Combining Image Regions and Human Activity for Indirect Object Recognition in Indoor Wide-Angle Views

Patrick Peursum Geoff West Svetha Venkatesh
Dept of Computing, Curtin University of Technology GPO Box U1987, Perth, Western Australia
{peursump, svetha, geoff}@cs.curtin.edu.au

Abstract

Traditional methods of object recognition are reliant on shape and so are very difficult to apply in cluttered, wide-angle and low-detail views such as surveillance scenes. To address this, a method of indirect object recognition is proposed, where human activity is used to infer both the location and identity of objects. No shape analysis is necessary. The concept is dubbed 'interaction signatures', since the premise is that a human will interact with objects in ways characteristic of the function of that object — for example, a person sits in a chair and drinks from a cup. The human-centred approach means that recognition is possible in low-detail views and is largely invariant to the shape of objects within the same functional class. This paper implements a Bayesian network for classifying region patches with object labels, building upon our previous work in automatically segmenting and recognising a human's interactions with the objects. Experiments show that interaction signatures can successfully find and label objects in low-detail views and are equally effective at recognising test objects that differ markedly in appearance from the training objects.

1 Introduction

Object recognition has been the focus of considerable research interest over the past few decades due to its fundamental applicability in computer vision. However, it has not been widely used in applications such as smart homes or surveillance despite significant benefits that object information context would provide to analysing human behaviour. Researchers who have made use of object context for human action recognition have been forced to manually pre-define the locations of objects [12]. This is because the wide-angle views in monitored scenes such as a smart home are particularly challenging for traditional shape-based object recognition, which relies on high-detail views of objects in order to extract distinctive features for object shape analysis. Unfortunately, wide-angle indoor views typically contain multiple objects cluttered together where objects are low in detail, can be partially occluded and only a few objects are relevant to the application. Moreover, close-up views that are more suitable for object recognition are generally inappropriate for surveillance systems since the latter require a maximal field of view.

Rather than attempt to directly analyse the shape of objects under low-detail conditions, this paper proposes to *indirectly* find and label objects based on the manner in which a human interacts with the objects. No object shape analysis methods are employed in labelling objects. The approach is based on the premise that a human will manipulate functionally-similar objects (which should be labelled equivalently) in a similar manner, and the human's posture will also imply the location of the manipulated object. For example, a person sits in a chair and drinks from a cup. In this way, the human's actions are a contextual 'signature' implying both the class and location of objects, hence this approach to object recognition is referred to as the paradigm of *interaction signatures*.

The significance of the approach is threefold. (1) Interaction signatures are able to find and label objects in scenes where objects are too low in detail but humans are not. (2) Object classification is largely invariant to the shape and orientation of objects since objects of the same functional class will be interacted with similarly and so are equally recognisable regardless of their appearance. (3) The actions of the human in the scene guide the system towards the location of relevant objects within the clutter. This paper demonstrates that interaction signatures can successfully label objects without any shape-based analysis. However, interaction signatures should not be viewed as a panacea replacement for shape-based recognition methods. Rather, interaction signatures have strengths and weaknesses that are *independent of* and *complementary to* shape-based recognition, thus the two approaches could potentially be integrated to improve the robustness of both.

This paper evaluates the utility of interaction signatures in a household-type environment due to the fact that occupants tend to interact frequently and repeatedly with appliances and other household objects, thus maximising the amount of human-object interaction evidence. No shape

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 IEEE

analysis is used. This paper builds upon our previous work in fast human pose modelling, activity segmentation and semantically-descriptive action labelling using missing observation data in HMMs [9]. To facilitate labelling, this paper decomposes the scene into region patches and a Bayesian framework is developed to learn and apply interaction signatures to the task of labelling the regions in the scene with semantically-descriptive object labels. The activity of printing a document is analysed to find and label six object classes — *Floor*, *Chair*, *Keyboard*, *Printer*, spare *Paper* and *None* for irrelevant objects. Per-frame region labels are combined over time in a process of evidence accumulation to form a labelled map of the objects in the scene. Shape-invariance of labelling is demonstrated by replacing the training objects with two sets of alternative objects whose shape and colour differs from the training objects. The lab is also completely refurbished, testing the system on an unfamiliar scene. Results show that even significant changes in the appearance of an object (such as draping a large sweater over the back of a chair) have no effect on recognition accuracy.

This paper is organised as follows: Section 2 briefly describes similar work in computer vision literature, followed by an overview of the process of labelling via interaction signatures in Section 3. Section 4 describes the experiments and results for object labelling. Finally, Section 5 outlines the conclusions and potential future work.

2 Related Work

There is a very large body of literature on the topic of shape-based object recognition — the reader is referred to surveys such as [1, 8]. Traditional shape-based recognition has largely been constrained to relatively close-up views in order to provide enough detail for distinctive feature extraction, a requirement that is unsuitable for surveillance applications which require wide-angle scenes.

As an alternative, several researchers have explored using human actions to reason about the elements in a scene. Some applications are finding pathways by learning the trajectories and routes that people take through an outdoors scene [6, 11]. Koile *et al* [5] learn heavily-used areas (dubbed Activity Zones) in indoor scenes for triggering appropriate responses by a smart environment, but the areas must still be manually labelled with meaningful names. Similarly, Grimson *et al* [3] use human motion to produce a depth map of the scene, but could not label the scene.

In another approach, Moore *et al* [7] use the hand motions of a human to refine an initial shape-based classification of an object, demonstrating the benefits of combining action evidence with shape analysis. However, in order to perform shape-based recognition they are constrained to close-up, top-down and uniplanar views, such as a desk or a kitchen bench. Furthermore, objects are detected by assuming that they are introduced into the scene and segmented by background subtraction, or labelled *a priori* in the case of immobile objects (eg: keyboards or chairs).

To handle object recognition in wide-angle views driven primarily by human actions, we have proposed the concept of interaction signatures. Our previous work [10] is based on coarse measurements of human motion (bounding box features) and so can only detect gross interactions with large objects such as chairs or floors. Additionally, image information is not considered, hence objects were labelled heuristically by assuming an object will exist near a certain portion of the human's body (eg: a chair will exist somewhere within the fitted ellipse of the person's silhouette). This paper eliminates the need for such heuristic labelling by developing a Bayesian classifier that labels image regions based on pose estimation and action recognition sub-systems developed in [9].

3 Interaction Signatures for Labelling

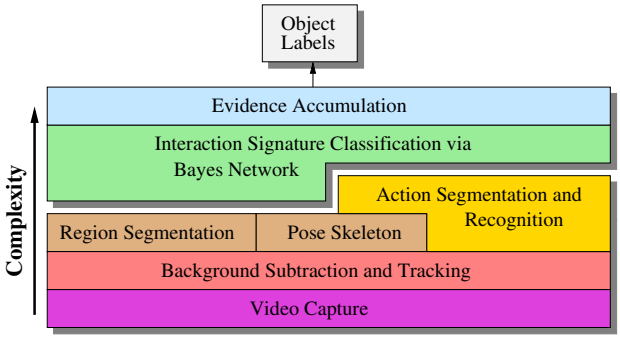

Figure 1. Hierarchy of processing necessary for labelling a scene using the interaction signatures paradigm. This paper addresses the top two layers — processing for lower layers is described in [9].

Figure 1 shows an overview of the process used to generate object labels from interaction signatures. Indoor human activity is captured at a resolution of 320×240 pixels and 25 frames per second from four ceiling-mounted cameras (one in each corner of the room) Video is processed to extract a real-time stick-figure skeleton of a person's silhouette to model the human's pose in 3D, as described in [9]. Figure 2 shows examples of the pose skeleton. Motion features are extracted from the pose skeleton and used for action recognition via Hidden Markov Models (HMMs). Recognition is robust to occlusions since the skeleton is allowed to mark occluded limbs as missing observation data in the HMM. Missing data in the HMM is also used to facilitate automatic segmentation of an activity (eg: printing a document in this paper) into its constituent actions [9].

Using the segmented actions, object labels are assigned to regions in a scene via the interaction signatures approach. For every single video frame where human activity is oc-

curring, each region is classified with an object label (or *None*) according to the (inter)action that the person is conducting, the posture of their skeleton and the location of the region relative to the person. To constrain processing, only regions that are 'within reach' of the person (ie: fall within the bounding box of the person's silhouette) are potential candidates for interaction. Regions are classified via a Bayesian network that is trained on the action being performed, the relative position of each region to the skeleton and the ground-truth of the object label for each region (Figure 4). Each frame may involve simultaneous interaction with more than one object (eg: when typing at a computer, the chair, floor and keyboard are all involved). Furthermore, each object may be split into several regions which are simultaneously and independently classified. Per-frame classifications are fed into a system of evidence accumulation. This maintains a set of weights relating to each possible object label for each region in the image, providing a relatively stable map of object labels in the scene that adapts as more interaction evidence is observed, thus being robust to per-frame errors in classification.

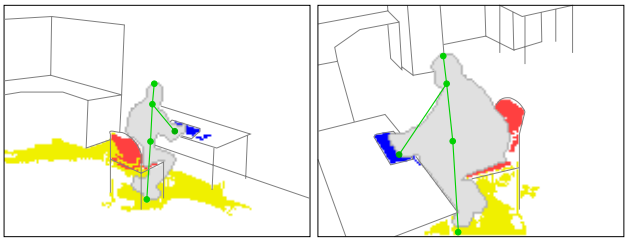

Figure 2. Examples of pose skeletonisation from silhouette, with the position of objects inferred from the skeleton. Scene outlines are manually added as an aid for the reader.

3.1 Higher-Level Activity as a Context for Lower-Level Action Recognition

Since labelling via interaction signatures depends on action recognition, the semantic descriptiveness of object labels is controlled by the semantic descriptiveness of action labels. However, interactions with different objects can involve visually identical motions, such as picking up a cup or picking up paper. Hence it is necessary to provide a context that can allow action recognition to distinguish between such visually-similar but semantically-distinct actions. Unfortunately, object context cannot be employed since this paper is concerned with the inverse problem of finding objects from actions. The alternative is to use the higher-level *activity* and temporal sequence of actions within this activity as a context for the lower-level actions taking place within the activity. For example, knowing that a person is conducting the activity 'print a document' restricts the expected set of actions and implies an ordering to the actions (eg: must pick up spare paper before loading the paper and getting the printout). A system for performing such context-enhanced action recognition is described in [9].

Note that it is not envisioned that activity context alone will suffice for all situations — to make a more robust action recognition system, all useful contexts should be considered (eg: shape-based object recognition, indoor *vs* outdoor scenes, time of day, type of room, etc). However, for the purposes of this research, activity and temporal sequence context provide sufficient information to distinguish between many visually-similar actions.

3.2 Bayes Classification of Regions

Regions are classified as belonging to a particular object (or not) by combining evidence of the relative position of the region with respect to the pose skeleton together with the action being performed. This combination is essentially what the term 'interaction signature' represents.

Figure 3. Example region segmentation (important objects are circled). Note that segmentation is quite rough, with the printer and chair broken up into several regions.

The scene image is decomposed into homogeneously-coloured region patches using a simple seeded region grower thresholded on colour [4]. Seeds are randomly chosen and recursively grown into the similarly-coloured neighbourhood. Resultant regions that are too small (less than 15 pixels) are discarded. Although the region segmentation is quite rough (Figure 3), interaction signatures are able to handle this level of noise. Regions are extracted from the background image rather than the video frame since any person moving in the scene will not show up in the background and so will not occlude object regions. Note that it is preferable that the regions are slightly over-segmented (ie: objects are split into multiple regions) since an interaction signature will be applicable to all constituent regions of an object but cannot label only part of a region.

Classification occurs for every frame for every region within 'reach' of the human (ie: overlapping the bounding box). Each camera view is processed independently for object labelling, avoiding the complication of having to fuse regions across near-orthogonal views into 3D. Note that the pose skeleton is generated in 3D for view-independent action recognition before being projected onto each 2D view for object labelling. The amount of data to be classified is quite large — for four views capturing at 25fps and typically involving 30 regions per frame, around 3,000 classifications

must be made per second of video. A Bayes classifier is employed since it can compactly represent the training data and continues the statistical approach of [9].

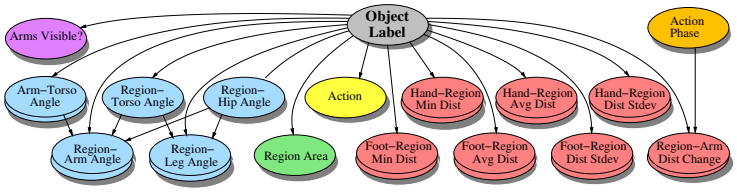

Figure 4. Bayesian classifier for per-frame labelling.

Figure 4 shows the Bayes network used to classify regions. Each node relating to an 'Arm' or 'Leg' actually represents two independent, identically-distributed (i.i.d.) nodes — one for each of the two arms (or legs). Nodes can be interpreted as fulfilling specific roles for interaction signature labelling. The leftmost group of five nodes (relating to angles) encodes the relative position of the region with respect to the skeleton. The expected position of a given region can be inferred via triangulation of the angles, hence the dependencies between the nodes. Distances to the region (seven lower-right nodes) also assist in defining the location of the region — for example, a *Keyboard* should exist near the person's hands. The *ArmsVisible?* node exists to constrain when it is appropriate to attempt to label certain objects since the pose skeleton may mark limbs as missing. For example, it is difficult to label a *Keyboard* when the arms are not detected. Constraints are also placed on the size and compactness of regions via the *RegionArea*, *AvgDist* and *DistStdev* nodes. Finally, the *Action* node defines the human's current action, a context that implies which object labels are valid for the current interaction.

Since all length and area features are in terms of 2D pixels, pixel measurements are scaled according to how far away the region is from the camera (in 3D). The only 3D measurement available is the human's position, so it is assumed that regions are approximately at the same location as the human for the purposes of scaling.

When defining the ground-truth for the training data, an object's regions must be labelled as that object only when the person is actually interacting with the object. For example, a patch of floor that is physically located on the other side of the room cannot be labelled as *Floor* since the person cannot be interacting with that region, even if the person's silhouette overlaps with that region due to the perspective projection onto the camera's 2D view. Only floor regions that are at the person's feet should be classified as *Floor*. Similarly, the keyboard, printer and paper cannot be labelled when the person's arms are not detected.

In addition to the object labels, *None* is used as an anti-label to classify regions that do not match an interaction signature. *None* is itself an 'interaction' that is learned along with the other object labels and allows the selection of the most-likely label at every frame as the classification, rather than having to define a minimum threshold for accepting an object label. *None* also allows the system to recover from errors in labelling by decaying incorrect object labels to *None*.

3.3 Evidence Accumulation

Classifications by the Bayesian network are on a per-frame basis, and so are sensitive to per-frame variations in the pose skeleton. Also, as soon as a particular interaction signature ceases, it is no longer possible to continue to label the objects that were involved in the signature. Hence a form of evidence accumulation is employed that builds a map of the object labels within the scene based on the per-frame classifications. The approach used is to maintain a list of weights for each region, one weight for every possible object label. Whenever a classification occurs on a region, the weight W of the classified label L is increased and all other labels' weights are decreased according to the following learning and decay functions:

$$W_L(t+1) = \frac{W_L(t)}{1-\ell_L} \qquad \text{if } W_L(t) < 0.5 \qquad \text{(1a)}$$

$$W_L(t+1) = \Big(W_L(t)\cdot(1-\ell_L)\Big) + \ell_L \qquad \text{if } W_L(t) \geqslant 0.5 \qquad \text{(1b)}$$

All other labels for that region are decayed via:

$$W_L(t+1) = W_L(t)\cdot(1-\delta_L) \qquad \text{(1c)}$$

where ℓ_L and δ_L are the learning and decay rates respectively from Table 1. Note that the learning function is split into two parts (below 0.5 and above 0.5 — Equations (1a) and (1b)). Figure 5 (right) depicts the weight learning and subsequent decay for one object label.

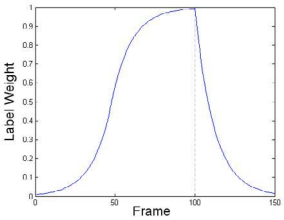

Figure 5. Idealised weight learning/decay function.

Label:	*None*	*Floor*	*Chair*	*Keyboard*	*Printer*	*Paper*
Learning Rate (ℓ)	0.005	0.02	0.03	0.06	0.1	0.25
Decay Rate (δ)	0.003	0.01	0.01	0.01	0.01	0.01

Table 1. Learning and decay rates for evidence accum.

The learning and decay rates in Table 1 were heuristically chosen by observing the duration of each action and how well each interaction signature was detected by the system (weaker signatures are not always recognised for every frame of the interaction). Interactions that have a very short duration (eg: getting paper) have a higher learning rate than longer, more prominent actions (eg: walking) to compensate for the difference in the amount of evidence available

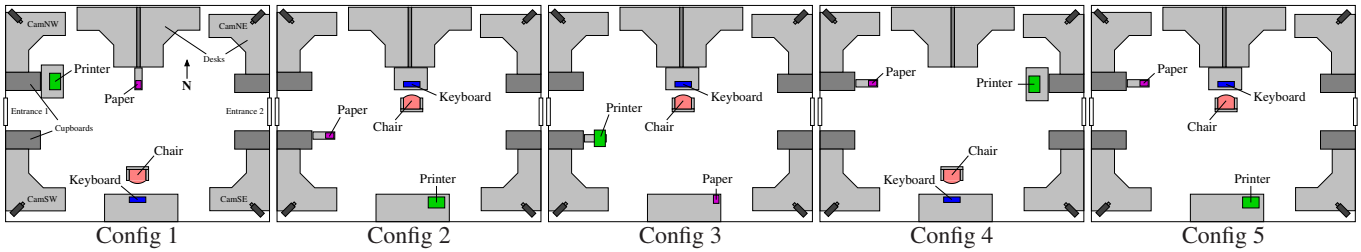

Figure 6. Floor plans of the five room configurations used. Coloured objects are targets of interest. Grey objects are not targets for labelling — greyscale intensity indicates their height. See Figure 3 for an example of the SE view for Config 1.

for each interaction signature. The rates for *None* are very low since it is the 'default' label and preference should be given to other, true, object labels.

The function enforces both a slow start to learning so that isolated misclassifications do not create much weight, and a fast decay so that label conflicts are resolved quickly.

4 Experiments and Analysis

Evaluation focuses on demonstrating the robustness to low detail and shape-invariance of interaction signature labelling. Objects are arranged in five configurations of relative positions and orientations (see Figure 6). Four cameras monitor the scene, one in each corner, hence at least half of the views for each object are from across the room and low in detail. One activity is modelled for the purposes of finding and labelling manipulated objects within the scene. It involves printing out a document, finding that the printer is out of paper, obtaining some spare paper to load the printer and finally retrieving the printout and returning to the computer. Several interaction signatures are defined within this activity, including the acts of typing, sitting, getting paper and retrieving the printout. From these signatures, five objects are labelled — *Floor*, *Chair*, *Keyboard*, *Printer* and spare *Paper* (with *None* making a sixth label). Note that objects such as tables or the computer's monitor are not labelled since there is no direct interaction with these objects and so the interaction signature would be too weak to use given the available features. Additional features would be necessary to recognise these subtle interaction signatures, such as estimating the direction of the person's gaze to find the computer monitor.

For this research, 50 sequences of the printing activity were captured, with each sequence lasting 1-2 minutes. Every 10 sequences, the locations of the relevant objects in the scene were changed, resulting in five different topographical configurations for object positions. This was done to prevent the system from learning the relative positions of objects and thus artificially assisting labelling. Similarly, the Bayes classifier is trained with data from all views to ensure that the scene's appearance from a particular view is not a factor in classification. During five-fold cross-validation testing, object labelling is performed independently on each view using the trained classifier. No fusion of information across views takes place due to the difficulty in reliably matching regions between very wide baseline views.

4.1 Region Labelling Evaluation

Figures 7, 8 and 9 show examples of region labelling. Since region segmentation is not perfect, regions are subjectively judged to be part of an object if they mostly fit within the object. Note that when an object is fractured into multiple regions, each region can still be labelled correctly. For example, although there are several segments to the chair object in each view, the majority of them are still labelled as *Chair* (easiest to see in SW and SE views of Figure 7).

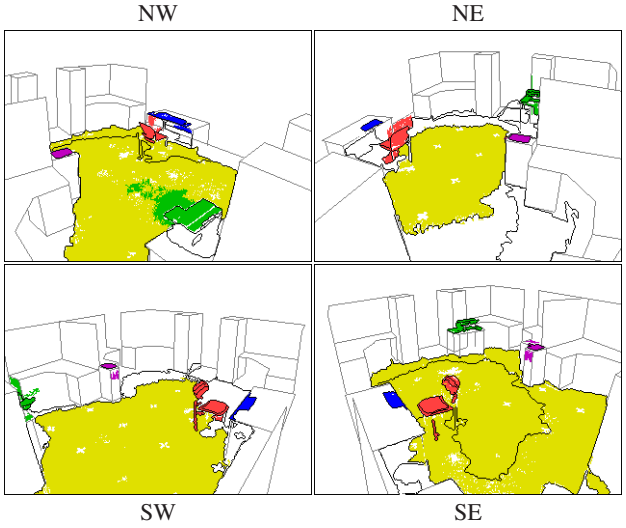

Figure 7. Final labels for each camera view of a sequence. Objects such as floors and chairs are split into multiple regions (outlined in black). Light outlines for walls and other obstacles are manually defined as an aid for the reader.

Regions that are directly adjacent to an object tend to be the main source of errors in mislabelling regions with the object's label. This is particularly an issue for *Printer* and *Paper* since their associated interaction signatures have the person reaching towards the object rather than physically touching it (as is the case for the other objects). Hence

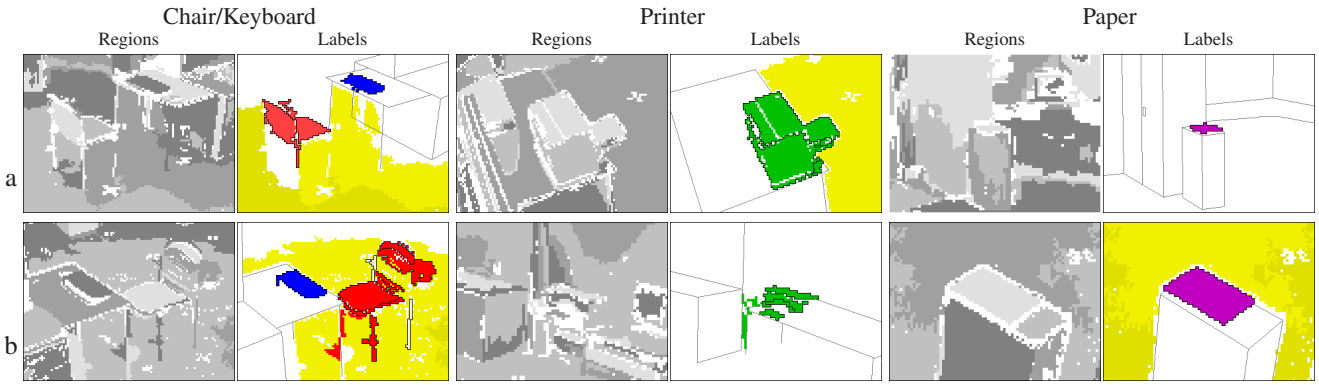

Figure 8. Final labelling with corresponding region segmentations from two sequences (zoomed to objects). Note that labelling can deal with objects that are fractured into multiple regions and embedded within clutter, such as the chairs and printers.

estimating the position of the *Printer* and *Paper* objects is highly sensitive to inaccuracies in the pose skeleton. In some cases, this can mean that the object is completely missed. The problem can also affect *Chair* labelling — in the NE view of Figure 7, the system believes that the chair's back extends much higher than is truly the case because it labels a region that is actually a couple of metres behind the chair. These errors are caused by clutter, both apparent clutter (due to noisy region segmentation and 2D perspective projection) and true clutter (due to the physical contents of the scene). Higher levels of clutter tend to degrade labelling precision since regions in close proximity to the object of interest are mislabelled as part of that object. Apparent clutter could be reduced by better region segmentation. However, physical clutter cannot be mitigated, hence it is important to note that although precision is degraded by clutter, labelling is still feasible.

Labelling accuracy also tends to become more error-prone for objects that are located further away because object detail decreases with distance. Moreover, a distant object will have several 'gaps' in its region coverage due to the minimum size constraint on regions. Distance also means that errors in the pose skeleton become more significant and adversely affect the accuracy of labelling. Problems with distance could be minimised by using higher resolutions, better region segmentation and more accurate pose estimation, but distance will always negatively affect labelling.

Finally, one of the limitations of labelling is the narrow definition of 'interaction' — only regions overlapping with the bounding box of the person are potential candidates for interaction signatures. Thus the printer in the NW view of Figure 7 is only partially labelled since the lower region of the printer rarely intersects with the person's bounding box and so has insufficient evidence for a label.

4.2 Error Robustness of Evidence Accumulation

Table 2a shows the confusion matrix for per-frame region classifications, summed over all views and all sequences. Regions that are never interacted with (ie: never

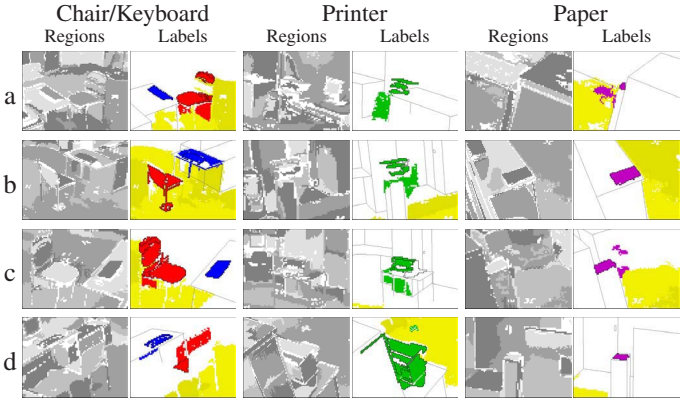

Figure 9. More final labels (zoomed to objects)

overlap the bounding box) are omitted to avoid artificially inflating the accuracy of the *None* label. Table 2b shows the confusion matrix for final region labelling of regions where a region is classified with the highest-weighted object label for that region as derived from evidence accumulation, thresholded so that extremely weak labels (less than 0.05) are instead labelled as *None*.

In general, the precision of final region labelling is better than that of the per-frame classifications. The most improved object classes are *Printer* and *Paper*, whose precisions nearly double from a very low 27.3% and 28.4% to a more reasonable 48.1% and 47.5% respectively. This is because final region labels are generated from the weighted combination of all the per-frame classifications. In some ways, evidence accumulation could be viewed as an 'ensemble of *classifications*' as opposed to ensembles of *classifiers*, such as bagging and boosting [2]. The difference is that boosting uses multiple classifiers to classify a single instance, whereas evidence accumulation uses a single classifier to classify multiple instances separated temporally but all relating to the same physical object.

On superficial inspection, the precision results seem merely adequate. However, since cameras are orthogonally-located, at least half of the cameras will view an object from across the room (up to seven metres distant). The distance

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

	Per-Frame Classifications							Final Region Classifications						
	None	*Floor*	*Chair*	*Keybrd*	*Printer*	*Paper*	*Recall*	*None*	*Floor*	*Chair*	*Keybrd*	*Printer*	*Paper*	*Recall*
None	4,467,866	24,480	232,823	46,180	57,264	8,281	92.4%	14,713	24	362	88	567	148	92.5%
Floor	29,929	283,120	821	0	2	0	90.2%	185	373	2	0	8	0	65.7%
Chair	55,256	0	722,785	2,623	0	0	92.6%	75	0	1,029	0	0	0	93.2%
Keyboard	8,580	0	0	118,001	0	0	93.2%	38	0	0	198	0	0	83.9%
Printer	9,320	0	0	0	21,314	0	69.6%	101	0	0	0	544	0	84.3%
Paper	1,449	0	0	0	264	3,295	65.8%	52	0	0	0	11	134	68.0%
Precision	97.7%	92.0%	75.6%	70.7%	27.0%	28.5%		97.0%	94.0%	73.9%	69.2%	48.1%	47.5%	

(a) (b)

Table 2. Confusion matrices for (a) per-frame classifications and (b) final region labelling after evidence accumulation.

also tends to produce more apparent clutter since objects further away from the camera are more difficult to separate and exist closer together due to the perspective projection onto the 2D view. Hence over half of the classifications are performed under conditions of low object detail and relative clutter. In light of this, the precision results and examples of labelling show that interaction signatures are relatively robust to clutter and low detail, although these conditions do of course degrade the accuracy of labelling.

4.3 Invariance to Object Shape and Colour

Experiments were also conducted to show that interaction signatures are relatively invariant to the exact shape or colour of the objects being labelled. Each training object for *Chair*, *Keyboard*, *Printer* and *Paper* was substituted with two sets of alternative objects and 12 new printer sequences were captured using these alternatives. These alternative objects were then labelled using the interaction signature classifier trained from the original objects. Note that the alternatives all had a different shape and colour to the original objects. To emphasize the shape-independence of the system, one of the alternative chairs also had a large sweater draped over the back. This covered most of the chair and would make it virtually impossible to recognise the chair by using shape and colour.

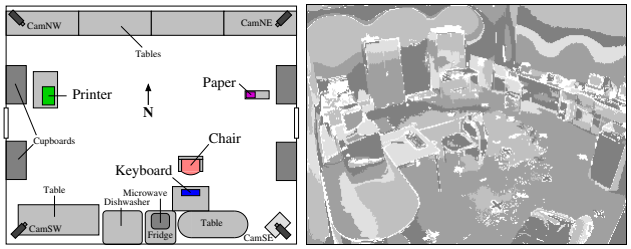

Figure 10. New room layout, with SE view segmentation.

Coincidentally, the lab itself had been completely refurbished between the time of the training data and the new sequences. This provided the opportunity to test the system on an unfamiliar scene — see Figure 10 for a view of the new room and an example of region segmentation for the new layout. The new scene tends to have more clutter around the walls but less clutter in the centre of the scene.

Table 3 shows the results of labelling with the alternative objects, and Figure 11 shows some examples of final region labelling. In general, results are comparable to the results produced from the original (training) objects. However, *Paper* (and to some extent, *Printer*) is significantly less accurate due to its unfortunate positioning near an area with high clutter (especially in the SE view). The amount of clutter 'blurs' the area that the interaction signature labels, causing precision to fall.

	Final Region Classifications						
	None	*Floor*	*Chair*	*Keybrd*	*Printer*	*Paper*	*Recall*
None	3705	0	109	27	185	123	89.3%
Floor	74	78	0	0	0	0	51.3%
Chair	62	0	398	0	0	0	86.5%
Keybrd	6	0	0	69	0	0	92.0%
Printer	21	0	0	0	123	0	85.4%
Paper	53	0	0	0	0	58	52.3%
Precisn	94.5%	100.0%	78.5%	71.9%	40.0%	32.0%	

Table 3. Confusion matrix for final region labelling of objects that are dissimilar in shape and colour to the training database objects.

Note that the sweater draped over the chair in Figure 11 (NE view) is only partially labelled. This is because the bounding box of the person never intersects the unlabelled half of the sweater whilst the person was sitting, thus no evidence exists to label that half. This effect is a consequence of restricting region labelling to the bounding box area, as discussed at the end of Section 4.1.

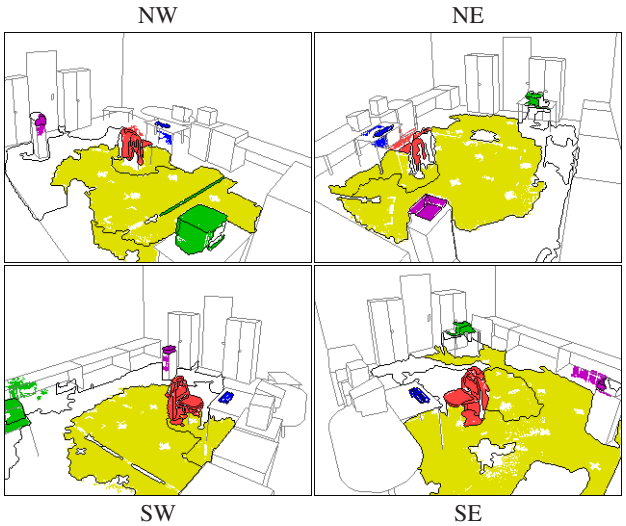

Figure 11. Final labels for a sequence with objects whose shape and colour is dissimilar to the training objects.

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

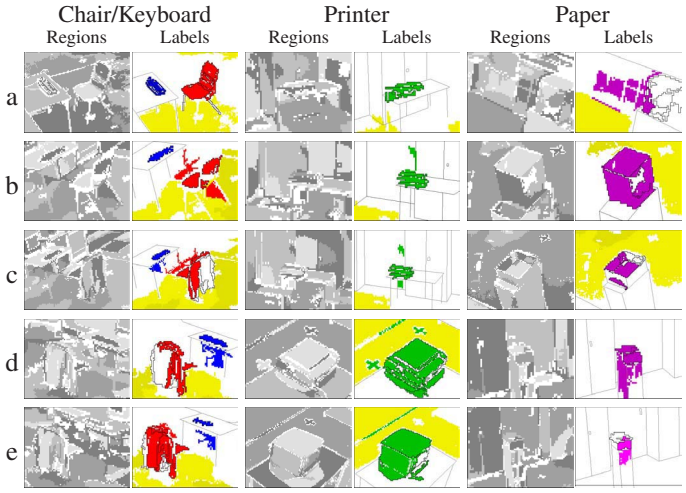

Figure 12. Region segmentations and final labels for several sequences with alternative objects (zoomed to objects). Note that the chairs in 12c–12e are covered by a sweater.

5 Conclusions

This paper has presented the concept of human-object interaction signatures to indirectly recognise objects by using the actions of humans to label an object's regions in wide-angle, real-world scenes. Per-frame classifications of image regions are used as evidence to generate a map of the scene showing where objects are and what regions are likely to be part of each object. Although labelling is by no means perfect, it is robust enough to produce a reasonable map of object labels despite the noisy real-world data it must deal with — sub-optimal segmentation of the image into regions, the fracturing of objects into multiple regions and an incomplete and often inaccurate pose skeleton. Importantly, the appearance-independent nature of interaction signatures means that the system is able to recognise objects when their shape and/or colour differs from the training objects. Even a sweater draped over the back of a chair does not affect the accuracy of labelling the chair.

Errors tend to occur in mislabelling regions that surround an object as being part of the object itself. This is due to several factors, including clutter, inaccuracy of the pose skeleton and the fact that regions are labelled independently of one another. The latter issue arises because constraints exist at the region level. This could be solved by adding additional constraints on the inter-region size and compactness for an object. Clutter is perhaps the most difficult problem to address since it is often a physical property of the scene, so better region segmentation algorithms will not be able to fully alleviate this issue.

Several extensions to this research are possible, including the integration of interaction signatures with traditional shape-based object recognition, the use of better underlying algorithms (for image segmentation, pose estimation and action recognition) and the fusion of regions into 3D via stereo matching to eliminate apparently adjacent regions that are in fact physically separated. Reducing the dependence of object labelling on accurate action classification is particularly crucial. Currently, if the action is misclassified the subsequent object labelling will also be incorrect, although evidence accumulation is able to minimise the effect of occasional errors in labelling. A more practical solution is to use many contexts in addition to action (such as an initial pass with shape-based object recognition [7]), improving robustness to errors in any single context.

References

[1] 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.

[2] T. G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *Workshop on Multiple Classifier Systems*, LNCS, pages 1–15. Springer-Verlag, 2000.

[3] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–29, 1998.

[4] R. M. Haralick and L. G. Shapiro. Image segmentation techniques. *Computer Vision, Graphics and Image Processing*, 29(1):100–132, 1985.

[5] K. Koile, K. Tollmar, D. Demirdjian, H. Shrobe, and T. Darrell. Activity zones for context-aware computing. In *Int'l Conference on Ubiquitous Computing*, 2003.

[6] D. Makris and T. Ellis. Finding paths in video sequences. In *British Machine Vision Conference*, pages 263–272, 2001.

[7] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *IEEE Int'l Conference on Computer Vision*, volume 1, pages 80–86. IEEE CS Press, 1999.

[8] C. G. Perrott and L. G. C. Hamey. Object recognition, a survey of the literature. Technical report, Macquarie University, NSW Australia, January 1991.

[9] P. Peursum, H. H. Bui, S. Venkatesh, and G. West. Robust recognition and segmentation of human actions using HMMs with missing observations. *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems*, 2005. To Appear.

[10] P. Peursum, S. Venkatesh, G. West, and H. H. Bui. Using interaction signatures to find and label chairs and floors. *IEEE Pervasive Computing*, 3(4):58–65, Oct–Dec 2004.

[11] M. Teal and T. Ellis. Spatial-temporal reasoning based on object motion. In *British Machine Vision Conference*, 1996.

[12] M. Xu and T. Ellis. Partial observation vs. blind tracking through occlusions. In *British Machine Vision Conference*, 2002.