher than the chair’s back. Additionally, the offset of the person from the chair further degrades the precision of labelling.

The use of occlusion to localise the extent of the chair was found to have significant benefits for precision. The effectiveness of occlusion can be explained by the fact that it is particularly useful in detecting and reducing one of the primary causes of over-labelling; that is, a person’s head and shoulders rising above the chair’s back. Unfortunately, occlusion was not fully exploited due to the limited number of occlusions that the experiments contained. To illustrate this, we only considered the chair views that experienced occlusions — precision for these was fairly high at 70% in contrast to the overall precision of 49%.

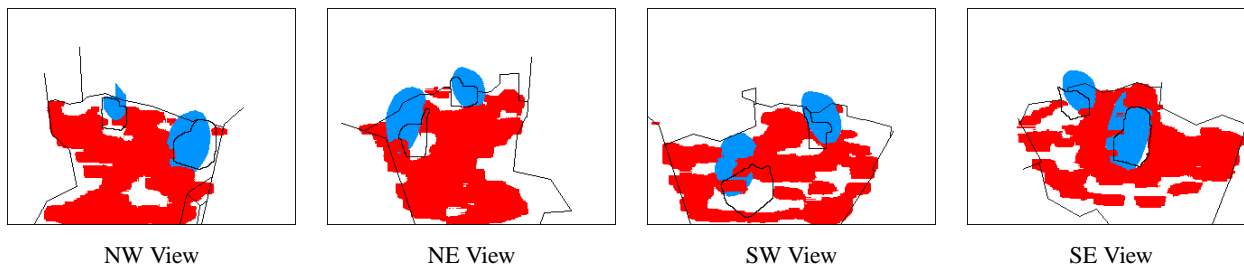


Figure 3. Floor (red) and chair (blue) labels for all views of the same test run, thresholded to remove weak labels. Edges show manually-defined ground truth for chairs, floors and occluding objects in the scene. Note that floor labelling is reasonably adept at detecting edges of occluding objects such as walls and chairs. Note also the effects of occlusion in the NW view (top-left chair) and SE view (lower-right chair) for finding chair boundaries.

Though Table 2a shows that the precision of floor labels is extremely good (94%), this is misleading since the open floor space extends over a large proportion of the view. Equally misleading is the recall figure for the floor, which seems quite low (67%). The failure here is that some portions of open floor space were not actually walked over during the experiments, so gaps exist in the coverage and adversely affect the recall. To illustrate this, Table 2b, shows the classifications without the ‘Other’ labels — floor recall improves markedly from 67% to 94%. In light of these issues, we did not attempt to analyse floor labelling numerically as was done for chairs. Instead, floor label evaluation was limited to visually inspecting the labelled images for floor labels that incorrectly spilled into chair areas or occluding walls and partitions — see Figure 3. Overall, floor labelling manages to detect occluding edges reasonably well with only minimal overflow. Over-labelling into chair spaces is also minimal, due both to the success of floor labelling and the fact that chair labels tend to overpower the floor labels.

4.4 Handling the Relocation of Chairs

To demonstrate the possibilities of interaction signatures in relation to handling the relocation of objects, an additional experiment was performed where we examined the effect on labelling of moving chairs around the scene. Since we assume only chairs are transferable, only chair labels are erased at the chair’s former location (not floors). If we dealt with more object types, the strongest weighted label type would be the label to erase.

Figure 4 shows a progressive example of labelling for one view, comparing the effect on labelling with and without object relocation detection. Note that as soon as the chair is picked up by the person in the upper image of Figure 4b, the system immediately recognises this and removes all pixel labels relating to the chair. Conversely, in the lower image where pick-up detection has been disabled, the labels are left unchanged. Similarly, when the chair is put down in Figure 4c, the floor labels are removed from the area the chair now occupies. Referring to 4d provides a good indication of the benefits of object relocation detection — in the lower image, there still seems to be a chair in the original position, and even the new position has quite a large component labelled as floor space. No such problems affect the upper image.

However, object relocation detection is not foolproof — if we consider all four views independently, object relocation events were only detected in 34 out of the total of 48 events (12 physical events × four views), or 70% of the time. Fortunately,

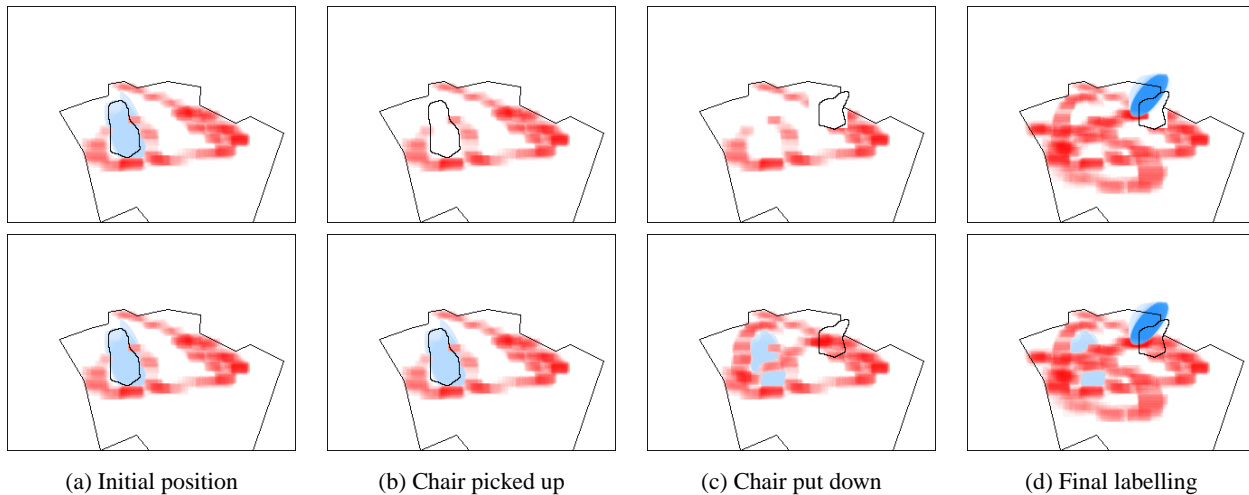


Figure 4. Example sequence with chair relocation. Upper row shows the effect of detecting chair relocations on labelling, lower row shows labelling with relocation detection disabled.

since a relocation event is nearly always detected in at least one view, we can often update the views that did not detect the event. Still, there are cases where this cross-view fusion is unable to update a view, often because one view of the relocation event is blocked by the person for too long a period — this view is out of sync with the other views and so cannot take advantage of cross-view fusion. Thus accuracy improves to 42 of 48 (87%) when using fusion, but 6 events are still missed. In the missed cases, the system is still able to recover somewhat by relying on the fact that evidence accumulation will tend to erase incorrect chair labels over time (see Figure 4d).

5 Conclusions and Future Work

Whilst this is a fairly early investigation into the concept of interaction signatures as a means of action-centred object labelling, it is encouraging that we have obtained reasonable results with crude measurements and are able to adapt to the relocation of objects within the scene. However, given that we have only dealt with chairs and floors any judgements on the general effectiveness of our approach would be premature. Moreover, there are too many heuristics and thresholds used in the labelling process and these must be eliminated in order to make the system more robust and portable to new environments.

To address this, we intend to improve the research along multiple paths. First is to refine our measurements of human foreground blobs so that we can determine more interesting information such as where the person’s limbs are. This will provide us with several benefits, such as the ability to handle more complex interactions and extend our very limited range of objects to include a larger variety. Similarly, no image information has been taken into account up until now. Even simple image segmentation techniques to divide the image into similarly-coloured areas would provide a wealth of information and allow us to move from pixel-level labelling into region-level labelling. We expect that this could be used to significantly improve labelling accuracy by using the image segmentation as secondary evidence in determining an object’s boundary. Furthermore, it is feasible that complementary sensors such as microphones can be used to provide further evidence for interaction signature recognition. However, our approach does have some definite limits — objects that are never interacted with, such as walls and ceilings, can never be detected by our system. Other objects such as tables are also difficult to detect since humans do not normally interact *directly* with the table, but rather with the objects on top of it.

References

- [1] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12):833–846, October 2001.
- [2] A. F. Bobick and Y. A. Ivanov. Action recognition using probabilistic parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 196–202. IEEE Computer Society Press, 1998.
- [3] M. Brand and V. Kettner. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):844–851, August 2000.
- [4] R. J. Campbell and P. J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, February 2001.
- [5] H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletonization. In *Proceedings of the IEEE Workshop on Application of Computer Vision*, pages 15–21. IEEE Computer Society Press, October 1999.
- [6] Georgia Institute of Technology. AwareHome research initiative. <http://www.cc.gatech.edu/fce/ahri/>.
- [7] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–29. IEEE Computer Society Press, 1998.
- [8] K. Koile, K. Tollmar, D. Demirdjian, H. Shrobe, and T. Darrell. Activity zones for context-aware computing. In *Proceedings of the 5th International Conference on Ubiquitous Computing*, pages 90–106. Springer-Verlag, 2003.
- [9] M. W. Lee, I. Cohen, and S. K. Jung. Particle filter with analytical inference for human body tracking. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 159–165. IEEE Computer Society Press, 2002.
- [10] D. Makris and T. Ellis. Finding paths in video sequences. In *Proceedings of the British Machine Vision Conference*, pages 263–272, 2001.
- [11] Microsoft Research Vision Group. EasyLiving project. <http://research.microsoft.com/easyliving/>.
- [12] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 80–86. IEEE Computer Society Press, 1999.
- [13] P. Peursum, S. Venkatesh, G. A. West, and H. H. Bui. Object labelling from human action recognition. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, pages 399–406. IEEE Computer Society Press, March 2003.
- [14] B. C. Sanders, R. C. Nelson, and R. Sukthankar. A theory of the quasi-static world. In *Proceedings of the IEEE International Conference on Pattern Recognition*, volume 3, pages 1–6. IEEE Computer Society Press, 2002.
- [15] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):1097–1104, October 1991.
- [16] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, August 2000.