Egocentric Activity Monitoring and Recovery

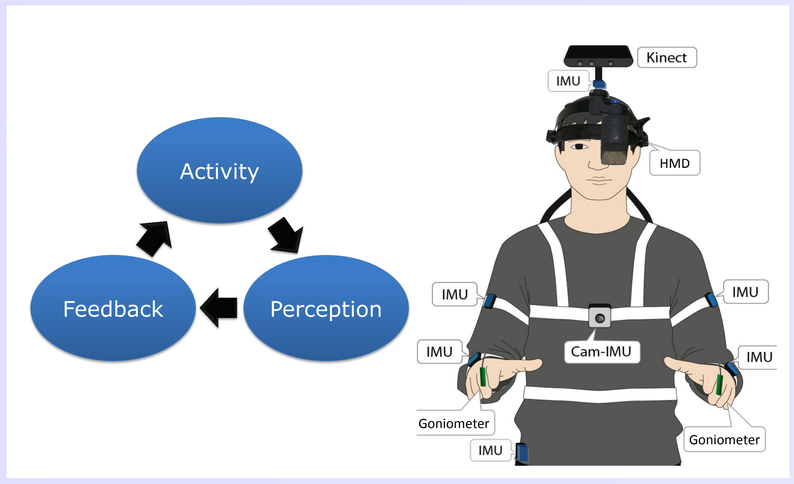
A. Behera, D. C. Hogg and A. G. Cohn
School of Computing, University of Leeds, UK
{A.Behera|D.C.Hogg|A.G.Cohn}@leeds.ac.uk

UNIVERSITY OF LEEDS

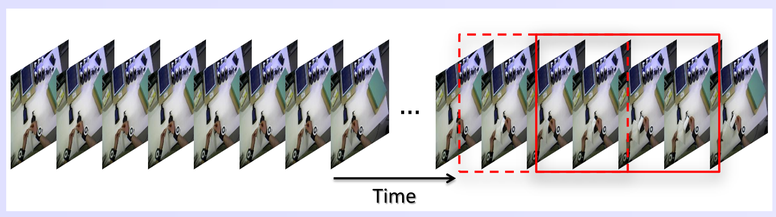
Cognito

SEVENTH FRAMEWORK PROGRAMME

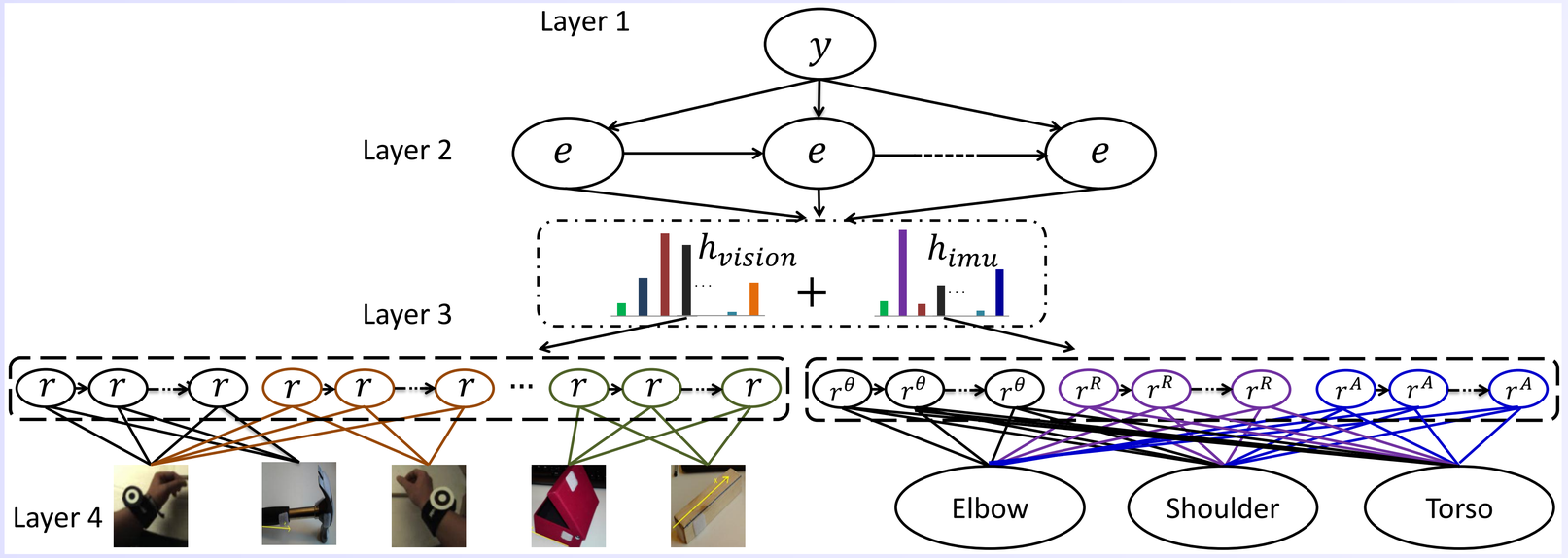
Motivation



In *bag-of-features* approaches, detection accuracy is often dependent on spatio-temporal distributions of features [1]. Additionally, these approaches are focused for classifying activities after fully observing the entire sequence.



Assignment of atomic event and activity labels to the live image sequences.



Overview of our hierarchical framework: atomic events e are inferred using spatiotemporal pairwise relations r from observed objects and wrists, and relations r^{θ}, r^{R} and r^{A} between body parts (*elbow-shoulder* and *shoulder-torso*) using inertia sensors. Activities y are represented as a set of temporally-consistent e.

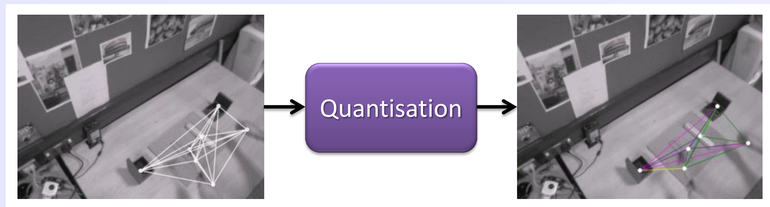
In the proposed hierarchical framework, we address the following principal contributions:

1. Recognition of atomic events using on-body sensors in order to assist users by providing *on-the-fly* instructions.
2. A learnt representation for the spatial and kinematic relationship between pairs of objects.
3. A histogram-based representation that summarises the relational structure between sets of objects within a temporal window, and provides the basis for atomic event classification.
4. Demonstrates the viability of the approach within an industrially motivated setting.

Pairwise Relational Features

At each time step t, the spatiotemporal relation between the objects o_m and o_n is represented by

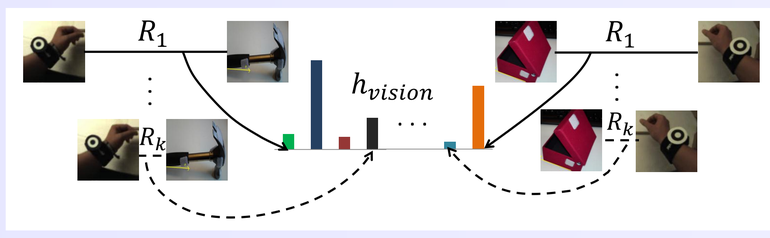
$$\mathbf{r} = (d_{m,n}, \frac{\dot{d}_{m,n}}{d_{m,n} + \epsilon}) \in \Re^2$$



Category-specific bin assignment to BoR histogram.

We describe the relational feature $\mathbf{r}$ with K possible relational words $\alpha_1 \dots \alpha_K$ by creating the relational vocabulary.

$$\alpha(r) = \arg\min_{\forall \alpha} D(\alpha, r)$$



Category-specific bin assignment to BoR histogram.

The relational words are encoded along with object category to represent a unique bin in the histogram of our *bag-of-relations* i.e.

$$b_i = ||\mathcal{A}_i||$$
$$\mathcal{A}_i = \{(\alpha_k, p)\}, k \in \{1 \dots K\} \text{ and } p \in \{1 \dots P\}\}$$

Learning and Inference

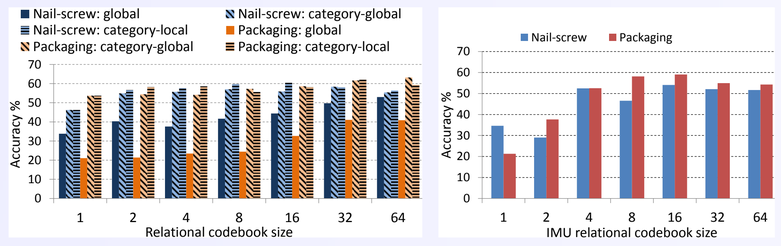
$h_{imu,t}$ and $h_{vision,t}$ are generated using *bag-of-relations*. Then a discriminative function $e = f(h_{imu,t}, h_{vision,t})$ is learned using a multi-class SVM. Then, $P(e_t|e_{t-1})$, $P(e_1)$ and $P(e_t|e_1 \dots e_{t-1}, y)$ are learned from the training examples. During prediction at time t, the most probable atomic event $\bar{e}_t$ and activity ($\bar{y}_t$) label corresponding to the observed histogram $\bar{h}_t$ are computed as:

$$P(\bar{e}_t|\bar{h}_t) \propto P(\bar{e}_t)P(\bar{e}_t|\bar{h}_t, f(\bar{h}_t))P(\bar{e}_t|\bar{e}_{t-1})$$
$$\bar{e}_t = \arg\max\{P(\bar{e}_t|\bar{h}_t)\}$$
$$\bar{y}_t = \arg\max\{P(\bar{y}_t|\bar{e}_t)P(\bar{e}_t|\bar{h}_t)\}$$
$$\bar{e}_{t+1} = \arg\max\{P(\bar{e}_{t+1}|\bar{e}_1 \dots \bar{e}_t, y)\}$$

Experiments

We have evaluated our hierarchical framework on two datasets: 1) *hammering nails and driving screws* and 2) *labelling and packaging bottles*. We used a sliding window of 2-seconds duration with 50% overlap and 'one-vs-all-subject' evaluation strategy.

Snapshots from the *hammering nails and driving screws* and the *labelling and packaging bottles* datasets.



Average performance with varying relational codebook size for 'one-vs-all-subjects' experiments using only h_{vision} (left) and h_{imu} (right).

Results and Discussion

All of our results are presented as classification accuracy over all windows.

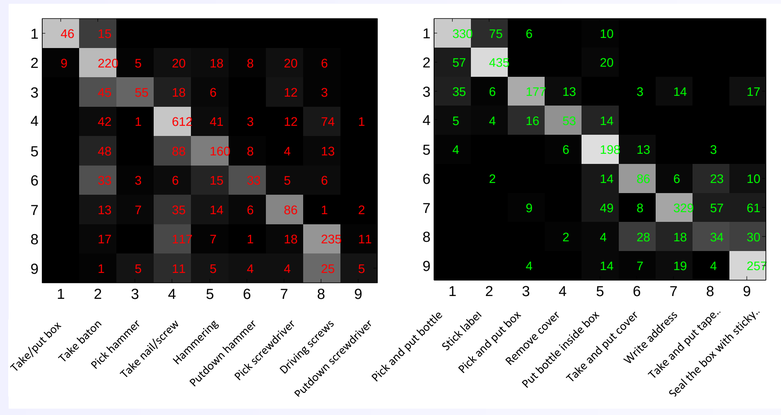
Hammering nails and driving screws.

	Vision	IMU	STIP	Vision IMU	Vision STIP	IMU STIP	Vision IMU STIP
s_1	**65.7**	65.2	65.9	73.4	78.1	70.7	75.4
s_2	64.5	**67.5**	67.2	72.3	73.4	77.5	77.2
s_3	61.7	53.5	**73.1**	62.0	72.2	64.9	68.4
s_4	**38.0**	10.3	9.2	25.9	35.6	11.0	18.7
s_5	72.5	**74.0**	77.4	80.3	82.1	84.7	86.6
Avg	**60.5**	49.8	58.6	62.8	68.3	61.7	65.3

Labelling and packaging bottles.

	Vision	IMU	STIP	Vision IMU	Vision STIP	IMU STIP	Vision IMU STIP
s_1	**61.3**	38.2	31.4	62.4	64.5	36.3	65.4
s_2	**53.5**	71.2	50.5	67.7	64.5	75.4	78.3
s_3	**63.0**	59.9	61.9	80.8	66.6	69.4	82.1
s_4	**76.1**	74.8	56.0	85.6	84.5	71.3	89.4
s_5	**56.5**	51.3	56.5	70.4	65.3	66.6	70.4
Avg	**62.1**	59.1	51.3	73.4	69.1	63.8	77.1

For both datasets, vision (60.5%, 62.1%) performs better than the other two individual representations IMU (49.8%, 59.1%) and STIP (58.6%, 51.3%).



Confusion matrix using $h_{vision+imu}$ for a) *hammering nails and driving screws* and b) *labelling bottles and packaging* dataset.

Acknowledgment

This research is supported by EU FP7 grant to the COGNITO (www.ict-cognito.org, ICT-248290) project. We also thank our collaborators in the COGNITO partners.

References

[1] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: *CVPR* (2008).