Training Activity Recognition Systems Online Using Real-time Crowdsourcing

Young Chol Song, Walter S. Lasecki, Jeffrey P. Bigham, Henry Kautz
University of Rochester Computer Science
Rochester, NY 14627 USA
{ysong,wlasecki,jbigham,kautz}@cs.rochester.edu

ABSTRACT

Automated activity recognition (AR) aims to provide context-aware information and support in a wide variety of applications, such as prompting systems and monitoring public spaces for danger. This type of recognition is typically easy for people, but difficult for automated systems due to the need for understanding the context and high-level reasoning behind actions. We present Legion:AR, a system that uses real-time crowdsourcing to generate reliable sets of activity labels. Legion:AR enables online training of existing learning models in deployed recognition systems. Our evaluations with 24 workers show Legion:AR is able to generate labels in real-time, even in the presence of privacy-protecting measures such as adding a veil to hide the user's identity.

Author Keywords

Activity Recognition, Crowdsourcing, Human Computation

ACM Classification Keywords

H.5.2 Information interfaces and presentation: Miscellaneous.

INTRODUCTION AND BACKGROUND

Recognizing human activities is typically easy for people, but difficult for automated systems due to the need for understanding the context and high-level reasoning behind actions. We present Legion:AR, a system that uses the crowd to generate reliable set of activity labels in real-time, then uses an active learning to train automatic activity recognition systems on-demand, using existing learning models, to recognize activities in the future. Our evaluations of Legion:AR using volunteer workers show that the crowds are able to effectively generate labels in real-time using our system, even in the presence of privacy-protecting measures such as adding a veil to obscure the user's identity.

Crowdsourcing aims to integrate groups of people into computational processes to solve problems too difficult for computers. Most work in crowdsourcing 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 or the find-fix-verify pattern of Soylent [1]. Unfortunately, this takes time, which makes these approaches unsuitable for real-time tasks. Real-time and nearly real-time crowdsourcing has been explored and shown useful in a variety of assistive domains, such as visual question answering[2] and real-time audio transcription[3].

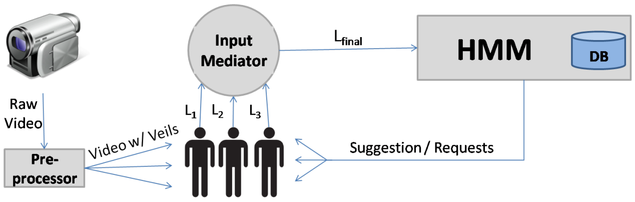

Figure 1. System figure for Legion:AR. When the automated system cannot recognize an action, the system forwards privacy-veiled video and guesses from the AR system to the crowd. The crowd generates labels and segments the video in real-time, which is returned for immediate use and to train the automated system.

We define users as people whose activities are being identified and workers as members of the crowd helping to generate label data. *The crowd* is a dynamic pool of potentially anonymous workers of varying reliability that can be recruited on-demand, in which no single worker can be relied upon to provide a completely correct answer.

Legion:AR

Legion:AR consists of three components: (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 AR system to identify activities.To recruit workers quickly, Legion:AR uses a variant of quikTurkit [2] to recruit crowd workers within seconds.

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 be time shifted by varying amounts, so we cannot simply associate tags based on timing data. Instead, we merge the partially complete label sets based on where they agree with other workers, using an online multiple sequence alignment (MSA) algorithm [3]. This results in a single, more complete, final stream than any one worker could generate individually.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

Generating Tasks

When the automatic system does not recognize an activity that is being performed, it streams video to be labeled by the crowd by Legion:AR. Users can select to maintain their privacy by having the system first obscure their identity using a *veil* before sending the video to workers (described later). Legion:AR begins to recruit workers into a waiting pool when started, so that workers can be ready on-demand. Once support is requested by the system, workers are immediately forwarded from the waiting pool to the labeling interface containing a video stream, text box and history of previous labels, and a field which displays the recent input of other workers performing the task and the system's prediction. Workers then choose to enter a new label for the activity or select an answer proposed by previous worker.

Training the Learning Model

Legion:AR is capable of training an AR system 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 with low confidence by the system. In our experiments we use a Hidden Markov Model (HMM)-based system. The HMM is given a sequence of labels and time segmentation intervals generated by the workers, and a stream of RFID tags recorded from an RFID reader on the user's wrist.

Privacy

Streaming live video of a person at home or even in a public areas clearly presents privacy issues. To help address this, we automatically generate an opaque colored region called a *veil* that hides the user's face or entire body. To prevent 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 video with lower-resolution or veils covering the user's face (or body when the face could not be confidently identified). Users can also select between using known or anonymous crowds (and set worker term limits), and choose to be prompted to either opt-in or opt-out of forwarding video to the crowd on a per-request basis. In the end, some information must be exposed to workers to allow them to accurately label activities, so user discretion must be used.

EXPERIMENTS

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 and taking medicine, and to monitor for situations where they may need assistance, such as a medical emergency.

Our crowd was composed of workers selected randomly from a set of 30 volunteers who had no prior experience with the system. We ran 8 tests, each with 5 of the 8 different activities performed in them. Each of these activities was deliberately performed quickly (around 10-30 seconds) to test the ability of the groups to agree in a short period of time. We conducted a leave-one-out cross validation of 8 activity sequences, comparing the output of an HMM trained with groups of workers, a single worker, and an expert labeler. The recognition system trained from crowd-generated labels was able to achieve high average precision (90.2%) as well as recall all the conducted activities, while the system trained from individual workers provided relatively noisy results (averaging 66.7% precision and 78% recall).

We also tested using the system as an additional worker. We used a new set of workers using the home monitoring data, and a system trained on each non-matching trial collected from the first run. The HMM forwarded predictions to the interface as a worker would, allowing the crowd to agreed with it if correct. 80% of workers found that these suggestions made it easier to quickly agree with others on a label.

Legion:AR can automatically augment the video with a privacy veil covering each user's face as a means of helping to protect user privacy from workers. However, since this veil obscures part of the video image, it potentially makes recognizing activities more difficult for workers. Our tests also showed that workers were not significantly affected by this alteration, with only one additional missed activity resulting from the addition of veils to the original tests.

CONCLUSIONS AND FUTURE WORK

In this paper, we presented 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, allowing systems to be trained online, 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. Further work has also shown that Legion:AR can use crowds recruited from sources such as Mechanical Turk to affordably label complex scenes involving fine-grained actions and multiple actors, while maintaining high accuracy and recall – enabling on-demand, activity labeling for context-aware systems.

REFERENCES

1. 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. *UIST 2010*, pp. 313–322, 2010.
2. 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. *UIST 2010*, pp. 333-342, 2010.
3. 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 *UIST 2012*, *To Appear*.