Review of : Jianxin Wu et al. *A scalable approach to activity recognition based on object use.* In International Conference on Computer vision (ICCV) 2007.

Jesse Hoey
Review for CS793

Introduction and Claims

The paper presents an approach for recognizing the activities of humans by using the objects they are handling. Activity recognition is proposed as a method with broad applications in

> "computer-assisted care and workfow optimisation, human-centered computing, and surveillance" [1].

The basic premise is that human activity could be recognized by simply looking at what objects are being used. If the person is holding the kettle, for example, they are likely making tea. The approach uses Radio-Frequency Identification Tags (RFID) to bootsrap learning models of objects in videos. A Dynamic Bayesian network (DBN) is defined that links RFID readings and SIFT features to object categories. The object categories, in turn, depend on the activity being performed. The parameters of the object model are learned from data using the expectation-maximization (EM) algorithm. The method is tested on three videos and results show that the method can leverage the RFID readings to learn accurate models of video frames, and that activities could be recognised at 80% and above.

The main claims made by the paper are as follows:

- Human activities can be recognised by identifying objects being used.
- Object and Activity models can be acquired automatically.
- RFID sensors can be used to bootstrap learning of computer vision models.

The assumptions made in the paper are as follows:

- Single activity per video. One human is moving and doing one thing in each video. They need this so they can reliably associate sensor readings with one person' activities.
- Single person per activity. Again, same reason as above. Avoids modeling more than one person (which would need to be done in the general case)
- Single object per video frame. This lets them avoid having multiple object models, all conditioning the same set of SIFT features, hence causing a data association problem.

[1] *quote from the paper*

- No or little occlusion (object is always visible). They don't explicitly state this, but it seems that their segmentation would need to pick *something* up for anything reasonable to happen.

Method outline

Here is an outline of the method.

Specify object-activity models (Ontology) This is the K-D part. A set of activities are specified. They are given names. A set of objects are specified. They are given names. Associations are drawn between the activities and the objects.

Specify object and activity dynamics models The dynamics models are defined, again from *a priori* information (prior knowledge of the domain).

Instrument the environment with RFID (noisy, faulty) The RFID tag reader is wearable as a wristband. The passive tags are located on objects. One inherent tension in the paper is the tradeoff between computer vision and RFID. The RFID tags are not very trustworthy (have lots of errors). Although these errors can be overcome, this is thought to be fraught with danger.

Segment all videos using change detection Simple back-ground subtraction method is used where pixels are grouped into super-pixels, and then blocks are identified in which change has happened. These regions are used for SIFT feature extraction and training in the next step.

Learn SIFT models of segmented regions SIFT models are extracted from the segmented regions. SIFT features are extracted and are binned into coarse clusters. The SIFT models are conditioned on the OBJECT index and all SIFT features are considered independent.

In training videos: given RFID tag readings, learn object-SIFT model Train the full model using EM. EM essentially estimates the parameters of $\theta \sim P(V|O)$ where V is unknown and O is known. It does this by first estimating $P(O|V,\theta)$ given the current guess at θ. This allows it to *fill in* the data - it now knows V and O (although O is actually not really known, but our estimate from θ is used, so we have a distribution over possible values of O). Now that V and O are both known, they can be used directly to re-measure θ' by counting co-occurrences of V and O in the data. Basically the method uses EM to learn $\theta \sim P(V|O)$

In test videos: recognise activities based on object use Testing is done using four different conditions as follows:

- Train and test on same video (are activities recognised under ideal conditions?)
- Train on one video, test on another (do the models generalise?)
- Train with ground truth object labels (are objects being learned properly?)
- Deal with missing RFID tags

Discussion

There is a tension between top-down *knowledge-driven* (K-D) approaches and bottom-up *data-driven* (D-D) approaches to human activity learning and classification. On the one hand, a system could be completely engineering using a K-D method, and simple sensors could reliably detect simple things. A complex enough ontology of everything will enable this to work. On the other hand, a D-D system could be undefined to start with and could *learn* models of activities from the data directly. The paper presents a well constructed blend of K-D and D-D approaches (a KD-D approach?) and shows how commonsense knowledge (of what activity is being performed) can help a D-D approach.

We can revisit the claims made, and see if they were borne out by the paper.

- Human activities can be recognised by identifying objects being used. This claim is verified by Tables 3 and 4 in the paper. The recognition can take place at rates of up to 80%. There is the concern that the presented rates of 80% are not enough. However, one must recall that 80% may be accurate enough to make a good decision, and this is all we really care about. So it depends what you plan to use the activity recognition results for. If it is for something more mission critical (e.g. for tele-operated surgery), then this 80% may not be sufficient and other methods would need to be consulted. If we look back at the author's claims, they say that activity recognition via object use is "viable", which certainly is borne out by the experiments, since "viability" just means "possible", without making any statements about at what level of recognition.

- Object and Activity models can be acquired automatically. This is tested in section 4.3 and rates of 90% for object models are reported. I would say this does, in fact, mean that this method works for automatic acquisition of these models.

- RFID sensors can be used to bootstrap learning of computer vision models. The performance of the vision-only techniques seems to rival that of RFID+vision, meaning that all teh relevant information from RFID is captured in the computer vision models. Therefore, I would say this claim is validated.

The take home message is that activity recognition is a challenging problem, but that using an elegant combination of top-down and bottom-up approaches, along with some solid probabilistic sensor fusion techniques, can be an accurate method. However, more accuracy would be desirable. Next steps would be to remove some of the assumptions from the system, such as a single object, person, or activity. This, however, poses challenges that could not be easily overcome using the current model, and more complex models would have to be proposed.