Real-Time Crowd Labeling for Deployable Activity Recognition

Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham
University of Rochester Computer Science
{wlasecki, ysong, kautz, jbigham}@cs.rochester.edu

ABSTRACT

Systems that automatically recognize human activities offer the potential of timely, task-relevant information and support. For example, prompting systems can help keep people with cognitive disabilities on track and surveillance systems can warn of activities of concern. Current automatic systems are difficult to deploy because they cannot identify novel activities, and, instead, must be trained in advance to recognize important activities. Identifying and labeling these events is time consuming and thus not suitable for real-time support of already-deployed activity recognition systems. In this paper, we introduce Legion:AR, a system that provides robust, deployable activity recognition by supplementing existing recognition systems with on-demand, real-time activity identification using input from the crowd.

Legion:AR uses activity labels collected from crowd workers to train an automatic activity recognition system online to automatically recognize future occurrences. To enable the crowd to keep up with real-time activities, Legion:AR intelligently merges input from multiple workers into a single ordered label set. We validate Legion:AR across multiple domains and crowds and discuss features that allow appropriate privacy and accuracy tradeoffs.

Author Keywords

Activity Recognition; Crowdsourcing; Human Computation

ACM Classification Keywords

H.4.m Information interfaces and presentation: Misc.

INTRODUCTION

Automatically recognizing human activities offers the potential for timely, task-relevant information and support in diverse domains, such as prompting systems that keep people with cognitive disabilities on track [34], smart homes that detect when to summon help so that older adults can live independently longer [29], and surveillance systems that detect troublesome behavior in busy environments [21]. Unfortunately, recognizing activities in the real world is difficult. The same high-level activity may be performed in very different ways by different people, with different tools, or in different contexts. As a result, identifying activities robustly can require both commonsense knowledge and high-level reasoning. Due to this, automatic approaches to recognizing these activities remain brittle, with the performance of systems highly dependent on the particular instances of an activity and context in which a system has been trained. People are able to recognize these activities well, but are generally not able to supply labels quickly or consistently enough to use in a real system. Moreover, monitoring services that employ trained human workers are prohibitively expensive and introduce privacy concerns. As a result, activity recognition systems are expensive to deploy and difficult to scale.

In this paper, we introduce Legion:AR, a system that combines the benefits of automatic and human activity labeling for more robust, deployable activity recognition. Legion:AR uses an active learning approach [34] in which automatic activity recognition is augmented with on-demand activity labels from the crowd when an observed activity cannot be confidently classified. In contrast to prior systems that have used human labelers to train activity recognition systems offline, Legion:AR is designed to work on-demand and in real-time, allowing the human labels to be integrated directly into the deployed system. By engaging a group of people (*the crowd*), Legion:AR is able to label activities as they occur more reliably than a single person can, especially in complex domains with multiple actors performing activities quickly.

Involving humans in activity recognition naturally raises privacy and confidentiality concerns. In addition to the crowd being a source of anonymous workers who can be completely unknown to the user, Legion:AR includes two features to provide application designers flexibility when choosing how to address these concerns. First, Legion:AR can automatically detect people and veil them by showing only a colored silhouette (Figure 1). We show that, in some common settings, hiding visual details does not significantly affect the quality of the labels generated by human workers. Second, Legion:AR allows designers to provide application users the ability to cancel crowd labeling of activities if they respond within a specified time window, e.g. opt-out via a prompt on a mobile device. Finally, in some domains, people may be willing to give up some privacy to receive the benefits of the system. For instance, older adults may make privacy sacrifices in order to independently live in their homes longer [20]. Because the crowd can be engaged only when automatic activity recognition fails, many of the privacy benefits of a fully-automated system are maintained using Legion:AR.

CSCW'13, February 23–27, 2013, San Antonio, Texas, USA.

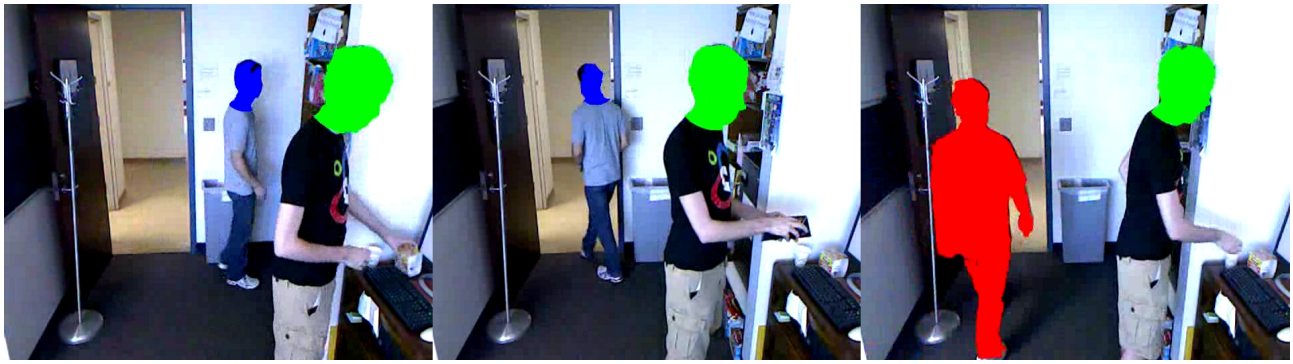

Figure 1. Worker view of a multi-actor scene from the monitoring domain in which people have been automatically identified and veiled in separate colors to preserve privacy. These veils can cover only the face (as seen in the left two panels), or the entire silhouette of the user ('red' user in the right most panel). Our experiments show that this veiling does not decrease workers' ability to accurately label the activities they are performing.

We explore Legion:AR in the context of two systems that use activity recognition. The first is a prompting system designed to help people with cognitive disabilities keep on track while performing activities of daily living, e.g. making dinner [8]. We test Legion:AR on a set of eight routine tasks in a home environment by training separate Hidden Markov Models (HMMs) using crowd labels generated in real-time and expert labels generated offline. We show that the crowd can outperform single users and that using the labels suggested by the HMM, crowd workers more quickly and accurately converge to consistent labels. The second is a surveillance system that uses a Microsoft Kinect 3D camera to gather visual data from self-service snack store in our department. We then used Legion:AR to label the activities of people near the store, e.g. choosing a snack, buying a snack, walking past. With this example, we demonstrate both that Legion:AR is able to label the activities of multiple actors simultaneously, and that the system can label activities accurately even when people are veiled in a colored silhouette to conceal their identity. We then show that workers are able to jointly generate accurate labels at different hierarchical levels, allowing Legion:AR to collect a richer sets of tags than even trained labelers.

Our contributions are the following:

- We introduce a system that enables the crowd to collectively label activities on-demand in real-time, and performs active learning from crowd-generated labels.
- We demonstrate that groups of workers (even anonymous web workers) are able to reliably generate *consistent* labels, in real-time, that are more accurate and complete than that of a single worker with the help of our system.
- We articulate the idea for creating more deployable activity recognition systems by supplementing automatic approaches with on-demand crowd labeling.
- We present features that application designers can employ to appropriately balance privacy and functionality in systems that use activity recognition.
- We demonstrate that workers can label activities of desired individual actors in scene containing multiple people, with varying granularity, even while preserving the privacy of users by automatically veiling people in video.

THE CROWD

We identify users as people whose activities are being identified and workers as members of the crowd helping to generate label data. We define *the crowd* as a dynamic pool of possibly anonymous workers of varying reliability that can be recruited on-demand. Because the pool is dynamic, workers come and go, and no specific worker can be relied upon to be available at a given time or to continue working on a job for a set amount of time. Workers cannot be relied upon to provide high-quality work of the type one might expect from a traditional employee for various reasons including misunderstanding of task directives, laziness, or even maliciousness. Finally, workers may experience delays that are beyond their control, such as network bandwidth variability.

For enabling real-time activity recognition, the dimensions of the crowd that are most relevant are (i) the time each recruited worker continues working on the task and (ii) the quality of the worker's output. These can be measured empirically for a specific crowd source, but are expected to be task-dependent [30]. A related dimension is the latency required to recruit workers to a particular job. Our system builds on prior work that has shown workers can reliably be recruited in seconds by using pre-recruiting and retainer pools [3, 4].

In this paper, we run experiments using pools of workers drawn both from local study participants and from Amazon's Mechanical Turk. For some applications that require real-time activity recognition, companies may find value in offering vetted pools of workers [2]; in others, family members or other trusted individuals may form the crowd. As a web-based framework, Legion:AR is compatible with diverse worker pools, and users may select the types of crowds that will power the system. Even when workers are trusted, Legion:AR is still a benefit because it allows workers the flexibility to come and go, and allows the group to collectively provide better and more complete labels than the average constituent labeling individually.

RELATED WORK

Legion:AR builds on work in both activity recognition and real-time human computation. In recent years, the development of small unobtrusive sensors has made it practical to create smart environments that can both recognize our ac-

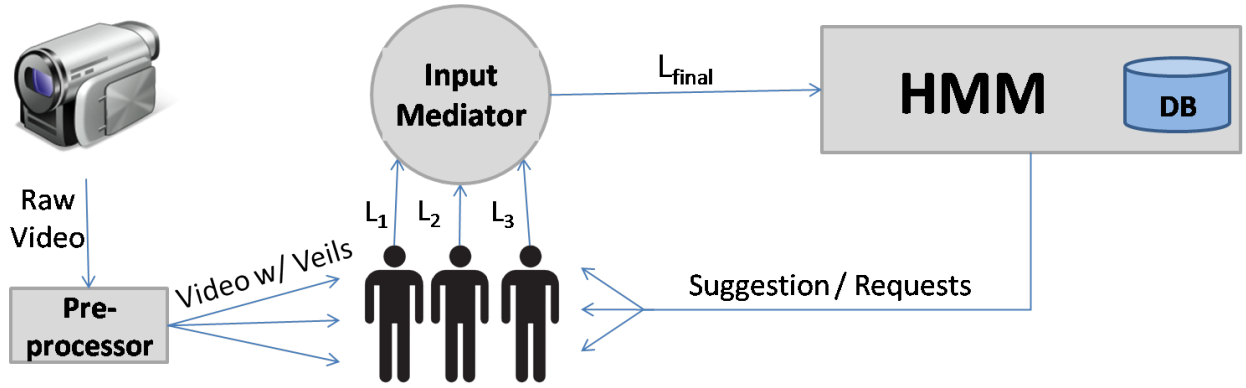

Figure 2. The Legion:AR system. When the automatic system (HMM) is unconfident about an activity label, it requests labels from the crowd. Crowd workers are able to see the system's predicted activity label and either validate it or propose a new label of their own. The labels from the crowd workers (L_i) are sent back to Legion:AR, which then determines a final label and uses it and the associated segment of sensor data to train the HMM. The HMM adds this to its model in order to identify the activity later and forward that prediction back to the crowd.

tions and act to improve our experience in the environment. The importance of these environments has emerged especially in the area of automated health monitoring, where there is tremendous need to promote aging in place and improve quality of life for the elderly [19]. However, accurate detection of activities in the home is challenging because of the wide range of possible activities. Crucial events in health monitoring, such as user errors, and critical actions, such as falling, occur randomly and are rarely observed. While an automated activity recognition system may perform accurately in well-trained situations, it is impossible to train all possible activities ahead of time in a deployed system. Legion:AR fills in the gap left by automated systems by leveraging the crowd.

Prior work in activity recognition has addressed the need to gather instances of activities in diverse contexts by engaging groups of people. A common use of human effort in activity recognition is for offline labeling, in which people label the activities in a video and this data is used to inform the learning model [34, 35]. These approaches have shown that both small controlled crowds and larger ones, such as Amazon's Mechanical Turk, are capable of providing effective training data. Another common approach is to ask users to perform and annotate activities in the contexts in which they want them to be recognized [1]. Crowdsourcing has been used to gather labels from many users, associate new users with existing personas, and then allow individuals to correct mislabeling of their activities [13]. In contrast to these approaches, Legion:AR enables the crowd to label activities in real-time, allowing it to correctly label novel events on their first occurrence. For home monitoring, this means that potentially harmful events, such as an individual cutting oneself with a knife, can be identified the first time they happen. Likewise, for public monitoring domains, criminal acts, fires, or other dangerous events can be correctly reported even if the system was not trained specifically for the event. To our knowledge, Legion:AR is the first system to employ the crowd in real-time to augment activity recognition systems so that they can respond interactively to novel activities and contexts.

Legion:AR is an example of human computation: integrating people into computational processes to solve problems too difficult for computers. Human computation has been shown useful in a variety of domains, e.g. writing and editing [6], image description and interpretation [3, 31], and protein folding [9]. Most work in human computation has focused on obtaining quality work from an unreliable pool of workers, and has generally introduced redundancy and layering into tasks so that multiple workers contribute and verify results at each stage. For instance, guaranteeing reliability through answer agreement [31] or the find-fix-verify pattern of Soylent [6]. Unfortunately, these approaches take time, making them less well-suited to real-time tasks such as activity recognition.

Several systems have explored how to make human computation interactive. As an example, VizWiz [3] answers visual questions for blind people quickly. It uses quikTurkit to prequeue workers on Amazon's Mechanical Turk so that they will be available when needed. Adrenaline uses a retainer model to recruit workers in just a few seconds to enable real-time photography assistance [4]. The Legion system enables real-time collaborative control of an existing user interface by allowing the crowd to collectively act as a single operator [15]. Individual crowd workers input suggestions as to what commands should be executed next, and the Legion system uses an 'input mediator' to intelligently choose which to forward to the actual interface based on such metrics as "crowd agreement." Legion:AR builds on this framework and workflow, and introduces the idea of forwarding not only select single inputs, but also sequences of inputs that are generated by merging multiple different workers contributions. This effectively expands the model introduced in Legion to handle situations in which the problem is not only the reliability of individual workers, but also the inability to provide answers at a sufficient rate, such as providing natural language activity, action, and object labels for complex actions. Legion:AR includes an HMM to learn from those labels, and introduces a number of other features specific to this domain, e.g. those features included specifically to address privacy concerns.

Prior work has looked into using multiple camera angles to allow automatic systems to use silhouettes in addition to RFID tags to recognize activities while preserving the anonymity of users in a home activity domain [22]. Other systems have

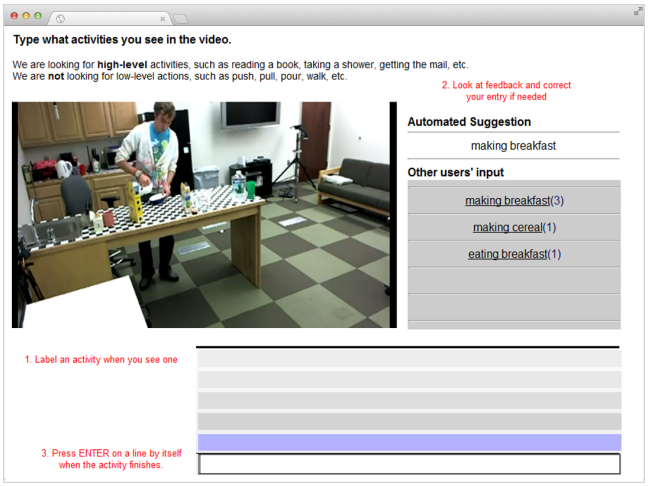

Figure 3. The Legion:AR worker interface. Workers watch live video of an activity and enter open-ended text labels into the text field below. They can see labels from other workers and the learning model (HMM) on the right, and can agree with them by clicking on them. Workers also indicate when activities have ended in order to segment the video.

used simple anonymous sensors such as motion detectors and pressure mats along with RFID tags to maintain a certain amount of privacy when detecting activities in the home [32]. We use a similar approach with Legion:AR, using RFID tags and automatically generating privacy veils to cover each user. Since people are able to gain more information from context, we show that it is not always necessary to use multiple cameras at different angles (though improvement could still be seen when more are available). Reducing the amount of information introduces a tradeoff between levels of privacy and activity awareness. Some of these tradeoffs have been explored in a workplace setting by [12].

Legion:AR

Legion:AR consists of three primary components (Figure 2): (i) the worker interface for displaying data and collecting real-time labels, (ii) an input mediator that collects inputs from multiple crowd workers and merges them into a final activity segment and label set, and (iii) an automatic activity recognition system that uses an HMM to identify activities. Applications using Legion:AR for activity recognition provide a data stream that is used by the HMM to generate an activity label. If it cannot confidently derive a label from the stream, then it asks the crowd to label the activity. Currently, it uses a variant of quikTurkit [3] to recruit crowd workers within a few seconds. Crowd workers see the worker interface (Figure 3) and either type the activities they see in the included video or agree with activity labels provided by others. These labels are sent back to Legion:AR, which merges the input of multiple workers by associating overlapping labels and inserting new ones into an ordered stream. Legion:AR then forwards the labels and corresponding segmentation to the HMM to be used for training so that the activity can be recognized later.

Requesters begin by starting the Legion:AR application and providing a data stream. Our current implementation of Legion:AR uses a HMM that can be trained using either video and RFID data, or visual data from a Microsoft Kinect. RFID data is collected from a powered, wrist-mounted RFID reader worn by users. The visual data is collected via cameras mounted in the area in which activities are to be observed.

These sensors provide a fine-grained context for recognizing individuals and their activities. RFID tags provide a rich source of context information through object interactions. Activity recognition using RFIDs are shown to be extremely accurate if users and environments are correctly instrumented [24]. However, in some situation this may be difficult and accuracy can fall due to some interactions not being picked up [17]. The tradeoffs involved in selecting a source of feature data for the automated system when using Legion:AR are the same as in traditional AR. In our kitchen lab environment, multiple RFID tags are instrumented on each object, and each RFID tag is uniquely identified for the object it is attached. Users then wore an RFID sensor on their wrists. We also use a Kinect, which is capable of detecting distinct individuals via its capability to differentiate multiple people from the background, making it possible to easily create separate per-user streams for each individual in the scene. User tracking is provided by the Kinect, which enables us to easily distinguish each person in the scene and veil their silhouettes accordingly. In our current system, we focus on indoor activities with relatively few ($n \leq 6$) people. However, our method of real-time crowd labeling can also be extended to outdoor activities and crowded environments using different sensors such as video cameras [11], accelerometers [1] and GPS [26].

Hidden Markov models and their variations have previously been successfully used for accurately modeling activities for recognition, and has been shown effective for modeling high-level activities such as the ones recognized in this paper [25]. Given sensor data from RFID tags or from the Kinect, and the activity from the previous time step, we try to classify what the current activity is. In this task, we use a HMM where each activity is represented as a single state. While this model is sufficient for our evaluations, we can also easily extend the model to recognize more complex activities [10, 24]. A simplified state diagram of our model is shown in Figure 4.

Generating Tasks

Legion:AR uses active learning to increase its knowledge of potential actions. Using multiple annotators who remain present for the entire task for active learning has shown to be effective [33]. However, prior work has not yet investigated this problem when workers are dynamic and change from moment to moment, meaning that the "best" worker to ask a specific question cannot be learned in most cases.

When Legion:AR does not recognize an activity that is being performed, it streams video to be labeled by the crowd. The stream is first fed through a pre-processor (custom for each input type) then identifies the number of actors visible in a scene and adds privacy veils when appropriate (discussed below). If multiple actors are detected in the scene, each one is marked using a different colored outline on the video to differentiate them for workers, then an individual labeling task is generated for each. This way, workers never have to focus on

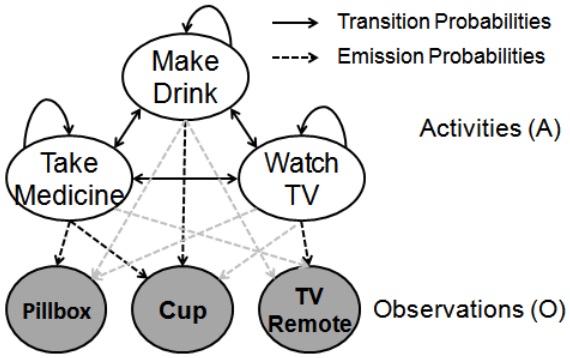

Figure 4. State Diagram of a single state per-activity HMM.

more than one user's actions and coordination between multiple crowd workers is easier, keeping the task simple enough for workers from sources such as Mechanical Turk to complete for a low price (under 10 cents per 90 second segment in our experiments). In this paper, we perform this segmentation using the Kinect SDK to reliably isolate each user.

Legion:AR begins to recruit workers into a waiting pool when started, so that workers can be ready on-demand. Workers can be recruited from any source of participants, but in this paper we considered only local volunteers and workers from Mechanical Turk. Once workers enter the waiting pool, they are paid a small amount of money to play a simple game. This game keeps workers focused on the current window, reducing the time needed to begin labeling once a task is issued. When the HMM encounters an activity it does not recognize with sufficient confidence, it makes a request to issue a new task.

Worker Interface

Once a task is generated, workers are immediately forwarded from the waiting screen to the labeling interface (Figure 3) which contains a video stream, text box, history of previous labels, and a field which displays the recent input of other workers performing the task and the HMM's best guess. Workers can either choose to enter a new label for the activity if the correct one has not yet been proposed, or they may select a previous answer proposed by another worker.

Once workers enter a label it is displayed in the current options viewer, making it visible to the other workers completing the task. Workers can change their answers at any time before the end of an activity in an attempt to agree with the best one. This style of interface provides two benefits: i) providing examples encourages workers to use similar phrasing in their answers, and ii) people are generally able to identify correct answers better than they can generate them [28]. Thus, this interface helps promote consistent labels that allow for more reliable learning from multiple sessions, and gives workers a better chance to be accurate. When a worker believes the current activity has ended, they signal this by submitting a blank label, or selecting 'Activity Complete'. The worker's input is then sent to Legion:AR's input mediator.

Combining Input

In order to recombine the input from participating workers, we assume workers each submit ordered, but incomplete, sets of labels for a task. These labels may correspond to slightly different definitions of a task with different beginning and ending times, and may be time shifted by varying amounts, so we cannot simply associate tags based on timing data.

Instead, we match labels entered by a single worker with the similar inputs from others in the same time window (we use 15 seconds). We use the label viewer on the worker interface to encourage workers to use similar terms, then use a relaxed measure of equality based on the Damerau-Levenshtein distance (similar to edit distance) which measures how many corrections must be made to a word or phrase to match another to account for typos. We set a dynamic cutoff of approximately 1 error per 5 letters typed, then consider any two words within this distance as equivalent. A structured language corpus such as WordNet [18] could also be used to consider semantic similarity between terms.

Worker labels and their equivalents are then merged into single nodes in a graph, given one is not already an ancestor of the others. This prevents cycles in the graph and leverages the proper-ordering assumption we made earlier. As nodes are added and merged, a directed acyclic graph is formed (shown in Figure 5). The nodes are weighted with the number of matched inputs they contain, and edges are weighted using the number of inputs in both the parent and child that were generated sequentially by the same user. For example, if 8 workers said "making coffee" (parent node), 6 workers said "making breakfast" (child node), and 3 of those workers said both in a row, then "making coffee" (weight 8) is linked to "making breakfast" (weight 6) by an edge of weight 3. As multiple partial sequences are aligned, we can use a most-likely traversal to recreate the original (correct) sequence of labels. Based on the tuning parameters of this traversal, more complete or more accurate label sequences can be recovered. Furthermore, since the algorithm works even when we limit the size of the current graph to only recent events, Legion:AR can be used to label video streams of unbounded length.

The power of this approach is that it runs fast enough to combine input without adding any additional delay (few milliseconds), and branches can be merged to increase the total recall of the final output. This can be done using a statistical model of actions that is learned over time, so as more activities are seen by the system, the better it becomes at predicting the order of actions. Alternatively, the graph can be greedily traversed to find the best sequence of labels to use. This approach scales better, but cannot account for ambiguity in the graph created by worker's input as well. After the sequence is reconstructed, it is then paired with the feature data recorded during the segment by either the RFID sensor or Kinect, and used to train the HMM on new activities, in real-time.

Multiple Actors

To label activities in scenes with multiple users, such as public spaces, we can use the same process described above, but with a separate crowd, task, and input mediator instance for each user in the scene. Users are made identifiable by coloring the veil that hides their identity, then asking workers to provide labels for the user of a specific color. Individual sets of labels can then be provided by workers for each individual.

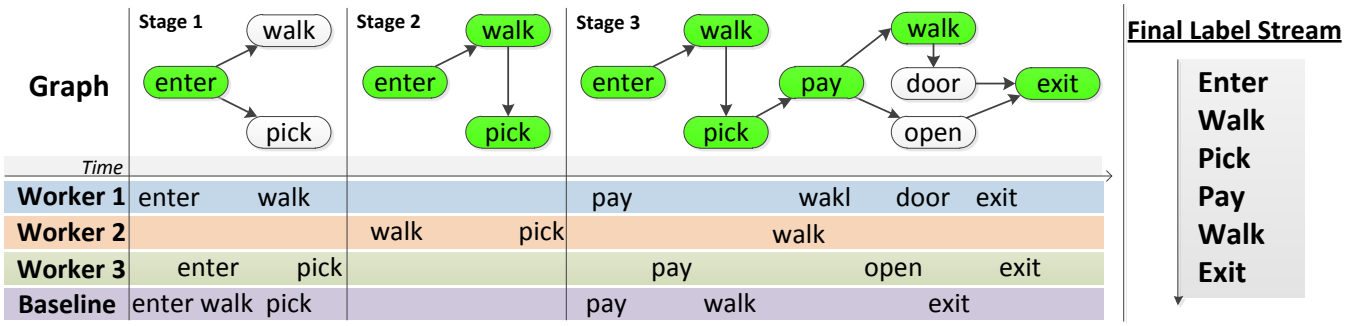

Figure 5. An example of the graph created by the input mediator. Green nodes represent sufficient agreement between multiple workers (here $N = 2$). The final sequence is correct despite overly-specific submissions by 2 out of the 3 workers, and a spelling error by one worker on the word 'walk'.

The system can also select a subset of users that it is unsure of the activities of rather than all of them, which helps to reduce cost. For instance, in the case where Legion:AR is being used to monitor an entire public square, there may potentially be hundreds of actors, but if only a single one is performing an unknown activity, we would only generate one task.

Complex Labels

Using multiple sets of workers each belonging to a different task can also independently label a different aspect of the activity such as the objects being used, the fine-grained actions being performed, or the effect of the current action. This can then be used to build (from a pre-defined set of relations) a more complex label consisting of each of these aspects being described at every point in time. For instance, we can label what agents, objects, and actions a given activity consists of by using sets of workers to each gather agent, object, action and activity labels. Providing the relationships between the data (and thus between the labels provided by workers) allows us to automatically recombine the labels to generate tags of a specific form. This works because in our tests, we found that workers were able to divide tasks into consistent segments given the same instructions, even if their labels differed when no special measures were taken.

Training the Learning Model

The HMM is given a sequence of labels and time segmentation intervals generated by the workers, and a stream of RFID tags recorded from the RFID reader on the user's wrist. Each interval has a matching label and a set of RFID tags that the HMM can train. Legion:AR is capable of training the HMM online, by using the sequence of labels and corresponding tags just as they would be if provided by an expert, as they are finalized by workers. Additionally, the learning model can be modified to provide feedback to Legion:AR such as its prediction of the current action, which is then forwarded to workers as a suggestion which they can agree with if correct. Using the underlying model to suggest answers can greatly reduce the amount of effort that volunteer workers will have to put in for actions that have already been learned (but still have low confidence) by the system.

Privacy Protection

In activity recognition, privacy is always a key topic. This is especially true in the case of Legion:AR since it can potentially be used to stream live video of a person's home or specific public location to a work force comprised of anonymous people on the web. Since some level of information must be exposed to workers in order to allow them to accurately label activities, users must use discretion in their use of the system and system developers must make users aware of the tradeoff between privacy versus recognition accuracy. We discuss some of the ways Legion:AR tries to help mitigate potential privacy concerns below.

Video Adjustments

To help obfuscates the identities of users, Legion:AR automatically hides their face (or even whole body) with a virtual *veil* (shown in Figure 1). However, this is does not preserve the anonymity of the surroundings in applications such as the home monitoring domain. To prevent detailed information from sources such as bills or other paperwork being forwarded to the crowd, we can use a low video resolution between 300×200 and 640×480. In initial tests, workers showed no difference in their ability to identify actions if given a higher resolution video. Legion:AR can also ensure no single worker from the crowd stays longer than a short period of time to prevent any one worker from collecting too much information about a person's routine or surroundings.

While our results indicate that veils and low resolution do not significantly reduce the recognition quality, it's clear that there is still a tradeoff. For example, if a user wants to hide their race, gender, or clothing from workers, covering their entire silhouette with a veil would be effective. However, to hide information such as height or weight, the system could veil the user entirely with an ambiguous shape (i.e. circle or rectangle). However, we expect this would have a significant impact on workers' ability to identify what action is being performed since little more than position would be available.

Opt-In/Out System

Legion:AR also uses a mobile app to prompt users before displaying video to the crowd, instead of letting the system decide by itself. When Legion:AR identifies a segment of actions it does not recognize, an alert to use the crowd to label the segment of video is sent to the user's phone or other mobile device. These alerts can be set to either opt-in, which may be useful when privacy is the primary focus, or op-out which is better suited to monitoring for critical events such as accidents in the home, where the user may not be able to answer the alert afterwards. In public setting where many users might be present, the system is not able to use opt-in feature.

Future implementations could use facial recognition to allow the ability for users opt-out of content being forwarded to the crowd by contacting the system with an image of themselves. This registration can be done once, and the system would remember their preferences when it recognizes them again later.

EXPERIMENTS

Legion:AR is capable of training activity recognition systems online. To validate this, we show that groups of workers can accurately generate labels in real-time, even while preserving the privacy of the actors in the scene. We then demonstrate that the system is able to automatically generate tasks that are manageable for single workers on crowd markets such as Amazon's Mechanical Turk, and that workers will label the correct aspect of the scene (the actions of the individual actor they were assigned to). Finally, we extend the idea of assigning groups of crowd workers to different sub-tasks, and demonstrate that, using Legion:AR, workers can label tasks at even fine granularities, allowing the system to collect more detailed activity labels than would be possible for a single worker to generate in real-time.

Real-time Activity Label Generation

To test the ability of groups of workers to generate labels for activities in real-time, we ran Legion:AR on a video stream of a person performing various household tasks in our lab, which is configured as a kitchen and living room. This recreates a scene which may be observed when monitoring elderly or cognitively disabled individuals to ensure they are performing necessary activities such as eating, taking medicine, toileting, and to monitor for situations where they may need assistance, such as a medical emergency.

Our crowd was composed of volunteers who had no prior experience with the system and were given a brief verbal description of how to label the data. In a real use-case, members of the crowd would differ from day to day, and having different workers at different times will likely effect the terminology used to label the same activities from one session to the next. We recreated these factors by scheduling tests on different days, with workers selected at random from the available set of 30 volunteers. Trials were not excluded for workers who misunderstood the directions or had connectivity issues, as this is expected in the real-world case. We ran 8 tests, each with 5 of the 8 different activities performed in them. Each of these activities was deliberately performed quickly (on the order of 10-30 seconds) in order to test the ability of the 5-worker group to agree in a short period of time. As activity times get longer, there is more time for the workers to converge on an answer and agreement becomes easier. We compared the results gathered from both a group of workers collaborating in real-time via our system and a single worker, to an expert labeling by training an HMM using the approach and comparing the resulting performance.

We conducted a leave-one-out cross validation of 8 activity sequences. The HMM was first trained online with worker generated labels and RFID observations from the training set, and then tested using the RFID sensor data from the test set. The output of the automated system was then compared with

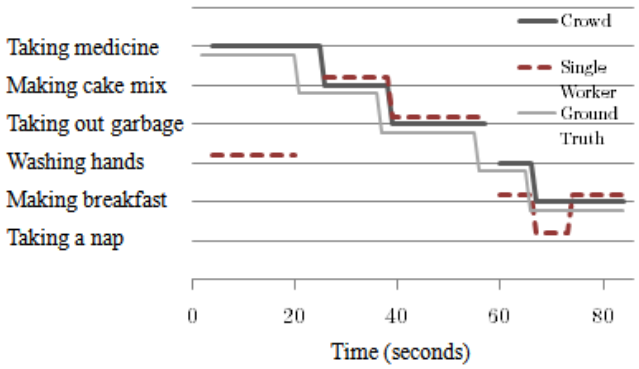

Figure 6. Automated recognition results. Inference results for a single activity sequence trained from crowd-generated labels (thick solid line) compared with the single worker labels (dotted line) and ground truth (thin solid line). Crowd-generated labels are more complete, allowing the automated system to do better when trained on them.

labels from the expert. However, since the expert labels differ from crowd generated labels, we manually associated each crowd label with a corresponding expert label for comparison. Our results show that the automated recognition system trained from crowd generated labels was able to achieve high average precision (90.2%) and was also able to recall all of the conducted activities, while the system trained from individual workers provided relatively noisy results (66.7% and 78% average precision and recall, respectively).

These labels were also highly consistent over multiple labels for the same actions, generating 8 unique labels for the activities shown. Since the system would automatically propose a label for actions it had seen already, the crowd was able to see previous best answers and agree with them, rather than generating a new name for each occurrence. This consistency is a key element of the labels produced by Legion:AR because it allows the system to accurately associate repeated events when it sees them again later, possibly performed in a slightly different way. For examples, without strong agreement the HMM might think that 'preparing breakfast' and 'making breakfast' are different activities, increasing the difficulty of the learning problem. The labels were also generated quickly enough to forward back to the system within seconds – fast enough to respond to critical events effectively.

Automatic Suggestions

We also tested what effect having the system propose possible answers as if it were an additional worker would have on the overall task performance, and the amount of effort required from the workers. We ran a new test with a different set of workers using the home monitoring data, and a system trained on each trial collected from the first run except for the one being tested on. The HMM constantly provided its prediction of what the current action was, and workers could accept this prediction if they agreed it was correct. While the results of the segmentation did not improve significantly, 80% of workers mentioned that using automatic suggestions made it easier to quickly agree with others on a shared label for the task being observed (between multiple runs). This is because it reduces the amount of search that must be done for an appropriate label.

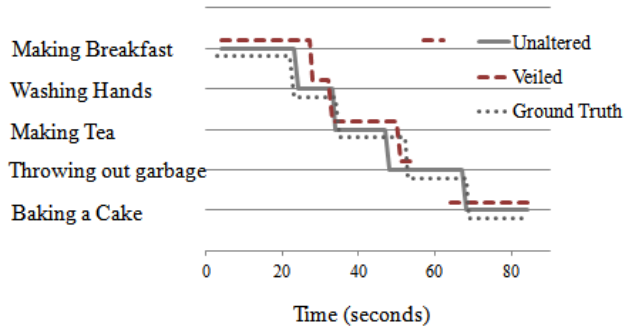

Figure 7. Crowd labeling results for a single activity sequence with veiled user (dashed red line) compared with labels of workers viewing the original unaltered video (solid line) and ground truth (thin dotted line).

Privacy

Since Legion:AR's automatically generated privacy veils obscure part of the video image, potentially making it more difficult for some activities to be identified, we ran tests to confirm that workers were still able to identify activities in the augmented video.

We re-ran a set of trials from the initial tests, again with 5 new workers. The results are shown in Figure 7. Only one label was missed due to the obfuscation of the user. This was caused by the veil generation on basic video being more imprecise than it was with the Kinect, resulting in covering all of the objects in front of a user, instead of being able to identify and cover only a person's body, or, if desired, even just their head and face. These findings agree with previous work using blur and pixelation filters that showed hiding identity in video clips could be done without having a detrimental effect on people's ability to identify properties of the scene such as position, number of actors, and level of activity [7].

Multi-actor Scenes

Next, we looked at how Legion:AR could be used in a more complex environment with multiple actors and activities occurring on at once. We mounted a Kinect near our department snack store to monitor people using this public space and watch for critical events such as theft, or accidents.

Because public space monitoring occurs on a much larger scale than home monitoring, there is a smaller group of concerned potential volunteers relative to the number of observed domains. As such it is important to show that other work forces beyond just volunteers, can effectively generate labels such as these.

We first evaluated the ability of workers recruited from Mechanical Turk to label specific workers based on their veil color. The results showed that over 11 trials, workers got on average 85% correct. There was no significant correlation between the number of actors in the scene and the quality of the labels generated by workers.

Complex Labels

In many domains, it is useful to have more fine grained labels including specific actions and what objects are being used. To produce these complex labels, we extend the procedure used to label scenes with multiple actors by using multiple crowds instead of just one. By first defining the structure of a label, then dividing it into smaller labeling tasks, multiple groups of workers can provide independent sets of labels that can then be automatically merged.

Entering	Stealing	Purchasing
walking	going back	counting money
standing	picking up	paying
walking into the room	walking	removing purse
leaving	grabbing something	pays for item

Figure 8. Activities and corresponding worker labels excerpts from the multi-actor scene.

Watching TV	*Actions*	Placing, Clicking, Pointing, Carrying, Watching
	Objects	TV, Television, Remote, Controller, Couch
Making Tea	*Actions*	Making, Drinking, Holding, Opening, Picking up
	Objects	Tea, Mug, Cup, Packet, Kettle
Taking Medicine	*Actions*	Sipping, Swallowing, Drinking, Taking, Opening
	Objects	Pills, Bottle, Medicine, Water, Cup

Figure 9. Selected subset of action and object tags generated by five workers for three activities.

We tested this by using two groups of volunteer workers. One group was instructed to label the actions being performed at each step, and the other was asked to list the objects being used in the current action. There were 25 actions and 28 objects that the user interacted with in two video sequences. Figure 9 shows the results from 5 workers generating tags.

As a baseline, we used an expert labeler who was experienced in annotating video for activity recognition. They were allowed to label and segment the video entirely offline and could replay and pause the video at will.

Individually, workers labeled an average of 48% of the objects used in the videos, while 5 workers were able to identify 90% (Figure 10). Additionally, workers correctly identified 12 unique objects that did not appear in the expert-labeled baseline. For actions, individual workers labeled an average of 32% of the actions used in the video, while 5 workers were able to identify 66%, with 24 unique actions that were not included in our baseline. For example, the crowd correctly identified both 'picking up' and 'clicking' on the remote, where the expert labeler only included 'picking up'.

The results can then be combined with the activity data using the timestamps, resulting in statements of the form:

`(clicking)` using `(remote control)` as part of `(watching TV)`

Using this type of fine grained labeling, we can provide a system trying to reason about the actions with more detailed information, such as hierarchical structure and the roles of different objects. We can also forward RFID data to workers to allow them to select which objects being read by the scanner rather than enter new labels which must be associated with the tags. Providing this information may also reduce the number of workers required to obtain high-quality results.

DISCUSSION

The results described in the previous section demonstrate that crowds of workers, both volunteer and paid, can accomplish the task of real-time label generation faster and more effec-

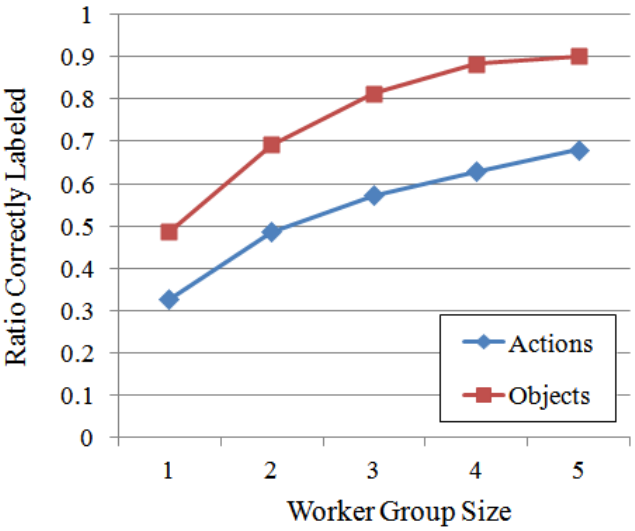

Figure 10. Average number of objects and actions correctly labeled by worker groups of different sizes over two different activity sequences. As the group size increases, more objects and actions are labeled.

tively than an individual could. In the fine-grained tests, even when compared to our experienced labeler (who generated labels offline), workers were able to find additional actions that had been overlooked in our baseline. This improvement demonstrates the potential for crowds to generate labels that are more comprehensive, as well as available quicker.

Collaborative Label Generation

We observed that users of Legion:AR seemed more comfortable in their task than those working alone. Exit interviews revealed that this was in large part because they received validation from the system, through either the automatic AR system or other workers, that they were providing the correct types of answers. Additionally, since each worker is required to do less work on average to contribute a single label, we can reduce the number of workers that are needed to attain agreement on every segment label.

Interface Improvements

Worker feedback also showed that most workers were able to quickly learn and easily use the labeling UI, but many found the change in modality between proposing a label using the keyboard and agree with a label using the mouse to be a hinderance. Because of this, several workers just retyped the content they saw in the label set from other workers, which the system then counted as a vote. In future versions we will address this issue by providing a optional number-based selection scheme that can quickly and easily be operated via keyboard, or adding auto-complete to reduce the worker's required effort. Some workers also noted that the positioning of the answer box relative to the video display made it more difficult to view the responses of other while still paying attention to the video. This can be addressed by relocating the suggestion list directly below the video so that the content is more in-line with the label entry and selection fields.

FUTURE WORK

Legion:AR enables a wide range of interactive applications and collaborative tasks that could benefit from activity recognition. For example, in game playing, a system may want to interact with a human player by observing their the actions, without having to first learn a set of actions before being functional. Streaming video of a video game or other desktop application instead of a camera is supported by Legion:AR— all that needs to be defined is a method for the HMM to receive feature data from the environment. Other applications such as public virtual assistants may seek to leverage knowledge about the user's actions even during one-off interactions.

Intent Recognition and Modeling Failed Attempts

Prior work has presented models of intent recognition by using a model that looks at both movement and interactions with other actors in the scene [26], and that modeling failures can improve classification accuracy when using specific learning models [27]. We plan to explore how the crowd can work collaboratively with the system to leverage these models by labeling subtle events, such as when two actors interact or when a user fails to complete an attempted action, then letting the system use the labels to infer actions. This could enable models that are able to identify suspicious actions before they occur, even before fully trained — further reducing response times and even helping prevent critical events.

Merging Crowd Input

Many tasks require special skills to complete, even if no special knowledge is required. In most cases, specially trained individuals exist who can perform these tasks, such as a professional captionist hired to transcribe audio to text in real-time. However, due to the requisite skill and training involved, these individuals are rare and come at a high cost.

In this work, we use a general framework for allowing groups of non-expert workers to collectively generate complete input even when no individual member of the group is capable of doing so individually. This is done by intelligently merging the input of all contributing workers. We used this approach in two forms: in the first, we averaged the workers' segment boundary times, and in the other we merged their labels.

Future systems using this framework need only to define the answer stream they want from the system, then present appropriate information to workers. For example, Legion:Scribe is a system that enables groups of non-expert workers recruited from local sources [16] or the crowd [14] to caption audio in real-time. Scribe builds of off of the same basic model as Legion:AR, but uses a fully automated approach to synthesize worker input into a single caption stream based on implicit agreement. On the other hand, in cases where there is no shared 'true' answer as there is with captioning, we have shown that the answers generated by groups of workers on more open-ended problems can be made consistent between sessions by displaying the group's answer proposals along the way. This makes the training process more tractable for the underlying learning model.

CONCLUSION

In this paper, we outlined a framework for using groups of non-expert workers to contribute information that would usually need to be generated by a trained expert. We demonstrated the utility of this approach using Legion:AR, a system for training an arbitrary activity recognition system in real-time using a crowd of workers. Legion:AR enables workers to collectively provide labels even in situations where a single user cannot, allowing systems to be trained online even in complex domains, thus avoiding the risk of missing an important event that requires immediate response the first time it is observed. We have shown that groups of workers can successfully label tasks in real-time using Legion:AR, even when no individual worker provides completely correct input. Furthermore, we found that the labels generated by groups of workers can outperform any single online worker, and even generate correct tags that were missed by an expert offline labeler. We also demonstrated methods for dividing complex tasks into forms that workers from multiple types of crowds, such as those available from micro-task marketplaces including Mechanical Turk, can accomplish collectively. Legion:AR provides the foundations for systems that can be deployed in a wide range of real-world setting with no prior training, and enables new applications of AR in situations where identifying user actions with only one-off observations is important.

REFERENCES

1. L. Bao and S. S. Intille Activity Recognition from User-Annotated Acceleration Data. *In Proc. of Pervasive 2004*, v. 3001/2004, pp. 1-17, 2004.
2. J.P. Bigham, R.E. Ladner, and Y. Borodin. The Design of the Human-Backed Access Technology. *In Proc. of the ACM Intl. Conf. on Computers and Accessibility (ASSETS 2011)*, pp. 3-10, 2011.
3. J.P Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R.C. Miller, R. Miller, A. Tatrowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. *Proc. of the ACM Symp. on User Interface Software and Technology (UIST 2010)*, pp. 333-342, 2010.
4. Michael S. Bernstein, Joel R. Brandt, Robert C. Miller, and David R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proc. of the 24th Symp. on User Interface Software and Technology (UIST 2011)*, 2011.
5. M. S. Bernstein, D. R. Karger, R. C. Miller, and J. Brandt. Analytic methods for optimizing realtime crowdsourcing. In *Proc. of Collective Intelligence (CI 2012)*, To Appear, 2012.
6. M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. *Proc. of the ACM Sypm. on User Interface Software and Tech. (UIST 2010)*, pp. 313–322, 2010.
7. M. Boyle, C. Edwards, and S. Greenberg. The effects of filtered video on awareness and privacy. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pp. 1–10, 2000.
8. Y. Chu, Y.C. Song, H. Kautz, and R. Levinson. When Did You Start Doing That Thing That You Do? Interactive Activity Recognition and Prompting. *AAAI 2011 Workshop on Artificial Intelligence and Smarter Living*. 2011.
9. S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and FoldIt Players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.
10. T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. *Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. I*, pp. 838–845, 2005.
11. L. Fiore, D. Fehr, R. Bodor, A. Drenner, G. Somasundaram, and N. Papanikolopoulos. Multi-camera human activity monitoring. *Journal of Intelligent and Robotic Systems, vol. 52*, pp. 5-43, 2008.
12. S. Hudson, and I. Smith. Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pp. 248–257, 1996.
13. N. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A. Campbell, and F. Zhao. Enabling Large-scale Human Activity Inference on Smartphones using Community Similarity Networks *Proc. of Ubicomp 2011*, 2011.
14. W.S. Lasecki, J.P. Bigham. Online Quality Control for Real-time Crowd Captioning. In *Proc. of the ACM International Conference on Computers and Accessibility (ASSETS 2012)*, pp. 143-150, 2012.
15. W.S. Lasecki, K.I. Murray, S. White, R.C. Miller, J.P. Bigham. Real-Time Crowd Control of Existing Interfaces. In *Proc. of the ACM Symp. on User Interface Software and Technology (UIST 2011)*, pp. 23-32, 2011.
16. W.S. Lasecki, C.D. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, J.P. Bigham. Real-Time Captioning by Groups of Non-Experts. In *Proc. of the ACM Symp. on User Interface Software and Technology (UIST 2012)*, pp. 23-34, 2012.
17. B. Logan, J. Healey, M. Philipose, E. Tapia, and S. Intille. A Long-term Evaluation of Sensing Modalities for Activity Recognition. *Proceedings of the 9th international conference on Ubiquitous computing*, pp. 483–500, 2007.
18. G. Miller. Wordnet: A lexical database for English. *Comm. ACM 38, 11*. pp. 3941, 1995.
19. J. Modayil, R. Levinson, C. Harman, D. Halper, and H. Kautz. Integrating Sensing and Cueing for more effective Activity Reminders. *Proc. of the AAAI Fall 2008 Symp. on AI in Eldercare: New Solutions to Old Problems*. pp. 60-66, 2008.
20. E.D. Mynatt, A.S. Melenhorst, A. D. Fisk, W. A. Rogers Aware technologies for aging in place: understanding user needs and attitudes. *Pervasive Computing, IEEE*, v3-2, pp. 36-41, 2004.
21. W. Niu, J. Long, D. Han, and Y-F. Wang. Human activity detection and recognition for video surveillance. *Proc. of ICME 2004*, pp. 719-722, 2005.
22. S. Park and H. Kautz Privacy-Preserving Recognition of Activities in Daily Living From Multi-View Silhouettes and RFID-Based Training. *AAAI Symp. on AI in Eldercare: New Solutions to Old Problems*. 2008.
23. A. Shinsel, T. Kulesza, M. Burnett, W. Curran, A. Groce, S. Stumpf, and W-K. Wong. Mini-Crowdsourcing End-User Assessment of Intelligent Assistants: A Cost-Benefit Study. *Proc. of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC '11)*. pp. 47-54, 2011.
24. D. Patterson, D. Fox, H. Kautz, M. Philipose. Finegrained activity recognition by aggregating abstract object usage. *In ISWC 05: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*. pp. 4451. 2005.
25. W. Pentney, et al. Sensor-based Understanding of Daily Life via Large-Scale Use of Common Sense. *Proceedings of AAAI*. 2006.
26. A. Sadilek, and H. Kautz. Recognizing Multi-Agent Activities from GPS Data. In *24th AAAI Conference on Artificial Intelligence (AAAI 2010)*. 2010.
27. A. Sadilek, and H. Kautz. Modeling and Reasoning about Success, Failure, and Intent of Multi-Agent Activities. In *Proc. of UbiComp 2010*.
28. Y. Sun, S. Roy, G. Little. Beyond Independent Agreement: A Tournament Selection Approach for Quality Assurance of Human Computation Tasks. *AAAI Workshop on Human Computation*. 2011.
29. E. Tapia, S. Intille, and K. Larson. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. *Pervasive Computing*, pp. 158-175, 2004.
30. M. Toomim, T. Kriplean, C. Prtner, and J. A. Landay. Utility of human-computer interactions: Toward a science of preference measurement. In *Proc. of the ACM Conf. on Human Factors in Computing Systems (CHI 2011)*, pp. 2275-2284, 2011.
31. L. von Ahn, and L. Dabbish. Labeling images with a computer game. *Proc. of the Conf. on Human Factors in Computing Systems (CHI 2004)*, pp. 319–326, 2004.
32. D. Wilson, and C. Atkeson. Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. *Proc. of PERVASIVE 2005*, pp. 62-79. 2005.
33. Y. Yan, R. Romer, F. Glenn, and D. Jennifer. Active Learning from Crowds. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1161–1168, 2011.
34. L. Zhao, G. Sukthankar, and R. Sukthankar. Robust Active Learning Using Crowdsourced Annotations for Activity Recognition. *AAAI 2011 Workshop on Human Computation*. pp. 74-79, 2011.
35. L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations. *IEEE Intl. Conf. on Social Computing*. 2011.