A COMPARATIVE STUDY OF COLLABORATIVE VS. TRADITIONAL MUSICAL MOOD ANNOTATION

Jacquelin A. Speck, Erik M. Schmidt, Brandon G. Morton and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-Lab)
Electrical and Computer Engineering, Drexel University
{jspeck, eschmidt, bmorton, ykim}@drexel.edu

ABSTRACT

Organizing music by emotional association is a natural process for humans, but the ambiguous nature of emotion makes it a difficult task for machines. Automatic systems for music emotion recognition rely on ground truth data collected from humans, and more effective methods for collecting such data are being continuously developed. In previous work, we developed MoodSwings, an online collaborative game for crowdsourcing dynamic (per-second) mood ratings from multiple players within the two-dimensional arousal-valence (A-V) representation of emotion. MoodSwings has proven effective for data collection, but potential data effects caused by collaborative labeling have not yet been analyzed. In this work, we compare the effectiveness of MoodSwings to that of a more traditional data collection method, where annotation is performed by single, paid annotators. We implement a simplified labeling task to run on Amazon's crowdsourcing engine, Mechanical Turk (MTurk), and analyze the labels collected with each method. A statistical comparison shows consistencies between MoodSwings and MTurk data, and we produce similar results using each as training data for automatic emotion production via supervised machine learning. Furthermore the new dataset collected via MTurk has been made available to the Music Information Retrieval community.

1. INTRODUCTION

The problem of automated emotion (mood) recognition within music has recently received increased attention within the music information retrieval (Music-IR) research community [1]. The perceptual nature of emotion necessitates that such systems be trained on ground truth data collected from humans, and the Music-IR community could benefit from further development and evaluation of methods for collecting such data. In prior work, we created MoodSwings, an online collaborative game for collecting per-second labels of music, based on the two-dimensional arousal-valence (A-V) model of emotion [2, 3]. MoodSwings captures emotion changes in synchrony with music and collects a distribution of multiple players' labels for each moment in a song. These quantitative labels are well suited to computational parameter estimation and supervised machine learning [4–6].

Initial studies of the game's effectiveness found that annotators settle upon their final ratings faster when playing against a partner (as opposed to random AI, which simulates a partner's participation when an odd number of players are online) [7]. However, the effects of collaborative annotation on the quality of ratings have yet to be established. In this work we compare MoodSwings to a more traditional data collection method via Amazon's Mechanical Turk (MTurk), [1] an online crowdsourcing engine. Through the construction of Human Intelligence Tasks (HITs), MTurk connects researchers with human subjects from all over the web and provides a means for payment. We design a traditional A-V labeling task, employing MTurk workers to label a dataset consisting of 240, 15-second clips previously annotated via MoodSwings [4]. We examine the collected labels to comparatively analyze our game versus traditional data collection.

The monetary incentives of MTurk unavoidably inject noise into our labels. This becomes an issue for data quality as it is undesirable to pay for unsatisfactory work. MTurk allows us to deny workers payment if they do not properly complete the task, and we develop an outlier detection algorithm to automatically detect such workers. In an attempt to reduce bias, the system relies on the use of expert annotators' labels as a baseline when trying to validate annotations. It filters out workers who demonstrate unwillingness to correctly perform the task. We compare the "clean" MTurk dataset to labels from our game statistically, and with

[1] http://mturk.com

respect to automatic mood prediction accuracy. The new dataset collected via MTurk has been made available to the Music Information Retrieval community. [2]

2. BACKGROUND

The natural language processing (NLP) [8] and machine vision [9, 10] communities have utilized MTurk extensively, but machine listening and Music-IR have been slow to adopt its use. Lee found crowdsourcing music similarity judgments on MTurk to be less time-consuming than collecting data from experts in the research community [11]. The experiment cost $130.90 and produced 6,732 similarity judgements, less than $0.02 per rating. HITs were rejected if workers rated songs too quickly or failed to assign high similarity to identical songs. While nearly half of all HITs were rejected, the dataset was obtained an order of magnitude more quickly than in their previous attempts. Comparing the datasets yields a Pearson's correlation coefficient of 0.495, consistent with previous NLP work involving MTurk [8]. As the previous data collection was assembled for MIREX, Lee returned the submitted systems using MTurk data as ground truth and found no significant alterations to the outcome, scoring a 5.7% difference on the Friedman test.

Mandel *et al.* employed MTurk for collecting free form tags to study relationships between audio tags and content [12]. The group collected 2,100 unique tags across 925 clips, for a reported cost of approximately $100. To ensure data quality, they rejected a HIT if any tag had more than 25 characters, if less than 5 tags were provided, or if less than half of tags were contained in a dictionary of commonly applied tags (Last.fm). All HITs by a particular worker were rejected if the worker used too small a vocabulary, if they used more than 15% "stop words" (e.g., "music" or "nice"), or if half of their individual HITs were rejected for other reasons. The authors then trained a support vector machine (SVM) classifier for content-based autotagging. With smoothed labels, the MTurk version increased performance to 63.4% versus 63.09% with MajorMiner.

3. DATA COLLECTION METHODS

In previous work we designed MoodSwings, a collaborative online game that leverages crowdsourcing to collect mood ratings [2]. The game board is based on the A-V space, where the valence dimension represents positive versus negative emotions and arousal represents high versus low energy [3]. Anonymously-partnered players label song clips together during each round, scoring points based on the overlap between their cursors, which encourages consensus. Bonus points are awarded to a player whose partner moves towards him/her, encouraging competition and discouraging players from blindly following their partners to score points. We recently initiated a redesign effort, investigating gameplay improvements suggested by an analysis of collected labels [7]. However, we have not addressed concerns about the game structure biasing annotations.

We designed a simplified labeling task, shown in Figure 1, for MTurk. Single workers provide A-V labels for clips from our dataset, consisting of 240 15-second clips, which are extended to 30 seconds to give workers additional annotation practice [4]. As in MoodSwings, we collect per-second labels, but no partner is present and no points are awarded. Workers are given detailed instructions describing the A-V space. They navigate to a website which hosts the task and label 11 randomly-chosen clips. The first clip is a practice round, omitted from our analysis. The third and ninth are identical, randomly chosen from a set of 10 "verification clips," which are evaluated to identify unsatisfactory work. Workers are given a 6-digit verification code to enter on the MTurk website as proof of completion which, if successful, earns workers $0.25 per HIT.

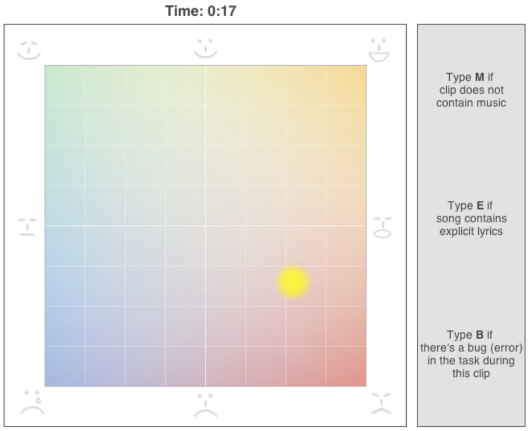

Figure 1. Screenshot of labeling task deployed on MTurk, depicting the A-V space and a yellow orb as the annotator's cursor. A sidebar provides additional instructions, e.g. workers may type "B" if they encounter bugs in the task.

4. FILTERING OF MECHANICAL TURK DATA

As previously discussed, quality control is an important issue with data collection on MTurk. In the labeling task, our interactions with workers are extremely limited and workers cannot ask for clarification of instructions during the HIT. It is difficult to gauge workers' understanding, and to determine if they were blindly moving the cursor to earn $0.25 for entering a verification code. Figure 2 shows examples of "good" and "bad" data collected for two song clips. We obtained annotations from 272 unique workers, an average of 5 HITs each. To determine which of the over 1,000 complete

[2] http://music.ece.drexel.edu/research/emotion/moodswingsturk

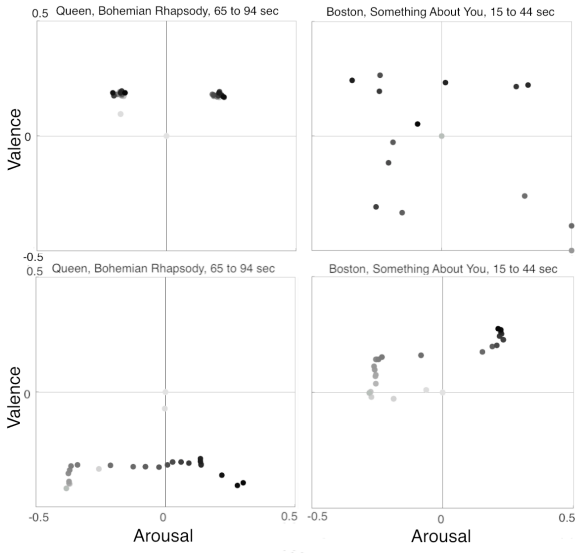

Figure 2. "Bad" (top) and "good" (bottom) worker data for two 30-second song clips. Labels get darker to show the progression of time. Both "bad" sets of labels indicate lack of understanding or attention to the task.

labeling sessions are valid, we utilize an automatic filtering system, which is trained on experts' annotations of our verification clips.

4.1 Baseline for Validity: Expert Annotations

We evaluate workers' verification clip labels to determine if they completed the task correctly. The 10 verification clips were handpicked for their obvious mood transitions, e.g., from low to high valence. Transitions occur near the middle of each clip. To provide a baseline for validity, ~10 Music-IR researchers labeled the verification clips twice each, during two sessions (one week apart). To demonstrate that the experts' labels provide a good baseline for validity, we measure the consistency of their ratings between sessions. Consistent ratings indicate attentiveness and understanding of the task, which characterize our expectations for correctly completed MTurk HITs.

The blue line in Figure 3 shows the normalized distances between the label distributions' means for each verification clip in the two annotation sessions, averaged over time. (e.g., normalized mean distance between clip one in session one and clip one in session two). As a baseline, the dashed line indicates the average distance over all individual clips from session 1 when they are compared to the combination of all remaining clips in session 2 (e.g., clip 1 from session 1 compared to the combined labels from session 2 clips 2-10). For all clips, the normalized mean distances between sessions 1 and 2 are well below the average, demonstrating consistent expert annotations over multiple trials.

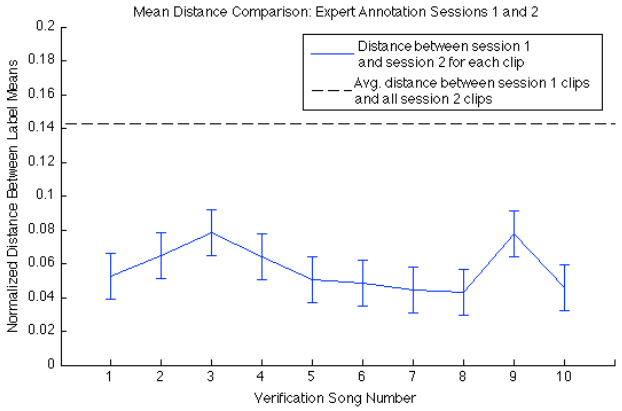

Figure 3. Distance between labels' means for each clip in expert annotation sessions 1 and 2, with error bars indicating ± 1 standard deviation. Dashed line indicates the average distance between individual clips from session 1 and the combination of all remaining clips from session 2.

4.2 Automatic Filtering System

We wish to reject data from workers who move about the A-V space without paying attention or who misunderstand the meanings of the A-V axes, but avoid rejecting valid ratings simply because they differ from our own subjective opinions. A one-class SVM for every second of each verification clip is trained on the expert labels, then used to detect invalid worker data. The experts' labels differ enough between individuals to account for many valid mood ratings, but to avoid penalizing workers for differences in opinion we only require that workers' verification clip labels fall within the decision boundary of the one-class SVM on average.

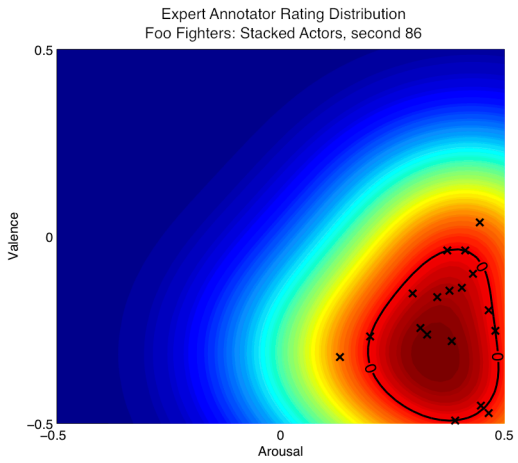

Figure 4. One-class SVM trained on expert data for one second of a verification clip. Expert labels (x), support vectors (0), and decision boundary are shown.

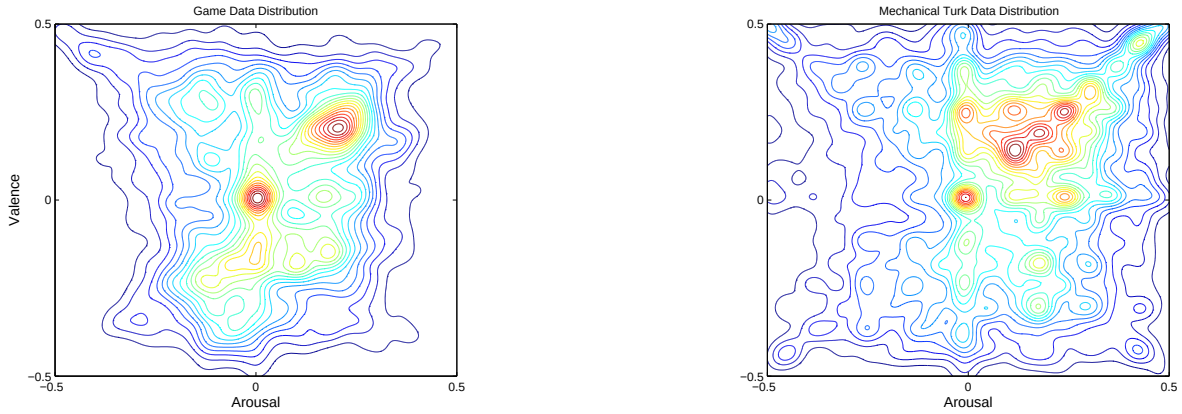

Figure 5. A-V distribution of data shown as a contour map. MoodSwings (left) and MTurk (right).

4.2.1 Novelty Detection and One-Class SVM

As our data is unlabeled, we cannot formulate the identification of valid labels as a traditional binary classification problem. The experts' labels exemplify how such labels may be clustered, a "positive class," but we encounter an unknown number of "negative classes." We use outlier (novelty) detection, training a supervised machine learning system on only positive examples [13]. Our system uses the one-class SVM implementation from the SVM-KM toolbox. [3] We use a Gaussian RBF kernel, tuning parameters to include most training data and exclude outliers. For a HIT to be approved, both verification clips must lie within our decision boundary on average. Workers must be approved for at least 60% of HITs completed, else all of their HITs are rejected. After automatic filtering, 113 workers had all HITs approved, and 88 had all HITs rejected.

Because emotions cannot be classified by machines with perfect accuracy, we use human judgments to measure the effectiveness of our automatic system. Plots of individual workers' labels for verification clips were visually examined by the authors. Annotations were classified "approved" if they followed a similar trajectory to that of the experts' labels over time, and "rejected" if they rapidly jumped between quadrants or moved in the opposite direction of expert labels. Ambiguous labels, for instance, those that did not follow a smooth trajectory, but moved towards the same quadrant as expert data, were labeled "unknown." Classification performance is shown in Table 1.

Manual Annotation	Number Accepted	Number Rejected
Approved	398	162
Rejected	147	527
Unknown	89	67
Precision	Recall	F-Measure
0.73	0.71	0.72

Table 1. Classification performance of automatic filtering system for HITs labeled "Approved," "Rejected," and "Unknown."

5. ANALYSIS OF COLLECTED DATA

The system collected 4,064 label sequences after two stages of filtering: first evaluating verification clip labels, and then removing labeling sessions of workers who kept the cursor at the origin for too long or consistently provided the same rating (e.g. consistently labeled all clips as angry throughout a game). We analyzed only the last half of each 30-second annotation round so that the first 15-seconds could give workers time to contemplate the mood of each clip. We assume that the relatively small number of workers who did not move after 15 seconds misunderstood the task, and thus filtered out their data. Table 2 shows statistics of the collected per-clip annotations in the dataset, before and after filtering.

Metric	Unfiltered Dataset	Verification Filtering	Stage 2 Filtering
Mean	49.79	18.20	16.93
St. Dev.	4.328	2.480	2.690
Max	72	24	23
Min	39	8	7

Table 2. Number of MTurk worker annotations for each clip before and after filtering.

[3] http://asi.insa-rouen.fr/enseignants/∼ arakotom/toolbox/index.html

Feature/ Topology	Average Mean Distance	Average KL Divergence	Average Randomized KL Divergence	T-test
MFCC	0.143 ± 0.007	1.501 ± 0.148	2.801 ± 0.294	20.68
Chroma	0.181 ± 0.008	3.555 ± 0.302	3.897 ± 0.313	21.08
S. Shape	0.158 ± 0.007	1.733 ± 0.172	2.501 ± 0.246	23.51
S. Contrast	$\mathbf{0.141 \pm 0.007}$	$\mathbf{1.486 \pm 0.158}$	$\mathbf{2.821 \pm 0.297}$	**21.17**
M.L. Combined	$\mathbf{0.130 \pm 0.006}$	$\mathbf{1.308 \pm 0.132}$	$\mathbf{2.928 \pm 0.310}$	**20.52**

Table 3. MLR results for short-time (one-second) A-V labels, repeating the experiments of [5].

5.1 Correlation Between Collected Labels

We compute Pearson's product-moment correlation between the datasets from MoodSwings and MTurk for each dimension. Pearson's correlation between example random variables X and Y is defined as:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (1)$$

To account for discrepancies between the number of annotations for each clip, we treat their per-second sample means as observations. We smooth both sets of labels to reduce noise between observations, as the mood in each second of a song clip cannot be assumed to be independent from that of previous seconds [6]. The results, 0.712 for Arousal and 0.846 for Valence, show more correlation between the two datasets than Lee's comparison of a MTurk-collected dataset to similarly crowdsourced data [11]. High correlation provides evidence that annotators' judgments are unaffected by collaborating with a partner during MoodSwings.

5.2 Overall Distribution Comparison

Figure 5 shows contour maps for the datasets collected with MoodSwings and MTurk. Both datasets have similar densities in the quadrant centers, though the MTurk dataset has higher densities along the spaces' extremities, which could be attributed to a larger sample. The MTurk dataset also contains small peaks throughout the distribution, whereas the MoodSwings set has more consistent clusters. It is difficult to pinpoint a cause for this difference, but multiple small peaks in the MTurk distribution may suggest that workers remain indecisive about their mood ratings throughout the duration of a clip. In previous work, we showed that it takes 7-8 seconds on average for players to reach 85% of the total distance from the origin to their final mood labels [7]. By contrast, it took MTurk workers 10-12 seconds to reach the same distance percentage. Faster convergence towards a mood decision in the game could imply that collaboration encourages annotators to re-evaluate their ratings earlier in the clips, perhaps improving the quality of collected data.

5.3 Performance in Emotion Prediction

To further establish correlation between the datasets, we use each as ground truth for the time-varying emotion prediction experiments of our previous work [5]. The prediction systems utilize supervised machine learning algorithms to map A-V labels to content-based audio features, e.g., mel-frequency cepstral coefficients (MFCCs), chroma, and statistical spectrum descriptors (SSDs), including spectral shape and contrast. Prediction performance for each feature, as well as combined performance using a multi-layer regression method for late-feature fusion, using multiple linear regression (MLR) is shown in Table 3. The results are similar: all features rank in the same order, and in terms of overall mean distance there is only slight improvement for the MTurk dataset. In terms of KL-divergence, the MTurk system performs significantly better. However, high KL values in [5] were later attributed to noisy distribution estimates at one-second intervals, taken independently from other time slices [6]. Increased performance on the MTurk set can be similarly attributed to the larger per-second sample sizes. Improvements based on the *quantity* of data collected are unrelated to the question of whether or not collaborative labeling biases the annotators' judgments.

6. DISCUSSION AND FUTURE WORK

The strong positive correlation between data from MoodSwings and MTurk provides evidence that collaborating with a partner does not bias annotators' mood judgments any more than participating in a traditional labeling task. We see similar mood prediction results between the label sets, although the MTurk set performs better with respect to KL-divergence. However, we attribute this increased performance to a larger sample size and propose that similar KL performance could be achieved if we collected a larger number of labels from MoodSwings. In terms of annotation quality, some evidence suggests that the game may be a superior data collection tool because it encourages participants to re-evaluate their ratings earlier in a labeling round. The set of labels from the MTurk annotation method is available to the Music-IR community for future research.

The logistics of each method are significant considera-

tions, particularly the pace of data collection and time spent on quality control. Monetary incentives can attract annotators very quickly, but researchers must determine how to separate anonymous paid annotators, e.g. MTurk workers, with good intentions from those who wish to obtain payment for as little work as possible. We advise researchers seeking to crowdsource subjective judgments from paid annotators to be wary of the complexity of quality control. Our one-class SVM system must be periodically retrained to account for varied mood judgments. As false approvals and rejections of mood labels are most accurately detected by humans, this requires manual labeling, which can be very labor intensive. Crowdsourcing the verification process may be a more viable solution [14]. Reliable workers may be identified through overall approval ratings available from MTurk and cold verify others' work in a separate task. However, paying a third party to verify results introduces further uncertainty to the filtering process. We prefer to deal with volunteer annotators, who are more likely to produce quality data without extensive filtering. Unpaid annotators do not benefit from producing a large quantity of low-quality annotations in a short amount of time. Because few people volunteer for tedious traditional labeling tasks, we hope that presenting the task as a fun game will attract annotators.

Further mood prediction work necessitates more data collection. In particular, some of our planned work requires annotations for a much larger, more varied song set. We have found the challenges of quality control for crowd-sourced data collection need to be considered when choosing a collection method. While using MTurk is a viable option, we plan to concentrate some of our future efforts on improving MoodSwings. We wish to attract annotators to our game as quickly as we attracted them with payment on MTurk. The redesign effort initiated with [7] made considerable strides towards improving the game. Continuing this effort by revamping the user interface, deploying the game on mobile platforms (e.g., iOS and Android) or a social networking website like Facebook.com and allowing participants to choose their own music will provide a more varied and enhanced gameplay experience. We hope these planned improvements will attract more annotators to our game. Dealing with certain paid annotators' attempts to earn money for unsatisfactory work makes a strong case for employing volunteer annotators, who we believe are less likely to "game the system."

7. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. of the 11th ISMIR Conf.*, Utrecht, Netherlands, 2010.

[2] Y. E. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. of the 9th Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, September 2008.

[3] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, U.K.: Oxford Univ. Press, 1989.

[4] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *MIR '10: Proc. of the Intl. Conf. on Multimedia Information Retrieval*, Philadelphia, PA, 2010, pp. 267–274.

[5] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. of the 11th ISMIR Conf.*, Utrecht, Netherlands, 2010.

[6] ——, "Prediction of time-varying musical mood distributions using Kalman filtering," in *Proceedings of the Ninth IEEE International Conference on Machine Learning and Applications*, Washington, D.C., December 2010, pp. 655–660.

[7] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, "Improving music emotion labeling using human computation," in *HCOMP 2010: Proc. of the ACM SIGKDD Workshop on Human Computation*, Washington, D.C., 2010.

[8] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," in *Proc. Empirical Methods in NLP*, 2008.

[9] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in neural information processing systems*. MIT Press, 2009.

[10] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *CVPR Workshops*, 2008.

[11] J. H. Lee, "Crowdsourcing music similarity judgments using mechanical turk," in *Proceedings of the 11th ISMIR Conferenceth International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, August 2010, pp. 183–188.

[12] M. I. Mandel, D. Eck, and Y. Bengio, "Learning tags that vary within a song," in *Proceedings of the 11th ISMIR Conferenceth International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, August 2010, pp. 399–404.

[13] L. M. Manevitz and M. Yousef, "One-class svms for document classification," in *The Journal of Machine Learning Research*, vol. 2, 2002.

[14] I. Sprio, G. Taylor, G. Williams, and C. Bregler, "Hands by hand: Crowdsourced motion tracking for gesture annotation," in *IEEE CVPR Workshop on Advancing Computer Vision with Humans in the Loop*, 2010.