Diamonds From the Rough: Improving Drawing, Painting, and Singing via Crowdsourcing

Yotam Gingold
Departments of Computer Science
Columbia University & Rutgers University
yotam@yotamgingold.com

Etienne Vouga and **Eitan Grinspun**
Department of Computer Science
Columbia University
New York, NY
evouga,eitan@cs.columbia.edu

Haym Hirsh
Department of Computer Science
Rutgers University
Piscataway, NJ
hirsh@cs.rutgers.edu

Abstract

It is well established that in certain domains, noisy inputs can be reliably combined to obtain a better answer than any individual. It is now possible to consider the crowdsourcing of physical actions, commonly used for creative expressions such as drawing, shading, and singing. We provide algorithms for converting low-quality input obtained from the physical actions of a crowd into high-quality output. The inputs take the form of line drawings, shaded images, and songs. We investigate single-individual crowds (multiple inputs from a single human) and multiple-individual crowds.

Introduction

The *wisdom of crowds* (Surowiecki 2004) suggests that it can be advantageous to aggregate information from many "low-quality" sources rather than relying on information from a single "high-quality" source. There may be several advantages: it may be difficult or impossible to access a high-quality source; it may be cheaper to obtain information from many low-quality sources; perhaps most surprising, *aggregation may consistently produce higher-quality output*. Galton (1907) presented one of the earliest examples of this surprising result, when he calculated the median of a crowd's estimates of the weight of a bull and found it to be within 1% of the truth.

We propose to draw on the wisdom of crowds to produce a single higher-quality output from a set of lower-quality inputs. We consider the scenario where many individuals contribute a single input, as well as the scenario where a single individual contributes many inputs. We focus on *creative tasks* such as drawing, painting, and singing.

Our approach may be framed in terms of crowdsourcing and aggregation. Technology makes it possible to crowdsource physical actions, e.g., using a touch-screen or microphone. To harness this data, we must address the question of *how to meaningfully aggregate* creative works. Unlike many examples of the wisdom of crowds, our input and output data are more complex than a single number or a vote from among a small finite set of choices.

Yu and Nickerson (2011) employed genetic algorithms and tournament selection to iteratively aggregate and improve the quality of a set of drawings; the algorithm assumes that a human is able to combine the best aspects of two creative pieces. By contrast, we consider settings in which this assumption does not hold.

We treat the case of *inherently low-quality* (ILQ) input. We assume that the initial human input is "as good as can be expected" for the available input hardware and software, and for the skill, level of focus, and allotted time of participating humans.

ILQ input can arise from multiple trials by single individuals (Vul and Pashler 2008), such as when a person with limited fine motor coordination makes repeated attempts to draw, write, or sign their name; the limitation may be due to disease (e.g., Parkinson's) or simply due to the limited form factor of the input device (finger-writing on a small screen). In another variation, the input may be reasonable, but an even better output is desired, such as when an average person sings or draws, but wishes they could do so better.

ILQ input can also arise from single trials across multiple individuals. For example, can we produce a great painting, if the humans and tools at our disposition limit us to only mediocre paintings? Even when we have humans and tools capable of painting expertly, economic conditions might favor participation of multiple less-skilled participants. Under a tight deadline, there may not be sufficient time for an expert to produce a great piece, but there may be sufficient time for a multitude of participants to produce mediocre pieces, or ILQ.

To explore this setting, we consider crowdsourcing and aggregation to produce better drawings, paintings, and songs from ILQ. We first analyze "smiley faces" sketched many times by the same individuals, we then aggregate similar paintings created by many individuals, and finally we analyze the same song sung many times by the same individuals.

Related Work

Crowdsourcing has been applied to algorithms and data collection in a variety of domains, including databases (Franklin et al. 2011), natural language processing (Snow et al. 2008), song identification (Huq, Cartwright, and Pardo 2010), and computer vision.

The problem of aggregating input from many (human) sources has been studied in the literature. This includes collaborative filtering (Goldberg et al. 1992; Adomavicius and Tuzhilin 2005), in which the preferences of many individuals are aggregated to generate reviews and recommendations (Goldberg et al. 1992; Adomavicius and Tuzhilin 2005). In computer vision, several projects (von Ahn and Dabbish 2004; von Ahn, Liu, and Blum 2006; Sorokin and Forsyth 2008; Spiro et al. 2010) have collected redundant input from many humans in order to ensure high-quality image labels or video annotations. Notably, Law and von Ahn (2009) also collected data on music. Typically, these approaches either filter the human input to select one output, concatenate it, or, for low-dimensional input such as a scalar quantity or a direction, average it. Dow et al. (2012) discuss feedback mechanisms to improve the quality of crowd-sourced product reviews. Ipeirotis et al. (2010) estimate worker quality in classification tasks. Little et al. (2010) divide the process of writing image descriptions, brainstorming company names, and deciphering blurry text into creation and decision tasks. Karger et al. (2011) present an algorithm for efficiently assigning tasks to workers and obtaining reliable answers in a binary classification task.

In computer graphics, several works have collected large quantities of data with the goal of aggregating them to achieve a "ground truth" benchmark (Cole et al. 2008; Chen, Golovinskiy, and Funkhouser 2009). Gingold et al. (2012) aggregated input from many users in order to enable image editing tasks.

Rohwer (2010) considered the question of aggregation for the creation of creative works in the context of fiction writing. He reported on unsuccessful attempts of a crowd to *self*-organize a fiction novel via wiki, contrasted with a successful process whereby an editor iteratively selected the next sentence among twitter-submitted candidates. In the editorial process, individual contributions were retained or discarded in whole, and those retained were concatenated.

Drawing

In this section, we study the question of whether multiple line drawings of the same object average to a better drawing. Note that two line drawings of the same object may contain a different number and arrangement of strokes. Finding a correspondence between two line drawings' strokes is an extremely challenging problem, unsolved in the general case.

Related work

The photographic average of human faces was first examined by Galton (1878), who commented on the effect of averaging but did not empirically evaluate attractiveness. More recently, Langlois and Roggman (1990) were the first to empirically evaluate averages of human faces; they report that average faces composed of 16 or more individuals were rated as more attractive than all but a few ($\approx \%5$) of the individual faces. While photographic averages of human faces smooth away blemishes and asymmetries, line drawings of general objects depict only what their creator chose to include, so we cannot assume that there are undesired blemishes or asymmetries to be smoothed away.

"The Sheep Market" (Koblin 2008) collected drawings of sheep from 1000 individuals on Amazon Mechanical Turk, though no aggregation or analysis was performed.

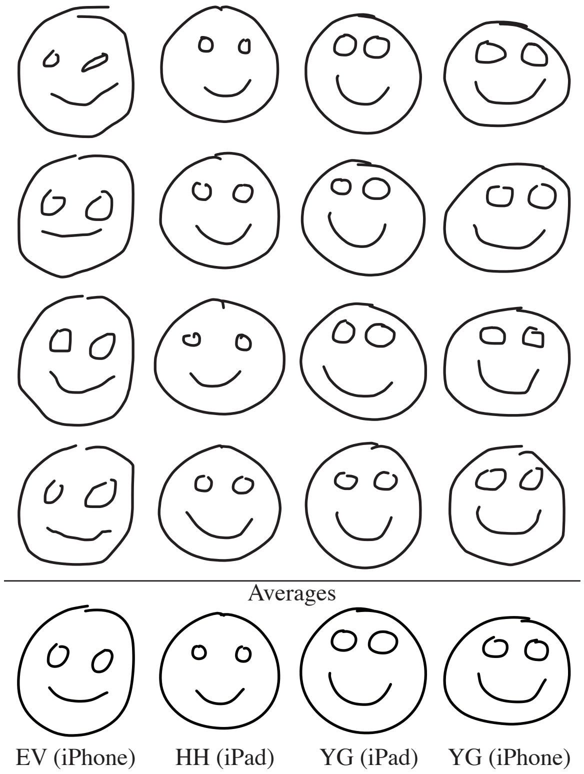


Figure 1: In each column: four of the 20 smiley faces drawn by a participant, as well as the average of all 20.

Cole et al. (2008) asked art students to draw line drawings of 3D models and used the data to evaluate computational line drawing algorithms. In the course of the evaluation, pixel-wise average images were created of the artists' line drawings. These do not depict averages of the drawings' individual lines; rather, they depict all drawings' lines together in one image.

Protocol

To sidestep challenging correspondence problems, we focus on simple "smiley faces" composed of four strokes: a head, two eyes, and a mouth. Three subjects, EV, HH, and YG, each drew a collection of 20 smiley faces using a vector-based drawing application for the iPhone (EV, YG) and iPad (HH, YG). Several input smiley faces are shown in Figure 1.

Averaging

To average a collection of smiley faces, we first resample all strokes at 100 evenly spaced locations, and then average the Cartesian coordinates of corresponding points along each curve. This produces the smiley faces shown in Figure 1, bottom row, and in Figure 2.

Evaluation

To evaluate the attractiveness of smiley faces, we conducted surveys asking evaluators to "Choose the most beautiful image" from among a gallery of smiley faces. The order of

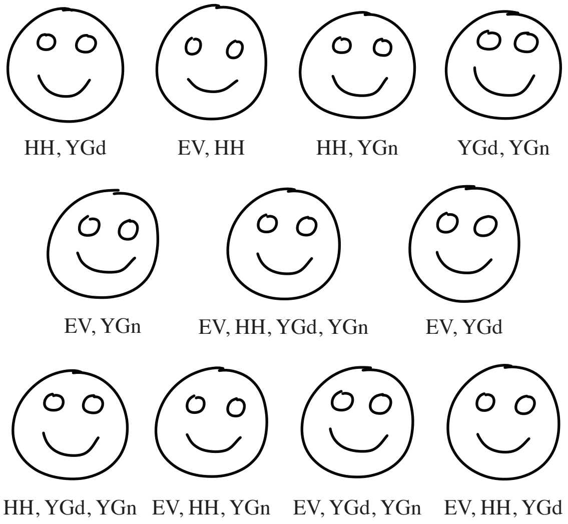


Figure 2: Average smiley faces over all possible multiple-subject combinations of EV-iPhone (EV), HH-iPad (HH), YG-iPad (YGd), and YG-iPhone (YGn).

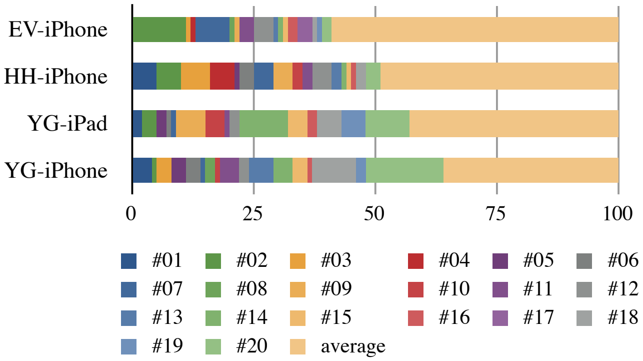


Figure 3: Vote share of each subject's individual smiley faces versus the average.

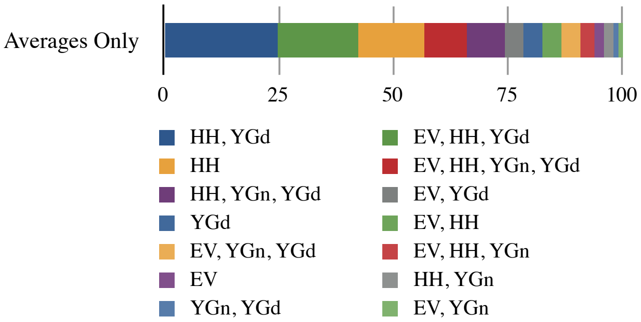


Figure 4: Vote share among average faces over all possible combinations of EV-iPhone (EV), HH-iPad (HH), YG-iPad (YGd), and YG-iPhone (YGn).

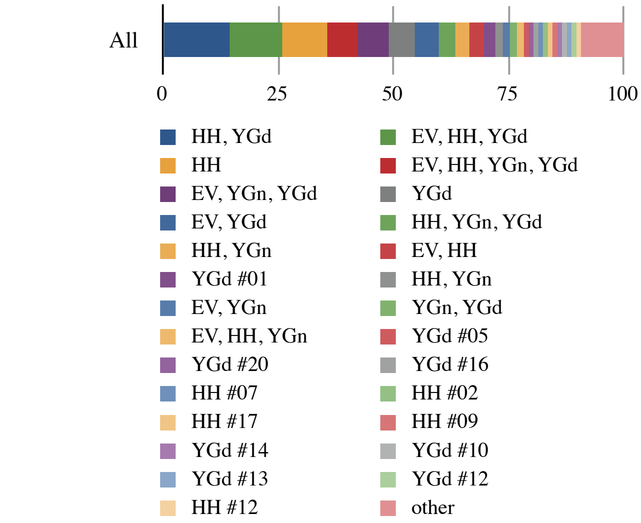


Figure 5: Vote share among all smiley faces from the experimental conditions.

smiley faces in the gallery was randomized across subjects. Experimental conditions (galleries) were: HH-iPad smiley faces and their average; EV-iPhone smiley faces and their average; YG-iPad smiley faces and the average; YG-iPhone smiley faces and the average; average faces over all possible combinations of EV-iPhone, HH-iPad, YG-iPad, and YG-iPhone; all smiley faces from the other experimental conditions ("all"). 100 evaluators were drawn from Amazon Mechanical Turk for each experimental condition, except for the "all" experiment, where 200 evaluators were used.

Discussion

In all experimental conditions, the most popular smiley faces were the ones computed by averaging (Figures 3 and 5). All results were statistically significant (EV-iPhone $\chi^2 = 648.66$, $p < 0.001$; HH-iPad $\chi^2 = 431.52$, $p < 0.001$; YG-iPad $\chi^2 = 322.40$, $p < 0.001$; YG-iPhone $\chi^2 = 215.17$, $p < 0.001$; "all" $\chi^2 = 534.03$, $p < 0.001$).

In the "all" experimental condition, we cannot say with confidence that the set of averages of multiple subjects' smiley faces performed better than the set of averages of single subjects' smiley faces ($\chi^2 = 2.48$, $p = 0.116$).

Averaging smoothes away noise and jitter from individual smiley faces. And while multiple-subject averages smooth away subjects' individual styles (Figure 4), averaging a single individual's smiley faces appears to preserves stylistic attributes, such as the elliptical shape and non-closedness of EV smiley faces' heads (Figure 1). Interestingly, HH smiley faces were quite popular (Figures 5); they were present in the smiley faces that received a combined 64% of the votes in the "all" experimental condition ($\chi^2 = 101.99$, $p < 0.001$), and the average of the HH smiley faces alone received 10% of those votes ($\chi^2 = 142.32$, $p < 0.001$).

Finally, because each subject drew 20 smiley faces, we investigated whether the repeated drawing itself led to an aesthetic improvement in the resulting smiley faces (a training bias). Figure 6 plots the fraction of votes received by the first 10 smiley faces drawn by each subject versus the fraction of votes received by the second 10 smiley faces.

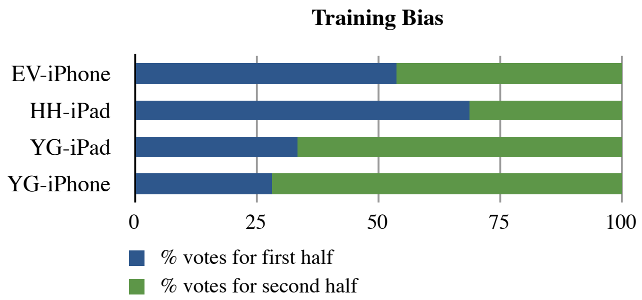


Figure 6: The share of votes received by the first half of an individual's drawn smiley faces versus the share of votes received by the second half.

We found that while YG exhibits a training bias (YG-iPad $\chi^2 = 6.33$, $p = 0.012$; YG-iPhone $\chi^2 = 12.25$, $p < 0.001$), HH exhibits a reverse training bias ($\chi^2 = 7.08$, $p = 0.008$), and EV does not exhibit any training bias ($\chi^2 = 0.22$, $p = 0.639$). Thus, we conclude that an individual cannot, in general, obtain a smiley face of comparable aesthetic quality to an average by training.

Shading

A naive human is physically able to apply paint to a canvas, yet, without practice, is unlikely to paint a pleasing portrait. In this section, we address the question of whether paintings of the same object created by multiple naive humans can be composited to create a better painting. (Specifically, we focus on greyscale paintings, which is perhaps more similar to drawing with charcoal than oil painting.)

Related Work

The photographic averaging of human faces (Galton 1878; Langlois and Roggman 1990) is more closely related to averaging paintings than drawings (previous section). In contrast to the domain of faces, where irregularities and blemishes are asymmetric, it is not *a priori* obvious that averaging paintings will produce better paintings.

The previously mentioned work of Cole et al. (2008) composited line drawings created by many skilled humans in a pixel-wise fashion. Neither the inputs nor the composited output resemble painting.

In "Ten Thousand Cents" (Koblin and Kawashima 2008), the image of a US dollar bill was divided into ten thousand squares and shown to individuals on Amazon Mechanical Turk, who were asked to digitally paint their own interpretation of the square. Each painting was arranged in a quilt-like fashion; no averaging or compositing was performed.

Protocol

A pool of 50 subjects were recruited using Amazon Mechanical Turk. Subjects accessed a web page which displayed a photograph of a still life (a pear). Subject were also given a canvas containing the outline of the pear and asked to "paint the object from the photograph into the canvas" using a paint brush tool with adjustable brush diameter and grey level (Figure 7, left). By initializing the canvas with the outline of the pear, we hoped to avoid the need to register subjects' paintings during analysis.

Aggregation and Discussion

A representative selection of paintings created by the subjects is shown in Figure 7, middle. 23 of the 50 subjects filled the entire pear with a single shade of grey (13 chose black). All but one subject generally adhered to the outline of the pear.

Assuming that subjects' paintings are already registered (due to the outline of the pear), it is natural to apply pixel-wise aggregation operations. The pixel-wise average and the pixel-wise median can be seen in Figure 7, right. As with drawing, aggregation has produced a result that is clearly superior to any of the inputs. The average is perhaps overly smoothed and produces paint outside the outline of the pear. The median is higher-contrast and has no such painting-outside-the-line artifacts.

Singing

The average person does not sing perfectly on key, but a chorus of such people can sound pleasing even when an individual solo would not. Singing is thus another domain where we might expect to produce higher-quality output from many low-quality inputs by applying some kind of averaging—in particular, by averaging base frequencies.

Related Work

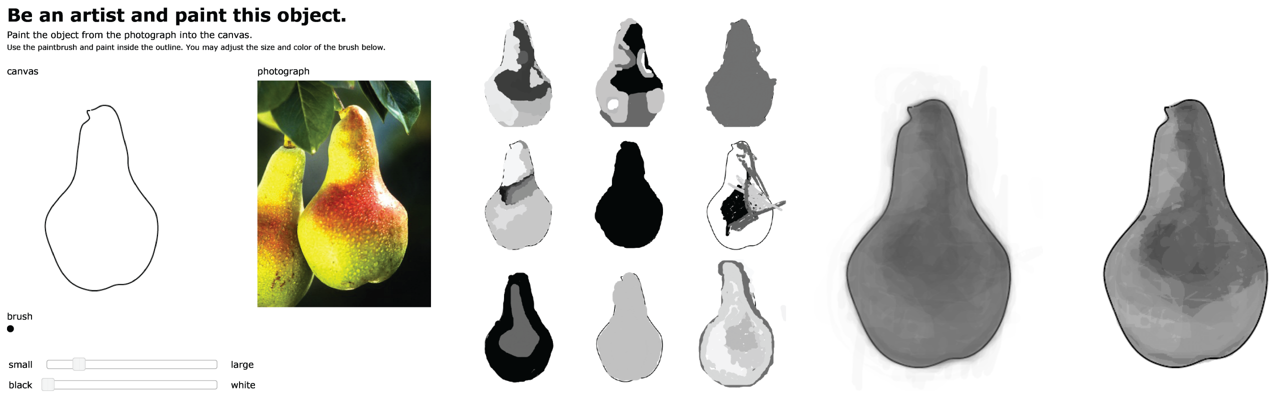
"Bicycle Built for Two Thousand" (Koblin and Massey 2009) collected 2088 recordings of humans imitating unrecognizably small pieces of the song "Daisy Bell" via Amazon Mechanical Turk. The result is a chorus (typically 36 humans at once), rather than an aggregate that resembles a single human singing.

Methodology

Each subject recorded himself singing "Happy Birthday To You" ten times, while simultaneously listening to a MIDI rendition of the song on headphones so that the tempo was the same accross recordings. From each recording, we extracted the $F0$ pitch frequency every 0.01 seconds using the software package *Praat* (Boersma and Weenink 2012). By inspecting the MIDI, we also determined the ground truth frequencies of each of the song's notes (which correspond to the song being played in F major). From this data we can compare how close any individual recording is to being on tune (see Figure 8, top-left).

To find the average frequency, we compute the geometric mean of each of the ten frequencies at each time sample, ignoring recordings for which *Praat* was unable to find an $F0$ frequency at that time. These pitches, for the set of recordings by subject EG, are plotted against the true pitches in Figure 8, top-middle. We also generated the average of all thirty recordings by all three subjects (Figure 8, top-right).

For each subject, we arbitrarily chose one recording to pitch-shift using the averaged frequencies. For this recording, we computed ratios $r_t = a_t/f_t$ at each time sample t, where a_t is the averaged $F0$ frequency at time t



Be an artist and paint this object.

Paint the object from the photograph into the canvas.
Use the paintbrush and paint inside the outline. You may adjust the size and color of the brush below.

Figure 7: The Amazon Mechanical Turk interface for shading *(left)*, and nine randomly selected generated shadings *(middle)*. The average *(near right)* and median *(far right)* of all of the shadings are visually pleasing despite the low quality of any individual shading.

and f_t is the recording's original $F0$ frequency. We then took the short-time Fourier transform of the recording using non-overlapping rectangular windows of size 0.01 seconds (equal to the $F0$ frequency sample rate), scaled all frequencies by r_t, and took the inverse transform to produce a retargeted recording that is more on-key. The new recording does contain some chirping artifacts, particularly at phrase transitions where the input recordings do not align temporally; we hope to address these artifacts in the future, perhaps by incorporating matching of the recordings in time (Dixon and Widmer 2005).

We quantify the improvement to the pitch gained by averaging input recordings as follows. We compute, for each note i of the song, the root mean squared error E of frequency:

$$E_i = \sqrt{\frac{1}{N}\sum(f_t - f_i)^2},$$

where the sum is taken over all $F0$ frequency samples f_t whose times fall within the duration of the note (actually, the middle third of this duration—shown in red), N is the total number of such samples, and f_i is the frequency of the note. We also calculated the root mean squared frequency error using the averaged frequencies a_t instead of f_t (Figure 8, bottom).

Discussion

Interestingly, across all three subjects and for most of the notes, the averaged frequency is as close or closer to being in tune than even the best individual recording. In other words, the same singer singing the same note tended to be flat about as often as sharp, instead of singing systematically off-key in the same direction.

Conclusion

The Internet has made it easier than ever to quickly and efficiently marshal a crowd, and to assign them simple, creative, physical actions like drawing, painting, or singing. We have shown ways to harness a crowd as a crucible for refining inherently low-quality input into higher-quality output. Moreover, we have shown that a single individual is capable of outperforming themself by generating a crowd's worth of data.

In all of our examples, the registration of inputs was crucial. Averaging drawn strokes works when the number and placement of strokes is consistent; this is the case for simple smiley faces, but not the case in general. Multiple line drawings of, for example, an apple are likely to be composed of different numbers and placements of strokes. We have investigated averaging line drawings at the pixel level, but it is difficult to output strokes from pixel-wise averaged drawings.

In our singing experiments, subjects sang "karaoke" while listening to the song on headphones, so that all recordings were more or less on tempo. We hope to explore whether more sophisticated notions of averaging music might yield pleasing results even in the presence of misalignment in time or systematic pitch bias. One possible approach to finding a mapping between pairs of recordings is by performing non-rigid image registration between their spectrograms; the "average deformation" can then be computed.

In the future, we hope to find more powerful and sophisticated averaging schemes capable of refining even more complex inputs: sketches of complex objects, drawn using arbitrary strokes; paintings of entire scenes, in color; and songs sung by several different people, at different tempos.

Acknowledgments

We are grateful to Tom Malone, Greg Little, Alex Rosmarin, and Julia Hirschberg for helpful discussions, and to Lara Kassoff for help with statistics. We also wish to acknowledge Dan Weld for recommending the book "Paths to Otherwhere" by James P. Hogan, which partially inspired this work. This research is supported in part by the Sloan Foundation, the NSF (CAREER Award CCF-06-43268 and grants IIS-09-16129, IIS-10-48948, IIS-11-17257, CMMI-11-29917, IIS-09-16845), and generous gifts from Adobe, Autodesk, Intel, mental images, Microsoft, NVIDIA, Side Effects Software, and the Walt Disney Company.

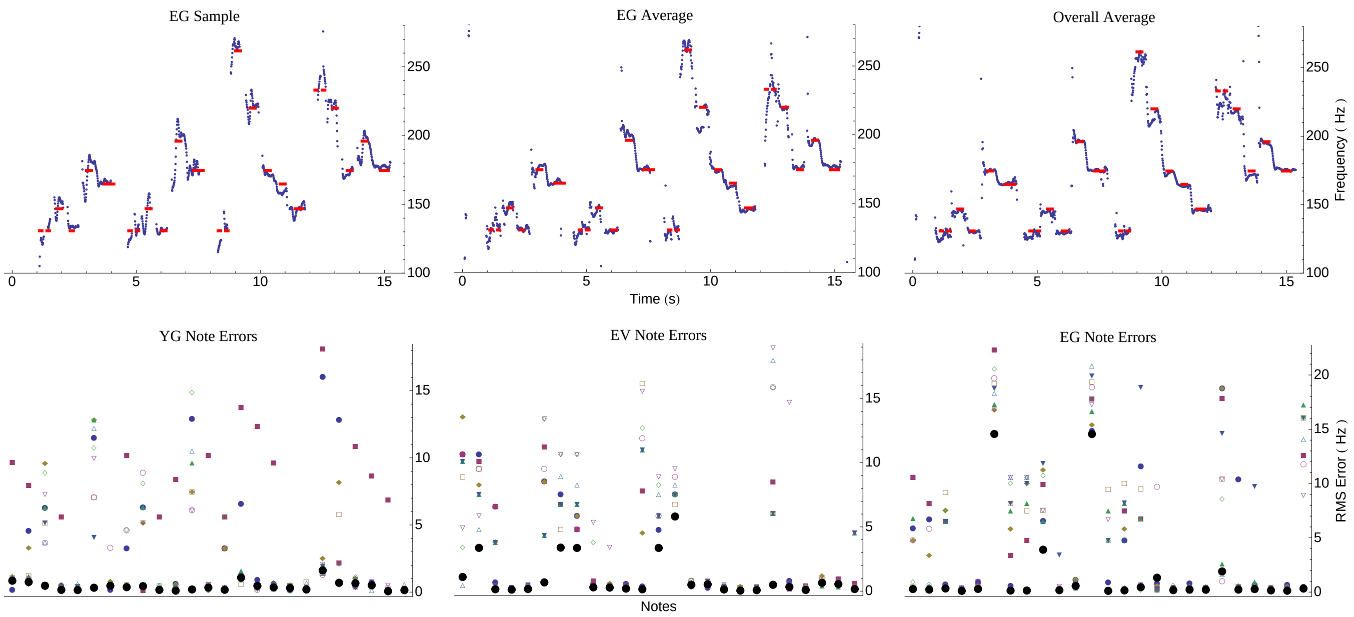


Figure 8: $F0$ frequency of subject EG singing “Happy Birthday To You” *(top-left)*, with ground truth pitches marked in red. Taking the geometric mean of these frequencies and the $F0$ frequencies of nine other recording by the same subject yields pitches that are closer to being on key *(top-middle)*; also including 10 recordings by each of two additional subjects further improves the accuracy of the average *(top-right)*. We quantitatively measure the improvement by plotting the root mean square frequency error for each note *(bottom)*. For each subject, different colors represents different recordings. The RMS error of the averaged frequencies is plotted in black. Notice that the average pitch is often better than the best individual pitch for that note, and, overall, the averaged pitches are closer to on-tune than any individual recording (color).

References

Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Trans. on Knowledge and Data Engineering* 17:734–749.

Boersma, P., and Weenink, D. 2012. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

Chen, X.; Golovinskiy, A.; and Funkhouser, T. 2009. A benchmark for 3D mesh segmentation. *ACM Trans. Graph.* 28(3).

Cole, F.; Golovinskiy, A.; Limpaecher, A.; Barros, H. S.; Finkelstein, A.; Funkhouser, T.; and Rusinkiewicz, S. 2008. Where do people draw lines? *ACM Transactions on Graphics (Proc. SIGGRAPH)* 27(3).

Dixon, S., and Widmer, G. 2005. MATCH: A music alignment tool chest. *ISMIR 2005, 6th International Conference on Music Information Retrieval.*

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *CSCW ’12: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM.

Franklin, M. J.; Kossmann, D.; Kraska, T.; Ramesh, S.; and Xin, R. 2011. CrowdDB: answering queries with crowdsourcing. In *Proceedings of the 2011 international conference on Management of data*, SIGMOD ’11, 61–72. New York, NY, USA: ACM.

Galton, F. 1878. Composite portraits. *Journal of the Anthropological Institute of Great Britain & Ireland* 8:132–142.

Galton, F. 1907. Vox populi. *Nature* 75:450–451.

Gingold, Y.; Shamir, A.; and Cohen-Or, D. 2012. Micro perceptual human computation. *ACM Transactions on Graphics (TOG)*.

Goldberg, D.; Nichols, D.; Oki, B. M.; and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35:61–70.

Huq, A.; Cartwright, M.; and Pardo, B. 2010. Crowdsourcing a real-world on-line query by humming system. *Proceedings of the 7th Sound and Music Computing Conference (SMC 2010)*.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, 64–67. New York, New York, USA: ACM Press.

Karger, D.; Oh, S.; and Shah, D. 2011. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, 284 –291.

Koblin, A., and Kawashima, T. 2008. Ten thousand cents. http://www.tenthousandcents.com/.

Koblin, A., and Massey, D. 2009. Bicycle built for two thousand. http://www.bicyclebuiltfortwothousand.com/.

Koblin, A. 2008. The sheep market. http://www.tenthousandcents.com/.

Langlois, J. H., and Roggman, L. A. 1990. Attractive faces are only average. *Psychological Science* 1(2):115–121.

Law, E., and von Ahn, L. 2009. Input-agreement: A new mechanism for data collection using human computation games. In *Proceedings of ACM SIGCHI*, 1197–1206.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*.

Rohwer, P. 2010. A note on human computation limits. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, 38–40. New York, NY, USA: ACM.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 254–263. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. *Proceedings of IEEE CVPR* 0:1–8.

Spiro, I.; Taylor, G.; Williams, G.; and Bregler, C. 2010. Hands by hand: Crowd-sourced motion tracking for gesture annotation. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 17–24.

Surowiecki, J. 2004. *The wisdom of crowds*. New York, NY: Random House.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of ACM SIGCHI*, 319–326.

von Ahn, L.; Liu, R.; and Blum, M. 2006. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, 55–64.

Vul, E., and Pashler, H. 2008. Measuring the crowd within. *Psychological Science* 19(7):645–647.

Yu, L., and Nickerson, J. 2011. Cooks or cobblers?: crowd creativity through combination. *Proceedings of the 2011 annual conference on Human factors in computing systems* 1393–1402.