Wearable Hand *Activity* Recognition for Event Summarization

W.W. Mayol

Department of Computer Science
University of Bristol
Woodland Road, BS8 1UB, UK
wmayol@cs.bris.ac.uk

D.W. Murray

Department of Engineering Science
University of Oxford
Parks Road OX1 3PJ, UK
dwm@robots.ox.ac.uk

Abstract

In this paper we develop a first step towards the recognition of hand activity by detecting objects subject to manipulation, and use the results to build a visual summary of events. The motivation is to extract information from hand activity without requiring that the wearer is explicit as in gesture-based interaction. Our method uses simple image measurements within a probabilistic framework and allows real-time implementation.

1 Introduction

Hands are a highly effective means for providing input to computers, and have been widely used to do so. For a wearable, hands can serve to recover two strands of information. In the most commonly found case, a wearer's hand gesture is used to signify an action or command, in the second and less studied case, hand activity itself could provide an extra cue to user context and intention.

Several camera-based methods have been devised for hand gesture recognition in computer interfacing, and recently they have been extended to the wearable domain e.g. the work of Kölsch *et al.* for outdoors hand gesture recognition [3], Starner *et al.* [5] who used a hat-mounted camera for sign language recognition, and the work of Kurata *et al.* [6] for detecting gestures within a cursor-and-click interface.

Consider now, for example, the task of building an instruction book from the subtle motions produced by an artisan as he makes a craft. The problem is how to build a summary of his actions when the involved hand gestures are not known in advance. Gesture recognition methods place the user in an imperative r*ô*le, demanding that he/she is explicit and attention-focused. Hand activity recognition is different from gesture recognition in that there is no explicit announcement of meaning: it is events (here associated to objects) that give cue to the detection of hand actions.

In this paper, a probabilistic framework that employs simple image measurements is developed, allowing the automatic selection of key video frames that detect manipulation events and summarise a span of wearer's hand activity.

The paper is organised as follows. Section 2 reviews methods for hand detection and object recognition, Section 3 describes the wearable sensor before presenting the method to filter attention in Section 4. Section 5 describes the methods used for detecting hands and recognizing objects of interest and a method to account for the spatial distribution of activity around the wearer. Section 6 introduce the methods for event detection and filtering. Section 7 describes the experiment before the paper ends with discussion and future work.

2 Visual detection of hands

The literature for detecting hands from images can be classified into two main groups. The first group of work uses a structural model of the hand, which can be three-dimensional as used by Rehg and Kanade [1] or a simpler two-dimensional contour as used by MacCormick and Isard [2]. These model-based approaches purse a correspondence between the observed hand and the model so that explicit degrees of freedom can be recovered. The second group of work, the view-based approach, uses a database of views of the hand and usually low-dimensional features are computed on them, such as in [3, 7].

When considering more complex hand activity such as object manipulation, the hand's high number of degrees of freedom and motion involved even in the simplest manual tasks, makes it currently unfeasible to establish direct correspondence between hand images and templates, particularly if the templates are articulated and three dimensional. In the same manner, holding a dictionary of multiple hand views is impractical due to the large number of variations involved. Instead of recovering hand configuration, an alternative possibility is to use a more invariant (and rougher) hand representation but concentrate instead on the recognition of the objects that are being subject to the manipulation.

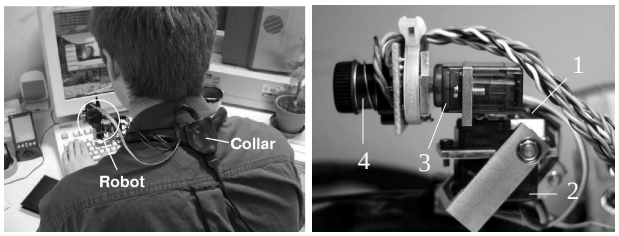


Figure 1: The Wearable Visual Robot. 1) elevation axis, 2) pan axis, 3) cyclotorsion axis, 4) CCD and optics.

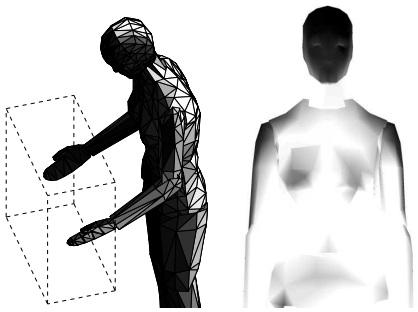
Here, this invariance comes from the use of colour-histograms for detecting both, hands and objects. In this respect, the approach is similar to the work of Swain and Ballard [8] who developed a system that used colour-histograms for identifying objects around a robotic arm equipped with vision. Another relevant work is by Schiele and Pentland [9], who developed a feature histogram based on Gaussian derivatives within a probabilistic framework. These example works clearly show that an object recognition system using feature histograms can perform well under challenging conditions such as substantial object occlusion, viewpoint and scale changes, and multiple objects within an image. However, two essential questions remain open: 1) where to look and 2) how to use the observed information usefully.

3 A wearable visual robot

A shoulder-mounted wearable active camera (Figure 1) first presented in [10] is used as the sensor to observe hand activity. We have used this device for a number of other applications that range from large-displacement image stabilization [10], gesture recognition [12] and real-time simultaneous localization and map building [13]. This device is intended as a front-end for a wearable computer, and because it has a larger degree of sensing autonomy than the one achieved by a passive camera, we prefer the term wearable visual robot. The current implementation uses three motors to compensate orientation in elevation, pan and cyclotorsion axis, has a 640x480 pixel non-interlaced image sensor with 42°of field of view (FOV) and a control interface connected to the computer. A speech synthesiser [14] provides feedback to the wearer on the state of the system.

The real advantage for an active camera is the concentration of sensor resolution in a small volume. In this work, the robot helps twofold, first towards a greater degree of independence from the posture of the wearer and second to attain a wider field of view, however the methods here described are equally applicable to a non-active camera.

In terms of wearability and sensor placement, the shoulder area is a good alternative to head-mounted devices



(a) (b)

Figure 2: Sensor Placement. (a) View of the handling space (dashed box) from different locations (polygons) around a humanoid figure, the darker the shade the better. (b) As before, a darker polygon is better, but here we consider field of view, view of the handling space and a walking motion as discussed in [11].

which are highly dependent on wearer's attention. Figure 2 show the relative weights obtained through simulation for camera locations when considering large field of view, view of the handling space and motion during walking for a humanoid-shaped model[1] (see [11] for details).

4 Attentional filtering

First, we have to select what to look at. In [15] Schiele and Pentland propose an *attentional filter* based on a motion cue that extracts image regions which, because that the camera is head-mounted, moves jointly with the wearer's attention. Image regions that remain static for a number of frames are considered of interest since the wearer may have turned his head towards them as he walks or followed them as the object moves. A similar idea is used in the work of Cheatle [16] that show impressive results for the summarization of wearer's attention, during say, a day out in the zoo.

In the case of an autonomous wearable camera observing hand manipulation in a workspace, the background is the one that is likely to remain static on the image and a global motion cue becomes less useful — objects of interest may remain relatively fixed to a wall or table. Furthermore, one of our aims is to remove the wearer's attention from non essential tasks, and object manipulation is a good example of where this is desirable as we do not gaze to supervise all manual activities.

For selecting the area of attention, we notice that there is a set of fine manipulation tasks that involves both hands working spatially close to the user's chest. Hands are detected using the skin detection method described in section 5.1, and the centre of mass (**COM**) derived for the de-

[1]Original VRML fi gure by Cindy Ballreich, 3Name3D

tected skin. When there is a single sleeved hand in view, the **COM** will usually lie within a skin patch, but when both hands are involved in fine handling or manipulating, the **COM** falls within or near the object subject to manipulation. This allows the wearable camera to centre and follow a single hand in view as it reaches for objects, or when both hands are involved in fine manipulation, it follows the area between them which is likely to fall within the object being handled. This simple procedure provides an active focus of attention.

5 Hand activity categorisation

5.1 Preprocessing

The first step in the detection of hands in colour imagery here consists of skin colour segmentation. Building a 2D colour histogram in UV-space provides a degree of illumination invariance, and, as the wearable camera delivers image pixels with separate luminance (Y) and colour (UV) channels, this is a convenient and cheap operation.

The histogram bins are populated by manually selecting regions from training exemplars of various classes C_i, allowing the conditional probabilities $p(c|C_i)$ that an arbitrary pixel colour $c(x, y)$ originated from class C_i to be determined. The classification likelihoods for this colour are then

$$p(C_i|c) = \frac{p(c|C_i)p(C_i)}{\sum_{j=1}^{N} p(c|C_j)p(C_j)} \quad (1)$$

and classification is determined from the largest likelihood. Here the simplest case of just two classes, skin and background, with uniform priors is considered.

After classifying image pixels, high frequency noise is removed by spatial filtering with a 5×5 mask. The resulting skin image $S(x, y) = 1, 0$, for skin and background respectively, is then passed on for centre of mass detection, $\mathbf{COM} = (\hat{x}, \hat{y})$ which is straight-forward to obtain

$$\hat{x} = \frac{M_{10}}{M_{00}}; \;\; \hat{y} = \frac{M_{01}}{M_{00}}$$

where

$$M_{mn} \;\; = \;\; \sum_i \sum_j x_i^m \; y_j^n \; (S(x_i, y_j)).$$

Figure 3 shows the main steps involved in image preprocessing.

5.2 Classification cues

The validity of the observation of a manipulation event is obtained by a combination of measurements that include the image's overall area of skin, object classification and event's spatial distribution. The joint likelihood

$$P_e = P_s P_m P_D \quad (2)$$

represents the validity of event classification. Following we introduce each one of these elements.

(a) (b) (c)

Figure 3: (a) View from the wearable camera. (b) Colour classification is applied though the segmentation obtained is noise. (c) Image after spatial filtering.

5.2.1 P_s: Overall area of skin

The likelihood of having enough skin area P_s is here simply computed as the ratio of detected skin pixels to image size. The rationale is that if there is little skin on the image, there is little prospect that a valid event is being observed.

5.2.2 P_m: Object classification

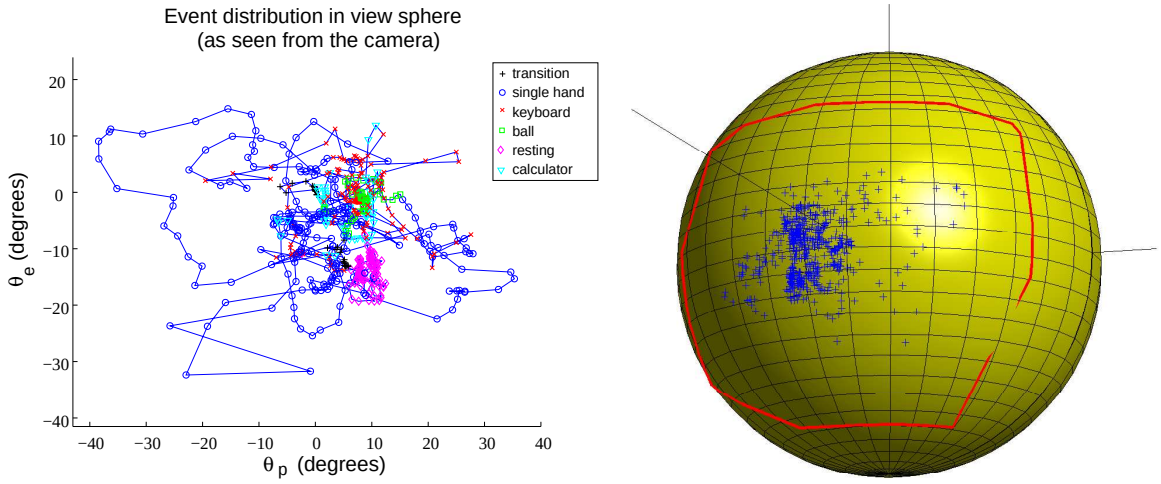
As we mention before, we use object recognition at the core of our event classification algorithm. For an object q (including the hand), a template colour histogram $\mathbf{H}_q$ with n_u bins is first derived offline from a rectangular sampling window of size n_v pixels placed over a training image region. Within the current colour image frame of an online sequence, a histogram $\mathbf{H}_k$ is computed in a sampling window V positioned at the centre of mass **COM** of skin pixels. The similarity between normalised histograms $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}_q$ is obtained via the sample of the Bhattacharyya coefficient

$$b_{kq} = \sum_{i=1}^{n_u} \sqrt{\hat{\mathbf{H}}_{ki} \hat{\mathbf{H}}_{qi}} \; . \quad (3)$$

This coefficient represents the cosine of the angle between histograms, and has been recently used with great success in the context of object tracking [17]. The class q that maximises this coefficient labels the classification. Other methods that have been used for histogram comparison include the χ^2 as used in [9, 15] and histogram intersection [8].

The use of colour histograms is favoured here as a way to gain robust classification under fast manipulations and view invariance, with the price being that we currently cant recover other object properties such as pose.

P_m, the membership to the assigned object class is directly given by Equation 3.



(a) (b)

Figure 4: (a) Event distribution around the view sphere as declared by P_m (most events lie to the right of the centre as the camera is worn on the left shoulder). (b) The view sphere (as seen from the workspace) with crosses indicating the position of the **COM** and the border line computed by the convex hull of all the observed positions during the sequence.

5.2.3 P_D: Spatial distribution of events

Further information can be obtained by incorporating some prior knowledge on the spatial distribution of events. Figure 4 show the event distribution around the camera's view sphere for a given manipulation sequence. Each symbol in Figure 4(a) represents a detected event at a given frame and the temporal path is represented by the line linking each pair of symbols. Fast single hand motions with velocities of several hundreds of degrees per second are observed (1 frame = 1/30s), and are spread within the FOV. This concentration of events has a level of resemblance to the visual exploration patterns observed by Yarbus [18].

The likelihood P_{Dk} of an observation window Vk at time k at position $(\theta_{e_k}, \theta_{p_k})$ in the view sphere belonging to class q_i, is obtained as the conditional probability

$$P_{Dk} = P(q_i|\theta_{e_k}, \theta_{p_k}) = \frac{P(\theta_{e_k}, \theta_{p_k}|q_i)P(q_i)}{\sum_j P(\theta_{e_k}, \theta_{p_k}|q_j)P(q_j)}.$$

Where θ_{e_k} and θ_{p_k} represent the angle in the view sphere (relative to the camera's placement) in the elevation and pan axis respectively. The probability $P(\theta_{e_k}, \theta_{p_k}|q_i)$ that the pair $(\theta_{e_k}, \theta_{p_k})$ belongs to q_i is drawn from a continuous density estimated by a normalised linear angular difference

$$P(\theta_{e_k}, \theta_{p_k}|q_i) = \frac{\sum_{i \neq k}(|\theta_{eVi} - \theta_{eVk}| + |\theta_{pVi} - \theta_{pVk}|)}{2(n - \frac{1}{2})}.$$

The incorporation of knowledge of where a manipulation event is expected to occur (and where it is unlikely), is used here to refine the response of event detection.

Figure 5 plots the instantaneous likelihoods P_s, P_d, P_m and the joint likelihood P_e for the manipulation sequence accompanying this paper (please see video at [20]).

Although, strictly, at some extreme conditions there is a degree of dependency between the elements composing P_e, we neglect it. If P_e is greater than a threshold, an instantaneous event is declared detected and the object class ratified, otherwise the frame is labelled as "transitional".

The response of the classification process is however noisy and filtering is necessary as described in the following section.

6 Event detection

This work is about detecting manipulation events and building visual summaries. However, this prompts the question on how to detect keyframe events?. Zelink and Irani [19] define a video event as something that usually extends for hundreds of frames, which is suitable for say a human figure crossing the FOV at a distance. However, when considering manipulation events, the spans of time involved are significantly smaller.

In this case, there are five different events to be recognised via their associated objects: Single hand (HS), hands resting on table (HR), handling a tennis ball (HB), hands operating a keyboard (HK) and hands operating a calculator (HC).

Figure 6 (a) shows the result of classification when a 0.5s mode filter that uses only frames up to the moment under analysis is used. The result is noisy, but, as the filter is causal (since we use information up to the moment of

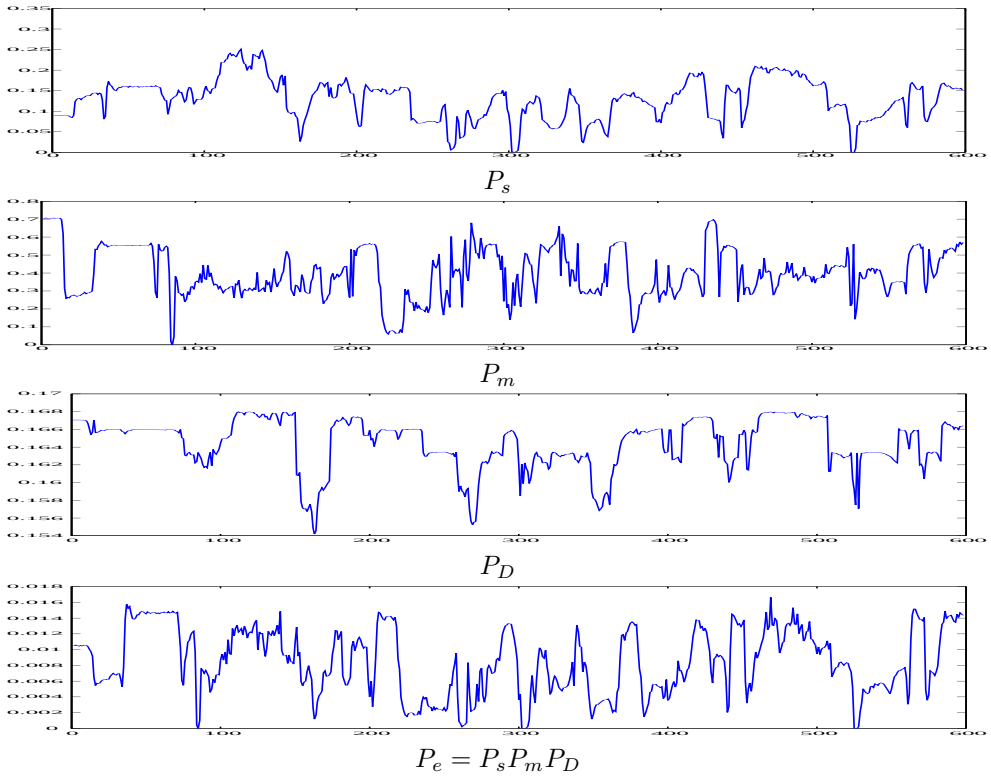


Figure 5: Likelihoods for different measurements of hand activity as a function of frame number. P_s indicates the likelihood that there is enough skin area in the image, P_m indicates the likelihood of the object belonging to the assigned class. P_D indicates the likelihood of observing a valid event based in the event distribution around the view sphere and P_e is the joint likelihood of all three. Notice that the class is assigned using Equation 3, and P_e is the confidence of such assignment.

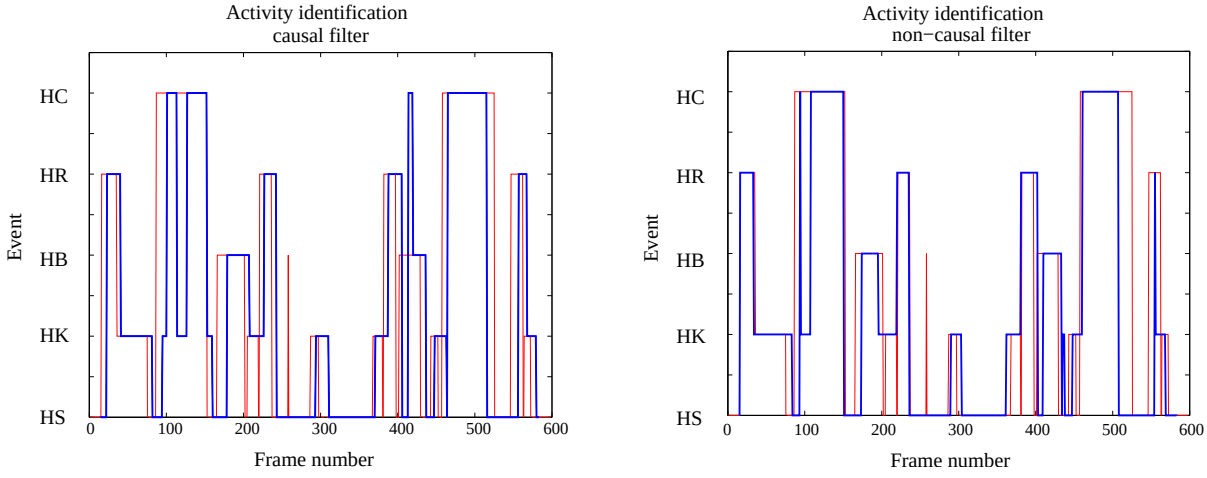


Figure 6: Activity identification results shown in bold blue as frames progress. Events are: observing single hand (HS), hands at keyboard (HK), handling a tennis ball (HB), hands resting on a table (HR) and hands operating a calculator (HC). (a) result of using a causal (on-time) mode filter. (b) results of using a non-causal (delayed) mode filter. Thin red line is the groundtruth.

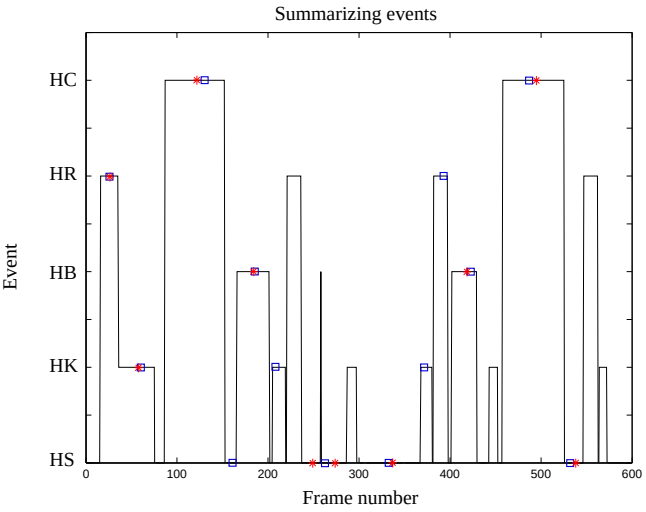


Figure 7: Results for the selection of keyframes that summarise the manipulation sequence. Stars indicate keyframes selected on the ground truth and squares those selected on the curve of detected events in Figure 6(b).

analysis), it is suited to real-time operation. Increasing the size of the window makes a cleaner recognition, however the larger this becomes the more out-of-phase the classification would be. If we relax the causality condition so that the filter can work delayed using events ahead of the moment under analysis, the filtering window can be larger and still produce in-phase smoother transitions. This is shown in Figure 6(b) for a $\pm$0.5s window. The fit to groundtruth is roughly the same for both filtering conditions (differing in about 3%), but a smoother state recognition simplifies further stages of processing. For example, it allows an event detector routine to summarise the manipulation sequence.

Figure 7 show the automatically selected *keyframe* manipulation events obtained by considering state continuity that lasted for a minimum of 0.5s. When an event lasts longer than this threshold, the midpoint gives the index to the selected keyframe. The algorithm detects all the peaks similarly detected for the groundtruth, and the extra "hallucinated" events correspond to real activity (they lie over the groundtruth) but activity that lasted for less than the 0.5s threshold.

7 Summarization experiment

For experiments we use various office objects that are manipulated by the wearer and observed from the wearable active camera producing a 600 frames sequence (please see video at [20]). In the experiments reported here, V has $n_v = 40^2$ pixels, corresponding to about 5° of FOV, and the total number of histogram bins is 256.

Images usually contain several of the recognisable objects, but the attentional filter based on the **COM** "illuminates" only the area of interest, area that is fed into the object recognition stage.

Figure 7 show when detected events occur, but of more immediate visual impact are the images themselves automatically selected to summarise the sequence. These are shown in Figure 8. The summarization algorithm considers the results shown in Figure 7 but it now enforces a threshold $P_e \geq 0.002$ to declare a valid event. Figure 9 show some of those frames that have $P_e < 0.002$ and are thus declared outliers or non-events.

8 Discussion and future work

Most work involving hands and computers (wearable or not) has been concentrated in recognizing *gestures*. In this work however, we are interested in the detection of hand *activity* as a cue to context and intention understanding.

The hand's high number of degrees of freedom and swift motions involved in common manipulation, demands the use of robust methods to detect hands and a way to detect the actions being performed. Here we have used a number of simple yet robust techniques within a probabilistic framework for the detection and summarization of hand activity. The focus of attention is linked to the centroid of skin colour which is tracked by a wearable active camera and the objects falling under a window driven by the focus of attention, categorized. The recognition of these events allow us to build a visual summary of wearer's hand activity.

Here, object templates are pre-learnt as a way to recover hand activity but without directly recognizing hand gestures. Certainly it would be possible to attempt to categorize objects automatically, and for this a number of well known clustering techniques such as KNN could be used. Also, the combination of the presented method and a more robust hand tracker e.g. the one presented in [4] would be desirable.

We believe that recognizing hand activity will open a number of applications in context recognition useful in the form of assistive devices and the building of models of interruptibility.

References

[1] J.M. Rehg and T. Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Workshop on Motion of Non-Rigid and Articulated Bodies*, pages 16–24, 1994.

[2] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conf. Computer Vision*, volume 2, pages 3–19, 2000.

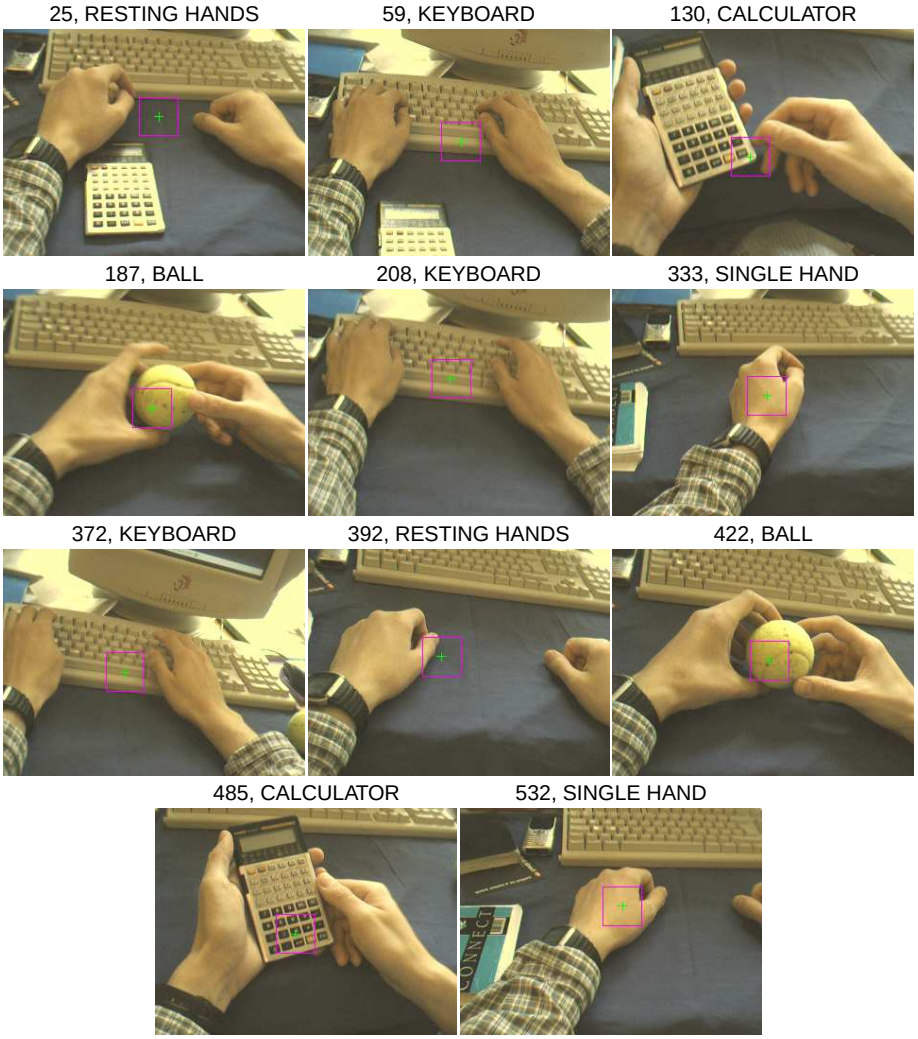


Figure 8: Automatically obtained keyframes that summarise a sequence of 600 frames of continuous manipulation of office objects. The wearable active camera follows the skin's centre of mass indicated by the cross and recognises objects within the square region via colour histograms. Labels indicate the state of the identified handling activity. Numbers indicate frame index for the video at [20].

Figure 9: Some of the detected outliers which have a joint likelihood P_e below threshold. Examples include errors due to colour misclassification, low skin area and events distant from their densities in the view sphere. Numbers indicate frame index and numbers in brackets their likelihood.

[3] M. Kölsch, M. Turk and T. Höllerer. Vision-based interfaces for mobility. In Proc. MobiQuitous '04 (1st IEEE Intl. Conf. on Mobile and Ubiquitous Systems: Networking and Services), pages 86-94, Boston, MA, Aug. 22-26, 2004.

[4] M. Kölsch and M. Turk. Robust hand detection. In *IEEE Intl. Conference on Automatic Face and Gesture Recognition*, pages 614–619, 2004.

[5] T. Starner, J. Weaver, and A. Pentland. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, December 1998.

[6] T. Kurata, T. Okuma, M. Kourogi, and K. Sakaue. The hand mouse: GMM hand color classifi cation and mean shift tracking. In *Second Int Workshop RATFG-RTS*, pages 119–124, 2001.

[7] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Filtering Using a Tree-Based Estimator. In Proc. 9th IEEE International Conference on Computer Vision, Vol. II, pages 1063-1070, Nice, France, October 2003.

[8] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

[9] B. Schiele and A. Pentland. Probabilistic object recognition and localization. In *International Conference on Computer Vision*, pages 177–182, 1999.

[10] W.W. Mayol, B. Tordoff, and D.W. Murray. Wearable visual robots. In *IEEE Int. Symposium on Wearable Computers*, pages 95–102, 2000.

[11] W.W. Mayol, B. Tordoff, and D.W. Murray. Designing a miniature wearable visual robot. In *IEEE Int. Conference on Robotics and Automation*, Washington DC, 2002.

[12] W.W. Mayol, A.J. Davison, B.J. Tordoff, N.D. Molton and D.W. Murray. Interaction between hand and wearable camera in 2D and 3D environments. In *Proc. British Machine Vision Conference*, BMVA. London UK, September 2004.

[13] A.J. Davison, W.W. Mayol and D.W. Murray. Real-Time Localisation and Mapping with Wearable Active Vision. In *Proc. IEEE International Symposium on Mixed and Augmented Reality*, IEEE Computer Society Press, Tokyo Japan. October 2003.

[14] A. Black and K. Lenzo. Flite: a small fast run-time synthesis engine. In *4th Speech Synthesis Workshop ISCA*, Scotland, 2001.

[15] B. Schiele and A. Pentland. Attentional objects for visual context understanding. Technical Report 500, MIT Media Lab, 1999.

[16] P. Cheatle. Media content and type selection from always-on wearable video. In *17th International Conference on Pattern Recognition ICPR*, volume 4, pages 979– 982, Aug 23-26 2004.

[17] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 25(5):564–577, 2003.

[18] A. L. Yarbus. *Eye movements and vision*. Plenum Press, 1967.

[19] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 123–130, December 8-14 2001.

[20] http://www.cs.bris.ac.uk/∼wmayol/videos/handactivity05.mpg