# Detecting Activities of Daily Living in First-person Camera Views

Hamed Pirsiavash Deva Ramanan
Department of Computer Science, University of California, Irvine
{hpirsiav,dramanan}@ics.uci.edu

## Abstract

*We present a novel dataset and novel algorithms for the problem of detecting activities of daily living (ADL) in first-person camera views. We have collected a dataset of 1 million frames of dozens of people performing unscripted, everyday activities. The dataset is annotated with activities, object tracks, hand positions, and interaction events. ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open). We develop novel representations including (1) temporal pyramids, which generalize the well-known spatial pyramid to approximate temporal correspondence when scoring a model and (2) composite object models that exploit the fact that objects look different when being interacted with. We perform an extensive empirical evaluation and demonstrate that our novel representations produce a two-fold improvement over traditional approaches. Our analysis suggests that real-world ADL recognition is "all about the objects," and in particular, "all about the objects being interacted with."*

## 1. Introduction

Activity recognition is a classic task in computer vision, but is relatively less well-defined compared to neighboring problems such as object recognition for which large-scale, established benchmarks exist [6, 5]. We believe this is so the following reasons: (1) It is difficult to define canonical categories of everyday behavior outside particular domains such as surveillance and sports analysis. (2) It is difficult to collect large-scale footage with rich intra-class variation. For example, unscripted surveillance footage tends to be repetitive, often dominated by scenes of people walking.

Traditionally, the above limitations have been addressed by using actor-scripted video footage of posture-defined action categories such as "skipping" or "jumping" [35, 11]. Such categories maybe artificial because they tend not be functionally defined, a core aspect of human movement [1].

We focus on the problem of detecting activities of daily living (ADL) from first-person wearable cameras. This formulation addresses many of the limitations described above, in that we use a natural list of daily activities developed from the medical literature on rehabilitation. These activities are chosen so as to capture the representative movements a person must undergo to perform everyday functions, such as eating and maintaining personal hygiene. Wearable cameras also provide a practical advantage of ease of capture; we have amassed a *diverse, 1 million-frame* dataset of people performing natural, everyday activities in diverse home environments. We argue that ease of data collection is one important benefit of wearable cameras.

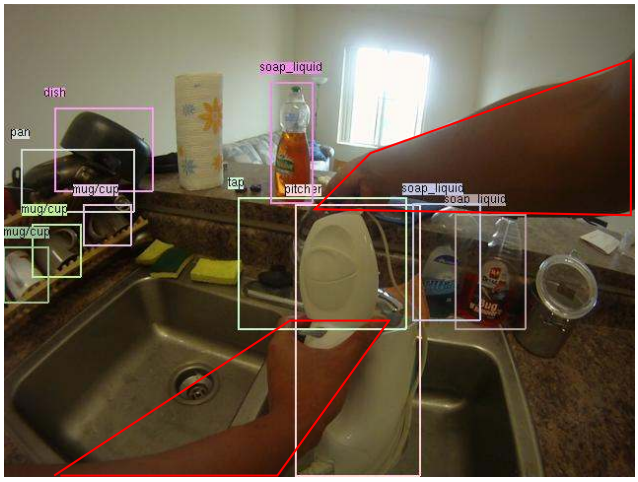



Figure 1: Activities of daily living (ADL) captured by a wearable camera.

**Application 1 (Tele-rehabilitation):** The medical literature on nursing and motor rehabilitation [21, 3] describes a variety of clinical benchmarks used to evaluate everyday functional activities such as picking up a telephone, drinking from a mug, and turning on a light switch, etc. We develop a taxonomy of everyday actions based on such medical evaluations (Fig.7). These evaluations are currently done in the hospital, but a computer-vision system capable of analyzing such activities would revolutionize the rehabilitative process, allowing for long-term, at-home monitoring.

**Application 2 (Life-logging):** A growing trend in ubiquitous computing is that of continual logging of visual personal histories [12, 17]. Initial work has shown promise for memory enhancement for patients with memory-loss [17]. However, as there has been limited algorithm development for processing and managing such massive records of daily activity, these systems currently suffer from behaving mostly as "write-only" memories. We believe the time is right for the vision community to consider such large-scale, "in the wild" activity recognition problems.

**Novel representations:** ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open). We develop novel representations including (1) temporal pyramids, which generalize the well-known spatial pyramid to approximate temporal correspondence when scoring a model and (2) composite object models that exploit the fact that objects look different when being interacted with.

**Dataset:** We introduce a fully-annotated dataset suitable for "egocentric" ADL-recognition. Our dataset is 1 million frames of 10 hours of video, amassed from 20 people performing non scripted ADL's in 20 different homes. Our dataset has been annotated with activity labels, bounding-box tracks of all objects in view, and annotations for which are being interacted with. With respect to existing egocentric datasets, our dataset is notable for its size and diversity of natural scenes. With respect to existing action datasets, our dataset is notable for its content of unscripted, everyday, activities collected in continuous (non-segmented) video streams. We use our dataset to perform a thorough investigation of state-of-the-art algorithms in both action and object recognition.

## 2. Related work

There has been a fair amount of work on everyday activity-recognition from the ubiquitous computing community [31, 39, 29] , much of it addressed from a "life-logging" perspective [2, 30] . Most approaches have ignored visual cues, and instead focused on alternate sensors such as RFID tags or accelerometers. This requires a fairly involved effort for instrumenting both the observer and the "outside world". One may argue that it is more practical to instrument the observer; for example, wearable cameras are easy to deploy, innocuous, and increasingly common [17].

There is a rich history of activity recognition in the vision community; we refer the reader to the recent surveys of [11, 40] for a detailed summary. Classic datasets tend to consist of scripted actions [24, 44], though recent work has looked at actions in televised and film footage [27]. Related work has also begun looking at the problem of everyday, at-home activity-recognition. Though there exists large body of work on recognizing actions from wearable cameras, most demonstrate results in a single scene, such as a kitchen or office [38, 8, 37, 15], possibly outfitted with actors in mocap suits [36]. Such a setting may allow one to assume *a priori* knowledge of the particular objects in the scene, which allows for instance-level visual recognition [16] or RFID-tagged objects [39, 43]. We focus on recognition in widely-varying, un-instrumented, "in-the-wild" environments.

Low-level features such as motion [9, 15, 20, 33] and skin-based color models [28] likely play a large role in analyzing wearable camera footage. We experimented with such cues, but found static image cues (such as image-based object-detectors) to be more stable, perhaps due to the unconstrained nature of our footage. Other researchers have examined unsupervised discovery of objects [19, 9] and actions [20] from wearable footage. We work with a list of semantically-driven actions and objects, as derived from the medical literature on ADL.

Our temporal pyramid representation is inspired by a large body of work on multiresolution models of video [18, 44]. Our model can be seen as a special case of a spatiotemporal pyramid [4]. However, we use interaction-based object models to determine spatial support rather than a spatial pyramid. Our model is also similar to the temporally-binned model of [23], but we use a weighted, multiscale, pyramid to approximate a coarse-to-fine temporal correspondence. Our interaction-based object models are inspired by studies from human vision [22] and are related to visual phrases [34], which capture visual composites of humans and objects in interactions. Our performance gains stem from the ability to capture large changes in object appearance (an open versus closed fridge) as well as the inclusion of a human in the composite model.

## 3. Temporal pyramids

In this section, we develop several simple but novel models of daily activities based on object-centric representations. We write $T$ for the set of frames to be analyzed using $K$ object models. We use these models to compute a score for object $i$ at a particular pixel location and scale $p = (x, y, s)$ in frame $t$:

$$\text{score}_i^t(p) \in [0, 1] \tag{1}$$

We use the object models of [10], which are not calibrated to return scores in $[0, 1]$. One may calibrate the models to return probabilities [32], but we divide the raw score of each object model by the maximum value found in the whole dataset. We then record the maximum value of each object model $i$ in each frame $t$:

$$f_i^t = \max_p \ \text{score}_i^t(p) \tag{2}$$

"Bag of features" is a naive way of aggregating these features by averaging them over time.

$$x_i^0 = \frac{1}{|T|} \sum_{t \in T} f_i^t \tag{3}$$

The above representation ignores any temporal structure; we may want to encode, for example, that "making tea" requires first boiling water and then (minutes later) pouring it into a cup. Such long-scale temporal structure is difficult to encode using standard hidden markov models. We develop a flexible model based on the spatial pyramid match kernel [25]. We represent features in a temporal pyramid, where the top level $j = 0$ is a histogram over the full temporal extent of a video clip (as in (3)), the next level is the concatenation of two histograms obtained by temporally segmenting the video into two halfs, and so on. We obtain a coarse-to-fine representation by concatenating all such histograms together:

$$x_i^{j,k} = \frac{2^{j-1}}{|T|} \sum_{t \in T^{j,k}} f_i^t; \quad \forall k \in \{1...2^j\} \tag{4}$$

where $T^{j,k}$ is the temporal extent of the $k$'th segment on the $j$'th level of the pyramid and $x_i^{j,k}$ is the feature for the $i$'th object detector on that segment. The scale factors define an implicit correspondence based on the finest temporal resolution at which a model feature matches a video feature [13]. We use $j \in \{0, 1\}$ levels. These allows us to encode long-scale temporal structure in a "soft" manner; one must touch a kettle at the beginning of a making tea action, but the precise time may vary a bit.

We use our models for activity recognition by learning linear SVM classifiers on features

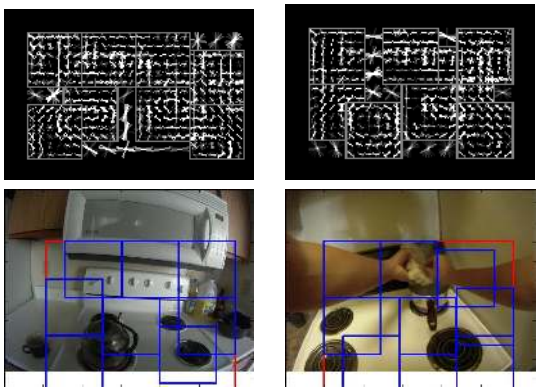
$$x = \min\Big( \begin{bmatrix} x_1^0 & \dots & x_i^{j,k} \dots & x_K^{L,2^L} \end{bmatrix}^T, 0.01 \Big)$$

with the public SVM implementation of [7]. We found an elementwise-minimum was useful to approximately "binarize" $x$, so that it softly encode the presence or lack thereof of object $i$ (inspired by the clipping post-processing step in SIFT [26]).We experimented with various histogram kernels [41], but found a simple linear kernel defined on an L1-normalized feature to work well.

## 4. Active object models

Recognizing objects undergoing hand manipulations is a crucial aspect of wearable ADL recognition (see Fig.2) [22]. Following recent work on human-object interaction models, one approach may be to detect objects and human body parts (hands) in frames, and then reason about their spatial relationship. However, this ignores the fact that objects may significantly change in appearance during interactions - an open fridge looks very different from a closed fridge.

Figure 2: Our dataset (**top row**) contains images of objects under different semantic states-of-use (e.g., a microwave with open or closed door). These semantic states are typically not captured in web-based photo collections **(bottom row)**. Our active/passive object models exploit such visual cues to determine which objects are being interacted with.



(a) passive stove (b) active stove

Figure 3: We visualize our passive and active stove models.

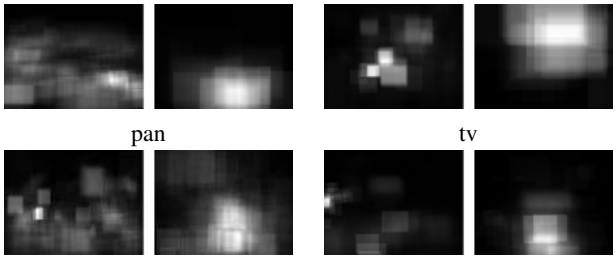



Figure 4: To visualize the average location of active vs passive objects in our ADL dataset, we make a rectangular mask for each bounding box and average them all for passive (on **left**) and active (on **right**) instances of annotated objects. Active images tend to have larger bounding boxes at the center of the image, indicating that active objects tend to occur at large scales near the center of the field of view.

Figure 5: We show the average passive (**left**) and active (**right**) TV remote in our ADL dataset. The left image is blurred due to the variation in viewing angle in our data, while the right image is more structured due to less pose variation for remotes-in-hand. The right image also contains more red-hued skin-pixels. We use such cues to build active object models.

**Active models:** Instead, we learn separate object detectors tuned for "active" objects being interacted with. Our approach is inspired by the recent "visual phrase" work of [34], which advocates detection of human-object composites rather than detectors built for each constituent object. In our case, we do not increase the spatial support of our detectors to explicitly include the human hand, but define an active object to be a visually disparate sub-category. We do this by training an additional object detector [10] using the subset of active training images for a particular object. We show an example for a "stove" object in Fig.3.

**Spatial reasoning:** While many objects can appear in the field of view, active objects tend to be at a consistent scale and location convenient for hand manipulation. We analyze the spatial bias of passive versus active objects in Fig.4. To exploit such cues, we augment our active object models to include the position and scale as additional features when detecting active objects:

$$\text{score}_i^t(p) = w \cdot \begin{bmatrix}\text{score}(p) & x & y & s & x^2 & y^2 & s^2\end{bmatrix}^T$$

Because we use linearly-parameterized templates as object detectors [10], we simply add the above spatial features to the local appearance features when learning active object models. We found this produced a small but noticeable improvement in our final results.

**Skin detectors:** Because active objects are manipulated by the hand, they tend to occur near skin pixels (as shown in Fig.5). We experimented with adding a skin detector feature to the above linear model, but failed to see consistent improvements. We hypothesize this was due to large variations in illumination in our dataset.

We augment the temporal pyramid feature from (4) to include $K$ additional features corresponding to active objects, as defined in this section. We refer to this model as "AO", for our object-centric model augmented with active objects. We refer to the original feature from (4) as "O", for our object-centric model.

# 5. Dataset

In the subsequent description, we refer to our dataset as the ADL dataset.

| action name | mean of length (secs) | std. dev. of length |
|---|---|---|
| combing hair | 26.50 | 9.00 |
| make up | 108.00 | 85.44 |
| brushing teeth | 128.86 | 45.50 |
| dental floss | 92.00 | 23.58 |
| washing hands/face | 76.00 | 36.33 |
| drying hands/face | 26.67 | 13.06 |
| laundry | 215.50 | 142.81 |
| washing dishes | 159.60 | 154.39 |
| moving dishes | 143.00 | 159.81 |
| making tea | 143.00 | 71.81 |
| making coffee | 85.33 | 54.45 |
| drinking water/bottle | 70.50 | 30.74 |
| drinking water/tap | 8.00 | 5.66 |
| making cold food/snack | 117.20 | 96.63 |
| vacuuming | 77.00 | 60.81 |
| watching tv | 189.60 | 98.74 |
| using computer | 105.60 | 32.94 |
| using cell | 18.67 | 9.45 |

Table 1: This table shows the statistics for the duration of each action. Some actions like "using cell" are shorter in time than other actions like "washing dishes". Many actions exhibit a rather large variability in duration, making action detection in continuous data difficult.

## 5.1. Collection and size

To collect our dataset, we used a GoPro camera designed for wearable capture of athletes during sporting events. We found a chest-mount easier than a helmet mount, both in terms of quality of data and ease of capture. The camera captures high definition quality video (1280x960) in the rate of 30 frames per second and with 170 degrees of viewing angle. A large viewing angle is important in capturing this type of footage to reduce object and hand truncation. We put together a list of 18 actions of daily activities and asked 20 people to do them all in their own apartment. In order to collect realistic and varied data, we didn't give users a detailed description of actions, and instead gave them the list of actions in Table 1 . Each capture session was roughly 30 minutes of unscripted morning activity. Our camera was equipped with sufficient battery and memory to allow for continuous capture. We collected more than 10 hours of first person video footage, with more than a million frames. The total collection process took one month of acquiring subjects and arranging capture sessions.

## 5.2. Annotation

We annotated every second (30 frames) with dense annotations of the form in Fig.1. We did so by assembling a team of 10 part-time annotators, working over a month span. The final dataset is annotated in terms of:

Figure 6: We show different kitchen scenes in our dataset. Unlike many other manually constructed action datasets, we exhibit a large variety of scenes and objects.

**Action label:** Our ADL dataset is temporally annotated in terms of 18 different actions. Table 1 shows the list of actions and also the statistics of their duration.

**Object bounding boxes:** Our ADL dataset is annotated in terms of 42 different objects, of which some listed in Table 2 . We asked the annotators to draw a tight bounding box around each known object and then track it and adjust the bounding box for every 30 frames.

**Object identity:** Our dataset is annotated with individual tracks of objects. We do not use such annotations for our current analysis, but it may be useful for evaluating tracking algorithms. Note that there is large amount of camera motion in this footage and can be considered a good benchmark for object detection and tracking algorithms.

**Human-object interaction:** We denote objects that are being interacted with as "active". We set a binary attribute flag for active objects in our annotation interface. We show that this knowledge is very helpful in action recognition.

### 5.3. Characteristics

In this section, we point out various distinguishing characteristics of our dataset. We refer to the following sets of figure captions for a detailed description, but we summarize the main points here. Our dataset contains large variability in scenes (Fig. 6) and object viewpoint and occlusion level (Fig. 2). In Fig. 4 and Fig. 5, we illustrate various biases (such as image location and skin color) which can be exploited to visually identify interacting objects.

**Functional taxonomy:** Many ADL actions are quite related to each other. We construct a functional taxonomy

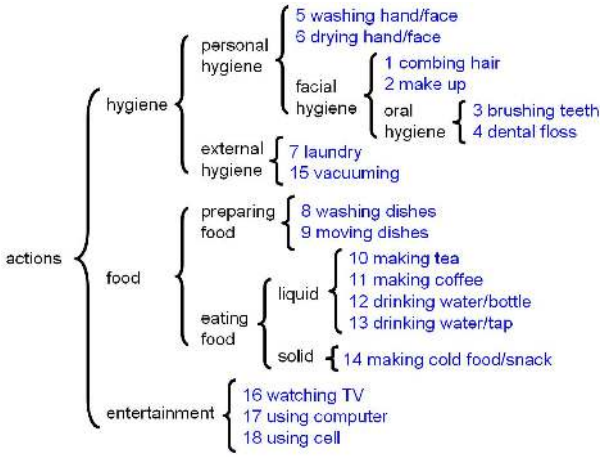



Figure 7: Our manually-designed functional ADL taxonomy.

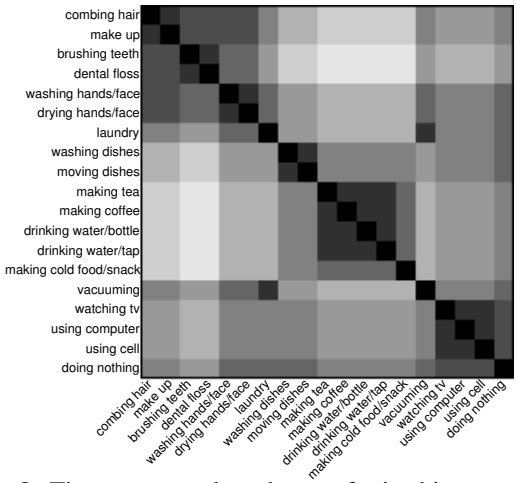



Figure 8: The taxonomy-based cost of mistaking one class for another is the (average) distance to their closest common ancestor in Fig.7 . Dark values correspond to a low cost; we pay less for mistaking "brushing teeth" with "flossing" as opposed to "making tea".

based on a bottom-up grouping of actions; at a high level, all ADL actions can be grouped into those based on personal hygiene, food, and entertainment. Fig. 7 shows the functional taxonomy we manually constructed. We can use this taxonomy in evaluating actions, meaning we penalize less for making mistakes between actions with similar functionalities. Following [14], we define the misclassification cost of two classes as the total distance to their closest common ancestor, divided by two. Fig. 8 illustrates this cost for all possible mistakes with brighter color for larger cost. We find the interesting phenomena that functionality correlates strongly with scene context; one both brushes their teeth and flosses in a bathroom. This suggests that approaches that rely on scene context might fare well under such a functional score.

Figure 9: Toothbrushes in our ADL dataset (**left**) look different than those found in web-based collections such as ImageNet (**right**). The latter typically contains large, iconic views of objects, often in displayed in isolation. Object detectors trained with the latter may not work well for our application, as we show in Table 2.

| Object | ADL | ImageNet |
|---|---|---|
| tap | 40.4 ± 24.3 | 0.1 |
| soap liquid | 32.5 ± 28.8 | 2.5 |
| fridge | 19.9 ± 12.6 | 0.4 |
| microwave | 43.1 ± 14.1 | 20.2 |
| oven/stove | 38.7 ± 22.3 | 0.1 |
| bottle | 21.0 ± 27.0 | 9.8 |
| kettle | 21.6 ± 24.2 | 0.1 |
| mug/cup | 23.5 ± 14.8 | 14.8 |
| washer/dryer | 47.6 ± 15.7 | 1.8 |
| tv | 69.0 ± 21.7 | 26.9 |

Table 2: Average precision results for part-based object detectors evaluated on our ADL dataset. We compare models trained on our ADL dataset versus ImageNet. Since the ADL-trained models are trained and evaluated across cross-validation splits, we report both the mean and standard deviation of average precision. The deviation is large because we have relatively few object instances in our dataset (people own a single tea kettle). Detectors trained on ImageNet perform poorly on our data because they fail to capture the large number of viewing angles and occlusion states present in our wearable data.

## 6. Experimental results

We implemented and evaluated our object-centric action models on our ADL dataset.

**Evaluation:** We use leave-one-out cross-validation, where we ensure that footage of the same person does not appear across both training and test data. We use average precision to evaluate object detection accuracy (following the standard convention [6]). We use class confusion matrices to evaluate action classification, scoring both classification error and the taxonomy-derived loss shown in Fig.8. We compute an overall classification rate by averaging the diagonal of this matrix, weighting all classes equally. Because we have 18 classes, chance performance corresponds to almost 5%.

**Co-occurring actions:** Some actions can co-occur in our dataset. In many cases, it may be more natural to think of the shorter action as interrupting the longer one; "watching TV" while waiting for water to boil while "making tea." Our annotations include such co-occurrences. For simplicity in our current evaluation, we assign only one label to our test frame, taking the shorter interrupting action when there is overlap.

**Training:** In training visual object detectors, we used off-the-shelf part-based model for object detection [10]. We use training data for 24 object categories with roughly 1200 training instances (with bounding-box labels) per category. In Table 2, we compare results using different training datasets. We show that models trained using web-based collections (such as ImageNet) tend to contain iconic viewpoints of images not present in our ADL dataset (Fig. 9). Additionally, wearable video contains images of objects under different states-of-use (an open microwave or fridge, as in Fig.2), also usually absent in online collections. When trained on data extracted from natural ADL footage, object detectors perform considerably better; for example, a faucet tap trained from ImageNet performs at 0.1% average precision, while faucet tap model trained from ADL data performs at 40% average precision. This suggests that, for our application, it is crucial to train on data with a large variety of viewpoints and scales. We find that there are certain objects in our labeled dataset for which current detection systems cannot model - e.g., they yield zero percent performance. We think this is due to small resolution and large geometric variation.

### 6.1. Action recognition results

Table 3 tabulates action classification accuracy for different versions of our system. We begin with a standard baseline; a SVM trained on a bag of quantized spatio-temporal interest points (STIPS) [42]. It performs fairly poorly, at 16.5% on classification of pre-segmented video clips. Adding our temporal pyramid model boosts performance to 22.8%, revealing the benefit of reasoning about temporal structure. Our bag-of-objects model (O) noticeably improves performance to 32.7%, which is further increased to 40.6% when augmented with the active-object model (AO). Our novel representations provide a factor of two improvement over contemporary approaches to action recognition.

To further analyze where future work should be focused, we evaluated our model with idealized perfect object detectors (IO), and augmented such idealized detectors with perfect knowledge of when objects are "active" (IA+IO). We do this by simply using the object and interaction annotations in our dataset. These dramatically increase performance to 77%, suggesting that for ADL recognition, "its all about the objects", and in particular, "its all about the objects being interacted with."

| | pre-segmented | | | |
|---|---|---|---|---|
| | segment class. accuracy | | taxonomy loss | |
| | pyramid | bag | pyramid | bag |
| STIP | 22.8 | 16.5 | 1.8792 | 2.1092 |
| O | 32.7 | 24.7 | 1.4017 | 1.7129 |
| AO | 40.6 | 36.0 | 1.2501 | 1.4256 |
| IO | 55.8 | 49.3 | 0.9267 | 0.9947 |
| IA+IO | 77.0 | 76.8 | 0.4664 | 0.4851 |

| | sliding window | | | |
|---|---|---|---|---|
| | frame class. accuracy | | taxonomy loss | |
| | pyramid | bag | pyramid | bag |
| STIP | 15.6 | 12.9 | 2.1957 | 2.1997 |
| O | 23.8 | 17.4 | 1.5975 | 1.8123 |
| AO | 28.8 | 23.9 | 1.5057 | 1.6515 |
| IO | 43.5 | 36.6 | 1.1047 | 1.2859 |
| IA+IO | 60.7 | 53.7 | 0.79532 | 0.9551 |

Table 3: Classification accuracy and taxonomy loss for action recognition using different representations. We compare results using both pre-segmented and temporally-continuous video clips. Please see the text for a detailed discussion, but our active-object (AO) model doubles the performance of typical action models based on space-time interest points (STIP). We also show that idealized, perfect object detectors (IO), augmented with the knowledge of which objects are being interacted with (IA+IO), dramatically increase performance.

We believe that accuracy is limited (even for the ideal case) due to genuine ambiguities in the data, as well as difficulties in annotation. Some actions such as "washing dishes" and "drinking water from tap" involve interactions with the same objects (mug and tap). Some objects are small and often occluded (e.g., dental floss) and so are not fully annotated. But it's likely that an ideal detector for such objects would be difficult to build.

One compelling aspect of ADL footage is that it is naturally collected as a continuous video stream, requiring one to solve a temporal segmentation problem. This temporal continuity is rare for action datasets, which tend to consist of pre-segmented clips. For each evaluation scenario, we train 1-vs-rest SVM classifiers on pre-segmented clips and test them on either pre-segmented or continuous videos. In the continuous case, we apply the detector within a temporal sliding window of 10 seconds and assign the best label to its center frame. We also add a background class label to the set of possible outputs in the continuous case. We score a model with its frame classification accuracy. As perhaps expected, performance decreases with respect to the pre-segmented case, but it is still reasonable.

We have constructed confusion matrices for all entries in Table 3 and will release them with the dataset. Due to space limitations, we show only a confusion matrix for our active-object (AO) model in Fig.10. Interestingly, many actions are mistaken for functionally similar ones - "make up", "brushing teeth", and "dental floss" are all mistaken for each other (and are instances of personal hygiene). We believe this holds because much of the functional taxonomy in Fig.7 is scene-based; people prepare food in a kitchen and maintain personal hygiene in a bathroom. Our bag-of-object representation acts as a coarse scene descriptor, and hence makes such functionally-reasonable mistakes.

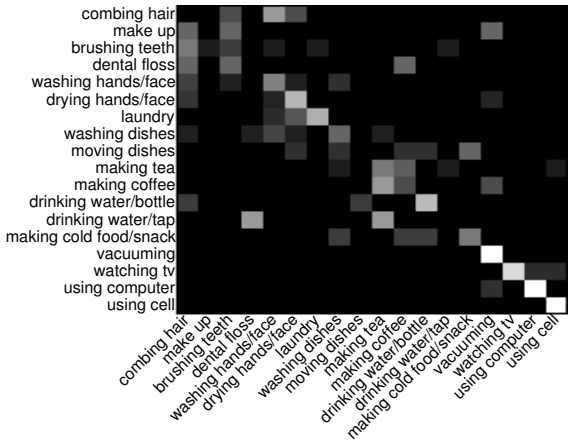



Figure 10: Confusion matrix for temporal pyramid with vision based active object detectors on pre-segmented videos. Segment classification accuracy = 40.6%.

**Conclusion:** We have presented a novel dataset, algorithms, and empirical evaluation for the problem of detecting activities of daily living (ADL) in first-person camera views. We present novel algorithms for exploiting temporal structure and interactive models of objects, both important for ADL recognition. To illustrate our algorithms, we have collected a dataset of 1 million frames of dozens of people performing, unscripted, everyday activities. We have annotated the dataset with activities, object tracks, hand positions, and interaction events. We have presented extensive experimental results that demonstrate our models greatly outperform existing approaches for wearable ADL-recognition, and also present a roadmap for future work on better models for objects and their interactions.

**Acknowledgements:** We thank Carl Vondrick for help in using his annotation system. Funding for this research was provided by NSF Grant 0954083, ONR-MURI Grant N00014-10-1-0933, and support from Intel.

## References

[1] M. Argyle and B. Foss. *The psychology of interpersonal behaviour*. Penguin Books Middlesex. England, 1967.

[2] M. Blum, A. Pentland, and G. Troster. Insense: Interest-based life logging. *Multimedia, IEEE*, 13(4):40–48, 2006.

[3] A. Catz, M. Itzkovich, E. Agranov, H. Ring, and A. Tamir. SCIM-spinal cord independence measure: a new disability

scale for patients with spinal cord lesions. *Spinal Cord*, 35(12):850–856, 1997.

[4] J. Choi, W. Jeon, and S. Lee. Spatio-temporal pyramid matching for sports videos. In *ACM ICMR*, 2008.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *JMLR*, 9, 2008.

[8] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[9] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE TPAMI*, 32(9), 2010.

[11] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion i: Tracking and animation. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):1–255, 2006.

[12] J. Gemmell, G. Bell, and R. Lueder. MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.

[13] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 2007.

[14] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[15] M. Hanheide, N. Hofemann, and G. Sagerer. Action recognition in a wearable assistance system. In *ICPR*, 2006.

[16] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of texture-less objects. *IEEE TPAMI*, 2011.

[17] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. SenseCam: A retrospective memory aid. *UbiComp*, 2006.

[18] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing: Image Communication*, 8(4), 1996.

[19] K. T. Kang H., Hebert M. Discovering object instances from scenes of daily living. In *ICCV*, 2011.

[20] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[21] B. Kopp, A. Kunkel, H. Flor, T. Platz, U. Rose, K. Mauritz, K. Gresser, K. McCulloch, and E. Taub. The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. *Arch. of physical medicine and rehab.*, 78(6), 1997.

[22] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11), 1999.

[23] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[24] I. Laptev and P. Perez. Retrieving actions in movies. In *ICCV*, 2007.

[25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[26] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[27] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[28] W. Mayol and D. Murray. Wearable hand activity recognition for event summarization. In *International Symposium on Wearable Computers*. IEEE, 2005.

[29] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In *IEEE Int. Symp. on Wearable Computers*, 2005.

[30] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE TPAMI*, 22(1):107–119, 2002.

[31] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 2004.

[32] J. Platt. Probabilistic outputs for support vector machines. *Advances in Large Margin Classifiers*, 10(3), 1999.

[33] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.

[34] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[35] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[36] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Workshop on Egocentric Vision*, 2009.

[37] L. Sun, U. Klank, and M. Beetz. Eyewatchme3d hand and object tracking for inside out activity analysis. In *IEEE Workshop on Egocentric Vision*, 2009.

[38] S. Sundaram and W. Cuevas. High level activity recognition using low resolution wearable vision. In *IEEE Workshop on Egocentric Vision*, 2009.

[39] E. Tapia, S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, pages 158–175, 2004.

[40] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Trans on*, 18(11):1473–1488, 2008.

[41] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3539–3546. IEEE, 2010.

[42] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[43] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. In *ICCV*, pages 1–8. IEEE, 2007.

[44] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, 2001.