Eurographics Workshop on 3D Object Retrieval (2010)
I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp (Editors)

Fast Human Classification of 3D Object Benchmarks

A.P.Jagadeesan*, J.Wenzel[#], J.R.Corney*, X.Yan* A.Sherlock[✧], C.Torres-Sanchez*, W.Regli[+]

*University of Strathclyde, [#]University of Edinburgh, [✧]Shapespace, [+]Drexel University

Department of Design, Manufacture & Engineering Management (DMEM), University of Strathclyde, James Weir Building, 75 Montrose St, Glasgow, G1 1XJ
Tel: +44 (0) 141 548 2254; Fax +44 (0) 141 548 4870
E-mail: jonathan.corney@strath.ac.uk

ABSTRACT

Although a significant number of benchmark data sets for 3D object based retrieval systems have been proposed over the last decade their value is dependent on a robust classification of their content being available. Ideally researchers would want hundreds of people to have classified thousands of parts and the results recorded in a manner that explicitly shows how the similarity assessments varies with the precision used to make the judgement.

This paper reports a study which investigated the proposition that Internet Crowdsourcing could be used to quickly and cheaply provide benchmark classifications of 3D shapes. The collective judgments of the anonymous workers produce a classification that has surprisingly fine granularity and precision. The paper reports the results of validating Crowdsourced judgements of 3D similarity against Purdue's ESB and concludes with an estimate of the overall costs associated with large scale classification tasks involving many tens of thousands of models.

Categories and Subject Descriptors (according to ACM CCS): H5.2 Information interfaces and presentation (e.g., HCI) --- User Interfaces: Evaluation/methodology, Theory and Methods; H.5.3 [Information Interfaces]: Group and Organization Interfaces –Web-based interaction.

1. Introduction

There is growing evidence that the Internet can be used to efficiently distribute work to a global work force at almost zero cost. These labourers do not belong to a group, a corporation or a network and neither do they communicate among themselves. But because of Internet technology, they can access tasks, execute them, upload the results and receive various forms of payment using any common web browser. This is a labour market open 24/7, with a diverse workforce available to perform tasks quickly and cheaply. Such Crowdsourcing has the potential to revolutionise the way jobs requiring human judgement are performed by offering a 'virtual automation' of tasks that 'are simple for humans but extremely difficult for computers' (e.g. *"is there a dog in this picture?"*). For example, Crowdsourcing has been proposed as a way of segmentation benchmarking[CGF 09]. Aware of the above the authors have investigated the crowd's ability to make subjective judgments about the relative similarity of shapes. To do this workers have been asked to sort hundreds of "thumbnails" images of 3D shapes into family groups. The research was motivated by the desire to answer the following questions:

1) Can 3D classification tasks be described clearly enough so that a culturally diverse workforce can comprehend what is required in a few seconds? Workers might be located in China, India, Europe or USA.
2) How long will it take to classify collections of 3D models (hours, days or weeks) and does the approach scale?
3) Will the answers produce a consensus result, or simply a broad distribution generated by almost random clicking?

If Crowdsourcing is viable for 3D classification tasks retrieval researchers will have a powerful tool which can help to refine algorithms that are trying to compute the high subjective property of "shape similarity".

In previous papers we have reported the initial results and methodology of our research that is investigating the tial of Crowdsourcing for various Geometric reasoning

© The Eurographics Association 2010.

DOI: 10.2312/3DOR/3DOR10/055-062

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

tasks (i.e. identification of canonical views, similarity matching, and part nesting)[JWC*08] [JLW*09][JWC*09]. In this paper we present the result of our investigation into the feasibility of scaling the approach from sorting collections of around 100 models to almost 500 parts.

The paper has the following structure: Sections 2 briefly describes the Crowdsourcing process and describes how the shape classification task was posed, section 3 describes how the results were clustered to allow comparison with the published classification for the dataset, section 4 discusses the results and how the approach scales. Finally some conclusions are drawn in Section 5.

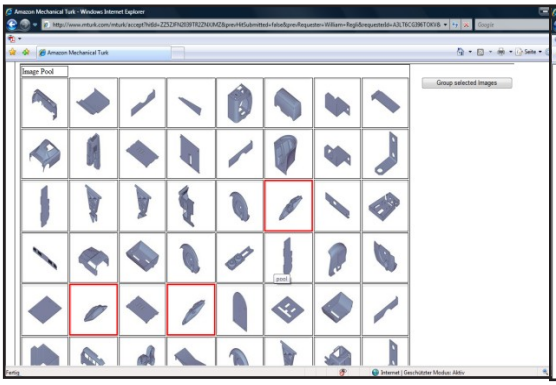
2. Crowdsourcing Terminology and Method

The Crowdsourcing platform used in the series of experiments reported here was Amazon's Mechanical Turk (Mturk) [mTurk08]. In mTurk's terminology the 'requesters' designs and posts a task known as a HIT (Human Intelligence Task). Requesters can "accept" or "reject" the results generated by the workers. Rejection impacts on the worker's reputation on the mTurk system. Payments for completing tasks can be redeemed as Amazon gift certificates or as cash transferred to a worker's bank account.

2.1 The 3D similarity HIT

To investigate if 3D similarity could be effectively Crowdsourced, an mTurk HIT was created containing the entire class (107 parts) of "flat-thin wall components" in Purdue's Engineering Shape Benchmark (ESB) [JKI*06].

Figure 1 Selecting images from the pool



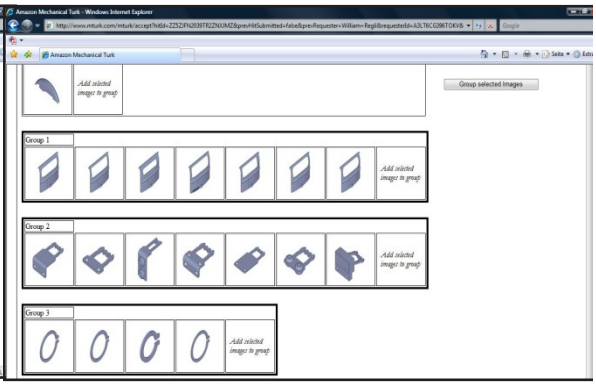
The educational background, previous exposure to 3D CAD and other personal data (such as gender, nationality and age) were also requested (although interestingly no strong correlation between previous experience and the quality of the results could be established). Details of the mTurk workers are summarized in section 4.

The 14 HITs were made available in two rounds: for the first one, 10 HITs were posted and all were accepted within 8min 48sec. The completion time for all the workers to submit their results was 37min 18sec. 70% of the results were valid. For the second round, 4 HITs were offered to workers who accepted the task within 2min 30sec. The completion time was 24min 18sec. There were 3 results valid and 1 invalid. There was a payment reward of $4 per valid HIT. 10 of the HITs were completed in a correct format, and 4 of them rejected. It should be noted that 1 of the rejected ones was due to browser compatibility problems (as tasks had to be carried out in Internet Explorer) and only 3 of the 14 workers actually failed to understand the task or submitted the results of apparently random clicking.

The HIT presented a "pool" of 107 images showing isometric views of the individual CAD-models in the collection. Workers were asked to "put similar looking models together into groups" (see Figure 1) by clicking first on their image and then in one of the rows below the pool. In this way every image selected appeared below the initial pool of images (see Figure 2) in a row (i.e. family) of part images judged similar by the individual mTurk worker. The workers were asked to continue this process until there were no images left in the pool. Facilities were available for workers to edit the contents of clusters during the process, create new rows (i.e. clusters) and move parts between rows. Further details can be found in [JWC*09].

The relationships identified by the workers were summed in a single similarity matrix, S, where the number of times each pair of models was grouped together in the same family of similar shapes was recorded. Given n different CAD-models, the matrix contained $n*(n-1)/2$ values.

Figure 2 Families of Similar Parts below the pool

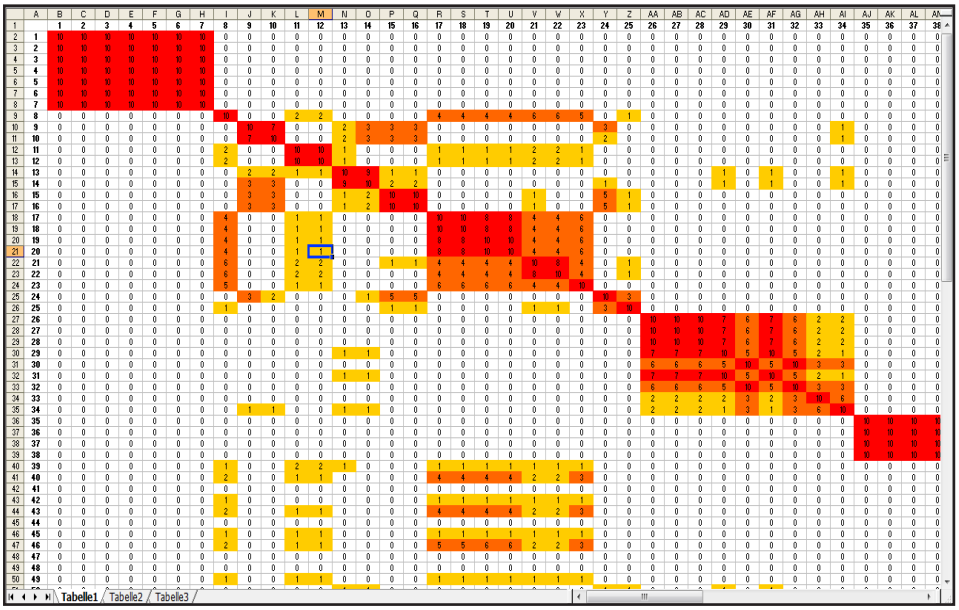


The cells of the matrix held a similarity measure (ranging from 0 to 10) for each pair of CAD-models. In other words, two parts could reach a maximum similarity of 10 in case where all the mTurk workers grouped them together and zero when they had never been associated (Figure 3).

The 107 models in ESB's "flat-thin wall components" dataset are divided into nine sub-groups. To enable comparison with these nine families "clusters" were generated from the similarity matrix (S) defined by mTurk workers.

© The Eurographics Association 2010.

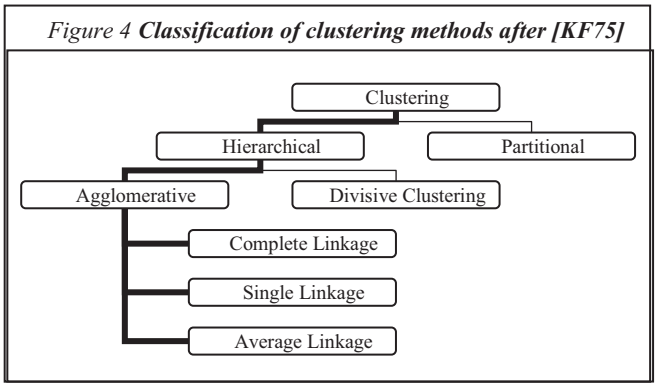
Figure 3 ***Excerpt from the similarity matrix***



2.2 Overview of Clustering Methods

Clustering can be defined as "the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait" [JMF*99].

Figure 4 ***Classification of clustering methods after [KF75]***



Broadly speaking there are two kinds of clustering algorithms used in data analysis 1) hierarchical clustering approaches which produce a nested series of partitions; 2) partition methods which generate binary divisions (e.g. k-means algorithm) [KF75] of data. Such divisive (or "top-down") methods start with all items together in one single cluster and perform binary splitting operations until a stopping criterion is met [KF75]. This study is concerned with the first type, the so called agglomerative method of the hierarchical approach. This "bottom-up" method [JMF*99] starts with each element as a single cluster (i.e. containing only one item), and successively merges clusters together until a stopping criterion is satisfied [KF75]. There are different approaches to determining exactly how individual items or clusters are selected for incorporation into clusters [JOH67]. Figure 4 illustrates the hierarchy of clustering methods (bold line indicates the methods used in this work).

The basis for all the different clustering algorithms is a distance measure that determines how similar or dissimilar the different elements are to each other [KF75]. This study has investigated three different methods known as "single-linkage", "complete linkage" and "average linkage" clustering. All linkage methods have in common that the clustering starts with the two most similar elements [JMF*99]. As soon as two elements (or later on groups of elements) are clustered together the columns and rows of a similarity matrix associated with them are merged and the distances are updated [JMF*99]. The linkage methods differ in the way the distance or similarity between the clusters is calculated (compare Figure 5):

1.) Single linkage clustering (also called 'minimum method'): the distance between two clusters is the minimum distance between two elements of the different clusters. In terms of similarity, the minimum distance is equivalent to maximum similarity.

2.) Complete linkage clustering (also 'maximum method'): the distance between clusters is equal to the maximum distance between two members of the different

© The Eurographics Association 2010.

clusters. In terms of similarity, the minimum similarity between a pair of elements from different cluster is taken as the representative similarity between the clusters.

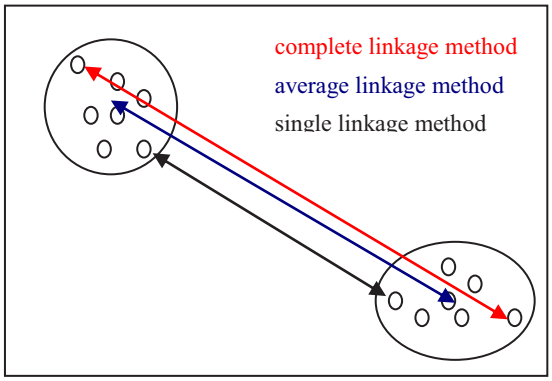
3.) Average linkage clustering: The average linkage method is a compromise between the two previous methods. The distance d between two different clusters A (having |A| elements) and B (having |B| elements) is the average distance taking all elements x and y of the clusters into consideration [JMF*99]. Mathematically the relationship is expressed as:

$$d(\mathrm{A},\mathrm{B}) = \frac{1}{|A|*|B|} * \sum_{x \in A} \sum_{y \in \mathrm{B}} d(x,y) \qquad (1)$$

A common way of presenting the different stages of agglomerative clustering processes is 'dendrograms', which show the order in which elements or groups of elements are clustered into a tree structure. Dendrograms illustrate the complete clustering process until all elements are merged to one all consuming cluster [KF75]. Figure 6 shows an example of a clustering process with its associated dendrogram. The process started with the clustering of the elements B and C. The dotted line shows the current stage of clustering.

Generally the clustering process can be halted at any stage and there are two different criteria, commonly, used to stop clustering processes; 1.) the distance criterion, when clustering is stopped if there are no more clusters within a certain distance; 2.) the number criterion, where the process of merging elements ends if a certain number of clusters is reached. However, because the dataset had only 107 items, the clustering process was simply halted when all shapes had been added (i.e. the root cluster).

Figure 5 Comparison of linkage methods



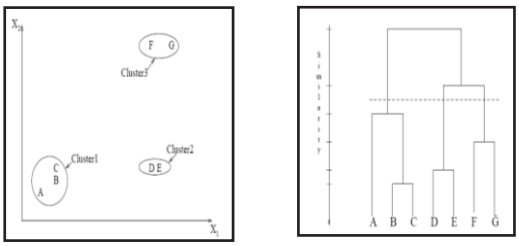


Figure 6 (a,b) **Dendrogram generated for clusters**

3 Implementation and Results

The MATLAB "Statistics Toolbox" was used to generate clusters. As the MATLAB clustering algorithms require a distance matrix instead of a similarity matrix (S) as input, the similarity matrix M was normalized to a distance matrix D. Equation (2) illustrates this process for 10 results:

$$\underline{D} = \frac{1}{10} * \left(\begin{bmatrix} 10 & 0 & \dots & 0 \\ 0 & 10 & & \dots \\ \dots & & \dots & 0 \\ 0 & \dots & 0 & 10 \end{bmatrix} - \underline{S} \right) \qquad 2)$$

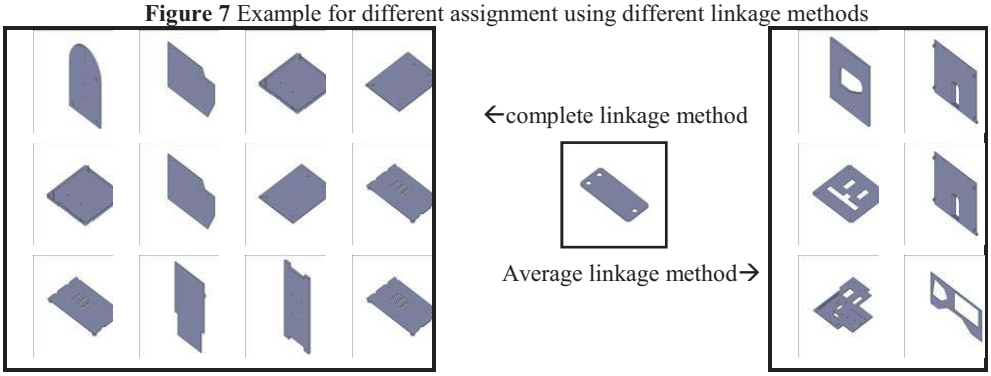
Dendrograms were plotted for each of the different linkage methods (see Figure 12 a,b,c) and results contrasted. When the results of the "average" to the "complete" methods are compared the differences are surprisingly small. For example at the level where the parts are clustered into 20 groups the major clusters differ by only a single part highlighted in Figure 7.

However the differences between these two linkage methods can be clearly seen in the structure of the higher levels of the dendrogram tree (i.e. near the root) in Figure 12(c). The average linkage method leads to a much richer structure in which clusters merge into larger clusters. By contrast the complete linkage dendrogram Figure 12(b) shows twelve clusters that are merged simultaneously at the root, this means for all combinations of clusters it is possible to find two parts that were never clustered together.

While the results of the "average" and the "complete" linkage methods are comparable, the results for the "single linkage" method are significantly different. Figure 12(c) shows that the relative range of cluster sizes is much more variable for the single linkage method. In "single linkage" the distances between the clusters are determined by the shortest distance between any two parts of the different clusters. This results in more parts being merged into aready big clusters, while smaller clusters stay isolated until the late stages of the linkage process. Consequently the authors decided to adopt the "average" linkage method because it resulted in dendrogram structures that showed most discrimination between families of similar models.

© The Eurographics Association 2010.

Figure 7 Example for different assignment using different linkage methods



The dendrogram generated by average linkage for the ESB's "flat thin wall components" is show in Figure 13.

3.1 Comparing ESB and Crowdsourced Clusters

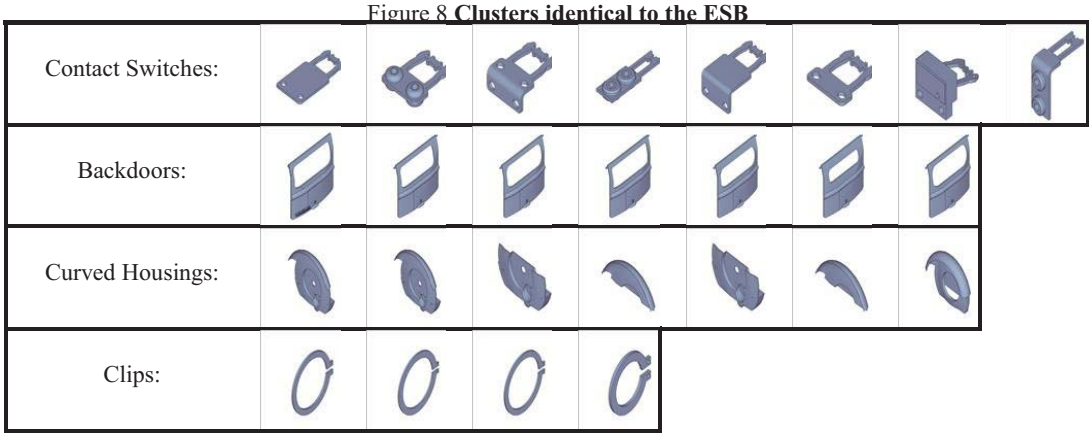
Despite the fact that all the 107 parts can be described as "flat-thin wall components", the ESB identifies 9 distinct families (or clusters) of models that can be identified within the overall classification. Focusing on the eight central clusters of ESB (and disregarding the "Miscellaneous" group) exact matches for half of them (Figure 8) can be identified at the clustering level illustrated in Figure 13. Close correlations can also be identified with two of the other ESB-clusters known as "thin plates" and "slender thin plates".

Even the 14 models in the "rectangular housings" grouping (the most complex parts in the ESB) can be seen in the results. Although the Turkers left out 2, of the 14 parts, they added 2 others from the "Miscellaneous" group, Figure 10. In summary the Crowdsourced classification appears to be comparable, and possibly better, than the published groupings. Having established the basic feasibility of the approach we investigated who the "Turkers" were and if the method would scale to groups of several 100 models.

3.2 Turkers' Background

At the beginning of this HIT the workers were asked several personal questions. Perhaps because of the time of day the tasks were posted the HITs were basically fulfilled by workers from the United States.

Figure 8 **Clusters identical to the ESB**



© The Eurographics Association 2010.

In Figure 13 (columns 3, 4 and 5), the "thin plates" cluster has three separated groupings. But at higher levels of tree, these clusters are merged into a single cluster whose only difference to the ESB group is in the classification of circuit boards. While average clustering assigns these "circuit boards" to the "slender thin plates", they were assigned to the "thin plates" group in the ESB (see Figure 9).

Against our expectations 50% of our tasks were done by female workers. We were even more surprised to learn that the age of the workers that fulfilled our HIT is nearly equally distributed between 20 and 60 years.

Figure 9 **Different assignments of circuit boards in ESB and our results**

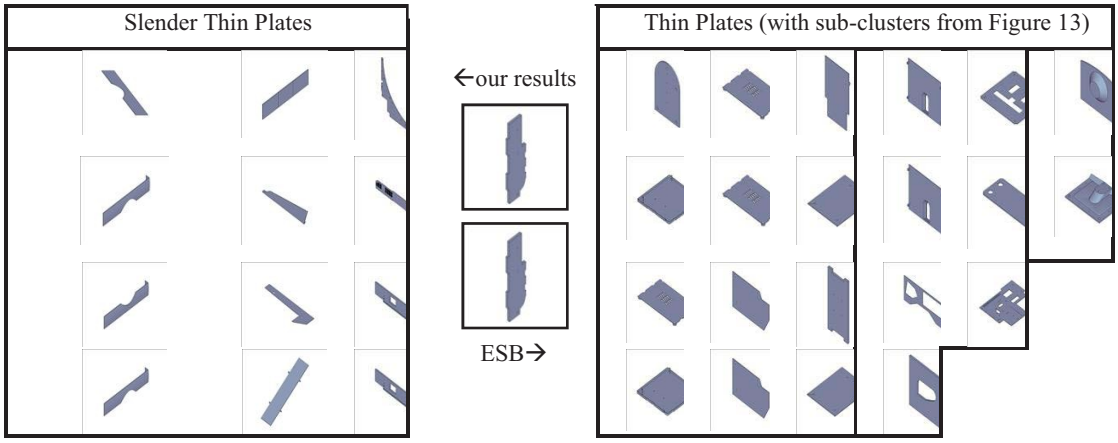


Figure 10 **Cluster "rectangular housings" in ESB and our results**

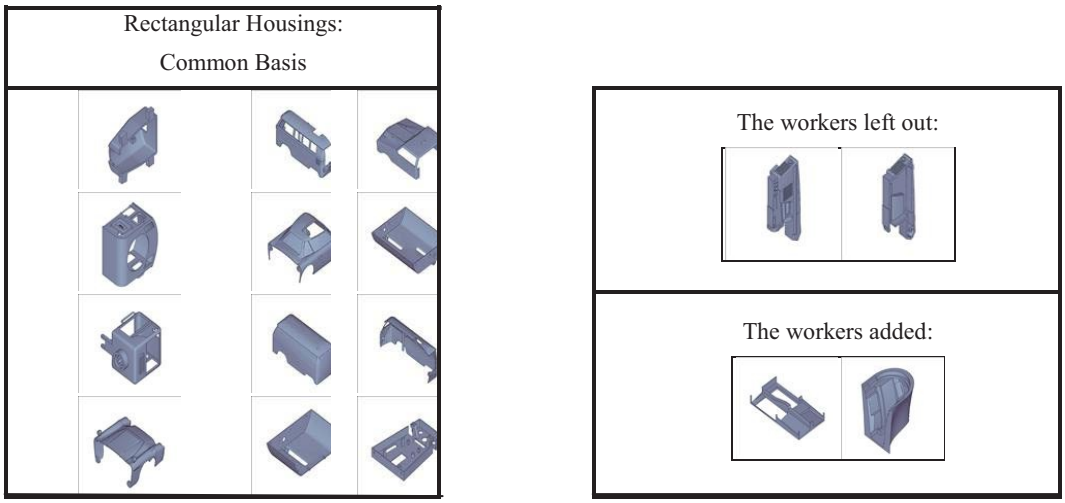
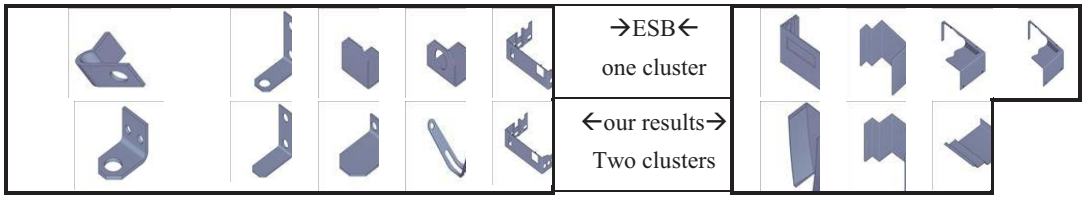


Figure 11 **Separation of "bracket like parts"**



© The Eurographics Association 2010.

Indeed there appears to be no stereotype of a MTurk worker. Considering our previous experiences we were not surprised that a majority of the workers (65%) spends more than 10 hours per week on MTurk. A quarter of the workers answered that they spend more than 30 hours per week on MTurk. Only one of 16 workers answered that he has experience with mechanical engineering. By contrast more than 31% of the workers answered that they have experience with CAD. Even more of them (44%) stated that they have experience with mathematics for analysis.

4.0 Scale-up to larger model collections

To discover if the approach can be scaled, a second class of the ESB was tested. The "rectangular-cubic prism" contains 281 parts and these were presented to workers as a single HIT. 15 HITs were posted and all the HITs were accepted in 14min26sec and returned in 1hr1min8sec. As there were many parts to cluster, workers had to scroll the HIT page up and down many times. Every worker with approved answer was rewarded $8.00. People who had done exceptional work (e.g. identified a large number of component families) were rewarded a bonus of $4.00. A dendrogram of the result is shown in Figure 14.

Similarly, the third class of the ESB, "solid of revolution" was tested. It contained 479 parts. 15 HITs were posted and all the HITs were accepted in 15min02sec and returned in 1hr49min40sec. For every approved answer, a reward of $12 was paid and for an outstanding work a bonus of $6.00 was granted. A dendrogram of the results is shown in Figure 15.

Again outstanding and exceptional work was defined as when a worker had identified a large number of clusters (i.e. a high resolution of sorting).

Table 1: mTurk Performance Summary

Number of 3D Models	Average Time to Classify minutes	Cost per HIT (including Bonus) $
107	22	4
281	51	12
479	68	18

From our experience it appears that posting around 10 to 15 HITs produces useful results and that 500 models can be sorted in around an hour. The payment level was determined by the UK's minimum hourly rate. However taking this as a maximum value (and the knowledge that once a type of work becomes established workers become more practised at the task as so reduces the costs) one could estimate that models could be classified in batches of 500 for around $10 a HIT.

Table 2 Percentage of difference in clusters. "ESB data" vs "Cluster-HIT"

ESB Benchmark Data	Approximate similarity between clusters (ESB vs HIT)
Flat-ThinWall Components	71.5%
Rectangular-Cubic Prism	85.8%
Solid Of Revolution	80.6%

5.0 Conclusions and Future Work

Validating the results of the Crowdsourced similarity clusters against the ESB's published groupings produces correlation of better than 70%. Examinations of the differences show they arise from inherently ambiguous parts, whose correct assignment is arguable. Our experience of the overall performance is summarized in Tables 1 and 2.

Future work will investigate if performance is improved by adopting an iterative approach, where each worker is allowed to incrementally improve the classifications previously returned by other workers. Recent work at MIT has reported that this approach has improved Crowdsourcing performance in a number of different applications [TurkIt09]

References

[CGF 09] CHEN X., GOLOVINSKIY A AND FUNKHOUSER T., 'A Benchmark for 3D Mesh Segmentation', *ACM Transactions on Graphics (*Proc. SIGGRAPH), (August 2009), Vol 28, No 3

[JWC*08] JAGADEESAN P., WENZEL J., CORNEY J.R., YAN X.T., SHERLOCK A., REGLI W., (2008), 'Geometric Reasoning With a Virtual Workforce (Crowdsourcing for CAD/CAM)', *Procs 2nd International Workshop Virtual Manufacturing VirMan 08 as part of the 5th INTUITION International Conference: Virtual Reality in Industry and Society*: From Research to Application, (October 6-8, 2008), Torino, Italy, ISBN 978-960-89028-7-9 276.

[JLW*09] JAGADEESAN P. , LYNN, A. , WENZEL, J., CORNEY, J.R. , YAN, X.T. , SHERLOCK, A. , TORRES-SANCHEZ, C. , REGLI, W., "Geometric Reasoning via Internet CrowdSourcing", *Procs. SIAM/ACM Joint Conference on Geometric and Physical Modeling*, (Oct 5th-8th 2009), San Francisco, California

[JWC*09] JAGADEESAN P. , WENZEL, J. , CORNEY, J.R. , YAN, X.T. , SHERLOCK, A. , TORRES-

© The Eurographics Association 2010.

SANCHEZ, C. , REGLI, W., "Validation of Purdue Engineering Shape Benchmark clusters by Crowdsourcing", *Procs International Conference on Product Lifecycle Management*, (July 4th-6th 2009), Bath

[JMF*99] JAIN A.K., MURTY M.N., FLYNN P.N, 'Data clustering: A review', *ACM Computing Surveys*, Vol. 31, (1999), No. 3, pp. 264-323

[JKI*06] JAYANTI, S. , KALYANARAMAN, Y. , IYER, N. & RAMANI, K. Developing An Engineering Shape Benchmark For CAD Models, *CAD*, Vol 38: 9, (Sept 2006), pp 939-953

[JOH67] JOHNSON S.C. 'Hierarchical clustering schemes', *Psychometrika*, Vol. 32, (1967), No. 3, pp. 241-254

[KF75] KUIPER F.K., FISHER L., 'A Monte Carlo comparison of six clustering procedures', *Biometrics*, Vol. 31, (1975), No. 3, pp. 777-783

[MTurk08] Amazon MTurk: http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk (Accessed 30th July 2008)

[SUR05] SUROWIECKI J., '*The wisdom of crowds*', Random House, New York. (2005), ISBN: 978-0-385-72170-7

[TurkIt09] TurkIt Website http://groups.csail.mit.edu/uid/turkit/ (Accessed 19th Jan 2010

[WEN09] WENZEL J., 'Feasibility study for the "crowdsourcing" of geometric reasoning in mechanical CAD', Thesis, (2009), Institut fuer Fertigungstechnik und Werkzeugmaschinen, Leibniz Universitaet Hannover (Institution).

Larger versions of the Figures on this page have been uploaded to the conference website as accompanying media

Figure 12: Linkage Methods

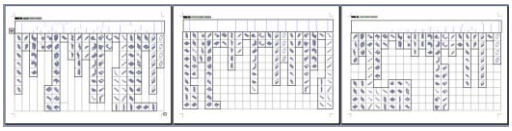


Figure 13. ESB "Flat thin walled component"

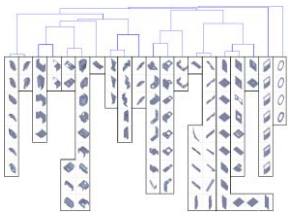
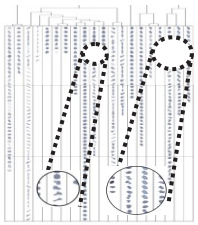


Figure 14. ESB "Rectangular-cubic prism" class Crowdsourced dendrogram (281 models)

Figure 15. Crowdsourced dendrogram for ESB "Solid OF Revolution" Class (479 Models)



© The Eurographics Association 2010.