Crowdsourcing Techniques for Affective Computing

(Handbook of Affective Computing chapter)

Robert R. Morris and Daniel McDuff

MIT Media Lab, MIT

ABSTRACT

In this chapter, we provide an overview of crowdsourcing, outlining common crowdsourcing platforms and best practices for the field as a whole. We illustrate how these practices have been applied to affective computing, surveying recent research in crowdsourced affective data and crowd-powered affective applications. We also discuss the ethical implications of crowdsourcing, especially as it pertains to affective computing domains. Finally, we look at the future of crowdsourcing and we discuss how new developments in the field might benefit affective computing.

GLOSSARY TERMS

affective computing, crowdsourcing, affective data, affective interfaces

INTRODUCTION

Crowdsourcing is a model of labor production that outsources work to large, loosely defined groups of people. In the past few years, crowdsourcing has proliferated widely throughout the fields of computer science and human-computer interaction, as exemplified by the dramatic increase in ACM citations for the term “crowdsourcing” (336 citations were reported in 2011, as compared to only four reported in 2007).

Moreover, despite being a relatively nascent discipline, crowdsourcing already has numerous high profile success stories to its name. For instance, in just 10 days, crowdworkers playing the *Fold-It* game managed to decipher the crystal structure of M-PMV –a feat of biochemistry that had previously eluded scientists for well over a decade (Khatib et al., 2011). M-PMV is a retrovirus that causes AIDS in chimps and monkeys and uncovering its protein structure could lead to the development of new antiretroviral drugs. Deciphering the complex three-dimensional shape of proteins like M-PMV has proven difficult for purely automated methods. Fold-It crowdsources this problem and leverages the natural three-dimensional problem-solving abilities of its players.

The ESP Game is another example of a wildly successful crowdsourcing endeavor. Remarkably, in just a few months of deployment, *The ESP Game* collected over 10 million image labels (von Ahn, 2006). In neither *The ESP Game* nor *FoldIt* were participants paid for their contributions.

Crowdsourcing techniques have a lot to offer the field of affective computing. Human expression of emotion is complex, multi-modal and nuanced. To design computer algorithms that accurately detect emotional expressions, large amounts of labeled data need to be collected under conditions that reflect those seen in real life. Collection and labeling using traditional lab based methods can be inefficient and sometimes impractical. By engaging a large workforce to collect, contribute and label emotional expressions (often via the Internet), the time and expense required to create databases for training expression recognition systems can be greatly reduced.

Furthermore, crowdsourcing offers exciting new ways to power emotionally intelligent affective computing applications. Technologies that provide personalized and

contextualized affective feedback require significant advances in natural language processing and commonsense reasoning. To sidestep this problem, crowds can be recruited on demand to support artificial intelligence, providing human-computation when automated methods alone are insufficient. In this chapter, we describe how crowdsourcing techniques can be applied to these and other research challenges within the field of affective computing.

CROWDSOURCING OVERVIEW

What is Crowdsourcing?

Crowdsourcing was first coined by Jeff Howe in a June 2006 article for *Wired* magazine. Following Howe's original conception of the term, we define crowdsourcing as a method for recruiting and organizing ad hoc labor, using an open-call for participation (Howe, 2006). Crucial to this definition is the notion of an extremely fluid workforce, one that is devoid of the managerial and contractual directives inherent to other, more traditional labor models, such as "outsourcing." The crowd has the freedom to do what it pleases, when it pleases, and it is often up to designers to find clever ways to recruit and retain this kind of ad hoc workforce.

Conversely, unlike the crowd itself, the entity requesting crowdsourced work is rarely fluid or loosely defined. Indeed, the request for work usually comes from a distinct organization or individual (known as the "requester", in crowdsourcing parlance). This

distinguishes crowdsourcing from other, more decentralized labor structures, such as commons-based peer production.

While crowdsourcing can take many forms, three dominant platforms have emerged in the past decade: (1) games with a purpose, (2) micro-task markets, and (3) open innovation contests. These platforms are also among the most relevant for affective computing research, so we will take a moment to describe how they work.

Games with a Purpose

Games with a Purpose (GWAPs) are games played online, ostensibly for fun, but the output of the gameplay is used to solve real-world problems. Perhaps the first, and best-known GWAP is the ESP game – an online, social game that has helped create millions of image labels for the web (von Ahn, 2006). While image labeling is ordinarily quite tedious, the ESP game uses social dynamics and game mechanics to make the experience fun and engaging. Image labels generated from the game are used as metadata for online images, improving image search and web accessibility for the individuals with visual impairments.

Other GWAPS have since been created to generate everything from simple image annotations to incredibly complex protein structures. GWAPs can be applied to many different problems, and they can be extremely powerful when done correctly. However, game design is challenging and it is not always possible to magically alchemize tedious or frustrating work into highly engaging gameplay. Also, GWAPs may require advertisements or marketing to attract sufficient numbers of players. Sometimes, it can be

easier to simply pay the crowd to complete work. In cases such as these, micro-task markets can be an attractive option.

Micro-task Markets

Micro-task markets are systems in which workers complete short jobs in exchange for monetary compensation. Workers are not obligated to work beyond the tasks they elect to complete, and employers are given the freedom to propose any price for completing the jobs they want done. To date, the largest and most popular micro-task market is Amazon's Mechanical Turk service (MTurk). MTurk hosts hundreds of thousands of tasks, ranging from image labeling to text transcription to spam filtering.

MTurk is an attractive resource because task completion is very fast and it has its own API, allowing employers to coordinate tasks programmatically. The cost of labor is also very low – in 2010, the median wage on MTurk was only $1.38/hour (Horton & Chilton, 2010). Unfortunately, MTurk's speed, low cost, and programmability are offset by its poor quality control. The system does not employ reliable reputation systems and it is very easy for workers to do the bare minimum, or even cheat, and still get paid. As of this writing, it is still fairly easy for workers to artificially boost their performance rankings on MTurk.

To use MTurk effectively, researchers need to carefully consider how they recruit workers, how they manage quality, and how they design their tasks. For tasks that require English language fluency (such as sentiment analysis), it is often necessary to restrict enrollment to individuals from English-speaking countries. Using MTurk's built-in

location filters can be a useful first-step, but the data should also be pruned post-hoc, by assessing IP addresses.

Task design also influences performance on MTurk (Kittur et al., 2008). Researchers should strive to make task instructions clear and provide examples of how to complete the task appropriately. Gold standard questions (i.e., questions with known answers) should also be placed within tasks, to help investigators identify problematic data submitted by careless or cheating workers. Another popular technique is to ask workers to pass qualification tests before offering them a chance to work on actual for-pay tasks. This technique slows down the recruitment process and reduces the pool of available workers, but it can help improve results dramatically.

Finally, for longer and more complex tasks, researchers might consider using other employment markets, such as oDesk, which tend to attract more specialized workers than those found on MTurk.

Open Innovation Contests

In addition to micro-task markets, open innovation contests provide yet another platform for conducting crowdsourced work. Unlike markets, in which compensation is given for all work, open innovation contests only compensate top performers. Challenges are posted online, either through company-wide initiatives – such as Nokia's Ideas Project, Dell's Idea Storm, and OpenIDEO - or through innovation hubs like Innocentive. Because these contests are crowdsourced and open, companies get the opportunity to

glean insights from thousands of people from many different backgrounds and disciplines.

The benefits from open innovation contests are not restricted to industry; academic research teams can use this approach to crowdsource interesting ideas and technologies that might not otherwise be considered by members of their own field. Recently, members of the affective computing community hosted the Facial Expression Recognition Analysis (FERA) challenge – a contest designed to see which research group could best detect facial expressions using automated methods. The FERA challenge was not crowdsourced in the traditional sense (the call for participation was not "open" and was primarily directed towards members of the affective computing community), but future iterations could aim for a larger, more diverse contestant pool. Indeed, Lakhani & Wolf (2005) find that most solutions on Innocentive – an open innovation hub - came from workers just outside the relevant discipline of the problem (i.e., a biology problem was more likely to be solved by a physicist, than a biologist). Individuals from adjacent disciplines may have sufficient training to understand the problem space, but they also have an outsider's perspective that helps them approach the issue from a radically new direction. An open innovation approach could attract fresh sets of eyes to difficult affective computing problems and might lead to exciting breakthroughs in the field.

Motivations for Participation

While many affective computing tasks can be crowdsourced using existing platforms, such as those described in the preceding sections, many situations require researchers to construct crowdsourcing applications of their own. When considering new crowdsourcing

systems for affective computing, perhaps the most important design questions pertain to incentive structures. Since crowdsourcing systems are typically open to anyone, and do not rely on contractual relationships, people are not bound to participate. Instead, they must feel somehow compelled to participate, and so the factors that make a given system compelling might also those that make it succeed or fail. There are many ways to attract crowdworkers, but researchers such as Malone et al (2009) describe three overarching motivations that govern most, if not all, crowdsourcing platforms: money, love and glory.

Money

The dominant incentive mechanism on micro-task markets is money. Different markets have different norms for remuneration rates, but higher pay generally gets more workers to do more work more quickly (Mason & Watts, 2009). In some cases, money can be the only motivator that will work. When the work is tedious and unpleasant (such as transcribing pages and pages of hand-written documents), crowdworkers are unlikely to participate unless fair monetary compensation is guaranteed. Money is not necessarily the best way to ensure quality, however. Researchers studying MTurk have found that higher rates of pay do not necessarily lead to higher quality work (Mason & Watts, 2009; Rogstadius et al., 2011).

Love

Of course, money is not always required to fuel crowdsourcing endeavors. If workers simply love the task itself, they will contribute their time for free. Some people love tasks that challenge them and encourage them to think creatively. Software developers cite the

creative challenges of coding as the primary reason they contribute to free and open source software projects (Lakhani & Wolf, 2003). Others enjoy tasks that offer them a chance to exercise skills and talents that they don't get to use in their ordinary working lives (Howe, 2009).

Others may be driven by idealism. They may contribute simply because they support the overall goals of the project. In the NASA Clickworkers project, for example, researchers needed thousands of individuals to help label and classify craters on mars. At the outset, the project designers could not be sure that people would contribute without being paid. However, because people were genuinely excited by the scientific goals of the project, they were willing to work for free, even though the task itself was somewhat repetitive.

Novelty can also be a powerful motivator. The chance to interact with new technology may be enough to encourage participation. For instance, participants may be willing to interact with affect recognition software simply because the technology is new and exciting and fun to experience first-hand. Affective computing researchers might consider using this as an angle to attract users to test out new designs or participate in research studies.

Glory

Finally, some crowdworkers are driven to compete for the recognition of their peers. Many programmers flock to coding contests for precisely this reason. Respect from one's peers can be a powerful motivator and many systems incorporate leaderboards and other

reputation signals to encourage participants to compete amongst their peers. Competition can also make a crowdsourcing task more game-like and social and can help enhance worker motivation.

Crowding Out

More often than not, many of the aforementioned motivational structures are combined together into one crowdsourcing system. For instance, the Fold-It project combines the intrinsically rewarding properties of video games with the lofty ideals of solving some of biochemistry's most challenging problems (Cooper et al., 2010). The game also allows players to compete for high scores, offering reputational incentives.

However, it is not always the case that extra incentive mechanisms are a good thing. Motivation does not increase monotonically with the number of incentive structures, and some motivations may in fact 'crowd-out' others (Benkler, 2007; Deci, 1971; Frey & Jegen, 2001). Most notably, extrinsic rewards, such as money, can overshadow a task's intrinsic rewards, causing people to put forth just enough effort to be paid. In some cases, additional incentive structures may have the opposite of their intended effect and may actually reduce motivation overall. When designing incentive structures for crowdsourcing applications, care should be taken to make sure different motivations don't conflict with one another.

Quality Control Techniques

Most of today's crowdsourcing systems are extraordinarily meritocratic. *Innocentive, the ESP Game,* and *Threadless,* and countless others are built on an open-call, such that

anyone, regardless of educational, geographical, or occupational background, can sign up and contribute. Huge crowds, comprised of people with diverse backgrounds, can often yield incredible results, if managed successfully. But, managing crowds successfully is not a trivial problem. The benefits of crowds can also be their drawback; while their size and diversity can bring new, innovative ideas, they can also bring about great variance in quality.

There are essentially two categories of quality management in crowdsourcing systems: input management and output management. On the input side, steps can be taken to ensure that, as information is collected, it is pruned for quality. On the output side, crowd contributions can be filtered such that only the best and most relevant material rises to the top.

Quality Control: Input Management

Granularity

In his seminal paper, "Coase's Penguin, or, Linux and the Nature of the Firm," Yochai Benkler outlines several components that underlie successful crowd-based systems (Benkler, 2001). Among them is the notion of 'granularity' – a way to decompose complex tasks down into simple, digestible components. Benkler also describes how tasks should be heterogeneously-grained, in order to accommodate different motivations among the participants. Many people have limited spare cycles and can only devote a little bit of time to crowd-based systems. Others may have more time, and they may be more incentivized to work for longer hours. *Wikipedia*, while not a crowdsourcing system

in the traditional sense of the definition, provides a nice illustration of heterogeneous granularity – the site offers the option to simply tweak one word as well as the option to write entire articles from start to finish.

Iteration

Quality can sometimes be improved when workers are allowed to build on existing work. *Wikipedia* for instance, uses the Wiki structure to let contributors build on the contributions of others. Also, in the crowdsourced Matlab programming challenge, participants constantly build upon each other's code throughout the contest. Rather than have everyone work in isolation until the contest is over, the event is structured so that participants can build upon the ideas of their peers throughout the contest.

With toolkits like *Turkit*, MTurk tasks can be coordinated via an iterative structure. In some cases, this approach can cause dramatic improvements in quality. For instance, Little et al. showed that, when MTurk workers iteratively improved the description of an image, the final result was preferred over other methods 9 of 11 times (Little, Chilton, Goldman, & Miller, 2009).

While iterative approaches can improve quality in some cases, they are not without their perils. Iterative conditions can sometimes lead to information cascades, wherein people follow the actions of others simply because they believe that those that came before them were well informed (Bikhchandani, Hirshleifer, & Welch, 1992). In cases such as these, crowdsourcing tasks should be parallelized instead.

Norms

In addition to the methods described so far, social norms can also affect the quality of contributions coming into crowdsourcing systems. Crowdsourcing news aggregator sites like *Slashdot* maintain norms in order to manage the quality of incoming news links. For sites such as these, there is a norm that governs the types of articles that should be submitted. In some cases, norms are made explicit, in writing, while in other cases they are implicit. Recently, the website Pinterest sent all new users an email explicitly reminding them to "be respectful, be authentic, and to cite all sources." Norms such as these can be a simple, yet powerful way to sculpt the contributions from the crowd. Also, crowdsourcing researchers and employers should take care to understand the norms of different platforms. MTurk, for instance, does not prevent employers from posting incredibly long tasks. However, the norms of the site revolve around small, micro-tasks and extremely long tasks are sometimes eyed with suspicion. For instance, Soleymani and Larson found that many MTurk workers were worried about accepting a HIT that involved 125 video annotations (Soleymani & Larson, 2010). Workers may not want to sign up for an incredibly long task, if they are not certain that they will be paid for their time.

Reputation

In many crowd-based systems, the potential for free-riding is considerble. In Amazon's Mechanical Turk service, for example, workers can get away with producing low quality work, largely because the punishments are low. Reputation structures, such as those found on *eBay.com*, reduce the prevalence of free-riders, thereby increasing the quality of the overall system. Indeed, studies on human cooperation show that when interactions have consequences that extend out into the future, defection and free-riding

drop dramatically (Axelrod, 2006). Individuals are more apt to cooperate when their interactions are recorded. These same principles can be applied to crowdsourcing systems for affective computing.

Quality Control: Output Management

It is not always possible to manage the quality of information that comes into crowdsourcing systems. While many of the approaches described above can improve the quality of content coming into the system, additional filtering may still be needed.

Statistical Techniques

When crowds are asked to make quantitative estimations, the group average can sometimes yields a more accurate result than any one person's estimation. This so-called "wisdom of crowds" effect was first described by Francis Galton in the early twentieth century and has since been replicated in countless other studies (Surowiecki, 2005). Thus, in some crowdsourcing domains, quality can be ensured through simple statistical techniques (in some cases, depending on the distribution of the responses, one may be able to average the opinions of the crowd to get the best answer). Unfortunately, not all crowd-based systems collect quantifiable information that can be neatly processed by simple parametric statistical techniques. In many cases, the information going into a crowd-based system is subjective and qualitative and not directly amenable to statistical manipulation. Oftentimes, the crowd is needed to rank contributions before statistics can be used. For instance, consider a case where crowds are recruited to contribute textual descriptions of affective images or movies. It may be hard to rank these descriptions or

know which are the most relevant, unless yet another set of crowdworkers is hired to curate them.

Crowd Voting

Crowds can generate massive amounts of information, and sometimes it takes the power of a crowd to sift through it all. Interestingly, just as we can use the crowd to gather information, we can also use the crowd to rate information for relevance and value (Benkler, 2007; Howe, 2009; Malone, Laubacher, & Dellarocas, 2009). This approach can be employed passively, such that the crowd does not even know that its behaviors are being used to rank information (e.g., Google's PageRank or Amazon's collaborative filtering algorithms). Or, it can be employed actively, such that the crowd is explicitly tasked to make objective ratings of crowd contributions, e.g., the verify step in Soylent's *find, fix verify* algorithm (see Bernstein et al., 2010).

CROWDSOURCING AFFECTIVE DATA

Data Collection

As with many domains of artificial intelligence, the performance of affective computing systems is dependent on the quality and quantity of training examples that are available. Crowdsourcing offers new ways to efficiently collect large amounts of affective data. However, there are a number of challenges in collecting data in relatively uncontrolled settings. In this section we discuss the potential for data collection via the crowd, along with the pitfalls and technological limitations.

We consider two forms of data collection; 1) using crowdworkers to generate original affective data, perhaps in response to a stimuli or by acting, and 2) using crowdworkers to source existing examples of affective data. Data could be in the form of text, audio and/or visual material.

The widespread availability of webcams and video platforms such as YouTube has facilitated the collection of large amounts of rich image and video data. For instance, Taylor et al. (2011) and Spiro (2012) describe ways to use webcams to crowdsource posed data for training gesture recognition systems. McDuff et al. (2012) present the first corpus of videos of naturalistic and spontaneous facial responses collected over the web. In this research, participants were asked to view a media clip in their web browser while their facial expressions were recorded using a webcam. Over five thousand video responses were collected in little over a month and none of the participants were paid.

Collecting naturalistic affective responses via the web raises some interesting issues. Opt-in participation is particularly important in cases where images from webcams as captured and stored. Another issue relates to data quality. High bandwidth data – such as real-time videos – may be limited in resolution and frame-rate, which can present challenges for data analysis and feature detection.

In addition to generating affective data, crowds can help curate and collect affective data. Naturalistic affective data can mined from various online repositories of videos, most notably YouTube.com. For instance, Morency, Mihalcea and Doshi (2011) created a corpus of 47 videos from YouTube for multimodal sentiment analysis. While their corpus was handpicked by the researchers themselves, future efforts could delegate

crowdworkers to help collect and curate even larger corpora of videos. Websites such as YouTube contain a plethora of videos showcasing naturalistic affective expressions and reactions. Finding these videos is a challenge in itself, however, and sifting through them all often requires the combined efforts of a large crowd of people.

Collecting Labels

Ground-truth labels are a fundamental component of datasets. One of the most popular uses of crowdsourcing in affective computing is in collecting ground truth labels of affective data. VidL was the first example of a distributed video-labeling tool specifically designed for labeling affective data (Ekhardt and Picard 2009). Games with a Purpose (GWAP) have also enabled the efficient collection of labels. For instance, Riek, O'Connor and Robinson (2011) present "Guess What?" - a GWAP with the intention of labeling affective video data of social situations. Soleymani and Larson (2010) sourced boredom annotations for a corpus of affective responses to videos. In this case the agreement between workers was low (Cohen's kappa = 0.01 for boredom labels and 0.07 for emotion word labels), which highlights the subjective nature of many of these tasks. Labeling motion tracks for training gestures recognition systems is another example of a crowdsourcing application that could help provide useful datasets for training affective computing systems (Spiro et al. 2010). Sign-language recognition is a particular area in which this may prove useful.

So far we have mostly considered non-verbal affective data. However, crowdsourcing has also been used to label emotional speech and text. Tarasov et al. propose ways to crowdsource emotional speech assets (2010). Sentiment labels for text

have been successfully crowdsourced (Hsueh et al. 2009) and Mohammad and Turney (2011) created a word-emotion lexicon using crowdsourced labels. GWAPs have been used for similar purposes (Pearl and Steyvers, 2010). Music can powerfully evoke affect and is frequently used as affective stimuli. Several examples of crowdsourced emotion labeling for music have been presented (Turnbull et al. 2007, Kim et al. 2008, Morton et al. 2010, Speck et al. 2011).

Evaluating Labeler Agreement

There are certain nuances to labeling affective data, not least the fact that in many cases there is no objective ground-truth. Rather, labels are often subjective judgments about perceived affective phenomena. Also, there is often inconsistency between multiple labelers. A number of methods have been proposed for evaluating agreement between multiple labelers. Cohen's kappa, κ, is the most commonly used statistic, although there is some disagreement over what thresholds indicate good and weak or bad agreement (Tarasov, 2010). As a guide $\kappa > 0.6$ might be considered good. However, in a number of cases labels which have $0.6 > \kappa > 0.3$ have been used even if this does not reflect strong agreement. Typically, greater numbers of labelers will increase the reliability of labels as random errors should begin to cancel out (Harrigan 2005, Soleymani and Larson 2010). In addition, techniques have been developed to identify good annotators over bad annotators and to identify biases that might exist in annotations (e.g. Tarasov, 2010).

As with most other crowdsourcing tasks, participants for affect labeling tasks can be recruited using micro-task markets (such as MTurk). However, social networking sites (such as Facebook) provide another efficient method of recruitment. Depending on the

platform, be it a GWAP or a more explicit, for-hire labeling job, participants may be volunteers or be paid workers.

Crowdsourcing labels for affective data raises many interesting questions about how to design crowdsourcing tasks. The optimal trade-off between expertise, number and diversity of labelers, time available and cost is likely to be dependent on whether data is naturalistic or posed, of multiple or single modalities and whether labeling requires training (e.g. facial action coding certification). For affect labeling in particular, designers need to consider whether labels will vary with different demographics (e.g. people from different cultural backgrounds). These differences, if undetected, could lead to unexplained heterogeneity in the labels.

CROWD-POWERED AFFECTIVE APPLICATIONS

Recently, human-computer interaction researchers have begun to explore user interfaces that utilize both automatic and human-powered processes. These "crowd-powered systems," as they are sometimes called, recruit human intelligence as needed, when automatic processes alone are insufficient. Crowd-powered applications have been developed to help people see (Bigham et al., 2010), edit Word documents (Bernstein et al., 2010), plan itineraries (Zhang et al., 2012), and even operate remote-controlled robots (Lasecki et al., 2011). In this context, crowds are not simply used to train algorithms. Rather, they are recruited on-demand, in response to the unique needs of the end-user, and they comprise a large part of the application's computational power. This approach is still quite novel, and it has yet to be used widely by researchers in the affective

computing community. That said, as of this writing, there are at least two affective computing applications that explore these on-demand, crowd-computing techniques.

Crowd-Powered Emotion Regulation

Morris & Picard (2012) describe ways to use crowd-powered techniques to power emotion regulatory technologies. Specifically, they outline ways to crowdsource elements of cognitive-based therapies, including cognitive reappraisal and cognitive restructuring, to help individuals regulate distressing emotions. In their design, crowds are recruited to help individuals reappraise and restructure emotion-eliciting thoughts and situations. Users submit short, 1-2 sentence descriptions of something causing them stress or untoward anxiety. These descriptions are sent to workers on MTurk, each of whom reframes the text in different ways. Some apply cognitive restructuring and examine the user's text for possible cognitive distortions (e.g., all-or-nothing thinking, overgeneralization). Others are asked to apply cognitive reappraisal – a technique that involves changing the meaning of a thought or situation to alter emotional experience (Gross & John, 2003). In all cases, crowdworkers are trained on-demand, and are given short 1-2 minute tutorials prior to completing the work. The crowd's work is coordinated programmatically, and a crowd-voting stage is implemented to help ensure only the best responses get returned to the user. The basic design of the system utilizes a "wisdom of crowds" approach, wherein the unique perspectives of many workers is used to generate novel and intriguing reappraisals that might not ordinarily be considered by a small set of skilled experts. Finally, there is also an empathy component, wherein crowdworkers are

taught to apply person-centered support to help the user know that they have been understood.

Analyses of the design revealed that, with minimal training, crowdworkers were able to classify cognitive distortions with 89% accuracy. The authors also tested the quality of the responses generated by their system. They found that responses with reappraisals were rated significantly higher than those generated by an open-response structure in which workers were simply asked to help the user feel better and in which contributions were not coordinated or filtered algorithmically.

Unfortunately, this application has yet to be thoroughly tested in long-term user studies, and it remains unclear how it will be received by real end-users. To be useful in real-world deployments, the system must be able to respond quickly and the quality of the responses will have to be high.

Crowd-Powered Social Stories

In addition to emotion regulation, crowd-powered design principles have also been applied to support individuals with autism spectrum disorder (ASD). To manage anxiety when faced with new situations, individuals diagnosed with ASD often rehearse behavioral repertoire using social stories – scripted routines that outline the steps involved in a given task or interaction (such as getting tickets at a movie theater). However, despite advances in common-sense reasoning, it is still impossible for purely computational processes to generate context-appropriate social stories for many different situations. Moreover, authoring social stories for individuals with autism can be complex

and it can be very time-consuming for any one person to exhaustively list the sequence of events needed to navigate a given social situation. It can be especially difficult to generate all the contingencies that must be considered in case a problem arises. To solve this problem, Boujarwah et al (2012) describe ways to crowdsource the creation of models for social stories for individuals with ASD. Specifically, crowdworkers are asked to brainstorm and classify steps involved in completing a particular task (such as "eating lunch"). Crowdworkers are then asked to brainstorm obstacles that an individual might encounter and ways to get around these obstacles. The general approach described by this work could potentially be applied to any individual facing a challenging new situation. In the future, the approaches described by Morris & Picard and Boujarwah et al might be combined, to help individuals navigate both the practical and emotional hurdles involved with stressful situations.

ETHICAL CONSIDERATIONS

Crowdsourcing is still an evolving field and many of the ethical implications it raises have yet to be resolved. As of this writing, MTurk does not impose minimum wage restrictions. Employers are free to offer any form of compensation, no matter how menial. While many people in the U.S. do not rely on MTurk as a primary source of income, many individuals in India consider their wages crucial for daily subsistence (Panos, 2010). In the future, greater oversight should be placed on wages, to ensure that crowdsourcing work does not evolve into a digital sweatshop, as some researchers fear (Fort, Adda, & Cohen, 2011). Moreover, more work should be done to help crowdworkers develop new, meaningful skills that generalize to other work domains. All

too often, crowdworkers are given tedious, rote tasks that contribute little in the way of new, marketable job skills.

Worker anonymity is another issue that can be particularly troublesome for crowdsourcing researchers, particularly those conducting affective computing studies. In most crowdsourcing systems, worker identities are kept hidden and it can be hard to know where they are from or how old they really are. Tasks that involve stress induction or exposure to challenging media (e.g., the International Affective Picture System (IAPs)) may be inappropriate for young persons and yet researchers may find it difficult or impossible to impose age restrictions on crowdsourcing platforms such as MTurk. While MTurk's requires users to be over 18 years of age, it is not clear how well this policy is enforced. An adult could easily register as a worker and then hand over the account to a child.

Another potential problem relates to liability issues. This issue is particularly important for assistive devices that rely on crowdsourced work, such as those described in our previous section on crowd-powered affective applications. If members of the crowd mislead the user or provide malicious feedback, it is unclear who is responsible. Should liability reside with the workers in the system or the designers of the system?

Finally, some have also considered how crowdsourcing design, by its very nature, can lead to malicious and dangerous applications. When work is parceled into tiny bits, it can be hard for crowdworkers to know whether their work, as a whole, is contributing to something virtuous or vicious. For instance, a despotic regime could easily crowdsource its efforts to identify dissidents in a large crowd of people. The regime's actual intent

could, in a sense, be laundered by decomposing the overarching goal into small, nondescript micro-tasks.

Future of Crowdsourcing

In general terms, crowdsourcing is simply a method of recruiting and organizing labor, and its basic framework has been around for many years, if not decades. Yet, in recent years the practice has evolved considerably and has proliferated rapidly. Advances in communication technologies, combined with new crowdsourcing platforms and techniques, have led to exciting new innovations for the field. And while the practice is still undergoing significant growing pains, particularly with regard to its ethical quandaries, it is likely to expand in the coming years. In this section, as we speculate on how the future of crowdsourcing will affect affective computing, we focus on three emerging trends in the field of crowdsourcing: (1) real-time crowdsourcing, (2) skilled crowdsourcing, and (3) offline crowdsourcing.

Real-time crowdsourcing

In most crowdsourcing situations, there exists a large gap between the time work is requested and the time work is completed. For micro-tasks that require mere seconds to complete, this latency is largely an effect of the time it takes to recruit and train new workers, not the time it takes to do the actual work. To solve this problem, Bernstein et al (2011) describe ways to place crowdworkers on retainer, so that workers are already recruited and trained by the time requests for work arrive in the system. In their model, workers get paid small amounts to wait for tasks, and are told to respond as soon as they

are notified (a javascript alert and audio chime is used to notify workers that a new job is ready to complete). Using this design, many workers can be recruited synchronously at a moment's notice, creating a sort of "flash mob" of workers. For affective computing technologies that require on-demand crowdsourcing, real-time crowdsourcing methods such as these will reduce latency dramatically and will pave the way for new types of interactive systems. Currently, most interactive affective technologies (e.g., social robots, emotional support systems) rely on automated algorithms and artificial intelligence. In the future, these technologies may be augmented by real-time crowdsourcing techniques, drawing on human intelligence when needed.

Skilled Crowdsourcing

For many affective computing applications, skilled workers are needed to label complex data or power sophisticated interactive systems. While some crowdsourcing platforms offer ways to train workers, it can be difficult to retain these trained workers for future tasks. In the future, crowdsourcing platforms will hopefully offer ways to target skilled workers, either by retaining and tracking those that have performed well in the past or by finding new workers that have the desired skills. Games with a purpose offer intriguing ways to find skilled workers, by allocating jobs only to players who have completed certain levels in a game. While this model has been explored somewhat in systems like Fold-It, more work can certainly be done to better understand how best to allocate work based on a player's achievement in a game or instructional program.

Offline Crowdsourcing

For the most part, crowdsourced work is situated online. That said, new services like taskrabbit.com and gigwalk.com are applying the crowdsourcing model to real-world tasks, such as moving furniture or conducting in-store audits. As more crowdsourcing platforms move offline, affective computing researchers can begin to take advantage of real-world data collection. Many individuals are already wearing biosensors as part of the quantified health movement. Given the proper incentives, some of these individuals might be willing to share subsets of their data, to help researchers develop more powerful affect detection systems. For instance, if enough people are wearing biosensors, and are willing to do upload their data, intriguing new datasets of real-world affective experiences can be crowdsourced. Just as twitter has helped researchers understand contagions and flu outbreaks, crowdsourced biosensor data might help us understand complex emotional and psychophysiological patterns across large groups of people. For instance, researchers could get a better understanding as to how groups of people react to traffic jams or other urban inconveniences. Such data might help guide new infrastructure and might be used to build new emotional support systems that intervene in extremely context-specific ways.

CONCLUSION

Although crowdsourcing is still relatively new, it already has the potential to dramatically accelerate affective computing research. Large datasets are crucial for improving the performance of affect recognition systems. Access to large groups of workers via crowdsourcing can make data collection and labeling much more efficient. Also, the

augmentation of computer systems with human intelligence can engender exciting new applications, such as emotionally intelligent assistive devices.

It is hard to predict how crowdsourcing will evolve in the coming years. Perhaps the best way for affective computing researchers to secure a future that benefits them is to create it themselves. Designing systems with effective and sustainable motivation strategies, creating methods for validating and verifying data collected and solving ethical issues associated with large-scale and distributed labor are the main areas that need to be addressed in the near future.

References

Axelrod, R. (2006). *The Evolution of Cooperation: Revised Edition* (Revised.). Basic Books.

Benkler, Y. (2002). Coase's Penguin, or Linux and the Nature of the Firm. *112 Yale Law Journal 112*, 369-446.

Benkler, Y. (2007). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. (2011). Crowds in two seconds: enabling realtime crowd-powered interfaces. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11 (pp. 33–42). New York, NY, USA: ACM.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., et al. (2010). Soylent. *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10* (p. 313). New York, New York, USA.

Bigham, J. P., White, S., Yeh, T., Jayant, C., Ji, H., Little, G., Miller, A., et al. (2010). VizWiz. *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10* (p. 333). New York, New York, USA.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Boujarwah, F., Abowd, G., & Arriaga, R. (2012). Socially computed scripts to support social problem solving skills. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12 (pp. 1987–1996). New York, NY, USA.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*, 105–115.

Eckhardt, M., & Picard, R. (2009). A more effective way to label affective expressions. *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1–2).

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, *37*(2), 413–420.

Frey, B. S., & Jegen, R. (2001). Motivation Crowding Theory. *Journal of Economic Surveys*, Journal of Economic Surveys, *15*(5), 589–611.

J. J. Gross and O. P. John, "Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being," *J. Pers. Soc. Psychol.*, vol. 85, no. 2, pp. 348–362, Aug. 2003.

Harrigan, J., Rosenthal, R., & Scherer, K. (2005). *New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press.

Horton, J. J., & Chilton, L. B. (2010). The Labor Economics of Paid Crowdsourcing. *Proceedings of the 11th ACM Conference on Electronic Commerce*. Available at SSRN: http://ssrn.com/abstract=1596874.

Howe, J. (2006, June). The Rise of Crowdsourcing. *Wired, 14*(6).

Howe, J. (2009). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (unedited ed.). Crown Business.

Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing* (pp. 27–35).

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. New York University Working Paper.

Khatib, F., DiMaio, F., Group, F. C., Group, F. V. C., Cooper, S., Kazmierczyk, M., Gilski, M., et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology, 18*(10), 1175–1177.

Kim, Y. E., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. *Proceedings of the International Symposium on Music Information Retrieval* (pp. 231–236).

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08 (pp. 453–456). New York, NY, USA:

Lakhani, K. R., & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In J. Feller, S. Fitzgerald, S. Hissam, & K. Lakhani (Eds.), *Perspective on Free and Open Source Software*. Cambridge, MA: MIT Press.

Lakhani, K.R., Jeppesen, L.B., Lohse, P.A., & Panetta.J.A. (2006). *The Value of Openness in Scientific Problem Solving*. Harvard Business School Working Paper No. 07-050.

Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., & Bigham, J. P. (2011). Real-time crowd control of existing interfaces. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11 (pp. 23–32). New York, NY, USA: ACM.

Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2009). TurKit: tools for iterative tasks on mechanical Turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09 (pp. 29–30). New York, NY, USA: ACM.

Malone, T. (2009). Harnessing Crowds: Mapping the Genome of Collective Intelligence. *MIT Sloan Resaerch*.

Mason, W., & Watts, D. J. (2009). Financial incentives and the "performance of crowds." *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '09* (p. 77). Presented at the the ACM SIGKDD Workshop, Paris, France.

McDuff, D. J., Kaliouby, R. E., & Picard, R. W. (2012). Crowdsourcing Facial Responses to Online Videos. *Transactions on Affective Computing, In Press*.

Mohammad, S. M., & Turney, P. D. (2011). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *59*(000), 1–24.

Morency, L. P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: harvesting opinions from the web. *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169–176).

Morris, R. R., & Picard, R. (2012). Crowdsourcing Collective Emotional Intelligence. *Proceedings of Collective Intelligence*, Cambridge, MA.

Morton, B. G., Speck, J. A., Schmidt, E. M., & Kim, Y. E. (2010). Improving music emotion labeling using human computation. *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 45–48).

Pearl, L., & Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 71–79).

Riek, L. D., O'Connor, M. F., & Robinson, P. (2011). Guess What? A Game for Affective Annotation of Video Using Crowd Sourcing. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective Computing and Intelligent Interaction* (Vol. 6974, pp. 277–285). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://www.springerlink.com/content/a40471253r44t616/

Rogstadius, J., Kostakos, V., Aniket Kittur, Smus, B., Laredo, J., & Vukovic, M. (2011). An Assessment of Intrinsic and Extrinsic Motivation in Crowdsourcing Markets. *ICWSM11*. Presented at the Association for the Advancement of Artificial Intelligence (AAAI), Barcelona, Spain.

Soleymani, M., & Larson, M. (2010). Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)* (pp. 4–8).

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation.

Spiro, I. (2012). Motion chain: a webcam game for crowdsourcing gesture collection. *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts* (pp. 1345–1350). Retrieved from http://dl.acm.org/citation.cfm?id=2212452

Spiro, I., Taylor, G., Williams, G., & Bregler, C. (2010). Hands by hand: crowd-sourced motion tracking for gesture annotation. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* (pp. 17–24). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5543191

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Tarasov, A., Cullen, C., Delany, S. (2010). Using Crowdsourcing for labeling emotional speech assets. *W3c workshop on Emotion ML*, Paris, France.

Taylor, G. W., Spiro, I., Bregler, C., & Fergus, R. (2011). Learning invariance through imitation. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 2729–2736).

Turnbull, D., Liu, R., Barrington, L., & Lanckriet, G. (2007). A game-based approach for collecting semantic annotations of music. *8th International Conference on Music Information Retrieval (ISMIR)*.

von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94.

Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D., & Horvitz, E. (2012). Human computation tasks with global constraints. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12 (pp. 217–226). New York, NY, USA: ACM.