






# A Multi-Annotator Survey of Sub-km Craters on Mars

Alistair Francis <sup>1,2,\*</sup> , Jonathan Brown <sup>2,†</sup>, Thomas Cameron <sup>2,†</sup>, Reuben Crawford Clarke <sup>2,†</sup>, Romilly Dodd <sup>2,†</sup>, Jennifer Hurdle <sup>2,†</sup>, Matthew Neave <sup>2,†</sup>, Jasmine Nowakowska <sup>2,†</sup>, Viran Patel <sup>2,†</sup> , Arianne Puttock <sup>2,†</sup>, Oliver Redmond <sup>2,†</sup>, Aaron Ruban <sup>2,†</sup>, Damien Ruban <sup>2,†</sup>, Meg Savage <sup>2,†</sup>, Wiggert Vermeer <sup>2,†</sup>, Alice Whelan <sup>2,†</sup> , Panagiotis Sidiropoulos <sup>1,3</sup>  and Jan-Peter Muller <sup>1</sup> 

<sup>1</sup> Mullard Space Science Laboratory, UCL, Holmbury Hill Rd, Dorking RH5 6NP, UK; panos@hummingbirdtech.com (P.S.); j.muller@ucl.ac.uk (J.-P.M.)

<sup>2</sup> The College of Richard Collyer, 82 Hurst Rd, Horsham RH12 2EJ, UK; jonathannyebrown@gmail.com (J.B.); thomas.j.cameron@btinternet.com (T.C.); reuben.crawfordwork@gmail.com (R.C.C.); romydodd13@gmail.com (R.D.); jhurdle54@yahoo.co.uk (J.H.); matthewneaveis@hotmail.co.uk (M.N.); jasminenowakowska@gmail.com (J.N.); virankp@outlook.com (V.P.); ap2573@bath.ac.uk (A.P.); oliverredmond12@gmail.com (O.R.); aaronruban@gmail.com (A.R.); damienruban46@gmail.com (D.R.); megsavage@rocketmail.com (M.S.); wiggert9@yahoo.co.uk (W.V.); alicewhelan13@gmail.com (A.W.)

<sup>3</sup> Hummingbird Technologies Ltd., 51 Hoxton Square, Hackney, London N1 6PB, UK

\* Correspondence: a.francis.16@ucl.ac.uk; Tel.: +44-1483-204926

† These authors contributed equally to this work.

Received: 30 June 2020; Accepted: 1 August 2020; Published: 3 August 2020



**Abstract:** We present here a dataset of nearly 5000 small craters across roughly 1700 km<sup>2</sup> of the Martian surface, in the MC-11 East quadrangle. The dataset covers twelve 2000-by-2000 pixel Context Camera images, each of which is comprehensively labelled by six annotators, whose results are combined using agglomerative clustering. Crater size-frequency distributions are centrally important to the estimation of planetary surface ages, in lieu of in-situ sampling. Older surfaces are exposed to meteoritic impactors for longer and, thus, are more densely cratered. However, whilst populations of larger craters are well understood, the processes governing the production and erosion of small (sub-km) craters are more poorly constrained. We argue that, by surveying larger numbers of small craters, the planetary science community can reduce some of the current uncertainties regarding their production and erosion rates. To this end, many have sought to use state-of-the-art object detection techniques utilising Deep Learning, which—although powerful—require very large amounts of labelled training data to perform optimally. This survey gives researchers a large dataset to analyse small crater statistics over MC-11 East, and allows them to better train and validate their crater detection algorithms. The collection of these data also demonstrates a multi-annotator method for the labelling of many small objects, which produces an estimated confidence score for each annotation and annotator.

**Keywords:** Mars; craters; remote sensing; object detection; planetary science

## 1. Introduction

Craters are formed by the impact of meteorites on a planet's surface. The size of a crater is correlated with the energy of the impact event and, thus, the mass of the associated impactor [1]. By measuring the density of craters on a planetary surface—taking into account the flux of impactors and the rate of erosion—we can estimate its age. In practice, measuring crater frequency as a function of size (known as a Crater Size-Frequency Distribution, or CSFD) allows for more precise age estimates than the total count. In CSFDs, we observe a clear inverse power law for larger craters [2], making

relative age estimates between regions straightforward. However, absolute age-dating of surfaces demands that several parameters related to both the production and erosion of craters be well constrained. The production rate is primarily affected by impactor flux, and the atmospheric and material characteristics of the planet. The erosion rate is the number of craters for a given size that are destroyed per unit time. This includes processes on Mars, such as fluvial and aeolian activity, volcanism, and obliteration via later impacts [3].

It has long been observed that all of these factors are easier to constrain for larger craters: fluxes of larger impactors are easier to measure, and many of the erosion effects (e.g., dust deposition) become negligible, because the craters are so large as to be impervious to them. The primary mechanism for erosion of large craters is in fact their obliteration due to even larger impacts imprinted on top of them. However, CSFDs at smaller size ranges are less applicable to age-dating, due to a less well bounded impactor flux, and erosion effects that are not dominated by a single cause, but are rather a mix of several non-negligible factors [3].

CSFDs are unreliable at smaller diameter ranges, thus it is tempting for scientists performing age-dating studies to simply ignore craters below, say, 1 km diameter. Unfortunately, larger craters are significantly rarer than smaller ones, meaning that, if we wish to measure surface age at higher spatial resolution, then we must rely on the statistics of smaller craters more heavily, as large ones are too sparse to draw reliable conclusions from. This also holds for young surfaces, which have seen relatively few impacts. Therefore, there will always be an interest in extending the range of crater diameters we can reliably use to age-date planetary surfaces. Additionally, small crater statistics could be used to help quantify those processes that are currently poorly constrained. For example, secondary cratering has been shown to be responsible for nearly a quarter of all craters with diameter  $\geq 1$  km globally on Mars [4], and dramatic examples, such as Zunil crater show rays with huge numbers of secondaries [5]. Clearly, any attempt to disentangle the complex statistics of primary and secondary crater populations must begin with a much larger survey of craters (e.g., [6]).

Manual surveys of Martian craters have successfully mapped all large craters, with a high degree of accuracy [7]. However, given that hundreds of millions of craters exist on the Martian surface, it will be impossible to conduct a manual survey globally, or indeed of any substantial portion of the surface. Eventually, automated detection is likely to lead to the most fruitful results. To this end, Machine Learning models—in particular, Convolutional Neural Networks (CNNs)—have been the state-of-the-art method for many Computer Vision tasks since 2012 [8]. Specifically, CNNs have been successfully employed for a huge variety of Object Detection applications (e.g., [9,10]). The power of these methods is revealed most spectacularly when the datasets used to train them are very large, because of their ability to generalise and interpolate in a hugely complex input space. These new techniques present a huge opportunity for the automation of regional or global crater surveys. In recent years, Crater Detection Algorithms (CDAs) for larger craters have been widely implemented: for example, [11] use a segmentation network to detect craters between 2 and 32 km in diameter in THEMIS imagery, whilst [12] used Digital Terrain Models as input. For High Resolution Stereo Camera (HRSC) imagery, which at 12.5 m resolution allows for sub-km crater detection, several CDAs have been proposed—either using more traditional Machine Learning algorithms, such as decision trees [13], or CNNs [14].

For Martian cratering, few datasets exist, and those that do (e.g., [7]) often do not include small craters. To date, the biggest openly available dataset of small Martian craters was released in 2012 [15]. This contains 3050 craters from the HRSC [16] over Nanedi Valles between 200 m and 5 km diameter. Our dataset extends and complements this prior effort, as it uses multiple annotators, a different camera system, is located in a different region of Mars, and it expands the diameter range down to tens of metres. Multi-annotator datasets have some precedent in the crater detection field: for example, the Moon Zoo project used several thousand citizen scientists to annotate Lunar craters over the Apollo 17 landing site [17]. The novel metric that we use to evaluate the consistency of our results

(Section 3.2) provides a framework for future crowdsourcing efforts in a variety of object detection problems, not restricted to crater counting.

## 2. Data Description

The dataset comprises 12 annotated scenes, each 2000-by-2000 pixels in size, from the Context Camera (CTX) [18] aboard the Mars Reconnaissance Orbiter. More specifically, the tiles have been selected in a pseudo-random fashion from the co-registered and orthorectified mosaic over the MC-11 East quadrant, the creation of which is detailed in [19,20]. These data are co-registered to a baseline Digital Elevation Model that uses HRSC data [21]. The resolution of the data is 6 m, and each image measures 12 km by 12 km. CTX is in a sun-synchronous orbit, meaning that all images are taken at 3 pm local time [22], leading to consistent illumination angles across the dataset. The 12 images cover a diverse set of locations, terrain types, and altitudes, as can be seen in Figure 1.

Each image comes in both png format and as a geo-coded tif image in data/images. Associated with each image is a folder in data/annotations/raw containing 6 PASCAL VOC annotation format xml files, each one from a different annotator. Finally, there is a clustered annotation file, also in PASCAL VOC format [23].

Table 1 breaks down the number of individual annotations, and inter-annotator agreement (as defined in Section 3.2), for each tile. There is a large variation in the number of craters found within each of the 12 images, with the least cratered having 223 individual annotations, and the most having 3954. We found there was a high variance in the number of labels produced by each annotator, with some annotators consistently marking fewer or more craters in all of the images they annotated. Generally there was a ratio of 2–4 between the most and least numerous surveys on a given tile, this provides a strong argument for a multi-annotator strategy, which reduces the impact of this variance on the accuracy of results.

**Table 1.** Scene-wise breakdown of annotations per annotator, including the number of labels, and agreement score (as defined in Section 3.2). The variation in agreement score is large, highlighting the difficulty of some of the scenes in comparison to others. For each tile, the set of six annotators (denoted ‘i’ to ‘vi’) is different, and the ordering of the columns is arbitrary.

SCENE	Labels per Annotator						TOTAL	Agreement Score (%)						MEAN
	i	ii	iii	iv	v	vi		i	ii	iii	iv	v	vi	
A	680	255	396	376	1111	413	3231	62.4	74.8	73.5	79	40.8	63.5	65.67
B	88	90	93	151	177	125	724	75.5	77.7	82.1	67.9	58.7	71.7	72.27
C	125	212	97	178	230	196	1038	82.6	69	78.6	74.6	68.9	75.4	74.85
D	72	43	32	94	73	85	399	75.7	77.1	82.3	63.9	75.7	62.9	72.93
E	277	503	690	778	869	837	3954	79.2	73	67.3	67.7	62.2	66.6	69.33
F	197	229	235	278	305	187	1431	75.1	74.7	80.6	69.7	65.3	80.4	74.3
G	45	22	45	45	60	51	268	74.9	83.2	80.0	79.5	66.2	75.7	76.58
H	24	36	43	37	40	43	223	86.5	78.6	74.8	83.7	72.3	77.6	78.91
I	147	135	174	209	183	262	1110	68.9	77.7	75.2	68.0	73.7	52.3	69.32
J	25	95	32	40	69	36	297	71.2	22.4	50.5	56.7	35.1	50.8	47.78
K	66	45	28	63	36	66	304	66.8	100	74.8	86.8	74.6	67.6	78.43
L	375	696	273	281	583	581	2789	74.1	46.4	71.7	73.9	63.6	47.1	62.78
<b>TOTAL</b>							<b>15,768</b>							<b>70.26</b>

In Table 2, we see information regarding the clustering of results, with average, minimum and maximum diameters for all images after clustering. During clustering, we first remove any annotations that have a diameter less than three pixels. Often, these very small craters are resolvable; however,

accurate labelling is difficult and may have resulted in unreliable final results, so were omitted. Therefore, many of the images have a minimum diameter of 18 m (three pixels), whilst the upper limit varies much more, given the sporadic distribution of larger craters in our images. The median average diameters are consistently below mean averages, suggesting a skewed distribution towards lower size ranges, which is to be expected due to the negative power-law distribution of CSFDs.

Each PASCAL VOC format xml file contains information about the image it is related to, including path information, and its dimensions. It then contains a list of object fields, which denote each crater label. The difficult field is used to denote when an annotator chose to express that they were unsure about a given annotation, whilst the bndbox field gives the boundaries of the crater. No geographical information is provided in the xml, with all dimensions in pixel coordinates. Many of the fields included in the xml file are somewhat redundant (e.g., pose, truncated), but are included to increase compatibility with software which expects PASCAL VOC format data as input. An example of these xml files is given below.

---

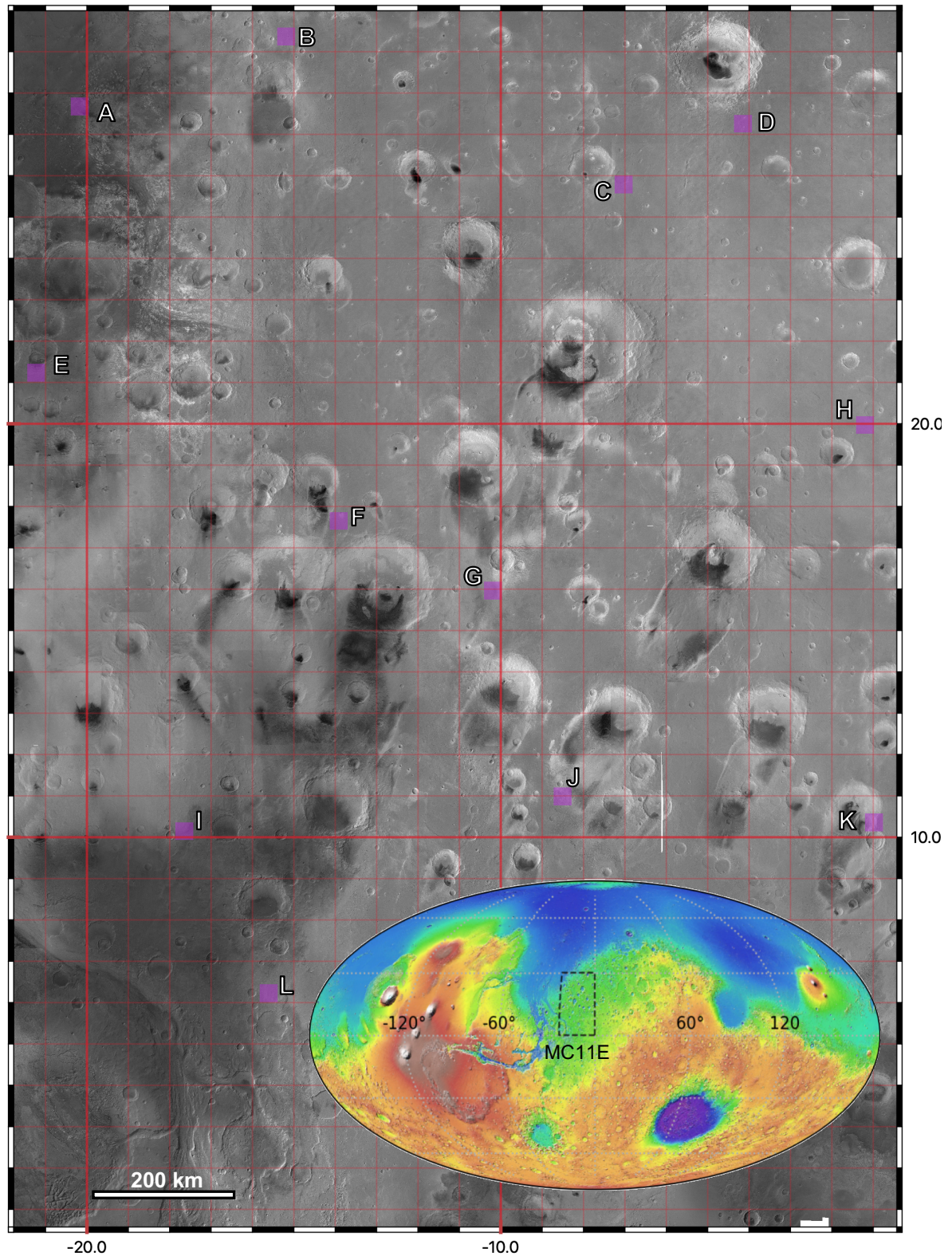
```

<annotation>
<folder>data/images/</folder>
<filename>MC11E-B.png</filename>
<path>data/images/MC11E-B.png</path>
<source>
<database>MSSL ORBYTS MCC</database>
<annotation>MSSL ORBYTS</annotation>
<image>NASA CTX / iMars</image>
</source>
<size>
<width>2000</width>
<height>2000</height>
<depth>1</depth>
</size>
<segmented>0</segmented>
<object>
<name>crater</name>
<pose>Unspecified</pose>
<truncated>0</truncated>
<difficult>0</difficult>
<bndbox>
<xmin>232</xmin>
<ymin>224</ymin>
<xmax>260</xmax>
<ymax>252</ymax>
</bndbox>
</object>
<object>
...

```

---

The same format is used for the clustered annotations, but they also contain an extra 'confidence' value in each object instance denoting the number of annotators who found that crater, between 1 and 6, held in data/annotations/clustered. This clustered annotation file is created using the *src/cluster\_surveys.py* script, with parameters set as default.



**Figure 1.** CTX Mosaic over MC-11 East quadrangle [20]. Purple squares indicate the 12 regions randomly sampled for annotations. The lower right shows a global elevation map of Mars from MOLA, with MC-11 East outlined.

### 3. Methods

#### 3.1. Collection

Our annotations were conducted by a group of 16 young ORBYTS (Original Research By Young Twinkle Students) [24] participants, at The College of Richard Collyer, Horsham, UK. In order to prepare, the annotators were given roughly 15 hours of seminars on Martian science, Machine Learning, and statistics. In addition to this, we worked collectively to closely define how annotations would be made, what was counted as a crater and what was not, and any possible edge-cases. This was all done in order to ensure a high accuracy and as low a variance as possible between the annotators. The tool used in annotation was a customised open-source image annotation tool *LabelImg*. Our version was made to restrict annotations to be equal width and height, as craters are usually most easily defined as a circle, rather than an ellipse. We also added a toggle for 'difficult' annotations, which could be triggered when annotators were unsure of their own marking.

After a training session using the software, annotations were intermittently conducted over a period of around eight weeks. Six annotators were selected using a random number generator for each tile and were then separated so as to not be influenced by one another whilst labelling. Once a tile was completed, an annotator would then be assigned another tile randomly from those still to be done. There is some variation in the quantity of annotations made by each annotator due to absences, differences in speed, and variation in the difficulty of the images.

Craters of all sizes that were visible were annotated. The lower diameter limit was determined by the resolution and quality of the CTX images. At the 6 m/pixel of CTX, we found that, in practice, few craters under around 30 m (5 pixels) in diameter were visible. However, given that some craters smaller than this could be resolved, we did not enforce a minimum diameter limit on individual annotations (the clustering process does filter results smaller than 3 pixels in diameter). If craters extended partially beyond the sides of the image they were ignored, as it could lead to ambiguity in size and position. Annotators began labelling larger craters and gradually zoomed in further, panning across the image systematically in order to comprehensively survey all sizes and locations.

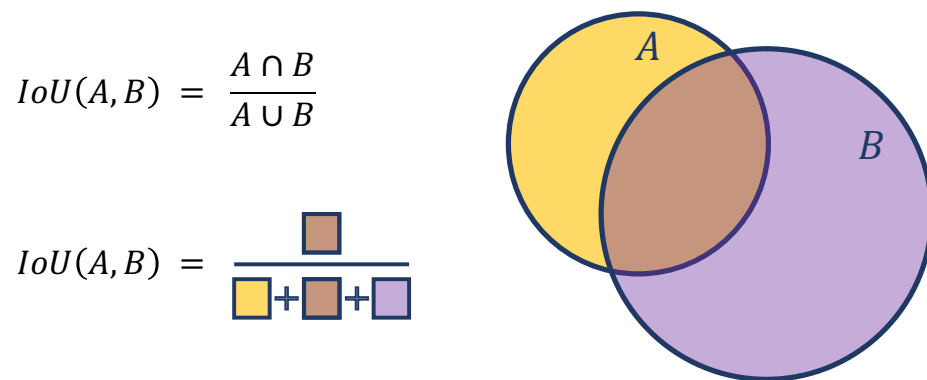
Combining individual annotations is done using agglomerative clustering, a technique used since the 1950s for a variety of clustering problems [25]. This has the benefit of only requiring two parameters: one is the function with which a distance is calculated between two annotations, and the other is the cut-off distance at which no more clustering is performed. First, all distances between different pairs of annotations is calculated; then, the closest two are paired, and the new point is calculated as the mean between those which were clustered. This process is repeated until no distances remain below the cut-off. For our application, we would like to ensure that the distance between any two markings by the same annotator is 1, meaning clusters can never contain more than one entry from a given annotator.

For two markings,  $L_i^{(n)}$  and  $L_j^{(m)}$ , by annotators  $n$  and  $m$ , we define the distance as

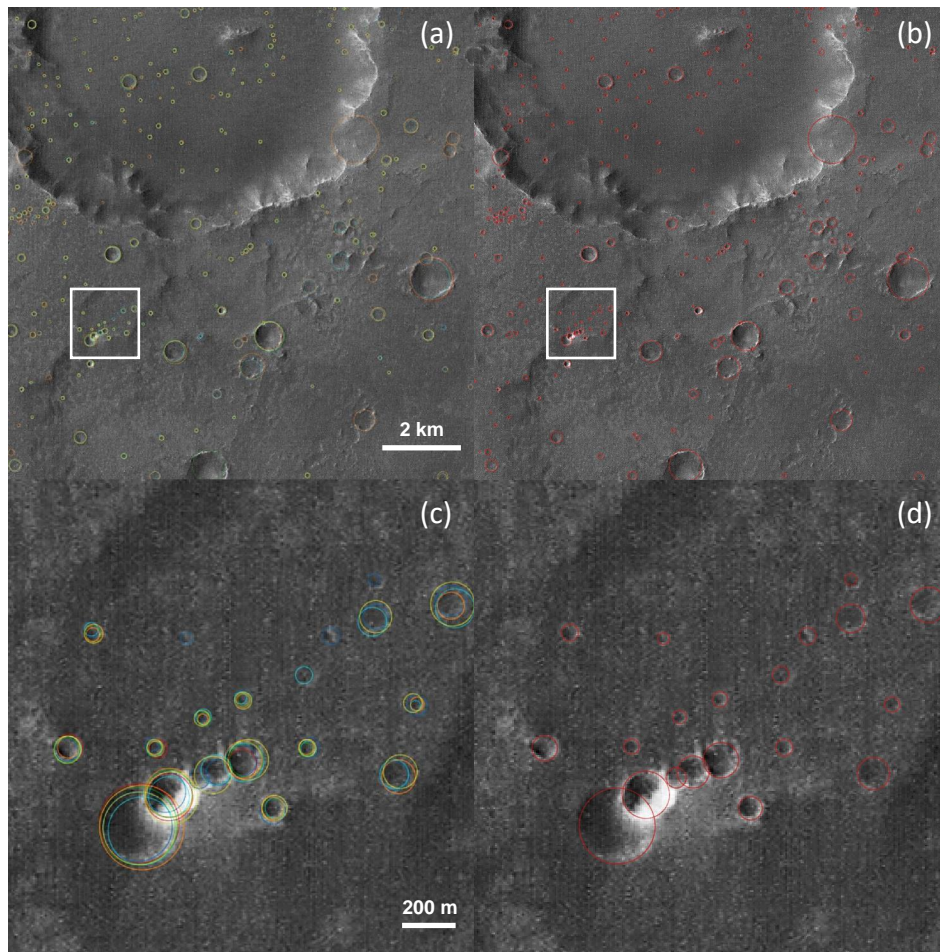
$$d(L_i^{(n)}, L_j^{(m)}) = \begin{cases} 1 - IoU(L_i^{(n)}, L_j^{(m)}) & \text{if } n \neq m \\ 1 & \text{if } n = m \end{cases} \quad (1)$$

where  $IoU(L_i^{(n)}, L_j^{(m)})$  is defined as the Intersection over Union of the two circles, also known as the Jaccard index (Figure 2).

The cut-off distance threshold was largely decided by visual inspection. Too low a value, and repeat entries for the same craters might be included in the final survey; too large and multiple craters might be combined erroneously. We found that a value of 0.9 (clustering any  $IoU > 0.1$ ) was close to optimal and, therefore, used this for all images. Table 2 gives details regarding the results of the agglomerative clustering, and a visual example of the results can be seen in Figure 3.



**Figure 2.** Graphical depiction of the Intersection over Union, or Jaccard index, for two annotations. The areas are calculated using circles with diameter equal to the width and height of the bounding box of the annotation.



**Figure 3.** Image 'C' from the dataset, with annotations displayed. In (a) we see the full image with unclustered labels, each colour representing a different annotator; (b) shows the output labels from agglomerative clustering; (c,d) show the same process, zoomed in on a portion of the image. As we see, the initial annotations are in close agreement with one another, and the clustering successfully separates craters, and accurately estimates their centres and diameters. The very large crater in the upper portion of the image was not labelled, as it extends past the bounds of the image.

**Table 2.** Scene-wise analysis of clustering process. On average, each crater was found by over three annotators, and has a mean diameter of less than 100 m. Annotations were filtered to disallow labels with  $D < 18$  m (3 pixels), hence the small discrepancy in individual counts between Table 1 and this. Tiles with fewer craters tend to have high average diameters, suggesting that either small craters are too difficult to see, or that they are preferentially eroded or filled by dust quicker in these areas, relative to those with lots of small craters.

SCENE	Valid Individual Annotations	Clustered Annotations	Average Annotations per Crater	Diameter (m)			
				Median	Mean	Min	Max
A	3230	1182	2.73	60.0	66.9	18.0	366.0
B	724	196	3.69	108.0	136.0	36.0	918.0
C	1038	269	3.86	78.0	118.8	30.0	1188.0
D	397	112	3.54	66.0	106.4	24.0	1152.0
E	3946	1042	3.79	42.0	57.7	18.0	1938.0
F	1430	372	3.84	78.0	105.4	18.0	642.0
G	267	72	3.71	105.0	146.4	18.0	642.0
H	223	60	3.72	222.0	263.3	66.0	774.0
I	1110	325	3.42	72.0	93.2	24.0	552.0
J	297	168	1.77	54.0	83.1	18.0	798.0
K	304	81	3.75	90.0	118.2	24.0	672.0
L	2780	884	3.14	60.0	64.6	18.0	630.0
<b>TOTAL</b>	<b>15,746</b>	<b>4763</b>	<b>3.31</b>	<b>60.0</b>	<b>81.1</b>	<b>18.0</b>	<b>1938.0</b>

### 3.2. Validation

We use statistical measures that quantify the amount of agreement between our annotators, and also between our annotators and an experienced crater counter in order to validate the accuracy of our dataset. It should be noted that these measures of agreement do not account for systematic errors (bias), but do tell us about variance to some extent. Ultimately, crater counting is an ambiguous pursuit—especially at smaller diameter ranges—because of the limited resolution of our instruments. To calculate each annotator’s agreement score on a given image, we perform the following calculations:

1. Let the  $i$ th label from annotator  $n$  be denoted as  $L_i^{(n)}$ .
2. For each  $L_i^{(n)}$  made by annotator  $n$ , compute the intersection-over-union of it with all labels from all other annotators.
3. Let the maximum of all these intersection-over-unions be  $MIoU_i^{(n)}$ . This is the highest intersection-over-union of one annotator’s label when compared to all other annotations from other annotators.
4. Take the mean average of  $MIoU_i^{(n)}$  across  $i$ , to calculate the  $n$ th annotator’s *Agreement Score*.

This score rests on the assumption that for any given annotator, the combined labels of all other annotators will be a more complete and reliable survey. However, this does become less reliable at extremely low numbers of annotators. The score also cannot factor in craters that were not found by any other annotators, however it does describe the consistency of detection within the group of annotators, and it tells us about the positional accuracy of markings. On top of this quantitative validation, all images were scanned visually to look for any obviously spurious annotations, which were



then removed from the dataset. These likely came about from accidental inputs and were easy to spot, although they were not common. Table 3 outlines some spatial statistics for the annotations, which detail the level of positional agreement between individual annotators. Overall, the positional agreement of clustered annotations is high, with a standard deviation in measured diameter of around 15%, and the central locations varying with a standard deviation of less than a pixel.

**Table 3.** Positional accuracy of annotations, by number of annotations per crater. Higher diameter craters tend to be annotated more, leading to a higher mean diameter for craters found by more annotators. The standard deviation of these annotations, in comparison to the final clustered value, is around 14–18% overall. The accuracy of the craters' centres is high, with a standard deviation consistently below one pixel (6 m).

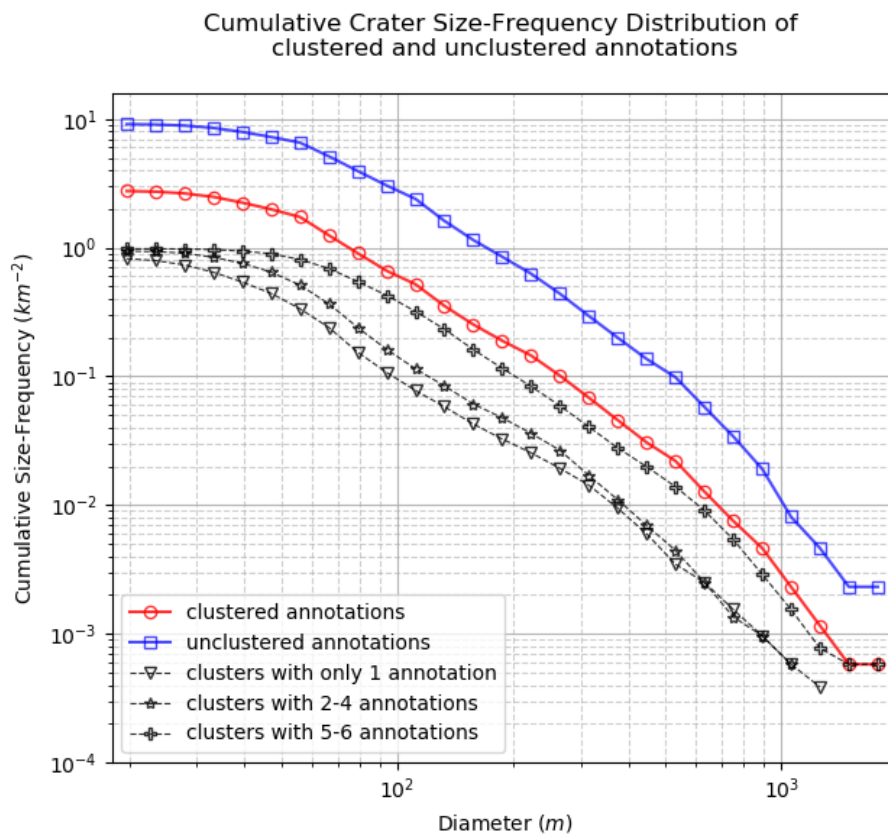
No. of Annotations for Crater	No. of Craters	Mean Diameter (m)	Standard Deviation of Diameter (%)	Standard Deviation of Centre (m)
1	1442	61.8	-	-
2	636	68.4	16.6	4.37
3	524	72.1	18.1	4.87
4	467	72.6	17.8	4.97
5	572	82.0	16.9	5.51
6	1122	120.5	13.8	5.49

We conducted a simple comparison exercise with an experienced crater counter to further measure the reliability of the multi-annotator labels. The 'expert' annotator labelled craters in 4 randomly selected tiles from the dataset (D, F, K, and L). Subsequently, the Agreement score between these expert labels were calculated against the set of other annotators. As we can see from Table 4, the expert achieved scores that exceeded the average inter-annotator score, and in the case of tiles D and F exceeded all of the other annotators. Given this—and the fact that the number of labels from the expert in each tile was always within the non-experts' range—we see that the expert's annotations are somewhat more reliable than an individual annotator's, but also that the agreement between the expert against the entire body of non-experts is high. This suggests that using six non-experts can provide results that match those of an expert considerably more closely than those of an individual non-expert.

The Agreement score and positional accuracy measures gives us a way to validate the consistency of individual annotations. However, it does not provide insight into whether the emergent statistics of our crater surveys are reliable. For this, it is helpful to consider the cumulative CSFD of our annotations (Figure 4). Cumulative CSFDs are generally defined as the number of craters above a given diameter per km<sup>2</sup>, and are used commonly in age-dating work [26]. Our data holds to the expected inverse power law found in cumulative CSFDs, showing a broadly constant gradient above 50 m diameters. Although no ground-truth exists by which to verify our results, we believe that with the Agreement score, positional accuracy metrics, comparison with experts, and the cumulative CSFDs all suggest that the annotations comprising our dataset are reliable both individually and in aggregate.

**Table 4.** Results for experiment using expert annotations on a subset of the dataset. In general, the expert annotator scores highly in comparison to the other annotators, and has a total count similar to the mean of the non-experts for each tile. This shows that when considered as a whole, the non-experts’ annotations have a high agreement with an expert’s, suggesting that multi-annotator methods can boost the performance of non-experts for Object Detection labelling tasks.

SCENE	Non-Expert Annotators						Expert Annotator	
	No. of Labels			Agreement Score (%)			No. of Labels	Agreement Score (%)
	Min	Max	Mean	Min	Max	Mean		
D	32	94	66.5	62.9	77.1	72.9	52	81.6
F	187	305	238.5	65.3	80.6	74.3	240	81.4
K	28	66	50.7	66.8	100	78.4	51	78.2
L	273	696	464.8	46.4	74.1	62.8	589	69.4



**Figure 4.** Cumulative CSFDs of the whole dataset. As expected, there are more unclustered annotations than clustered. At small diameter ranges (<50 m) we see that the gradient tails off, which is an expected symptom of the limited resolution of our data. At the other end of the size range, the statistics become more unreliable, as the number of craters becomes very low. In relative terms, there are more clusters with single annotations (craters with low confidence) at smaller diameter ranges, whilst most of the large craters were found by five or six annotators, which is to be expected, as they are much more visible.

#### 4. Discussion

This small crater survey offers a valuable training and validation set for researchers aiming to develop Crater Detection Algorithms. In the specific case of Deep Learning, we believe that a combination of this with other datasets (e.g., [7,15]) may provide enough data at many scales, locations and with different instruments to create a highly generalised CDA, which is robust to changes in surface features, noise, crater morphology, and crater size.

Many previous crater surveys have focused on annotating larger diameter (>1 km) craters, which brings with it different challenges. The 2014 global survey by Robbins et al. [7] used both Digital Elevation Models and infrared images, as opposed to high-resolution images. Resolving these larger craters was likely less difficult; however, the extreme erosion of some large craters, and their tendency to more frequently overlap with one another, posed difficulties that were less affected in our size range.

The multi-annotator approach we have used is scalable, and it can be adapted to many other Object Detection tasks, both within and beyond Remote Sensing. We believe a more nuanced approach to training and validating models, which accounts for confidence, is particularly useful for problems wherein ground-truth data is unreliable, because of issues, like noise and resolution. Framing an ambiguous problem, such as small crater detection as a binary one leads to model performance being measured in a way that does not reflect real-world capabilities.

We suggest, based on our experience in this project, that the extended training and seminar series conducted prior to labelling resulted in annotations with higher accuracy. In addition, the sense of genuine scientific collaboration (as opposed to labouring at a task one does not hold a stake in) between the ORBYTS students and academic team was a strong motivator that helped annotators to stay focused and productive during the labelling sessions. Future efforts that use this hybrid crowdsourcing approach should continue to optimise the balance between speed (fewer repeated annotations and less training) and accuracy (more repeated annotations and more training) on a case-by-case basis. For small craters, we believe that the community would benefit from future work that samples areas from across different parts of the planet, and provides more rich annotations of physical crater characteristics where possible (e.g., estimates of crater erosion, ellipticity, depth, etc.).

#### 5. User Notes

For access to the dataset, and associated python code to cluster, display and analyse the data, please visit: <https://doi.org/10.5281/zenodo.3946647>.

**Author Contributions:** Conceptualization, A.F., P.S., J.-P.M.; methodology, A.F., P.S., J.-P.M., J.B., T.C., R.C.C., R.D., J.H., M.N., J.N., V.P., A.P., O.R., A.R., D.R., M.S., W.V., A.W.; software, A.F.; validation, A.F., J.B., T.C., R.C.C., R.D., J.H., M.N., J.N., V.P., A.P., O.R., A.R., D.R., M.S., W.V., A.W.; formal analysis, A.F.; investigation, A.F., J.B., T.C., R.C.C., R.D., J.H., M.N., J.N., V.P., A.P., O.R., A.R., D.R., M.S., W.V., A.W.; data curation, A.F.; writing—original draft preparation, A.F., J.B., T.C., R.C.C., R.D., J.H., M.N., J.N., V.P., A.P., O.R., A.R., D.R., M.S., W.V., A.W.; writing—review and editing, A.F., P.S., J.-P.M.; visualization, A.F.; supervision, A.F., P.S., J.P.M.; project administration, A.F., P.S., J.-P.M.; funding acquisition, A.F., J.-P.M.

**Funding:** Partial funding was provided by STFC through grant no. 1912521 and the ORBYTS scheme.

**Acknowledgments:** The authors would like to thank those involved in the planning and administration of the ORBYTS scheme: William Dunn, Jonathan Holdship, Lucinda Offer, Marcell Tessenyi and Maria Niculescu-Duvaz, among others. We would also like to acknowledge the resources, space and support provided by The College of Richard Collyer, in particular by Matthew Horncastle and James Waller, whose enthusiasm and organisation helped the project tremendously.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

CDA	Crater Detection Algorithm
CNN	Convolutional Neural Network
CSFD	Crater Size-Frequency Distribution
CTX	ConTeXt camera
HRSC	High Resolution Stereo Camera
IoU	Intersection over Union
MC-11	Mars Chart-11
MIoU	Mean Intersection over Union
ORBYTS	Original Research By Young Twinkle Students
PASCAL VOC	Pattern Analysis, Statistical Modelling and Computational Learning - Visual Object Classes

## References

- Ivanov, B.; Neukum, G.; Wagner, R. Size-frequency distributions of planetary impact craters and asteroids. In *Collisional Processes in the Solar System*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 1–34.
- Barlow, N.G. Crater size-frequency distributions and a revised Martian relative chronology. *Icarus* **1988**, *75*, 285–305. [[CrossRef](#)]
- Williams, J.P.; van der Bogert, C.H.; Pathare, A.V.; Michael, G.G.; Kirchoff, M.R.; Hiesinger, H. Dating very young planetary surfaces from crater statistics: A review of issues and challenges. *Meteorit. Planet. Sci.* **2018**, *53*, 554–582. [[CrossRef](#)]
- Robbins, S.J.; Hynek, B.M. The secondary crater population of Mars. *Earth Planet. Sci. Lett.* **2014**, *400*, 66–76. [[CrossRef](#)]
- McEwen, A.S.; Preblich, B.S.; Turtle, E.P.; Artemieva, N.A.; Golombek, M.P.; Hurst, M.; Kirk, R.L.; Burr, D.M.; Christensen, P.R. The rayed crater Zunil and interpretations of small impact craters on Mars. *Icarus* **2005**, *176*, 351–381. [[CrossRef](#)]
- Robbins, S.J.; Hynek, B.M. Secondary crater fields from 24 large primary craters on Mars: Insights into nearby secondary crater production. *J. Geophys. Res. Planets* **2011**, *116*, E10003. [[CrossRef](#)]
- Robbins, S.J.; Hynek, B.M. A new global database of Mars impact craters  $\geq 1$  km: 1. Database creation, properties, and parameters. *J. Geophys. Res. Planets* **2012**, *117*, E05004. [[CrossRef](#)]
- Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, **2012**; pp. 1097–1105.
- Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, **2015**; pp. 91–99.
- Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
- DeLatte, D.M.; Crites, S.T.; Guttenberg, N.; Tasker, E.J.; Yairi, T. Segmentation Convolutional Neural Networks for Automatic Crater Detection on Mars. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2944–2957. [[CrossRef](#)]
- Lee, C. Automated crater detection on Mars using deep learning. *Planet. Space Sci.* **2019**, *170*, 16–28. [[CrossRef](#)]
- Urbach, E.R.; Stepinski, T.F. Automatic detection of sub-km craters in high resolution planetary images. *Planet. Space Sci.* **2009**, *57*, 880–887. [[CrossRef](#)]
- Cohen, J.P.; Lo, H.Z.; Lu, T.; Ding, W. Crater detection via convolutional neural networks. *arXiv* **2016**, arXiv:1601.00978.
- Bandeira, L.; Ding, W.; Stepinski, T.F. Detection of sub-kilometer craters in high resolution planetary images using shape and texture features. *Adv. Space Res.* **2012**, *49*, 64–74. [[CrossRef](#)]
- Jaumann, R.; Neukum, G.; Behnke, T.; Duxbury, T.C.; Eichertopf, K.; Flohrer, J.; Gasselt, S.; Giese, B.; Gwinner, K.; Hauber, E.; et al. The high-resolution stereo camera (HRSC) experiment on Mars Express: Instrument aspects and experiment conduct from interplanetary cruise through the nominal mission. *Planet. Space Sci.* **2007**, *55*, 928–952. [[CrossRef](#)]

17. Bugiolacchi, R.; Bamford, S.; Tar, P.; Thacker, N.; Crawford, I.A.; Joy, K.H.; Grindrod, P.M.; Lintott, C. The Moon Zoo citizen science project: Preliminary results for the Apollo 17 landing site. *Icarus* **2016**, *271*, 30–48. [[CrossRef](#)]
18. Malin, M.C.; Bell, J.F.; Cantor, B.A.; Caplinger, M.A.; Calvin, W.M.; Clancy, R.T.; Edgett, K.S.; Edwards, L.; Haberle, R.M.; James, P.B.; et al. Context camera investigation on board the Mars Reconnaissance Orbiter. *J. Geophys. Res. Planets* **2007**, *112*. [[CrossRef](#)]
19. Michael, G.; Walter, S.; Kneissl, T.; Zuschneid, W.; Gross, C.; McGuire, P.; Dumke, A.; Schreiner, B.; van Gasselt, S.; Gwinner, K.; et al. Systematic processing of Mars Express HRSC panchromatic and colour image mosaics: Image equalisation using an external brightness reference. *Planet. Space Sci.* **2016**, *121*, 18–26. [[CrossRef](#)]
20. Sidiropoulos, P.; Muller, J.P.; Watson, G.; Michael, G.; Walter, S. Automatic coregistration and orthorectification (ACRO) and subsequent mosaicing of NASA high-resolution imagery over the Mars MC11 quadrangle, using HRSC as a baseline. *Planet. Space Sci.* **2018**, *151*, 33–42. [[CrossRef](#)]
21. Gwinner, K.; Jaumann, R.; Hauber, E.; Hoffmann, H.; Heipke, C.; Oberst, J.; Neukum, G.; Ansan, V.; Bostelmann, J.; Dumke, A.; et al. The High Resolution Stereo Camera (HRSC) of Mars Express and its approach to science analysis and mapping for Mars and its satellites. *Planet. Space Sci.* **2016**, *126*, 93–138. [[CrossRef](#)]
22. Zurek, R.W.; Smrekar, S.E. An overview of the Mars Reconnaissance Orbiter (MRO) science mission. *J. Geophys. Res. Planets* **2007**, *112*, E05S01. [[CrossRef](#)]
23. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
24. Original research by Young twinkle students (ORBYTS): When can students start performing original research? *Phys. Educ.* **2017**, *53*. [[CrossRef](#)]
25. McQuitty, L.L. Elementary linkage analysis for isolating orthogonal and oblique types and typical relevancies. *Educ. Psychol. Meas.* **1957**, *17*, 207–229. [[CrossRef](#)]
26. Michael, G.; Neukum, G. Planetary surface dating from crater size–frequency distribution measurements: Partial resurfacing events and statistical age uncertainty. *Earth Planet. Sci. Lett.* **2010**, *294*, 223–229. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).