Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations

Liyue Zhao*   Gita Sukthankar*   Rahul Sukthankar†*‡

*Department of EECS, University of Central Florida, Email: {lyzhao,gitars}@eecs.ucf.edu
†The Robotics Institute, Carnegie Mellon University, Email: rahuls@cs.cmu.edu

Abstract—

Crowdsourcing has become an popular approach for annotating the large quantities of data required to train machine learning algorithms. However, obtaining labels in this manner poses two important challenges. First, naively labeling all of the data can be prohibitively expensive. Second, a significant fraction of the annotations can be incorrect due to carelessness or limited domain expertise of crowdsourced workers. Active learning provides a natural formulation to address the former issue by affordably selecting an appropriate subset of instances to label. Unfortunately, most active learning strategies are myopic and sensitive to label noise, which leads to poorly trained classifiers. We propose an active learning method that is specifically designed to be robust to such noise. We present an application of our technique in the domain of activity recognition for eldercare and validate the proposed approach using both simulated and real-world experiments using Amazon Mechanical Turk.

I. INTRODUCTION

Social computing is revolutionizing machine learning because it transforms the problem of generating a large corpus of labeled real-world data from a monolithic labor-intensive ordeal into a manageable set of small tasks, processed by thousands of human workers in a timely and affordable manner. Human computation, organized through services such as Amazon's Mechanical Turk (MTurk) have made it possible for researchers to acquire sufficient quantities of crowdsourced labels, enabling the development of a variety of applications driven by supervised learning models [1], [2].

However, crowdsourcing is not a panacea for data labeling for two reasons. First, the human labelers recruited by crowdsourcing typically lack specialized training and are thus limited to labeling data that requires little domain-specific expertise. For instance, in our eldercare application, it is infeasible to directly crowdsource labels for inertial measurement unit (IMU) traces generated by body-worn sensors. A popular solution to this issue [3] is for researchers to create a parallel, time-aligned channel of information (such as video) that can be provided to labelers. While this auxillary data would not necessarily be available in the deployed system, it serves a vital role in the classifier training process. While video does enable human workers to annotate IMU data, our studies show

that the acquired labels are still corrupted by an unacceptable level of noise. In this paper, we propose augmenting the action labeling task with an object recognition task (on images). By integrating information about both objects and actions (similar in spirit to [4]), we are able to better infer more accurate labels for the IMU traces.

Second, exhaustively labeling large corpora can be prohibitively expensive, motivating demand for approaches such as active learning [5] for affordably training supervised classifiers with a significantly smaller subset of labeled data. Active learning iteratively selects, from a pool of unlabeled data, those samples that would have the greatest impact on the classifier. Unfortunately, a direct application of traditional active learning algorithms to crowdsourced training data is problematic due to the problem of label noise. We propose a modified active learning paradigm that balances the traditionally myopic selection of instances in active learning (e.g., based on proximity of data to the decision boundary) with a term that also considers the distribution and local consistency of labels. In other words, we aim to automatically identify those unlabeled instances that are most valuable to label, but also samples that might benefit from relabeling because their initial labels seem suspect. Since acquiring multiple labels for each sample *en masse* would be impractical due to expense, our proposed method seeks to incrementally label only the most important samples. Additionally, we identify problematic samples (those for which we cannot get a sufficiently consistent set of labels) and withhold them from the training set. Thus, through the judicious use of inconsistency detection and incremental relabeling, significantly boost the ability of active learning to exploit crowdsourced data. Our experiments show that this enables us to affordably train classifiers on noisy crowd-sourced data.

II. RELATED WORK

Although active learning is a mature field [5], the majority of work assumes the absence of label noise (i.e., perfect oracles). Furthermore, most active learners incrementally select the next instance in a greedy and myopic manner, though efforts in pool-based active learning [6], [7] or hierarchical approaches [8] take a more global perspective. The main problem for active learning in our application is that uncertainty sampling is inherently "noise-seeking" [9] and thus the quality of learned classifiers degrades dramatically in the

IEEE computer society

presence of crowdsourced annotations. The serious issue of label corruption in social computing has been identified both for crowdsourced data collection [10] and in the context of attacks on collaborative tagging [11].

Our work is motivated by Nguyen and Smeulders [12] and Donmez et al. [13]. Philosophically, our work is most aligned with Sheng et al. [14], where relabeling is used to obviate the effects of noise and Vijayanarasimhan et al. [15], which identifies promising crowdsourcing annotation tasks given a specific data labeling budget. Proactive learning [16] attacks the related problem of selecting an optimal oracle using a decision theoretic approach. Finally, our application area of activity recognition using body-worn inertial sensors has become an important research topic, particularly in the context of eldercare and healthcare applications, as surveyed in [17].

III. Proposed Approach

Figure 1 presents an overview of our approach. Our goal is to accurately label large quantities of IMU data with activity labels, from which we can train activity recognition classifiers. Given temporally-aligned video and IMU data, such as those provided in the CMU-MMAC dataset [3], we assign MTurk workers to label short, automatically-segmented video clips of cooking activities with the label(s) corresponding to the current action(s). Unfortunately, labels predicted using the raw crowdsourced labels are unacceptably inaccurate. Inspired by ideas in multi-task learning [18], we ask each worker to solve a second related task — to identify visible objects. As detailed below, we train a Naive Bayes classifier to infer actions based on the combined annotations.

We then apply this framework in an active learning context to iteratively annotate data. The key idea is to select instances according to a mixture of two criteria: 1) Maxmin [19] and 2) based on the distribution of labels in local neighborhoods.

A. Inferring Actions from Visible Objects

Our experiments use the CMU-MMAC dataset [3], which consists of data collected from subjects performing unscripted recipes in an instrumented kitchen using a variety of sensors. Our goal is to perform reognition on the IMU data, but because this data is difficult to annotate directly, we also provide annotators with temporally aligned video from an egocentric camera. As discussed earlier, to compensate for the poor annotation accuracy of raw crowdsourced action labels, we ask MTurk workers to annotate which objects are visible in each scene. We use these as a secondary source from which to infer action labels, and the two sources are combined using a Naive Bayes formulation. Experiments show that while this significantly improves annotation accuracy, the label noise is still too high for traditional active learning methods.

B. Adding Robustness to Label Noise for Active Learning

Active learning seeks to iteratively obtain labels for data by identifying, at each iteration, the most uncertain sample (i.e., the instance, which if labeled would have the greatest impact on the classifier). Before describing our modification,

we briefly review active learning in the context of a support vector machine (SVM) framework. Let $\mathcal{T} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ be a set of initially labeled instances, with corresponding labels given by $\mathcal{L} = \{y_1, y_2, \ldots, y_l\}$. We also define a set of unlabeled instances as $\mathcal{U} = \{\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_n\}$. Our method extends Tong and Koller's *Maxmin Margin* strategy [7] for selecting samples that have the greatest impact on the loss: $\text{Loss}_{\text{uncertainty}}(\mathbf{x}_i) = \min_{\mathbf{x}_i \in \mathbf{U}}(V^+(\mathbf{x}_i), V^-(\mathbf{x}_i))$, where $V^+(\mathbf{x}_i)$ and $V^-(\mathbf{x}_i)$ denote the sizes of the version spaces resulting from labeling $\mathbf{x}_i$ as $+$ or $-$, respectively. This formulation extends naturally from binary to multiclass settings by computing the product of the loss function for the appropriate classification hyperplanes from all classes. The version space idea also applies to SVMs with non-linear kernels; in our work, we employ the popular Gaussian kernel.

Note that this standard formulation for SVM-based active learning is potentially ill-suited to our task for two important reasons. First, it assumes that the requested labels are perfect. This can be a dangerous assumption since it makes the SVM very sensitive to label noise. Second, a given sample, once labeled can never be selected for relabeling, which means that an error, once made, tends to become "locked in".

To address the first concern, we modify the loss function, as described below, to add a term characterizing the inconsistency of each labeled sample based on labels in its neighborhood. The intuition behind this idea is that one should request labels for samples that are in unexplored regions of the feature space or those near samples that may have been incorrectly labeled. To address the second concern, we also propose relabeling such points using crowdsourcing. Let $\mathbf{x}_i \in \mathcal{U}$ be the unlabeled sample with unknown label y_i and $\mathbf{x}_j \in \mathcal{T}$ be a labeled sample with known label y_j. Now, we can estimate the probability of labeling $\mathbf{x}_i$ as class c (based on labels in its neighborhood) as

$$p(y_i = c) = \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}_j \in \mathcal{T}} h(c, y_j) f(\mathbf{x}_i, \mathbf{x}_j). \quad (1)$$

Here, $h(c, y_i) = 1$ if the labels agree and 0 otherwise, and $f(.,.)$ is a Gaussian function (corresponding to the SVM RBF kernel). We formulate the inconsistency term of the modified loss simply as the entropy of the label distribution $p(y_i)$:

$$\text{Loss}_{\text{inconsistency}}(\mathbf{x}_i) = -\sum_{c \in \mathcal{C}} p(y_i = c) \ln(p(y_i = c)) \quad (2)$$

Finally, the modified loss consists of a linear combination of these two terms. The sample that we select is obtained using:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}_i \in \mathbf{U}} \text{Loss}_{\text{uncertainty}}(\mathbf{x}_i) + \lambda \cdot \text{Loss}_{\text{inconsistency}}(\mathbf{x}_i), \quad (3)$$

where λ denotes the mixing weight between the two criteria and depends on the expected annotation accuracy. Given *a priori* knowledge about annotation accuracy, one could set λ appropriately — with reliable labels, a lower value for λ suffices. In practice, since the expected accuracy of labels is unknown, we employ cross-validation with a hold-out set to determine λ. As suggested in [16], we initialize our labeled set by crowdsourcing k "representative samples" taken near cluster centers of our data.
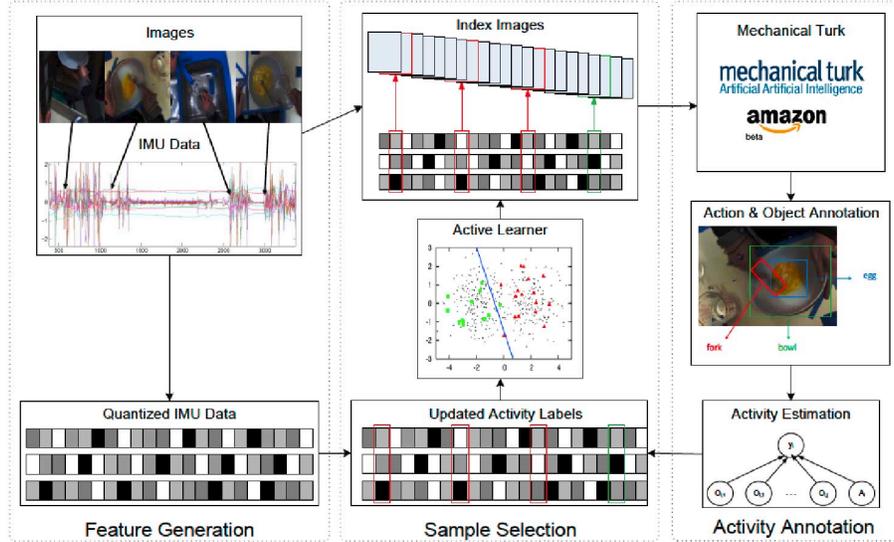
Fig. 1. Overview of the proposed method. Since labeling IMU data is difficult for untrained users, we crowdsource short activity labels for temporally aligned video clips supplemented with object labels for still images. The object labels generate more accurate activity labels. Our modified active learning approach uses both uncertainty and inconsistency, requests relabels for important samples, and rejects problematic samples. The resulting system can train accurate activity recognition classifiers from unreliable crowdsourced data.

C. Relabeling Strategy

As discussed above, the modified active learning scheme can crowdsource additional labels for samples that might have been incorrectly labeled. Given multiple labels, the most straightforward strategy for incorporating them is a voting scheme where the most popular label is assumed to be correct, termed *majority relabeling* (MR) in our experiments below. Unfortunately, this form of voting is no better than random when too many workers disagree (when it would be better to discard the sample) and does not distinguish between a clear and slight majorities.

For these reasons, we propose an alternate scheme, termed *absolute majority relabeling* (AMR) that incrementally adds labels until the sample is either accepted (with a clear majority) or deemed problematic due to inconsistent labeling and discarded. The intuition behind our proposed scheme is straightforward: given the distribution of labels for the sample, we accept the majority label iff it has received more than 50% of the votes. Otherwise, we incrementally request additional labels; if a clear majority fails to emerge after $N = 7$ labels, we mark the sample as problematic and remove it from the training set.

$$\text{relabeling accepts label } c \text{ iff:} \quad p(Y = c) > 0.5. \quad (4)$$

We can make a further modification to the AMR strategy by allowing workers to provide soft votes for multiple labels rather than a hard vote for a single label (analogous to cumulative voting). Specifically, each worker's vote consists of the label distribution $p(Y|O, A)$ generated using Naive Bayes (described in Section III-A). These histograms are accumulated and normalized. The decision criterion remains the same: if $\exists c \in \mathcal{C}$ s.t. $p(Y = c) > 0.5$ then we accept the

label c. If no consistent labeling emerges after N workers have relabeled the sample, we discard the sample from the training set. In the experiments below, we term this modified relabeling scheme as *absolute cumulative majority relabeling* (ACMR).

IV. DATA COLLECTION

CMU-MMAC [3] is primarily an unlabeled dataset. To test our annotation strategies, we use the labeled subset posted on the CMU-MMAC website[1] as ground truth labels. This consists of labels for 10 subjects who baked the brownie recipe, where the annotations were manually generated from the video data alone.

A. Synthetic Data

In addition to evaluation on CMU-MMAC, we study the performance of our proposed methods on a synthetic dataset under controlled noise conditions. Based on previous experience, we model label noise in crowdsourced annotations as a combination of two factors, *random error* and *systematic error*.

The *random error* comes from annotator carelessness, when workers occasionally provide an incorrect label for the given task. We treat this as i.i.d. noise on each label. Our model assumes that there is a fixed probability $r \in [0, 1]$ that a given sample will be corrupted by random noise. Such samples are assigned an incorrect label, chosen from among the remaining labels with uniform chance.

The second source, *systematic error* results from samples that are inherently difficult for any worker to annotate. For example, some videos are ambiguous or depict transitions between two actions and are thus challenging to annotate. The

---

[1]http://www.cs.cmu.edu/~espriggs/cmu-mmac/annotations/

label noise for these samples is correlated. We simulate this by *a priori* identifying n random samples as challenging. For each challenging sample, there is a high probability $s \in [0, 1]$ that workers labeling this sample will generate a label chosen with uniform probability from among the set of incorrect choices.

B. Real Data

The crowdsourced labels are collected using Amazon's Mechanical Turk, which allows developers to publish distributed annotation tasks to workers. Our task requires workers to annotate a short video (approximately 1–2 seconds in duration) taken from the CMU-MMAC dataset showing a subject preparing food in a kitchen environment. The work is asked to select a single most likely action and tag the set of objects visible in the video.

To enable us to study relabeling under repeatable conditions, we ask 7 different workers to annotate each video in the dataset. In the experiments, whenever an algorithm requests an additional crowdsourced label, we pull an appropriate label from the queue of 7 labels for the given point.

V. Experiments

We present a selection of results from three categories of experiments. In the first set, we perform a controlled study on synthetic data by corrupting labels with known quantities of noise to evaluate the robustness of different active learning selection criteria. The second set evaluates the three proposed relabeling strategies under controlled conditions. The third set examines the performance of the different active learning sample selection criteria and the relabeling strategies on the CMU-MMAC dataset, and examines the impact of combining object and action labels to infer more accurate annotations.

A. Robustness to Annotation Noise

The first set of experiments examines the robustness of various active learning strategies to annotation noise. Figure 2 compares four active learning sample selection strategies under four scenarios with increasing levels of random error ($r = 0\%$, 20%, 40%). The selection strategies are:

1) Maxmin [7] - the standard criterion that only considers uncertainty (black);
2) DWUS [12] - a global criterion that combines both uncertainty and data distributions (red);
3) Proposed - our criterion that combines both uncertainty and inconsistency criteria (blue).

In these experiments, there is no systematic error ($s = 0$).

We make several observations. First, in the noise-free case, the classifier error of all three strategies decreases sharply and converges to a low error rate. "Maxmin" and "Proposed" perform better than "DWUS" at the final phase of the learning. This is consistent with our expectations that the standard criterion, which mainly focuses on most uncertain samples is well suited for noise-free scenarios, and that our proposed criterion can match this.

Next, in the $r = 20\%$ noise case, we see that "DWUS" decreases quickly but converges to a higher error rate since

the representative term in its loss function forces the strategy select samples from high density regions. That means although "DWUS" benefits by selecting representative samples in the initial phase, it suffers in the final phase because the most uncertain samples remaining probably do not occur in the high-density regions. We observe that the "Proposed" scheme continues to have low errors in the final phase of learning since it focuses on inconsistent samples that have mislabeled neighbors.

Finally, under the challenging conditions of 40% noise, none of the three criteria perform particularly well. From this, we can conclude that improving the criteria for selection in active learning can effectively counter moderate noise but is not sufficient by itself when dealing with very noisy annotations. This validates our earlier observation that crowdsourced labels may require judicious relabeling in addition to improved active learning strategies.

B. Relabeling — Synthetic Data

In this section we evaluate the performance of relabeling strategies on synthetic data with both random and systematic noise. We employ two sets of synthetic data in this experiment. The first simulates annotation with random error $r = 20\%$. The second set of data has 50% of samples correctly annotated and the other 50% with systematic error $s = 40\%$. The relabeling strategy can request up to 7 labels to annotate a given sample. We applied Maxmin and our proposed active learning methods to select the most informative sample to be annotated. This experiment involves two relabeling strategies:

1) Majority relabeling (MR) - a sample is relabeled with the most popular action label. If multiple labels are equally popular, a random label from that set is used.
2) Absolute majority relabeling (AMR) - proposed strategy using only the raw action labels.

Figure 3 indicates the results by applying different relabeling strategies on synthetic data with random error ($r = 20\%$) only. The results shows that by asking multiple workers to annotate samples, the majority relabeling (MR) (dashed lines) effectively reduces the learning error to a low rate (approximately 5%) even when the label noise is 20%. Moreover, with the absolute majority relabeling (AMR) (solid lines), which identifies unreliable labels and asks additional workers to relabel them, the learning error converges rapidly to a much lower level for both of the active learning sample selection criteria.

Another experiment on the synthetic data with systematic error ($s = 40\%$) is shown in Figure 4. This experiment also validates the benefit of relabeling. AMR (solid lines) outperforms the MR (dashed lines).

C. Relabeling — CMU-MMAC Dataset

In this experiment, we compare the classifier error of different relabeling strategies (see Figure 5). Specifically, we examine the following conditions:

1) Majority relabeling (MR) - a sample is relabeled with the most popular action label (black dashed).

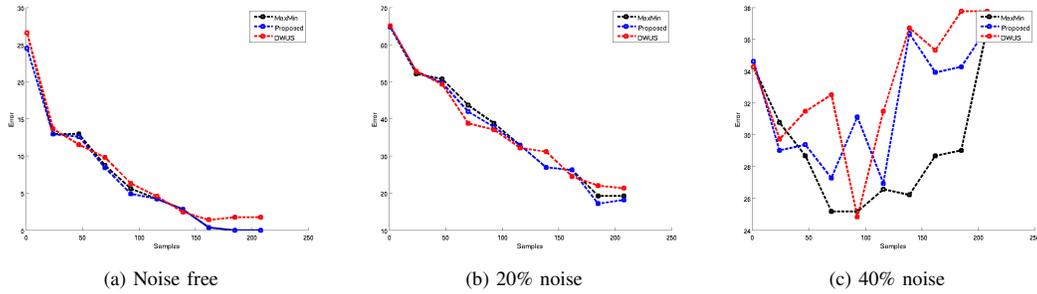(a) Noise free   (b) 20% noise   (c) 40% noise

Fig. 2. Impact of label noise on classifier error rates for different active learning selection criteria (lower is better). The figure key is as follows: 1) Maxmin [7] (red), 2) DWUS [12] (black) 3) Proposed (blue).
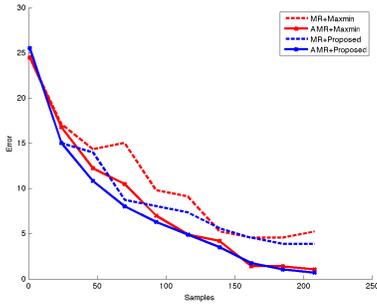

Fig. 3. The performances of different relabeling strategies on synthetic data with random error ($r = 20\%$).
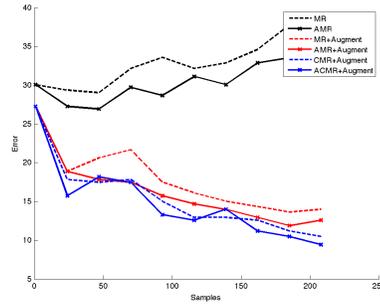

Fig. 5. The learning error by using Maxmin active learner combined with different relabeling strategies.
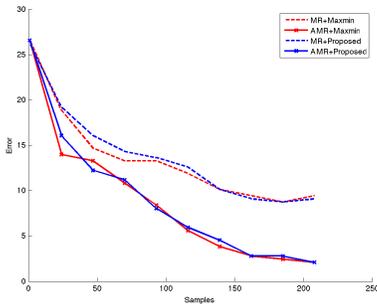

Fig. 4. The performances of different relabeling strategies on synthetic data with systematic error ($s = 40\%$). Absolute majority (AMR) dominates majority relabeling for both active learning selection criteria.

2) Absolute majority relabeling (AMR) - proposed strategy described above (black solid).
3) Majority + augmented labels (MR+Augment) - for each worker, we infer an augmented label using the Naive Bayes model based on both action and visible objects and select the most popular (red dash).
4) Absolute majority relabeling + augmented labels (AMR+Augment) - AMR applied to augmented labels (red solid).
5) Cumulative majority relabeling + augmented labels (CMR+Augment) - rather than accumulating peak votes

as in MR+Augment, we take the peak of the summed distributions (blue dashed)
6) Absolute cumulative majority relabeling + augmented labels (ACMR+Augment) - we accept the majority label from CMR only if it has an absolute majority, i.e., $> 50\%$ of probability mass (blue solid).

Note that in AMR and ACMR, samples where incremental crowdsourcing fails to obtain a label that can capture an absolute majority ($> 50\%$ after up to $N = 7$ independent votes) are discarded from the training set. To facilitate direct comparisons, all of these relabeling strategies are employed in conjunction with the standard uncertainty-based active learning sample selection criterion (Maxmin [7]).

As observed in Figure 5, using the raw action labels collected directly from MTurk (black dashed and solid) is problematic in conjunction with Maxmin because the label noise is severe enough to cause active learning to increase classifier error over time. This is true both for the straightforward majority relabeling strategy as well as our proposed AMR strategy (though the latter does better). Employing the augmented feature set, which combines action and object labels, addresses this shortcoming.

We also observe that employing cumulative voting with the augmented features (CMR+Augment, ACMR+Augment) during active learning significantly improves the classifier performance, and that the proposed absolute majority schemes
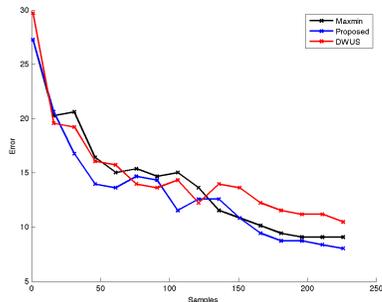
Fig. 6. Proposed relabeling strategy with different active learning methods on real data from MTurk.

are consistently better than their counterparts (AMR vs. MR, and ACMR vs. CMR), as seen by the contrast between dashed and solid lines. This validates our hypothesis that discarding inconsistently labeled samples from the training set leads to better performance. The best performance is achieved by the proposed method (ACMR+Augment), which combines absolute majority voting over label distributions with the augmented object+action features.

The final set of experiments (see Figure 6) explores the impact of the different active learning sample selection criteria on the CMU-MMAC dataset. In this experiment, all of the active learning criteria used the best relabeling strategy along with the augmented feature set (ACMR+Augment), with a pool of upto $N = 7$ independent relabeling opportunities. The active learning strategies are:

- Maximin [7], which focuses on sample uncertainty (black line);
- DWUS [12], which considers both unertainty and representative samples using the data distribution (red line);
- Proposed method that combines sample uncertainty with label consistency (blue line).

We see that DWUS initially improves faster than Maximin, since focusing on representative samples is an advantage. However, the strategy ultimtely converges to a higher error rate than the others because DUWS has reduced sensitivity to the uncertain samples. The proposed method performs much better than DUWS and slightly better than Maximin. The addition of the inconsistency term initially pushes the selection criterion towards samples far from the labeled data and then gradually picks up samples that have suspicious labels (those inconsistent with their local neighborhood).

VI. CONCLUSION

Although crowdsourcing annotations using active learning is an attractive and affordable idea for large-scale data labeling, the approach poses significant difficulties. Our study in the domain of wearable sensor-based activity recognition shows that a straightforward approach using the raw annotations obtained from Mechanical Turk in conjunction with standard margin criteria for SVM-based active learning would fail due to the high degree of annotation noise. This paper makes

three contributions that enable us to robustly train under these challenging conditions. First, we infer more accurate action annotations by combining objects with actions in a probabilistic framework. Second, we propose a new criterion for selecting instances in active learning that combines uncertainty and inconsistency measures. Third, we show that relabeling significantly improves the performance of active learning when the quality of annotations cannot be trusted. Our experiments using the CMU-MMAC dataset and Mechanical Turk confirm that the proposed approach improves active learning in noisy real-world conditions.

REFERENCES

[1] C. Vondrick, D. Ramanan, and D. Patterson, "Efficiently scaling up video annotation with crowdsourced marketplaces," in *ECCV*, 2010.

[2] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[3] F. De la Torre, J. Hodgins, A. Bargtell, X. Artal, J. Macey, A. Castellis, and J. Beltran, "Guide to the CMU Multimodal Activity Database," Carnegie Mellon, Tech. Rep. CMU-RI-TR-08-22, 2008.

[4] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, "A scalable approach to activity recognition based on object use," in *ICCV*, 2007.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *JAIR*, vol. 4, pp. 129–145, 1996.

[6] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM Multimedia*, 2001.

[7] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *JMLR*, vol. 2, pp. 45–66, 2002.

[8] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *ICML*, 2008.

[9] M. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *ICML*, 2006.

[10] P. Ipeirotis, "Crowdsourcing using Mechanical Turk: Quality management and scalability," in *Proceedings of Workshop on Crowdsourcing for Search and Data Mining*, 2011.

[11] M. Ramezani, J. Sandvig, T. Schimoler, J. Gemmell, B. Mobasher, and R. Burke, "Evaluating the impact of attacks in collaborative tagging environments," in *Proceedings of Social Computing*, 2009.

[12] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *ICML*, 2004, pp. 79–86.

[13] P. Donmez, J. Carbonell, and P. Bennett, "Dual strategy active learning," in *ECML*, 2007.

[14] V. Sheng, F. Provost, and P. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *ICKDM*, 2008.

[15] S. Vijayanarasimhan, P. Jain, and K. Grauman, "Far-sighted active learning on a budget for image and video recognition," in *CVPR*, 2010.

[16] P. Donmez and J. Carbonell, "Proactive learning: Cost-sensitive active learning with multiple imperfect oracles," in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008.

[17] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *ICACS*, 2010.

[18] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, 1997.

[19] S. Tong and D. Koller, "Active learning for parameter estimation in Bayesian networks," in *Proceedings of NIPS*, 2001.