Utility data annotation with Amazon Mechanical Turk

Alexander Sorokin, David Forsyth
University of Illinois at Urbana-Champaign
201 N Goodwin
Urbana,IL 61820
{sorokin2,daf}@uiuc.edu

Abstract

We show how to outsource data annotation to Amazon Mechanical Turk. Doing so has produced annotations in quite large numbers relatively cheaply. The quality is good, and can be checked and controlled. Annotations are produced quickly. We describe results for several different annotation problems. We describe some strategies for determining when the task is well specified and properly priced.

1. Introduction

Big annotated image datasets now play an important role in Computer Vision research. Many of them were built **in-house** ([18, 11, 12, 3, 13, 5] and many others). This consumes significant amounts of highly skilled labor, requires much management work, is expensive and creates a perception that annotation is difficult. Another successful strategy is to make the annotation process **completely public** ([24]) and even entertaining [26, 27]), at the cost of diminished control over what annotations are produced and necessary centralization to achieve high volume of participation. Finally, **dedicated annotation services** ([28]) can produce high volume quality annotations, but at high price.

We show that image annotation work can be efficiently outsourced to an online worker community (currently Amazon Mechanical Turk [2]) (sec. 2). The resulting annotations are good (sec. 2.3.2), cheap (sec. 2.3.1) and can be aimed at specific research issues.

2. How to do it

Each annotation task is converted into a Human Intelligence Task (HIT). The tasks are submitted to Amazon Mechanical Turk (MT). Online workers choose to work on the submitted tasks. Every worker opens our web page with a HIT and does what we ask them to do. They "submit" the result to Amazon. We then fetch all results from Amazon MT and convert them into annotations. The core tasks for a researcher are: (1) define an annotation protocol and (2) determine what data needs to be annotated.

Exp	Task	img	labels	cost USD	time	effective pay/hr
1	1	170	510	$8	750m	$0.76
2	2	170	510	$8	380m	$0.77
3	3	305	915	$14	950m	$0.41[1]
4	4	305	915	$14	150m	$1.07
5	4	337	1011	$15	170m	$0.9
Total:		982	3861	$59		

Table 1. **Collected data.** In our five experiments we have collected **3861** labels for 982 distinct images for only **US $59**. In experiments 4 and 5 the throughput exceeds 300 annotations per hour even at low ($1/hour) hourly rate. We expect further increase in throughput as we increase the pay to effective market rate.

The annotation protocol should be implemented within an IFRAME of a web browser. We call the implementation of a protocol an **annotation module**. The most common implementation choices will be HTML/JS interface, Java or Flash applet. The annotation module must be developed for every radically new annotation protocol. We have already built 4 different annotation modules(in Flash) for labeling images of people. As the design process is quite straightforward, we aim to **accomodate requests to build** annotation modules for various research projects.

Our architecture requires very little resources administered by the researcher (bash, python, Matlab and a web server or Amazon S3).

2.1. Quality assurance

There are three distinct aspects of quality assurance: (a) Ensuring that the workers understand the requested task and try to perform it well; (b) cleaning up occasional errors; (c) detecting and preventing cheating in the system. We discuss three viable strategies for QA: multiple annotations, grading

[1]This number includes around 30% of poor annotations.

and gold standard evaluation (with immediate feedback).

The basic strategy is to **collect multiple annotations** for every image. This will account for natural variability of human performance, reduce the influence of occasional errors and allow us to catch malicious users. However, this increases the cost of annotation.

The second strategy is to perform a separate **grading task**. A worker looks at several annotated images and scores every annotation. We get explicit quality assesments at a fraction of the cost, because grading is easy.

The third strategy is to build a **gold standard** - a collection of images with trusted annotations. Images from the gold standard are injected into the annotation process. The worker doesn't know if an image comes from the new data or from the gold standard. If the annotations provided by the worker significantly deviate from the gold standard, we suspect that the worker is not doing what we asked for. We reveal the gold standard annotation to the worker after they sumbit their own annotation. This immediate feedback clarifies what we expect and encourages to follow the protocol. This strategy is again cheap, as only a fraction of images comes from the gold standard.

It is most important to ensure that contributors with high impact understand the task and follow the requested protocol. As can be seen in fig 2, the bulk of annotation is produced by a few contributors. In our experiments we collected multiple annotations to study consistency. In only one experiment did we have a significant contributor providing poor annotations (Fig 2, experiment 3, see the low times among the first contributors. See also figure 5 experiment 3, example "G", yellow curve).

2.2. Annotation protocols

We implemented four annotation protocols (fig 1): two coarse object segmentation protocols, polygonal labeling and 14-point human landmark labeling. Object segmentation protocols show an image to the worker and a small image of the query (person). We ask the worker to click on every circle (site) overlapping with the query (person). Protocol one places sites on a **regular grid**, whereas protocol two places sites at the **centers of superpixels** (computed with [19, 17]).

The third protocol, **polygonal labeling**, is very similar to the one adopted in LabelMe[24]. We ask the worker to trace the boundary of the person in the image.

The fourth protocol labels the landmarks of the human body used for pose annotation in [23]. We ask the worker to click on locations of the **14 points** in the specified order: right ankle, right knee, right hip, left hip, left knee, left ankle, right wrist, right elbow, right shoulder, left shoulder, left elbow, left wrist, neck and head. The worker is always reminded what the next landmark is.

2.3. Annotation results

So far we have run five annotation experiments using data collected from Youtube (experiments 1, 2, 5), the dataset of people from [23] (exp. 3, 4) and small sample of data from LabelMe[24], Weizman [6] and our own dataset (exp. 5). In all experiments we are interested in people. As shown in table 1 we have a total of **3861** annotations for 982 distinct images collected for a total cost of **US$ 59**. This is very cheap as discussed in section 2.3.1. We describe the quality of annotations in section 2.3.2.

We present sample annotation results (fig 1,4,5) to show the representative annotations and highlight the most prominent failures. We are extremely satisfied with the quality of the annotations taking into account that workers receive no feedback from us. We are currently implementing QA strategies described above to provide feedback to workers so we can stop using the multiple duplicate annotations strategy.

2.3.1 Pricing

The work throughput is elastic and depends on the price of the task. If the price is too low, workers will participate out of curiosity and for entertainment, but may feel underpaid and will loose motivation. If the price is too high, we could be wasting resources and possibly attracting inefficient workers. As table 1 shows, the hourly pay in experiments 4 and 5 was roughly \$1/hour. In these experiments we had a comments field and some comments suggested that the pay should be increased by a factor of 3. From this we conclude that the perceived fair pricing is about **US \$3/hour**. The fact that our experiments 1-5 finished completely shows the elasticity of the workforce. We note that even at US \$1/hour we had a high throughput of 300 annotations per hour.

2.3.2 Annotation quality

To understand the quality of annotations we use three simple consistency scores for a pair of annotations ($a1$ and $a2$) of the same type. For protocols 1,2 and 3 we divide the area where annotations disagree by the area marked by any of the two annotations. We can think about this as XOR($a1$,$a2$)/OR($a1$,$a2$). For protocols 1 and 2 XOR counts of sites with the different annotations, OR counts the sites marked by any of the two annotations $a1$ and $a2$. For protocol 3, XOR is the area of the symmetric difference and OR is the area of the union. For protocol 4 we measure the average distance between the selected landmark locations. Ideally, the locations coincide and the score is 0.

We then select the two best annotations for every image by simply taking a pair with the lowest score, i.e. we take the most consistent pair of annotations. For protocol 3 we

Figure 1. **Example results** show the example results obtained from the annotation experiments. The first column is the implementation of the protocol, the second column show obtained results, the third column shows some poor annotations we observed. The user interfaces are similar, simple and are easy to implement. The total cost of annotating the images shown in this figure was **US $0.66**.

further assume that the polygon with more vertices is a better annotation and we put it first in the pair. The distribution of scores and a detailed analysis appears in figures 4,5. We show all scores ordered from the best (lowest) on the left to the worst (highest) on the right. We select 5:15:95[2] percentiles of quality and show the respective annotations.

Looking at the images we see that the workers mostly try to accomplish the task. Some of the errors come from sloppy annotations (especially in the heavily underpaid experiment 3 - polygonal labeling). Most of the disagreements come from difficult cases, when the question we ask is dif-

[2] 5 through 95 with step 15

ficult to answer. Consider figure 5, experiment 2, sample "G", leftmost circle. One annotator decided to mark the bat, while the other decided not to. This is not the fault of the annotators, but is rather a sign for us to give better instructions. The situation is even more difficult in experiment 4, where we ask to label landmarks that are not immediately visible. In figure 6 we show consistency of the annotations of each landmark between the 35th and the 65th percentile of figure 5. It is obvious from this figure that hips are much more difficult to localize compared to shoulders, knees, elbows, wrists, ankles, the head and the neck.

3. Related work

Crisp understanding of the purpose of annotated data is crucial. When it is clear what annotations should be made, quite large annotated datasets appear [16, 15, 4, 22, 25, 18]. Such datasets last for a long time and allow for significant advances in methods and theories. For object recognition, there isn't really a consensus on what should be annotated and what annotations are required, so we have a large number of competing datasets.

To build large scale datasets researchers have made people label images for free. **LabelMe**[24] is a public online image annotation tool. LabelMe has over 11845 images and 18524 video frames with at least one object labeled [24]. The current web site counter displays 222970 labeled objects. The annotation process is simple and intuitive; users can browse existing annotations to get the idea of what kind of annotations are required. The dataset is freely available for download and comes with handy Matlab toolbox to browse and search the dataset. The dataset is semi-centralized. MIT maintains a publicly-accessible repository, they accept images to be added to the dataset and they distribute the source code to allow interested parties to set up a similar repository. To our knowledge this is the most open project. On the other hand LabelMe has no explicit annotation tasks and annotation batches. The progress can only be measured in the number of images annotated. In contrast we aim at annotating project-specific data in well-defined batches. We also minimized the need for maintenance of a centralized database. An annotation project can run with only researcher's laptop and computing utility services easily accessible online.

The **ESP game** [26] and **Peekaboom** [27] are interactive games that collect image annotations by entertaining people. The players cooperate by providing textual and location information that is likely to describe the content of the image to the partner. The games are great success. They are known to have produced over 37 million [8] and 1 million [27] annotations respectively. The Peekaboom project recently released a collection of 57797 images annotated through gameplay. The game-based approach has two inconveniences. The first is centralization. To achieve proper scale, it is necessary to have a well-attended game service that features the game. This constrains publishing of a new game to obtain project-specific annotations. The second one is the game itself. To achieve reasonable scale one has to design a game. The game should be entertaining or else nobody will play it. This will require creativity and experimentation to create appropriate annotation interface. In contrast, our model serves as a drop-in, minimum effort, utility annotation.

Building in-house datasets was another common strategy. The most prominent examples here include: Berkeley segmentation dataset [18], Caltech 5/101 [11]/256 [12], Pascal VOC datasets [10, 9], UIUC car dataset [1], MIT [20] and INRIA [7] pedestrian datasets, Yale face dataset [4], FERET [22], CMU PIE [25] and (Labeled [13]) Faces in the Wild [5]. Every dataset above is a focused data collection targeted at a specific research problem: segmentation, car detection, pedestrian detection, face detection and recognition, object category recognition. The datasets are relatively small compared to those produced by large scale annotation projects.

Finally, dedicated annotation services can provide quality and scale, but at a high price. **ImageParsing.com** has built one of the world largest annotated datasets[28]. With over 49357 images, 587391 video frames and 3,927,130 annotated physical objects [28] this is a really invaluable resource for vision scientists. At the same time, the cost of entry is steep. Obtaining standard data would require at least US $1000 investment and custom annotations would require at least US $5000 [14]. In contrast our model will produce a 1000 images with custom annotations for under US $40. ImageParsing.com provides high quality annotations and has a large number of images available for free. It is important to note that [28] presents probably the most rigorous and the most varied definition of the image labeling task. Their definitions might not fit every single research project, but we argue that this degree of rigor must be embraced and adopted by all researchers.

4. Discussion

We presented a data annotation framework to obtain project-specific annotations very quickly on a large scale. It is important to turn annotation process into a utility, because this will make the researchers answer the important research issues: "**What data** to annotate?" and "What **type of annotations** to use?". As annotation happens quickly, cheaply and with minimum participation of the researchers, we can allow for multiple runs of annotation to iteratively refine the precise definition of annotation protocols. Finally, we shall ask "What happens when we get 1/10/100 million annotated images?".

We plan to implement more annotation protocols ([18, 3, 28, 9, 21], other **suggestions are welcome**) and the qual-

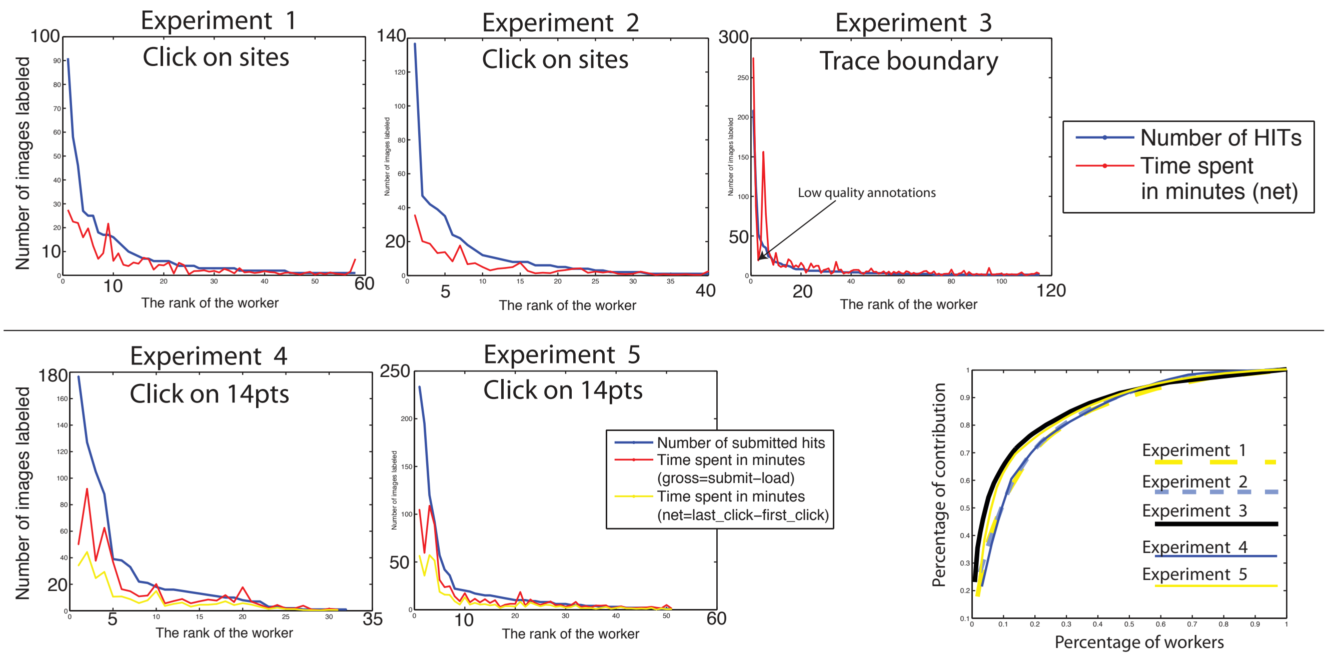

Figure 2. **Contributions.** The first five graphs plot the contribution and the time spent against the rank of the worker. The rank is determined by the total amount of the contribution by a particular worker. The lower the rank the higher the contributions. Note that the scales differ from experiment to experiment, because of different complexity of the tasks. The sixth graph plots the total contribution against the percentage of the top workers. It is really astonishing how closely the curves follow each other. These graphs give insight into the job distribution among the workers: (1) single top contributors produce very significant amounts spending hours on the task (2) top contributors are very effective in performing the tasks and (3) top 20% of annotators produce 70% of the data.

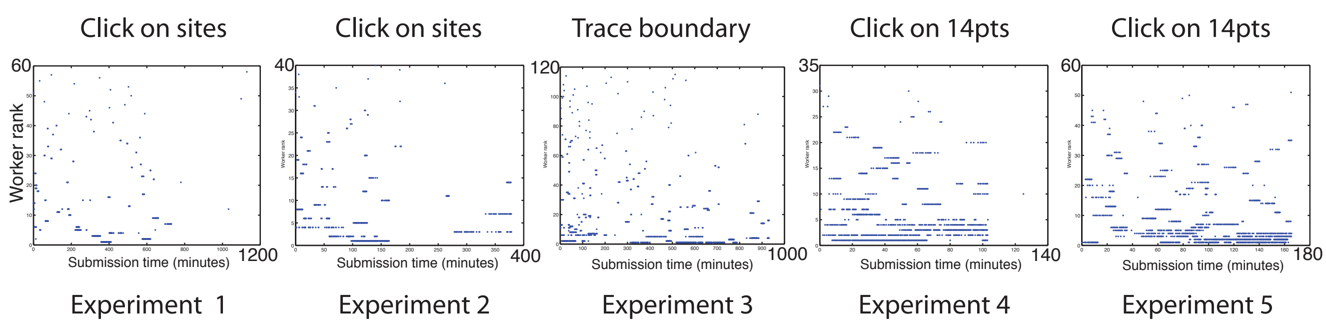

Figure 3. **Temporal structure of annotations.** We show a scatterplot of all submitted annotations. The horizontal axis is time in minutes when we receive the annotation. The vertical axis is the rank of the worker who produced the annotation. The bottom lines have many dots, as they show when the most significant contributors participated in the annotation process. Note the different scales of the scatterplots. The horizonal scale reflects the total time of the annotation while the vertical scale reflects the total number of people who participated in the annotation. The plots show how interesting the tasks are to the workers. In experiments 4 and 5 the workers start early and participate until the available tasks are exhausted - the dots all end at the same time, when no more tasks are left. In experiments 1,2 and 3 it takes much longer for significant annotators to come. This is a direct consequence of the task pricing (sec 2.3.1). Experiments 1 and 2 pay 30% less than experiments 4 and 5, while experiment 3 pays 50% less.

ity assurance strategies we discussed. We will make all the code and data available online.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, November 2004. 4

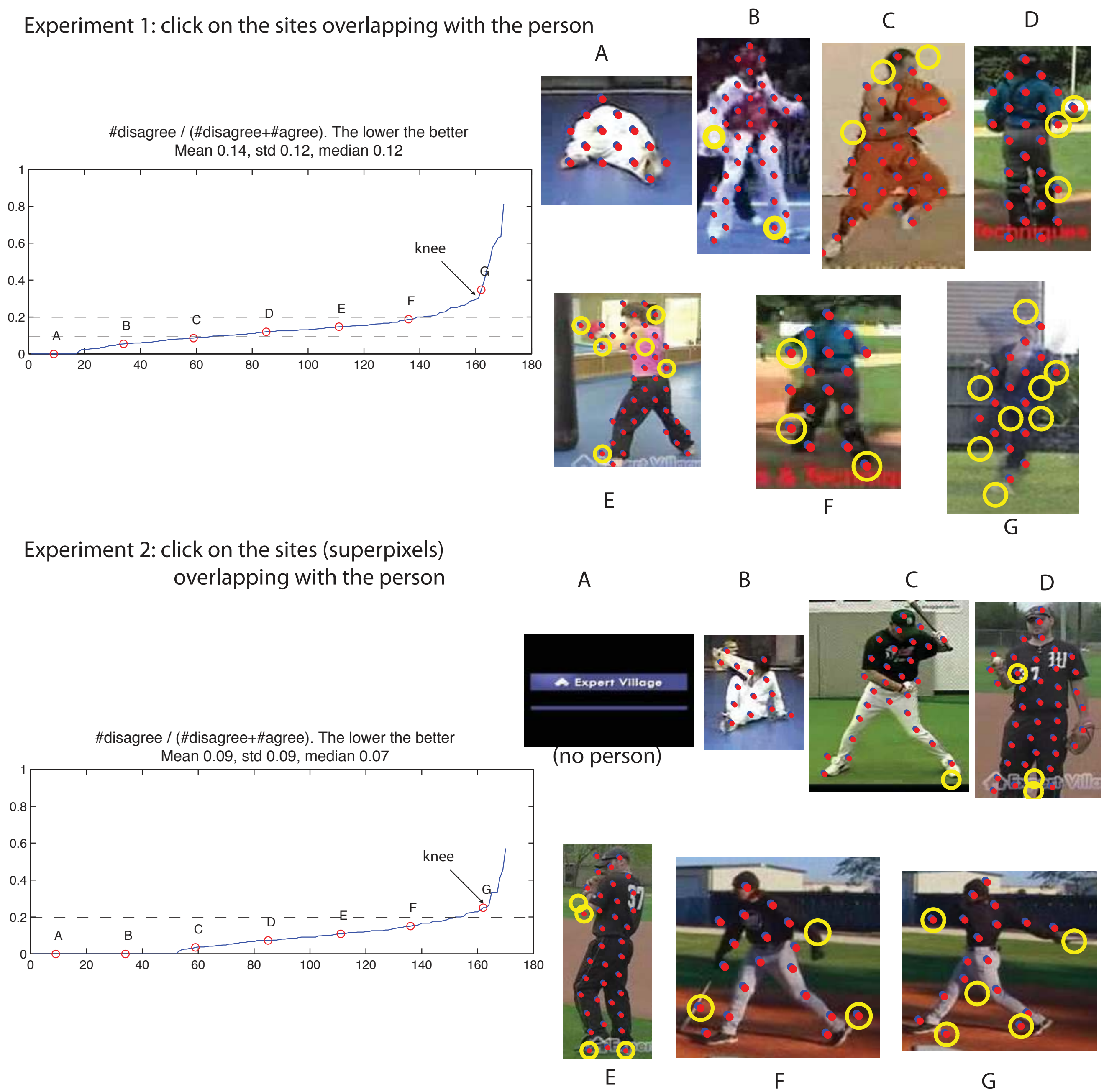

Figure 4. **Quality details.** We present detailed analysis of annotation quality for experiments 1 and 2. For every image the best fitting pair of annotations is selected. The score of the best pair is shown in the figure. We count the number of the sites where the two annotators disagree and divide by all sites labeled by at least one of the two annotators. The scores are ordered low (best) to high (worst). This is effectively a cumulative distribution function of the annotation scores. For clarity we render annotations at 5:15:95 percentiles of the score. Blue and red dots show annotations provided by annotator 1. Yellow circle shows the disagreement. Not surprisingly, superpixels make annotations more consistent compared to a regular grid.

[2] Amazon mechanical turk. http://www.mturk.com/. 1

[3] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold. Evaluation of localized semantics: Data, methodology, and experiments. *IJCV*, 2008. 1, 4

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *PAMI*, 19(7):711–720, 1997. Special Issue on

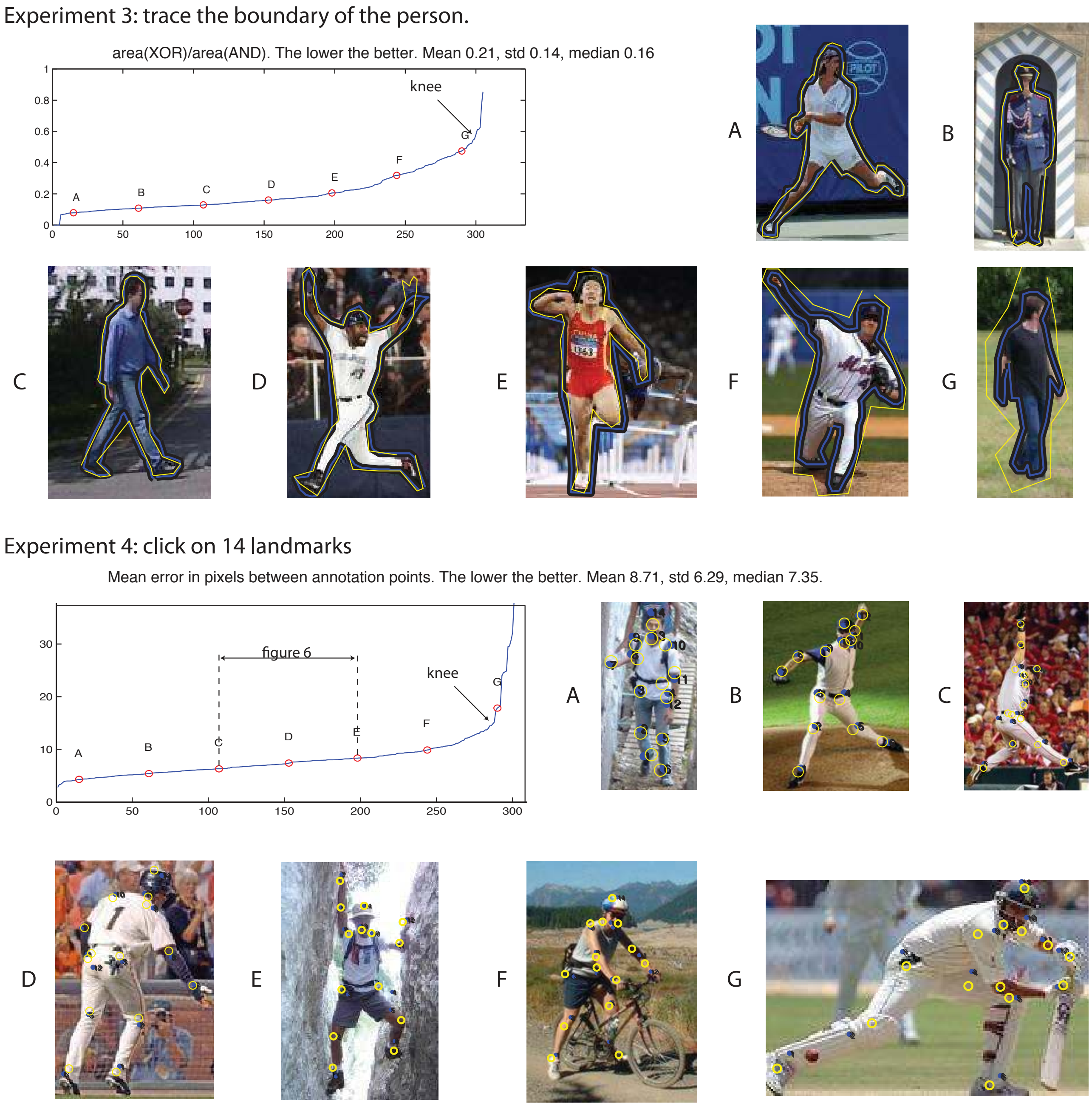

Figure 5. **Quality details.** We present detailed analysis of annotation quality for experiments 3 and 4. For every image the best fitting pair of annotations is selected. The score of the best pair is shown in the figure. For experiment 3 we score annotations by the area of their symmetric difference (XOR) divided by the area of their union(OR). For experiment 4 we compute the average distance between the marked points. The scores are ordered low (best) to high (worst). For clarity we render annotations at 5:15:95 percentiles of the score. Blue curve and dots show annotation 1, yellow curve and dots show annotation 2 of the pair. For experiment 3 we additionally assume that the polygon with more vertices is a better annotation, so annotation 1 (blue) always has more vertices.

Face Recognition. 4

[5] T. L. Berg, A. C. Berg, J. Edwards, and D. Forsyth. Who's in the picture? In *Proc. NIPS*, 2004. 1, 4

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *ICCV*, pages 1395–1402, 2005. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4

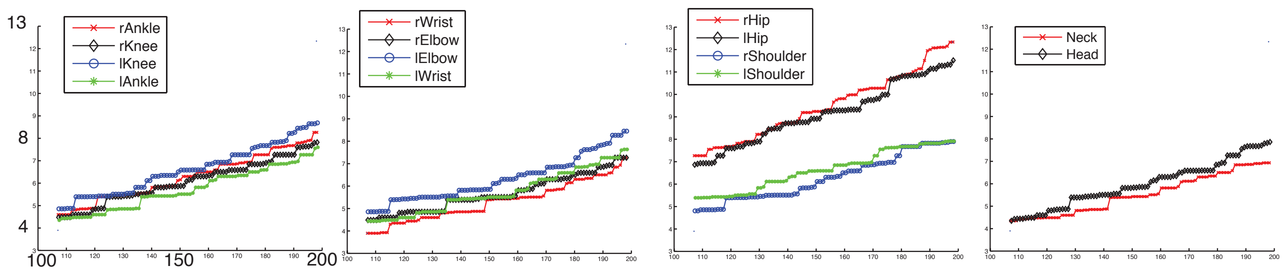

Figure 6. **Quality details per landmark.** We present analysis of annotation quality per landmark in experiment 4. We show scores of the best pair for all annotations between 35th and 65th percentiles - between points "C" and "E" of experiment 4 in fig. 5. All the plots have the same scale: from image 100 to 200 on horizontal axis and from 3 pixels to 13 pixels of error on the vertical axis. These graphs show annotators have greater difficulty choosing a consistent location for the hip than for any other landmark; this may be because some place the hip at the point a tailor would use and others mark the waist, or because the location of the hip is difficult to decide under clothing.

[8] Espgame. www.espgame.org, 2008. 4

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 4

[10] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf. 4

[11] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006. 1, 4

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 1, 4

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 1, 4

[14] Imageparsing. ImageParsing.com, 2008. 4

[15] Linguistic data consortium. www.ldc.upenn.edu/. 4

[16] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994. 4

[17] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 2004. in press. 2

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision*, 2001. 1, 4

[19] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004. 2

[20] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 2000. 4

[21] The pascal visual object classes challenge 2008. http://www.pascal-network.org/challenges/VOC/voc2008/index.html. 4

[22] P. J. Phillips, A. Martin, C. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *Computer*, 33(2):56–63, 2000. 4

[23] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2007. 2

[24] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008. 1, 2, 4

[25] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression(pie) database. In *AFGR*, 2002. 4

[26] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004. 1, 4

[27] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *ACM CHI*, 2006. 1, 4

[28] B. Yao, X. Yang, and S.-C. Zhu. Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks. In *EMMCVPR*, 2007. 1, 4